# THE VIRAL AUTHORS OF EVOLUTION

## FROM SEQUENCE TO STRUCTURE

by

William Stevens Henriques

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

of

Doctor of Philosophy

in

Microbiology and Immunology

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2024

©COPYRIGHT

by

William Stevens Henriques

2024

All Rights Reserved

# DEDICATION

This dissertation is dedicated to *E. coli* and elephants, and all the biology in between.

# ACKNOWLEDGEMENTS

Science is an adventure into the unknown that requires curiosity, boldness, and luck. But science is also coming back with a new kernel of knowledge, which requires grit, creativity, and clarity. I first want to thank my mentor Blake Wiedenheft for teaching me these lessons. Working with Blake has made me a better storyteller and scientist. Science is also filled with failures and missteps. I thank my co-mentor Harmit Malik, who has reminded me that failures and missteps are part of the process, and the most important thing is to return to the lab day in and day out with both compassion and enthusiasm.

No one ventures into the unknowns of science alone. Mentors and colleagues offer a helping hand in the squalls of failures, tedium, and rejection. I would like to thank Anna, Artem, Andrew, Royce, and Janet for their mentorship, as well as Murat, Tanner, and Calvin for their bioinformatic companionship and Nate for many shared hours at the microscope. A special thanks to Dr. Martin Lawrence and Colin Gauvin at Montana State's cryoEM Facility, the Research Cyberinfrastructure team and especially Coltran Hophan-Nichols, and my graduate committee for their encouragement and guidance.

Finally, I want to acknowledge my friends who kept me sane and singing, the wonderful crew at Blackbird kitchen who kept me connected to the Bozeman community, and my family, who was always on the other end of the phone during those long dark nights on Everest.

This work was funded in part by the National Institutes of Health (R35GM134867), the M.J. Murdock Charitable Trust, a young investigator award to Blake Wiedenheft from Amgen, and the Montana State University Agricultural Experimental Station (USDA NIFA).

TABLE OF CONTENTS

1. INTRODUCTION ..... 1

Conserved Viral Defense Systems ..... 2
Unique Viral Defense Systems ..... 2
Abundance Leads to Co-option For Defense ..... 4
Domestication Beyond Defense ..... 5
Sequence and Structure ..... 6
Dissertation Overview ..... 7
References ..... 12

2. GENOME SEQUENCE, PHYLOGENETIC ANALYSIS, AND STRUCTURE BASED ANNOTATION REVEALS METABOLIC POTENTIAL OF *CHLORELLA SP.* SLA-04 ..... 20

Contribution of Authors and Co-Authors ..... 20
Manuscript Information ..... 22
Abstract ..... 23
Introduction ..... 24
Methods ..... 25
Strain Information ..... 25
Strain Cultivation Conditions ..... 25
Axenic Isolation ..... 26
Genome Sequencing and Assembly ..... 26
High Molecular Weight DNA Extraction ..... 26
RNA Extraction ..... 27
DNA and RNA Sequencing ..... 28
Genome Assembly and Annotation ..... 28
Comparative Analyses ..... 30
Phylogenetic Analysis. ..... 30
Results ..... 31
General Genome Characteristics ..... 31
Chlorella sp. SLA-04 Genome Assembly Statistics ..... 31
Genome Completeness Analysis ..... 31
Annotation of Chlorella Sp. SLA-04 ..... 32
Annotation Using Protein 3D Structural Homology ..... 33
Phylogenetic Analysis of Chlorella Sp. SLA-04 ..... 35
Genomic Analysis of Metabolic Pathways in *Chlorella Sp*. SLA-04 ..... 35
Discussion ..... 37
Conclusion ..... 40
Data Access ..... 40
Acknowledgements ..... 40
References ..... 40

TABLE OF CONTENTS CONTINUED

Supplementary Figures ..... 48
Supplementary Tables ..... 52

3. CLARIFYING CRISPR: WHY REPEATS IDENTIFIED IN THE HUMAN GENOME SHOULD NOT BE CONSIDERED CRISPRS ..... 54

Contribution of Authors and Co-Authors ..... 54
Manuscript Information ..... 55
Abstract ..... 56
Introduction ..... 56
Methods ..... 57
CRISPR Predictions ..... 57
Identification of Cas Genes in the Human Genome ..... 59
Comparison of Repeat Length Distribution of CRISPRs from Prokaryotes and HCRISPRs ..... 59
Results and Discussion ..... 60
References ..... 65
Supplementary Figures ..... 68
Supplemental Information ..... 69

4. THE DIVERSE EVOLUTIONARY HISTORIES OF DOMESTICATED METAVIRAL CAPSID GENES IN MAMMALS ..... 73

Contribution of Authors and Co-Authors ..... 73
Manuscript Information ..... 74
Abstract ..... 75
Introduction ..... 76
Results ..... 80
Divergent Metaviral Capsid-derived Genes in the Human Genome ..... 80
Domesticated Metaviral Capsid Genes Experienced Distinct Evolutionary Retention Fates ..... 84
Domesticated Metaviral Capsid Genes are Retained under Purifying and Positive Selection ..... 89
Domesticated Capsid Genes are Structurally Diverse ..... 93
Diverse N-terminal Structures in Domesticated Genes Derived from Metaviral Ancestors ..... 98
A Novel Putative RNA-binding Domain in the PNMA Family ..... 101
Discussion ..... 105
Methods ..... 111
Building Capsid HMMs ..... 111
Identifying Capsid Genes in Vertebrate Genomes ..... 113

TABLE OF CONTENTS CONTINUED

Matching Human Capsid-like ORF Sequences to Previously Annotated Genes ..... 114
Analysis of Domain Architecture and Genetic Context ..... 114
Phylogenetic Analysis of Capsid-like Sequences ..... 114
Identification of Mammalian Orthologs ..... 115
Analysis of Evolutionary Selective Constraints ..... 119
Structure Predictions and Homology Modelling ..... 121
Phylogenetic Analysis of Putative RBD Sequences ..... 122
Expression Analysis ..... 123
International Mouse Phenotyping Consortium ..... 123
Data and Resource Availability ..... 123
References ..... 124
Supplementary Figures ..... 136
Supplementary Tables ..... 153

5. THE RETROTRANSPOSON-DERIVED CAPSID GENES *PNMA1* AND *PNMA4* MAINTAIN REPRODUCTIVE CAPACITY ..... 159

Contribution of Authors and Co-Authors ..... 159
Manuscript Information ..... 162
Abstract ..... 163
Introduction ..... 163
Results ..... 165
Expression of *PNMA1* and *PNMA4* Declines with Age in Human Ovaries ..... 165
Male and Female Mice Lacking *Pnma1* or *Pnma4* Show Premature Loss of Fertility ..... 168
*Pnma1* and *Pnma4* Mutants Acquire Abdominal Obesity but Appear Behaviorally Normal ..... 173
PNMA1 and PNMA4 Form Capsid-like Structures in Gonadal Tissue ..... 175
Human Variation at the *PNMA1* and *PNMA4* Loci is Associated with Reproductive Defects ..... 178
Discussion ..... 179
Methods ..... 183
Generation of Mutant Mouse Models ..... 183
Mouse Husbandry ..... 184
Mouse Genotyping ..... 185
Fertility Assay ..... 186
Body and Gonad Weight ..... 186
Immunofluorescence of Testis Sections ..... 186
Epididymal Sperm Counts ..... 188

TABLE OF CONTENTS CONTINUED

Periodic Acid-Schiff (PAS) Staining ..... 188
Blood Collection, Sample Preparation, and Hormone Analysis ..... 189
Ovarian Histology ..... 189
Oocyte Collection and Maturation ..... 189
Mouse Oocyte Isolation, Culturing, Microinjection and High-Resolution Confocal Microscopy ..... 190
Plasmid Creation ..... 191
HEK293T Cell Culture ..... 191
HEK293T Transfection ..... 192
Cloning, Expression, and Purification of Recombinant PNMA1 and PNMA4 ..... 193
Negative Stain Transmission Electron Microscopy ..... 194
Velocity Gradient Ultracentrifugation ..... 195
Western Blotting ..... 195
Quantitative PCR ..... 197
Extracellular Protein Analysis ..... 198
Analysis of Testicular Capsid-Like PNMA4 Protein ..... 198
Lysate Preparation ..... 198
Double Sucrose Gradient ..... 199
Iodixanol Step Gradient ..... 199
Northern Blot Analysis ..... 199
*PNMA1-5* Expression Analysis in Human and Mouse Ovary ..... 200
Analysis of Traits Associated with Human *PNMA1* and *PNMA4* Variation ..... 201
Statistical Analysis ..... 201
Neurobehavioral Tests ..... 201
Open Field Test ..... 202
Grip Strength Test ..... 202
Inverted Hang Test ..... 202
Y Maze Spontaneous Alternation ..... 203
Fear Conditioning ..... 203
References ..... 204
Supplementary Figures ..... 212
Supplementary Tables ..... 224

6. A VIRALLY-ENCODED TRNA NEUTRALIZES THE PARIS ANTIVIRAL DEFENSE SYSTEM ..... 228

Contribution of Authors and Co-Authors ..... 228
Manuscript Information ..... 231
Abstract ..... 232
Introduction ..... 232

TABLE OF CONTENTS CONTINUED

Results ..... 234
PARIS Forms a Propeller-shaped Complex ..... 234
AriA Sequesters AriB ..... 236
AriA Binds to Ocr and Releases AriB ..... 239
Activated PARIS Inhibits Translation ..... 244
Triggers and a Suppressor of PARIS ..... 245
Viral TRNA Resistant to AriB Cleavage ..... 246
Evolution of the PARIS Defense System ..... 249
Discussion ..... 250
Methods ..... 252
Bacterial Strains, Phages, and Plasmids ..... 252
Expression and Purification of the PARIS Complex ..... 253
*In Vivo* Protein Pull-down ..... 254
Glutaraldehyde Cross-linking Assay ..... 255
*In Vitro* Protein Pull-down ..... 255
Cryo-EM Sample Preparation and Data Collection ..... 256
Cryo-EM Data Processing of Inactive PARIS ..... 257
Cryo-EM Data Processing of Ocr Pulldown ..... 258
Model Building and Refinement ..... 258
ATPase Assays ..... 259
Efficiency of Plaquing Assay ..... 260
Liquid Culture Phage Infection ..... 261
Solid Media Toxicity Assay ..... 261
Liquid Culture Toxicity Assay ..... 262
Fluorescence Microscopy ..... 262
Extraction of Total DNA ..... 263
TUNEL Assay ..... 263
Hi-C Procedure and Sequencing ..... 264
Metabolic Labeling ..... 265
*In Vitro* Translation ..... 266
Isolation and Sequencing of T5 Escaper Mutants ..... 267
Extraction of Total RNA ..... 268
Preparation of Enriched Small RNA Fractions from the Total RNA Samples ..... 269
Cleavage Assay by AriB ..... 270
Mapping of the AriB Cleavage Site by Primer Extension ..... 270
Mapping the AriB Cleavage Site by Specific Reverse Transcription (RT) and Sanger Sequencing ..... 271
Northern Blot ..... 272
Defense System Detection ..... 272
Domain Annotation ..... 273

TABLE OF CONTENTS CONTINUED

AriA and AriB Tree ..... 273
Defense System ATPase AAA15/21 Tree ..... 274
Statistics and Reproducibility ..... 274
References ..... 274
Extended Data Figures ..... 280
Extended Data Tables ..... 291
Supplementary Tables ..... 292

7. STRUCTURE REVEALS WHY GENOME FOLDING IS NECESSARY FOR SITE-SPECIFIC INTEGRATION OF FOREIGN DNA INTO CRISPR ARRAYS ..... 312

Contribution of Authors and Co-Authors ..... 312
Manuscript Information ..... 314
Abstract ..... 315
Introduction ..... 315
Results ..... 318
Cryo-EM Structure of Type I-F CRISPR Integration Complex ..... 318
Foreign DNA Constrains Cas2/3 Linker against Cas1 ..... 322
Cas2 Homodimers Recognize and Bend Inverted Repeat Sequences ..... 324
Cas2 Homodimer is Surrounded by Four DNA Helices ..... 326
IHF and the Leader Direct Integration into Diverse Sequences ..... 326
5’ GT is Critical for Cas1-mediated Integration ..... 329
Discussion ..... 330
Methods ..... 335
Nucleic Acid Preparation ..... 335
Cas1 and Cas2/3 Mutagenesis ..... 336
Protein Purification ..... 336
*In Vitro* Integration Assays ..... 338
Assembly and Purification of I-F Integration Complex ..... 339
Cryo-EM Sample Preparation and Data Acquisition ..... 340
Cryo-EM Image Processing ..... 342
Model Building and Validation ..... 343
Cas1, Cas2/3, and Repeat Conservation Analysis ..... 344
Data Availability ..... 347
Extended Data Figures ..... 347
References ..... 358
Supplementary Tables ..... 366

TABLE OF CONTENTS CONTINUED

8. STRUCTURES REVEAL HOW A CRISPR INTEGRASE CAPTURES, DELIVERS, AND INTEGRATES FOREIGN DNA INTO THE CRISPR ARRAY ..... 374

Contribution of Authors and Co-Authors ..... 374
Manuscript Information ..... 375
Abstract ..... 376
Introduction ..... 376
Results ..... 378
Cas1-2/3 Forms a DNA Capture Complex ..... 378
Structural Basis for Cas3 HD Nuclease and RecA Helicase Inhibition ..... 382
Cas3 Blocks Foreign DNA Integration ..... 385
Foreign DNA Capture Exposes Genome Binding Sites ..... 386
Structure Precludes a Role for Cas3 in PAM Trimming ..... 390
Foreign DNA Binding Exacerbates Asymmetric Instability ..... 391
The PAM Changes the Stability of the Integration Complex ..... 391
The CRISPR Repeat Channel Stretches and Distorts The Repeat ..... 395
Discussion ..... 396
Materials & Methods ..... 399
Conservation Analysis of Cas2/3 and Cas1 ..... 399
Nucleic Acid Preparation ..... 400
Protein Purification ..... 400
Assembly and Purification of the Delivery Complex ..... 402
Assembly and Purification of the Integration Complex ..... 402
CryoEM Sample Preparation and Data Acquisition – Cas1-2/3 Capture Complex ..... 403
CryoEM Sample Preparation and Data Acquisition – Cas1-2/3 Delivery Complex ..... 403
CryoEM Sample Preparation and Data Acquisition – Cas1-2/3 Integration Complex ..... 404
Cryo-EM Image Processing ..... 405
Model Building and Validation ..... 405
References ..... 406
Supplementary Figures ..... 414
Supplementary Tables ..... 424

9. CONCLUSION ..... 426

From Sequence to Structure ..... 428
Structure Determines Function ..... 429
Concluding Remarks ..... 430

TABLE OF CONTENTS CONTINUED

References ..... 431

CUMULATIVE REFERENCES CITED ..... 433

APPENDIX: RAPID CRYOEM STRUCTURE DETERMINATION FROM CELL-FREE TRANSCRIPTION AND TRANSLATION ..... 493

Abstract ..... 494
Main Text ..... 494
References ..... 498

# LIST OF TABLES

Table Page

1. Supplementary Table 2.1. Genome statistics of *Chlorella sp.* SLA-04 genome. ..... 52
2. Supplementary Table 2.2. Summary of repetitive elements identified by RepeatMasker using Dfam and Repbase as a combined reference database. ..... 53
3. Supplementary Table 2.3. Active transcription of lipid synthesis genes within the Chlorella sp. SLA-04 genome. ..... 53
4. Supplementary Table 4.1. Model statistics for custom capsid Hidden Markov Models (HMMs) build from multiple sequence alignments (MSA) of capsid amino acid sequences. ..... 153
5. Supplementary Table 4.2. Gene names table. ..... 155
6. Supplementary Table 4.3. Table of genomes used in this study ..... 156
7. Supplementary Table 4.4. BLAST statistics for consensus Repbase metaviral sequences used as proxies for ancestral metaviruses ..... 157
8. Supplementary Table 4.5. Phenotypes associated with knockouts of domesticated metaviral genes in mice. ..... 158
9. Supplementary Table 5.1. Traits associated with PNMA1 and PNMA4 variants in the human population. ..... 227
10. Extended Data Table 6.1. Cryo-EM data collection, processing, and model validation ..... 291
11. Supplementary Table 1. Bacterial strains, phages, and plasmids used in the study ..... 299
12. Supplementary Table 2. Primers used for cloning ..... 311
13. Supplementary Table 3. Nucleotide sequence of the probes used for Northern blot and in primer extension ..... 311
14. Table 7.1. Cryo-EM data collection, refinement, and validation statistics ..... 321
15. Supplementary Table 1. Oligonucleotides used in this study. ..... 373
16. Supplementary Table 8.1. Refinement statistics for the Cas1-2/3 complex ..... 424

LIST OF TABLES CONTINUED

17. Supplementary Table 8.2. Refinement statistics for Cas1-2/3 bound to dsDNA fragment ..... 425

# LIST OF FIGURES

Figure Page

1. Figure 2.1. Complete genome of Chlorella sp. SLA-04. ..... 32
2. Figure 2.2. Annotation using protein structural similarity improves functional understanding of SLA-04 proteome. ..... 33
3. Figure 2.3. Phylogenomic analysis of green algae performed using orthologous genes indicates that SLA-04 is distinct from other Chlorella sorokiniana species. ..... 34
4. Figure 2.4. Flow of carbon from fixation to lipid synthesis in Chlorella sp. SLA-04. ..... 36
5. Supplementary Figure 2.1. DNA isolation and sequencing parameters for Chlorella sp. SLA-04. ..... 48
6. Supplementary Figure 2.2. Phylogenetic tree of LTR retrotransposon capsid domains places SLA-04 retrotransposons among Copia/Pseudoviridae LTR retrotransposons. ..... 49
7. Supplementary Figure 2.3. Phylogenomic analysis of green algae performed using 18S rDNA. ..... 50
8. Supplementary Figure 2.4. Presence of ATP citrate lyase and sphingolipid enzymes across algal genomes. ..... 51
9. Figure 3.1. Comparison of architectural features in CRISPRs and hCRISPRs. ..... 60
10. Figure 3.2. Repetitive elements in the human genome do not contain architectural features consistent with CRISPR loci ..... 62
11. Supplementary Figure 3.1. Previously reported CRISPR-like elements in the mitochondrial genome of a plant (Vicia faba) and the genome of a giant mimivirus (APMV) ..... 68
12. Figure 4.1. Full-length capsid-like ORFs in the human genome ..... 82
13. Figure 4.2. Metaviral-derived capsid genes show distinct evolutionary trajectories across placental mammals. ..... 85

LIST OF FIGURES CONTINUED

14. Figure 4.3. RTL3 and RTL10 exhibit domain-specific and lineage-specific patterns of conservation. ..... 87

15. Figure 4.4. Evolutionary rates of domesticated metaviral genes across placental mammals and primates. ..... 91

16. Figure 4.5. Four independent metavirus domestication events include structurally distinct N-terminal domains. ..... 94

17. Figure 4.6. Predicted RNA-binding domain in the PNMA family and related metaviruses ..... 102

18. Supplementary Figure 4.1. Constructing Hidden Markov Models for the capsid domain of LTR retrotransposons ..... 136

19. Supplementary Figure 4.2. Phylogeny of ARC proteins ..... 137

20. Supplementary Figure 4.3. Phylogeny of the putative RNA-binding domain of PNMA proteins. ..... 139

21. Supplementary Figure 4.4. Phylogeny of PNMA protein capsid domains. ..... 140

22. Supplementary Figure 4.5. PNMA6E/F has duplicated multiple times in distinct lineages. ..... 142

23. Supplementary Figure 4.6. Phylogenetic tree for SIRH/RTL family (half capsid alignment). ..... 144

24. Supplementary Figure 4.7. Phylogenetic tree for SIRH/RTL family (full capsid alignment). ..... 145

25. Supplementary Figure 4.8. Conservation and divergence of the capsid domain in the SIRH/RTL family. ..... 146

26. Supplementary Figure 4.9. Lack of structural conservation in the N-terminal domain of the SIRH/RTL family. ..... 148

27. Supplementary Figure 4.10. The evolutionary origins of the putative RNA-binding domain of the PNMA family. ..... 150

28. Supplementary Figure 4.11. Human expression of domesticated metaviral capsid genes. ..... 152

LIST OF FIGURES CONTINUED

29. Figure 5.1. PNMA1 and PNMA4 are expressed in human ovaries. ..... 166

30. Figure 5.2. PNMA1 and PNMA4 are conserved among eutherians ..... 168

31. Figure 5.3. Male mice lacking Pnma1 or Pnma4 prematurely lose reproductive capacity ..... 170

32. Figure 5.4. Female mice lacking Pnma1 or Pnma4 have age-dependent reproductive defects. ..... 172

33. Figure 5.5. Pnma1 and Pnma4 mutants gain abdominal fat with age. ..... 174

34. Figure 5.6. PNMA proteins form capsid-like structures that can exit human cells ..... 176

35. Figure 5.7. Model for PNMA1 and PNMA4 function. ..... 179

36. Supplementary Figure 5.1. PNMA2, PNMA3, and PNMA5 expression analysis in human ovaries. ..... 212

37. Supplementary Figure 5.2. DAZL has the capacity to post-transcriptionally regulate PNMA1 and PNMA4. ..... 213

38. Supplementary Figure 5.3. Diagram of Pnma1 and Pnma4 loci on M. musculus chromosome 12 ..... 214

39. Supplementary Figure 5.4. Defective testicular and ovarian morphology in Pnma1 and Pnma4 mutants ..... 216

40. Supplementary Figure 5.5. Analysis of undifferentiated spermatogonial cells by PLZF staining ..... 217

41. Supplementary Figure 5.6. Several ovarian features are not dramatically altered in Pnma1 and Pnma4 mutants. ..... 218

42. Supplementary Figure 5.7. Oocytes that enter meiosis are largely unaffected in Pnma1 and Pnma4 mutants. ..... 219

43. Supplementary Figure 5.8. Single-cell analysis of Pnma1 and Pnma4 expression in aging mouse ovaries ..... 220

44. Supplementary Figure 5.9. Pnma1 and Pnma4 are expressed in subpopulations of ovarian cells ..... 221

LIST OF FIGURES CONTINUED

45. Supplementary Figure 5.10. Mice lacking Pnma1 or Pnma4 exhibit normal neurobehavioral and muscular traits. ..... 222

46. Supplementary Figure 5.11. Size exclusion chromatography of recombinant PNMA1 and PNMA4. ..... 223

47. Supplementary Figure 5.12. PNMA1 and PNMA4 expressed in human cells forms capsid-sized particles. ..... 224

48. Figure 6.1. PARIS is a two-component system that assembles into a propeller-shaped supramolecular complex. ..... 235

49. Figure 6.2. Interaction between the AriA ATPase and AriB essential for PARIS-mediated defense. ..... 237

50. Figure 6.3. AriA interacts with Ocr in an ATPase-dependent manner, leading to the release of AriB from the PARIS complex ..... 240

51. Figure 6.4. The activation of PARIS leads to cell death and inhibition of translation ..... 242

52. Figure 6.5. PARIS cleaves E. coli tRNALys and T5 encodes a non-cleavable tRNALys. ..... 247

53. Figure 6.6. Phylogenies of PARIS and related ABC ATPase powered defense systems ..... 249

54. Extended Data Figure 6.1. AlphaFold2 structural predictions of AriA and AriB. ..... 280

55. Extended Data Figure 6.2. Structural comparison of PARIS homologs. ..... 281

56. Extended Data Figure 6.3. Workflow for structural determination of the PARIS complex ..... 282

57. Extended Data Figure 6.4. AriB toxicity is dependent on association with AriA ..... 283

58. Extended Data Figure 6.5. Purification of activated AriB. ..... 284

59. Extended Data Figure 6.6. Structure determination of AriA purified using a Ocr-Strep pulldown. ..... 285

LIST OF FIGURES CONTINUED

60. Extended Data Figure 6.7. AriA ATPase activity. ..... 286

61. Extended Data Figure 6.8. PARIS activation does not result in total RNA or DNA degradation ..... 287

62. Extended Data Figure 6.9. PARIS activation results in DNA compactization akin to inhibition of translation with chloramphenicol. ..... 288

63. Extended Data Figure 6.10. Identification of the T5 genomic region required for infection of PARIS+ cells. ..... 289

64. Extended Data Figure 6.11. tRNALys cleavage product and phylogenetics of PARIS ..... 290

65. Figure 7.1. Cryo-EM structure of the type I-F CRISPR integration complex ..... 319

66. Figure 7.2. Foreign DNA constrains the Cas2/3 linker against conserved Cas1 residues. ..... 323

67. Figure 7.3. The Cas2 homodimer simultaneously coordinates four dsDNA helices critical to CRISPR integration. ..... 325

68. Figure 7.4. Sequence motifs in the leader and IHF proteins facilitate Cas1-2/3-based integration into diverse repeat sequences. ..... 328

69. Figure 7.5. I-F CRISPR integration complex suggests a mechanism for primed acquisition and Cascade impact on integration. ..... 331

70. Figure 7.6. DNA is a flexible scaffold that controls DNA mobilization. ..... 333

71. Extended Data Figure 7.1. Cryo-EM sample preparation, imaging and processing for type I-F integration complex ..... 349

72. Extended Data Figure 7.2. Cas1-2/3 undergoes a large structural rearrangement during integration. ..... 350

73. Extended Data Figure 7.3. Conservation analysis of Cas1-2/3 residues involved in DNA binding and integration. ..... 351

74. Extended Data Figure 7.4. Cas1-2/3 predominantly recognizes IR motifs, CRISPR repeat and foreign DNA through non-sequence specific interactions. ..... 352

LIST OF FIGURES CONTINUED

75. Extended Data Figure 7.5. Purification of structure-guided mutants of Cas1-2/3 ..... 353

76. Extended Data Figure 7.6. Validation of Cas1-2/3 interactions with the foreign DNA and IR motifs..... 354

77. Extended Data Figure 7.7. PAM blocks Cas-mediated integration of foreign DNA into CRISPR repeat. ..... 355

78. Extended Data Figure 7.8. Control reactions for integration assay and generation of 32P-labelled ladder. ..... 356

79. Extended Data Figure 7.9. Validation of Cas1-2/3 interactions with the repeat. ..... 357

80. Figure 8.1. One side of the Cas1-2/3 integrase has a positively charged foreign DNA-binding channel. ..... 379

81. Figure 8.2. The Cas3 active sites are blocked in the DNA capture complex. ..... 384

82. Figure 8.3. Foreign DNA triggers a conformational rearrangement that exposes leader recognition sites and preserves Cas3 inhibition ..... 387

83. Figure 8.4. The PAM blocks the second transesterification by altering complex stability. ..... 392

84. Figure 8.5. The repeat channel distorts the CRISPR repeat..... 395

85. Supplementary Figure 8.1. Purification and data processing of the Cas1-2/3 capture complex. ..... 414

86. Supplementary Figure 8.2 Cas1 in complex with Cas2/3 preserves dimeric architecture. ..... 415

87. Supplementary Figure 8.3. Conservation of the four-bladed propeller Cas1-2/3 architecture across type I-F systems ..... 416

88. Supplementary Figure 8.4. The Cas2 C-terminus determines the extent of complex interactions ..... 417

89. Supplementary Figure 8.5. Positively charged foreign DNA binding residues are conserved in the Cas1-2/3 complex ..... 418

LIST OF FIGURES CONTINUED

90. Supplementary Figure 8.6. A conserved domain inserted into the HD active site explains nuclease inhibition ..... 419

91. Supplementary Figure 8.7. Three interfaces clamp Cas3 in place, inhibiting both interference and integration interactions..... 420

92. Supplementary Figure 8.8. Purification and data processing of the Cas1-2/3 capture complex ..... 421

93. Supplementary Figure 8.9. Purification and cryoEM data analysis of PAM-containing integration complex..... 422

94. Supplementary Figure 8.10. Positively charged residues in the repeat channel are conserved in type I-F CRISPR systems..... 423

95. Figure A.1. Sequence to structure in 24 hours ..... 495

96. Figure A.2. TxTl cryoEM produces high resolution structures ..... 497

# ABSTRACT

Viruses are the most abundant biological entities on the planet. Like any parasite, these genetic parasites require a host. Inside a host cell, viruses hijack the transcription and translation machinery to replicate. In response to the ubiquitous threat of viral predation, host cells have evolved an astonishing diversity of mechanisms to stop, control, and even repurpose pieces of viruses. In this dissertation, I use bioinformatic, biochemical, and structural approaches to show this ancient conflict shapes host evolution by constantly rewriting the host genome. Research conducted during this dissertation has spanned the domains of life with three major foci. First, I have used bioinformatics to identify genetic parasites and defense systems embedded within algal and human genomes. Second, I have conducted deep evolutionary analyses to identify viral genes that have been repurposed for host function. And third, I have used cryogenic electron microscopy to determine the structure of viral defense systems in bacteria.

# CHAPTER ONE

## INTRODUCTION

Viruses are the most abundant biological entities on the planet[1,2]. As obligate intracellular genetic parasites, viruses require a host cell which they invade and reprogram to produce more virus. Viral progeny exit the reprogrammed host cell and remain biologically inert until they encounter a new host, at which point the viral replication cycle begins again. Since the discovery of the tobacco mosaic virus in 1898, viruses have been found to infect cells from every domain of life[3–5]. This selfish reproductive strategy is the most successful on the planet: there are an estimated 10 viral particles for every cell on earth[1,2,6]. Viruses influence life from the cellular level to the global level. An estimated 20-40% of the ocean's biomass is lysed every day by viral infection, in a stunning illustration of how selfish genetic elements run riot in the world[1].

However, these estimates incompletely capture the extent to which selfish genetic elements dominate the biosphere. Some genetic parasites, viral-like elements called transposons and retrotransposons, embed in the host genome, replicate within the cell, and insert new copies directly into the genome of their original host. A surprising finding of the Human Genome Project was the discovery that approximately half of the human genome is made up of transposons, retrotransposons, and their decaying remnants[7–9]. Metagenomic studies indicate that selfish genetic elements may make up to 8% of bacterial genes[10]. In eukaryotes, selfish genetic elements can make up anywhere from 3% to 92% of the genome[11,12]. A survey of complete genomes and meta transcriptome data found that transposase genes are by far the most abundant genes in nature[13]. These proteins are used by selfish genetic elements to integrate selfish DNA

directly into the host genome[13]. Between extracellular viruses and intracellular genome parasites, selfish genetic elements exert tremendous pressure on their cellular hosts.

## Conserved Viral Defense Systems

In response to the ubiquitous threat of selfish genetic elements, host cells across all domains of life have evolved diverse defense systems. Some defense systems are conserved from bacteria to eukaryotes and interfere with critical and universal steps in viral replication[14–16]. The cGAS/CD-NTase and STING pathways sense viral infection and synthesize cyclic di-nucleotide secondary messengers to activate an anti-viral response[17–19]. The viperin systems alter ribonucleotide pools in response to viral infection, terminating RNA synthesis[20,21]. Activation of gasdermins upon viral infection triggers assembly of membrane-puncturing pores leading to cell death before the replication cycle can complete[22,23]. Argonaute proteins use short nucleic-acid guides to target and cleave selfish RNA or DNA[24,25]. Toll/Interleukin-1 receptor (TIR) domains associated with pathogen-associated molecular pattern receptors deplete cellular NAD+ upon sensing a viral molecular pattern, depleting cellular energy[26,27]. Defense mechanisms conserved across the domains of life emphasize the ubiquity of viral parasitism. They also underscore the truth in the saying coined by the inventor of the microscope, Antoni von Leeuwenhoek, and famously paraphrased by Jacques Monod and François Jacob: “what is true for bacteria is also true for elephants.”[28]

## Unique Viral Defense Systems

However, not everything that is true for bacteria is true for elephants. Some defense systems, such as the adaptive immune system of vertebrates, have no known homologues in

prokaryotes. Others, such as the CRISPR-Cas and restriction modification (RM) systems, have only been found in prokaryotes[29,30]. Studying these outliers have provided significant insights into the diversity of viral defense mechanisms and tools that have revolutionized medicine and molecular biology.

Early studies of bacterial immune systems revealed bacteria possess a mechanism to distinguish between self and non-self[31,32]. Now called restriction-modification (RM) systems, these two-component systems generally encode a methyl-transferase and a nuclease. Methylation of the genome creates a distinct signature of self. Invading nucleic acids from viruses are recognized and cleaved at specific target sites by nucleases, shutting down the viral infection[33,34]. In a clear example of 'molecular arms race'[35–38], viruses have evolved mechanisms to shut down these first-line of defense systems. For example, the "overcoming classical restriction" (Ocr) protein jams up restriction modification systems by mimicking viral DNA and blocking the active site of the restriction endonuclease.[39] By blocking nuclease activity, viruses carrying Ocr circumvent this first-line defense system.

CRISPR-Cas systems are prokaryote-specific immune systems that also distinguish between self and non-self. However, a key distinguishing feature of the CRISPR-Cas systems is that they are both adaptable and heritable. In CRISPR-Cas systems, clustered regularly interspaced short palindromic repeats (CRISPRs) catalogue foreign DNA to be read out by CRISPR-associated (Cas) proteins which target and degrade complementary foreign nucleic acid. [40,41]. Though structurally diverse, CRISPR-Cas systems generate adaptive immunity through three universally conserved steps: biogenesis, interference, and adaptation (reviewed in [42–45]). In CRISPRs, conserved direct repeat sequences flank unique virally-derived sequences called

spacers[46–48]. During biogenesis, Cas proteins process the transcribed CRISPR array into individual guide RNAs that contain a single unique spacer sequence[49,50]. These guide RNAs are loaded into RNA-guided endonucleases that identify and cleave foreign nucleic acid during interference[51,52] In the adaptation stage, Cas1 and Cas2 assemble into a heterohexameric complex that binds fragments of foreign nucleic acid and inserts them as new spacers through a transposition-like mechanism[53].

## Abundance Leads to Co-option For Defense

The Cas1 protein that mediates foreign DNA insertion derives from the transposase of an ancient selfish genetic element called a casposon[51,54,55]. Cas1 represents a striking example of how selfish genes may be repurposed for the host as 'guns for hire' [37,51,55]. In CRISPR immunity, Cas1 performs a very similar function to its transposase ancestor, inserting short fragments of foreign DNA into the CRISPR by transposition[53,56–59]. However, Cas1 no longer services the replication of a selfish genetic element, but rather functions to protect the cellular host from selfish threats by providing the CRISPR interference machinery with updated guide RNAs.

Intriguingly, adaptive immunity in eukaryotes also relies on a domesticated transposase. Vertebrates possess an adaptive immune system that is capable of generating molecular memories of pathogens[60]. However, most of vertebrate adaptive immunity is not heritable. Memories of infections are not transmitted from generation to generation, though they are preserved for months to years in the form of memory B cells. Rather, the V(D)J recombinase generates fragments of DNA that are ligated back together to create variable antigen recognition sites in antibodies in a mechanism that is quite similar to transposition[54,61]. Indeed, domestication of an ancient DNA transposon gave rise to this form of eukaryotic adaptive immunity. These two

striking examples of transposases domesticated for defense in prokaryotes and eukaryotes begs a deeper analysis of parasitic genes embedded within host genomes to identify parasites repurposed for host function.

## Domestication Beyond Defense

In mammalian genomes, the most common class of selfish elements are retrotransposons, which generate new copies using a 'copy-and-paste' mechanism[62–65]. Retrotransposons are classified into two broad groups: long-terminal-repeat (LTR) retrotransposons and non-long-terminal-repeat (non-LTR) retrotransposons, which include both autonomous LINE-1 retrotransposons and non-autonomous Alu and SVA elements that rely on LINE-1 replication machinery. LTR retrotransposons (which include endogenous retroviruses, ERVs) encode structural group-specific antigen (*gag)* genes and enzymatic polymerase (*pol*) genes, flanked by LTRs. Their 'exogenous' retroviral counterparts also encode an envelope protein (*env*)[66,67], which mediates membrane fusion needed for infection, and can subvert immune defenses via an immunosuppressive domain[68].

Multiple retroviral envelope genes have been domesticated for retroviral defense in mice[69,70], humans[71], sheep[72], cats[73,74], and chickens[75]. However, *env* genes have also been domesticated for non-defense function as a class of genes called *syncitins*. Syncitin proteins mediate membrane fusion to form multinucleated syncytial trophoblast cells in the placenta[76,77]. *Syncytin* gene domestication is a remarkable example of convergence, with at least seven independent events in different mammalian orders[77,78], and in unusual lineages of lizard[79] and fish[80]. These *env* domestications illustrate two key points. First, viral genes can be substrate for

novel host functions beyond their viral function. Second, the homologous viral domains may be repurposed for completely different functions.

Gag genes of LTR retrotransposons and retroviruses encode a polyprotein that includes capsid, matrix, and nucleocapsid domains that together package the viral genome during the virion assembly process[65,81]. Like *env*, *gag* has been domesticated both for retroviral restriction in mice (*e.g.,* the *Fv1* gene[82–84]), as well as for other critical host functions[85–89]. At least four domesticated capsid genes have been shown to assemble into capsid-like structures[90–94], which perform essential functions in host reproduction[87] and neuronal function[90,95]. One of these *gag*-derived genes is vertebrate *Arc (A*ctivity-*R*egulated, *C*ytoskeletal-associated), which originated from an ancient LTR retrotransposon. ARC protein forms capsid-like structures and functions in the brain to regulate learning and memory as a signaling hub and messenger RNA shuttle in neurons[90,95]. An independent domestication event led to *dArc1* in *Drosophila* species, which also forms capsid-like structures and packages mRNA for intercellular neuronal signaling[92,96]. Together with Cas1 and the V(D)J recombinase, *Arc* and *dArc1* elegantly demonstrate how functional homologues from selfish genetic elements have been repurposed for similar host functions across massive evolutionary distances.

## Sequence and Structure

Genetic conflict plays out in the ecosystem of genomes[97]. Two major technological advances have facilitated the study of these ecosystems by rendering previously invisible molecular processes visible. Genome sequences contain the set of instructions that build proteins, organelles, cells, and organisms. Fred Sanger changed the world when he developed his method for efficiently sequencing DNA[98], because for the first time in human history, the genetic code

was readable. This has had profound implications for evolutionary biology, as it is now possible to reconstruct the evolutionary history of entire kingdoms of life from DNA sequence. It has also revolutionized our understanding of genetic conflict. The first full length genomes ever sequenced were the MS2 viral RNA genomes and the bacteriophage ØX174 genome[98,99]. The explosion of sequenced genomes has led to a corresponding explosion in the number of genetic parasites and defense systems against those parasites[100].

While genomic information provides the blueprint for a biological entity, structural techniques are critical for understanding the structure and function of the machines encoded by nucleic acid instructions. Here again, viruses have played a critical role in developing a new way of seeing. Because most viruses are not visible in the visible light spectrum, the development of new microscopy techniques was necessary to understand the physical basis for viral infections. Some of the first electron micrographs ever recorded were images of viral infections of bacteria[101] and the first three-dimensional reconstruction from electron microscopy data was the reconstruction of a T4 bacteriophage tail[102].

## Dissertation Overview

This dissertation examines both sequence and structure to understand the evolutionary outcomes of genetic conflict. The first three chapters deal with the assembly and annotation of defense systems, genetic parasites, and domesticated genes in two complete, telomere-to-telomere eukaryotic genomes. The fourth chapter begins to explore the functions beyond defense of domesticated genes in placental mammals. The final three chapters shift into the prokaryotic realm and use a combination of bioinformatics and cryogenic electron microscopy to study the ongoing molecular arms race in bacteria.

**Chapter 2** presents telomere-to-telomere sequence of the genome of the algae SLA-04 derived from long-read sequencing technology. Algae are a broad class of photosynthetic eukaryotes that are phylogenetically and physiologically diverse. Most of the phylogenetic diversity has been inferred from 18S rDNA sequencing since there are only a few complete genomes available in public databases. We used ultra-long-read Nanopore sequencing to determine a gapless, telomere-to-telomere complete genome sequence of *Chlorella sp.* SLA-04, previously described as *Chlorella sorokiniana* SLA-04. *Chlorella sp.* SLA-04 is a green alga that grows to high cell density in a wide variety of environments – high and neutral pH, high and low alkalinity, and high and low salinity. SLA-04's ability to grow in high pH and high alkalinity media without external $CO_2$ supply is favorable for large-scale algal biomass production. Phylogenetic analysis performed using ribosomal DNA and conserved protein sequences consistently reveal that *Chlorella sp.* SLA-04 forms a distinct lineage from other strains of *Chlorella sorokiniana*. We complement traditional genome annotation methods with high throughput structural predictions and demonstrate that this approach expands functional prediction of the SLA-04 proteome. Genomic analysis of the SLA-04 genome identifies the genes capable of utilizing TCA cycle intermediates to replenish cytosolic acetyl-CoA pools for lipid production. We also identify a complete metabolic pathway for sphingolipid anabolism that may allow SLA-04 to readily adapt to changing environmental conditions and facilitate robust cultivation in mass production systems. Collectively, this work clarifies the phylogeny of *Chlorella sp*. SLA-04 within Trebouxiophyceae and demonstrates how structural predictions can be used to improve annotation beyond sequence-based methods.

**Chapter 3** presents an analysis of the first human telomere-to-telomere genome in search of CRISPR-Cas adaptive immune systems. CRISPR-Cas systems are found in about 40% of bacterial and 85% of archaeal genomes, but not to date in eukaryotic genomes. Recently, a paper published in *Communications Biology* reported the identification of 12,572 putative CRISPRs in the human genome, which they call "hCRISPR". We attempted to reproduce this analysis using the telomere-to-telomere human genome and show that repetitive elements identified as putative CRISPR loci contain neither the repeat-spacer-repeat architecture nor the *cas* genes characteristic of functional CRISPR systems.

**Chapter 4** presents an in-depth evolutionary analysis of viral genes domesticated for host function in placental mammals. Using a complete human genome assembly and 25 additional vertebrate genomes, we re-analyzed the evolutionary trajectories and functional potential of capsid genes domesticated from *Metaviridae*, a lineage of retrovirus-like retrotransposons. Our study expands on previous analyses to unearth several new insights about the evolutionary histories of these ancient genes. We find that at least five independent domestication events occurred from diverse *Metaviridae*, giving rise to three universally-retained single-copy genes evolving under purifying selection and two gene families unique to placental mammals with multiple members showing evidence of rapid evolution. In the *SIRH/RTL* family, we find diverse amino-terminal domains, widespread loss of protein-coding capacity in *RTL10* despite its retention in several mammalian lineages, and differential utilization of an ancient programmed ribosomal frameshift in *RTL3* between the domesticated capsid and protease domains. Our analyses also reveal that most members of the *PNMA* family in mammalian genomes encode a conserved putative amino-terminal RNA-binding domain both adjoining and independent from

domesticated capsid domains. Our analyses lead to a significant correction of previous annotations of the essential *CCDC8* gene. We show that this putative RNA-binding domain is also present in several extant *Metaviridae,* revealing a novel protein domain configuration in retrotransposons. Collectively, our study reveals the divergent outcomes of multiple domestication events from diverse *Metaviridae* in the common ancestor of placental mammals.

**Chapter 5** demonstrates that two PNMA family genes, PNMA1 and PNMA4, support reproductive capacity during aging. Analysis of donated human ovaries shows that expression of both genes declines normally with age, while several PNMA1 and PNMA4 variants identified in genome-wide association studies are causally associated with low testosterone, altered puberty onset, or obesity. Six-week-old mice lacking either Pnma1 or Pnma4 are indistinguishable from wild-type littermates, but by six months the mutant mice become prematurely subfertile, with precipitous drops in sex hormone levels, gonadal atrophy, and abdominal obesity; overall they produce markedly fewer offspring than controls. These findings expand our understanding of factors that maintain human reproductive health and lend insight into the functional consequence of the domestication of retrotransposon-derived genes.

**Chapter 6,** explains how viruses compete for limited cellular resources by delivering defense mechanisms that protect the host from competing genetic parasites[1]. PARIS is a defense system, often encoded in viral genomes, that is composed of a 55 kDa ABC ATPase (AriA) and a 35 kDa TOPRIM nuclease (AriB)[2]. However, the mechanism by which AriA and AriB function in phage defense is unknown. Here we show that AriA and AriB assemble into a 425 kDa supramolecular immune complex. We use cryo-EM to determine the structure of this complex which explains how six molecules of AriA assemble into a propeller-shaped scaffold that

coordinates three subunits of AriB. ATP-dependent detection of foreign proteins triggers the release of AriB, which assembles into a homodimeric nuclease that blocks infection by cleaving host $tRNA^{Lys}$. Phage T5 subverts PARIS immunity through expression of a $tRNA^{Lys}$ variant that is not cleaved by PARIS, and thereby restores viral infection. Collectively, these data explain how AriA functions as an ATP-dependent sensor that detects viral proteins and activates the AriB toxin. PARIS is one of an emerging set of immune systems that form macromolecular complexes for the recognition of foreign proteins, rather than foreign nucleic acids[3].

**Chapter 7** describes how bacteria and archaea acquire resistance to viruses and plasmids by integrating fragments of foreign DNA into the first repeat of a CRISPR array. The mechanism of site-specific integration in diverse CRISPR systems remains poorly understood. Here, we determine a 560 kDa integration complex structure that explains how *Pseudomonas aeruginosa* Cas (Cas1-2/3) and non-Cas proteins (integration host factor) fold 150 base-pairs of host DNA into a U-shaped bend and a loop that protrude from Cas1-2/3 at right angles. The U-shaped bend traps foreign DNA on one face of the Cas1-2/3 integrase, while the loop places the first CRISPR repeat in the Cas1 active site. Both Cas3s rotate 100-degrees to expose DNA binding sites on either side of the Cas2 homodimer, that each bind an inverted repeat motif in the leader. Leader sequence motifs direct Cas1-2/3-mediated integration to diverse repeat sequences that have a 5'-GT. Collectively, this work reveals new DNA binding surfaces on Cas2 that are critical for DNA folding, and site-specific delivery of foreign DNA.

In **Chapter 8**, I expand on the work of Chapter 7 by purifying the Cas1-2/3 fusion of *Pseudomonas aeruginosa* and determining multiple cryoEM structures at distinct stages of CRISPR adaptation. Collectively, these structures reveal a prominent channel on one face of the

integration complex that captures short fragments of foreign DNA. Foreign DNA binding triggers a series of conformational changes that reposition Cas3, exposing the Cas1 active sites and new DNA binding surfaces that are necessary for homing the DNA-bound integrase to the CRISPR locus. Upon docking to the CRISPR, the integrase catalyzes two sequential transesterification reactions that covalently link a discrete fragment of foreign dsDNA to the first repeat of the CRISPR locus. Taken together, these cryoEM structures clarify how the Cas1/2-3 fusion proteins orchestrate foreign DNA capture, site specific delivery, and integration of new DNA into the bacterial genome.

## References


1. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).

2. Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proceedings of the National Academy of Sciences* **96**, 2192–2197 (1999).

3. Torsvik, T. & Dundas, I. D. Bacteriophage of Halobacterium salinarium. *Nature* **248**, 680–681 (1974).

4. d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysentétriques. *Comptes rendus de l'Académie des sciences* **165**, 373–375 (1917).

5. Beijerinck, M. W. On a Contagium vivum fluidum causing the Spotdisease of the Tobacco-leaves. *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings Series B Physical Sciences* **1**, 170–176 (1898).

6. Suttle, C. A. The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol* **28**, 237–243 (1994).

7. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

8. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

9. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).

10. Brazelton, W. J. & Baross, J. A. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *The ISME Journal* **3**, 1420–1424 (2009).

11. Shao, C. *et al.* The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. *Cell* **186**, 1279-1294.e19 (2023).

12. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res* **8**, 464–478 (1998).

13. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Research* **38**, 4207–4217 (2010).

14. Culbertson, E. M. & Levin, T. C. Eukaryotic CD-NTase, STING, and viperin proteins evolved via domain shuffling, horizontal transfer, and ancient inheritance from prokaryotes. *PLOS Biology* **21**, e3002436 (2023).

15. Wein, T. & Sorek, R. Bacterial origins of human cell-autonomous innate immune mechanisms. *Nat Rev Immunol* **22**, 629–638 (2022).

16. Ledvina, H. E. & Whiteley, A. T. Conservation and similarity of bacterial and eukaryotic innate immunity. *Nat Rev Microbiol* **22**, 420–434 (2024).

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

18. Wu, J. & Chen, Z. J. Innate Immune Sensing and Signaling of Cytosolic Nucleic Acids. *Annual Review of Immunology* **32**, 461–488 (2014).

19. Wu, J. *et al.* Cyclic GMP-AMP Is an Endogenous Second Messenger in Innate Immune Signaling by Cytosolic DNA. *Science* **339**, 826–830 (2013).

20. Bernheim, A. *et al.* Prokaryotic viperins produce diverse antiviral molecules. *Nature* **589**, 120–124 (2021).

21. Rivera-Serrano, E. E. *et al.* Viperin Reveals Its True Function. *Annual Review of Virology* **7**, 421–446 (2020).

22. Broz, P., Pelegrín, P. & Shao, F. The gasdermins, a protein family executing cell death and inflammation. *Nat Rev Immunol* **20**, 143–157 (2020).

23. Johnson, A. G. *et al.* Bacterial gasdermins reveal an ancient mechanism of cell death. *Science* **375**, 221–225 (2022).

24. Zaremba, M. *et al.* Short prokaryotic Argonautes provide defence against incoming mobile genetic elements through NAD+ depletion. *Nat Microbiol* **7**, 1857–1869 (2022).

25. Anzelon, T. A. *et al.* Structural basis for piRNA targeting. *Nature* **597**, 285–289 (2021).

26. Ofir, G. *et al.* Antiviral activity of bacterial TIR domains via immune signalling molecules. *Nature* **600**, 116–120 (2021).

27. Bayless, A. M. *et al.* Plant and prokaryotic TIR domains generate distinct cyclic ADPR NADase products. *Science Advances* **9**, eade8487 (2023).

28. Lane, N. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philos Trans R Soc Lond B Biol Sci* **370**, 20140344 (2015).

29. Koonin, E. V. Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol Direct* **12**, 5 (2017).

30. Dimitriu, T., Szczelkun, M. D. & Westra, E. R. Evolutionary Ecology and Interplay of Prokaryotic Innate and Adaptive Immune Systems. *Current Biology* **30**, R1189–R1202 (2020).

31. Arber, W., Hattman, S. & Dussoix, D. On the host-controlled modification of bacteriophage λ. *Virology* **21**, 30–35 (1963).

32. Arber, W. Host specificity of DNA produced by *Escherichia coli*. *Journal of Molecular Biology* **11**, 247–256 (1965).

33. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A., Wilson, G. G. & Murray, N. E. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Research* **42**, 3–19 (2014).

34. Bickle, T. A. & Krüger, D. H. Biology of DNA restriction. *Microbiol Rev* **57**, 434–450 (1993).

35. Elde, N. C., Child, S. J., Geballe, A. P. & Malik, H. S. Protein kinase R reveals an evolutionary model for defeating viral mimicry. *Nature* **457**, 485–489 (2009).

36. Elde, N. C. & Malik, H. S. The evolutionary conundrum of pathogen mimicry. *Nature Reviews Microbiology* **7**, 787–797 (2009).

37. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nature Reviews Genetics* **21**, 119–131 (2020).

38. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol* **18**, 113–119 (2020).

39. Krüger, D. H., Reuter, M., Hansen, S. & Schroeder, C. Influence of phage T3 and T7 gene functions on a type III (EcoP1) DNA restriction-modification system in vivo. *Mol Gen Genet* **185**, 457–461 (1982).

40. Lee, H. & Sashital, D. G. Creating memories: molecular mechanisms of CRISPR adaptation. *Trends in Biochemical Sciences* **47**, 464–476 (2022).

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

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

43. Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, eaal5056 (2017).

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

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

46. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561 (2005).

47. Mojica, F. J. M., Díez-Villaseñor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology* **36**, 244–246 (2000).

48. Jansen, Ruud., Embden, Jan. D. A. van, Gaastra, Wim. & Schouls, Leo. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* **43**, 1565–1575 (2002).

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

50. Brouns, S. J. J. *et al.* Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* **321**, 960–964 (2008).

51. Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D. & Koonin, E. V. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biology* **12**, 36 (2014).

52. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLOS Computational Biology* **1**, e60 (2005).

53. Nuñez, J. K., Lee, A. S. Y., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature* **519**, 193–198 (2015).

54. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat Rev Genet* **16**, 184–192 (2015).

55. Krupovic, M., Béguin, P. & Koonin, E. V. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. *Current Opinion in Microbiology* **38**, 36–43 (2017).

56. Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**, 1113–1118 (2017).

57. Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P. & Wiedenheft, B. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. *Current Biology* **31**, 3515-3524.e6 (2021).

58. Fagerlund, R. D. *et al.* Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. *Proceedings of the National Academy of Sciences* **114**, E5122–E5128 (2017).

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

60. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet* **11**, 47–59 (2010).

61. Liu, C., Zhang, Y., Liu, C. C. & Schatz, D. G. Structural insights into the evolution of the RAG recombinase. *Nat Rev Immunol* **22**, 353–370 (2022).

62. Krupovic, M. *et al.* Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. *Journal of Virology* **92**, e00515-18 (2018).

63. Finnegan, D. J. Retrotransposons. *Curr. Biol.* **22**, R432-437 (2012).

64. Burns, K. H. & Boeke, J. D. Human Transposon Tectonics. *Cell* **149**, 740–752 (2012).

65. Dodonova, S. O., Prinz, S., Bilanchone, V., Sandmeyer, S. & Briggs, J. A. G. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *PNAS* **116**, 10048–10057 (2019).

66. Doolittle, R. F., Johnson, M. S. & McClure, M. A. Origins and Evolutionary Relationships of Retroviruses. *The Quarterly Review of Biology* **64**, 1–30 (1989).

67. Hayward, A. Origin of the retroviruses: when, where, and how? *Current Opinion in Virology* **25**, 23–27 (2017).

68. Ashkenazi, A., Faingold, O. & Shai, Y. HIV-1 fusion protein exerts complex immunosuppressive effects. *Trends in Biochemical Sciences* **38**, 345–349 (2013).

69. Ikeda, H., Laigret, F., Martin, M. A. & Repaske, R. Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. *J Virol* **55**, 768–777 (1985).

70. Taylor, G. M., Gao, Y. & Sanders, D. A. Fv-4: identification of the defect in Env and the mechanism of resistance to ecotropic murine leukemia virus. *J Virol* **75**, 11244–11248 (2001).

71. Frank, J. A. *et al.* Evolution and antiviral activity of a human protein of retroviral origin. *Science* **378**, 422–428 (2022).

72. Varela, M., Spencer, T. E., Palmarini, M. & Arnaud, F. Friendly Viruses. *Annals of the New York Academy of Sciences* **1178**, 157–172 (2009).

73. Ito, J. *et al.* Refrex-1, a Soluble Restriction Factor against Feline Endogenous and Exogenous Retroviruses. *Journal of Virology* **87**, 12029–12040 (2013).

74. Ito, J., Baba, T., Kawasaki, J. & Nishigaki, K. Ancestral Mutations Acquired in Refrex-1, a Restriction Factor against Feline Retroviruses, during its Cooption and Domestication. *Journal of Virology* **90**, 1470–1485 (2016).

75. Robinson, H. L., Astrin, S. M., Senior, A. M. & Salazar, F. H. Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. *Journal of Virology* **40**, 745–751 (1981).

76. Kim, F. J., Battini, J.-L., Manel, N. & Sitbon, M. Emergence of vertebrate retroviruses and envelope capture. *Virology* **318**, 183–191 (2004).

77. Dupressoir, A., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* **33**, 663–671 (2012).

78. Lavialle, C. *et al.* Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120507 (2013).

79. Cornelis, G. *et al.* An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proceedings of the National Academy of Sciences* **114**, E10991–E11000 (2017).

80. Henzy, J. E., Gifford, R. J., Kenaley, C. P. & Johnson, W. E. An Intact Retroviral Gene Conserved in Spiny-Rayed Fishes for over 100 My. *Molecular Biology and Evolution* **34**, 634–639 (2017).

81. Olson, E. D. & Musier-Forsyth, K. Retroviral Gag protein - RNA interactions: Implications for specific genomic RNA packaging and virion assembly. *Semin Cell Dev Biol* **86**, 129–139 (2019).

82. Bénit, L. *et al.* Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J Virol* **71**, 5652–5657 (1997).

83. Yap, M. W., Colbeck, E., Ellis, S. A. & Stoye, J. P. Evolution of the Retroviral Restriction Gene Fv1: Inhibition of Non-MLV Retroviruses. *PLOS Pathogens* **10**, e1003968 (2014).

84. Young, G. R., Yap, M. W., Michaux, J. R., Steppan, S. J. & Stoye, J. P. Evolutionary journey of the retroviral restriction gene Fv1. *PNAS* **115**, 10130–10135 (2018).

85. Brandt, J., Veith, A.-M. & Volff, J.-N. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. *Cytogenet. Genome Res.* **110**, 307–317 (2005b).

86. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends in Genetics* **22**, 585–589 (2006).

87. Ono, R. *et al.* Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet* **38**, 101–106 (2006).

88. Sekita, Y. *et al.* Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nat Genet* **40**, 243–248 (2008).

89. Kokošar, J. & Kordiš, D. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol Biol Evol* **30**, 1015–1031 (2013).

90. Pastuzyn, E. D. *et al.* The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **172**, 275-288.e18 (2018).

91. Abed, M. *et al.* The Gag protein PEG10 binds to RNA and regulates trophoblast stem cell lineage specification. *PLOS ONE* **14**, e0214110 (2019).

92. Erlendsson, S. *et al.* Structures of virus-like capsids formed by the Drosophila neuronal Arc proteins. *Nat Neurosci* 1–4 (2020) doi:10.1038/s41593-019-0569-y.

93. Segel, M. *et al.* Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science* **373**, 882–889 (2021).

94. Xu, J. *et al.* PNMA2 forms immunogenic non-enveloped virus-like capsids associated with paraneoplastic neurological syndrome. *Cell* **187**, 831-845.e19 (2024).

95. Nikolaienko, O., Patil, S., Eriksen, M. S. & Bramham, C. R. Arc protein: a flexible hub for synaptic plasticity and cognition. *Seminars in Cell & Developmental Biology* **77**, 33–42 (2018).

96. Ashley, J. *et al.* Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172**, 262-274.e11 (2018).

97. Brookfield, J. F. Y. The ecology of the genome — mobile DNA elements and their hosts. *Nat Rev Genet* **6**, 128–136 (2005).

98. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

99. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).

100. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).

101. Luria, S. E. & Anderson, T. F. The Identification and Characterization of Bacteriophages with the Electron Microscope. *Proc Natl Acad Sci U S A* **28**, 127-130.1 (1942).

102. De Rosier, D. J. & Klug, A. Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature* **217**, 130–134 (1968).

# CHAPTER TWO

# GENOME SEQUENCE, PHYLOGENETIC ANALYSIS, AND STRUCTURE BASED ANNOTATION REVEALS METABOLIC POTENTIAL OF *CHLORELLA SP.* SLA-04

## Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Calvin L.C. Goemann

Contributions: Conceptualization, methodology, investigation & data collection, genomics & bioinformatics analysis, writing – original draft, writing – review & editing

Co-Author: Royce Wilkinson

Contributions: Conceptualization, methodology, investigation & data collection, writing – review & editing

Co-Author: William Henriques

Contributions: Genomics & bioinformatics analysis, writing – review & editing

Co-Author: Huyen Bui

Contributions: Genomics & bioinformatics analysis, writing – review & editing

Co-Author: Hannah M. Goemann

Contributions: Investigation & data collection, genomics & bioinformatics analysis, writing – review & editing

Co-Author: Ross P. Carlson

Contributions: Conceptualization, writing – review & editing

Co-Author: Srihar Viamajala

Contributions: Conceptualization, writing – review & editing

Co-Author: Robin Gerlach

Contributions: Conceptualization, writing – review & editing

Co-Author: Blake Wiedenheft

Contributions: Conceptualization, methodology, writing – original draft, writing – review & editing

Manuscript Information

Calvin L. C. Goemann, Royce Wilkinson, William Henriques, Huyen Bui, Hannah M. Goemann, Ross P. Carlson, Sridhar Viamajala, Robin Gerlach, Blake Wiedenheft

*Algal Research*

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

Elsevier
Volume 69
January 2023

Submitted: 17 February 2022
Published online: 13 December 2022
https://doi.org/10.1016/j.algal.2022.102943

## Abstract

Algae are a broad class of photosynthetic eukaryotes that are phylogenetically and physiologically diverse. Most of the phylogenetic diversity has been inferred from 18S rDNA sequencing since there are only a few complete genomes available in public databases. Here we use ultra-long-read Nanopore sequencing to determine a gapless, telomere-to-telomere complete genome sequence of *Chlorella sp.* SLA-04, previously described as *Chlorella sorokiniana* SLA-04. *Chlorella sp.* SLA-04 is a green alga that grows to high cell density in a wide variety of environments – high and neutral pH, high and low alkalinity, and high and low salinity. SLA-04's ability to grow in high pH and high alkalinity media without external $CO_2$ supply is favorable for large-scale algal biomass production. Phylogenetic analysis performed using ribosomal DNA and conserved protein sequences consistently reveal that *Chlorella sp.* SLA-04 forms a distinct lineage from other strains of *Chlorella sorokiniana*. We complement traditional genome annotation methods with high throughput structural predictions and demonstrate that this approach expands functional prediction of the SLA-04 proteome. Genomic analysis of the SLA-04 genome identifies the genes capable of utilizing TCA cycle intermediates to replenish cytosolic acetyl-CoA pools for lipid production. We also identify a complete metabolic pathway for sphingolipid anabolism that may allow SLA-04 to readily adapt to changing environmental conditions and facilitate robust cultivation in mass production systems. Collectively, this work clarifies the phylogeny of *Chlorella sp*. SLA-04 within Trebouxiophyceae and demonstrates how structural predictions can be used to improve annotation beyond sequence-based methods.

## Introduction

Algae are diverse photosynthetic eukaryotes that assimilate six times more nitrogen than terrestrial plants, perform 45% of global oxygenic photosynthesis, and 50% of total $CO_2$ fixation[1–4]. The phylogenetic and physiological diversity of algae is mirrored by the diversity of aquatic ecosystems where algae are found (e.g., temperatures 0-56ºC, pH 0-11.5, and salt concentrations 0.02-3M)[5,6]. However, it is currently unclear how selective pressures from disparate aquatic environments shape algal genotypes and corresponding phenotypes.

Genomic analysis of algae across different environments is essential to understand how algae survive and adapt to changing conditions. As of September 2022, there are 170 green algae genomes in the National Center for Biotechnology Information (NCBI) database. Most of these genomes are highly fragmented and only 16 contain defined chromosome sequences. This resource limitation restricts efforts to confidently identify genotypes that drive functional diversity. Understanding the phylogenetic and functional diversity of algae will benefit from a broader collection of complete algal genomes that can be used for comparative analyses.

To date, most algal genomes have been sequenced using short-read sequencing platforms[7,8]. While these platforms are accurate, PCR bias and short read lengths frequently result in incomplete assemblies and omission of repetitive regions, such as retroelements that can represent up to 90% of a eukaryotic genome[9–12]. In contrast, ultra-long-read sequencing platforms generate megabase-length reads that span repetitive regions, and these methods do not rely on PCR[13,14]. While the error rates for long-read technologies are higher, consensus correction software significantly increases accuracy as a function of read depth[15,16]. Long-read sequence data with >100x coverage often results in highly accurate chromosome level assemblies[17].

Here we determine a complete genome sequence of *Chlorella. sp.* SLA-04, previously described as *Chlorella sorokiniana* SLA-04, using ultra-long-read Nanopore sequencing. *Chlorella sp.* SLA-04 is an oleaginous green alga that was isolated from a brackish, alkaline lake (100mM $HCO_3^-$, pH 10) in eastern Washington (USA)[18]. SLA-04 has growth and composition characteristics that are attractive for biofuel production. Phylogenetic analysis performed using either ribosomal DNA or conserved protein sequences reveal that SLA-04 forms a distinct lineage from other strains of the same species. While sequenced-based annotation methods confidently assign putative function to a majority of the SLA-04 genes, 44.2% of the genes were classified as "hypothetical". Using a high-throughput structure prediction software (i.e., Alphafold) we predict the 3D protein structures encoded by the "hypothetical" proteome. Structural predictions are queried against the protein data bank (PDB) to identify functional homologs and expand functional predictions for 677 SLA-04 genes that are widely conserved in algae and a subset of genes that are unique to *Chlorella sp.* SLA-04. Collectively, this work expands our phylogenetic and functional understanding of algae.

## Methods

### Strain Information

Strain Cultivation Conditions Chlorella sp. SLA-04 cultures were grown in Modified Bold's Basal Medium pH 8.7 (Bolds) containing NaNO3 (2.94 mM), K2HPO4 (1.43 mM), KH2PO4 (1.43 mM), MgSO4·7H2O (0.30 mM), CaCl2·2H2O (0.17 mM), NaCl (0.42 mM), 1 mL/L of EDTA solution, 1 mL/L iron solution, and 2 mL/L trace metal solution. The EDTA solution contained 50 g/L Na2EDTA and 31 g/L KOH. The iron solution contained 4.98 g/L FeSO4·7H2O and 1 mL/L concentrated H2SO4. The trace metal solution contained H3BO3 (9.7

mM), MnCl2·4H2O (1.26 mM), ZnCl2 (0.15 mM), CuCl2·2H2O (0.11 mM), Na2MoO4·2H2O (0.07 mM), CoCl2·6H2O (0.06 mM), NiCl2·6H2O (0.04 mM), V2O5 (0.01 mM) and KBr (0.08 mM)19. Media was titrated to pH 8.7 with NaOH. Cultures were grown at 23oC with 400 µmol·m-2·s-1 light intensity cycling with a 14:10 hour light/dark cycle using T5 fluorescent tubes.

Axenic Isolation An axenic culture of Chlorella sp. SLA-04 was generated through antibiotic treatment modified from Mustapa et al.20. Briefly, clonal isolates of SLA-04 were streaked onto plates containing Bolds agar with 700 µg/mL ampicillin and 200 µg/mL cefotaxime. Individual colonies were picked and grown in liquid Bolds. Plate antibiotic treatment and isolation was repeated a second time and algal colonies were picked and grown in liquid Bolds with 10 µg/mL tetracycline. Axenic cultures were confirmed using amplicon sequencing of the V4 region of the 16S rRNA on an Illumina MiSeq platform (Sup. Fig. 2.1)21. A small percentage of 16S rRNA reads within the axenic culture mapped to Ralstonia species, but Ralstonia sequences are common contaminants of DNA purification columns22. Axenic stock cultures of SLA-04 were maintained on agar slants and in liquid culture of Bolds with 10 µg/mL tetracycline.

Genome Sequencing and Assembly

High Molecular Weight DNA Extraction Chlorella sp. SLA-04 was grown in Bolds to log phase. Cells from 100 mL of culture were spun down at 3,000 x g for 10 minutes and transferred to a liquid N2 chilled mortar. The pellet was ground for 30 minutes with liquid N2 added as needed. Ground algal biomass was transferred to 15 mL of lysis buffer (20 mM EDTA pH 8.25, 10 mM Tris, 500 mM guanidine HCl, 200 mM NaCl, 1% w/w Triton X-100, and 9 mg

hemicellulase) and incubated at 37oC for 1 hour. To remove RNA, RNase A was added to a final concentration of 20 µg/mL and the mixture was incubated for an additional 30 minutes at 37°C. To remove proteins, Proteinase K (12 µg) was added and the mixture was incubated at 50°C for 2 hours. Samples were pelleted at 15,000 x g at 4°C for 20 minutes and the supernatant was added to an equilibrated Qiagen Genomic-tip 100/G gravity column (Cat. No.10243). DNA was washed and eluted per the manufacturer's instructions. DNA was precipitated with 3.5 mL molecular grade isopropyl alcohol (IPA). The DNA pellet was collected by centrifugation (5850 x g) at 4°C for 30 minutes and resuspended in molecular grade water.

RNA Extraction *Chlorella sp.* SLA-04 RNA was extracted using the Qiagen RNeasy powerplant RNA isolation kit (Cat. No.13500-50) with modifications. SLA-04 culture was grown to log phase ($2x10^7$ cells/mL) in Bolds. Cells from 200 mL of culture were pelleted at 3,000 x g for 5 minutes at 4°C and the pellet was ground for 10 minutes in a liquid $N_2$ chilled mortar using a pestle before being transferred to 1 mL Trizol in a PowerBead tube provided in the Qiagen RNeasy powerplant kit. Potassium acetate (1 M) was added to a final concentration of 0.2 M to improve polysaccharide precipitation. Phenolic separation solution (50 µL) from the Qiagen powerplant kit was added. The sample was shaken at 6.5 meters per second for 60 seconds using a MP Biomedicals FastPrep-24 Classic bead beater, incubated at room temp for 20 minutes, and centrifuged at 13,000 x g for 10 minutes at 4°C. Protein and cell debris was removed by adding 200 µL chloroform, shaking for 30 seconds, centrifuging at 13,000 x g for 10 minutes at 4°C and transferring the aqueous phase to a new tube. RNA was extracted and purified from the aqueous phase using the RNeasy powerplant RNA isolation kit per

manufacturer's instructions starting at the inhibitor removal step (step 6). All glassware was washed with 2% diethylpyrocarbonate (DEPC) in water (v/v) and autoclaved before use.

DNA and RNA Sequencing The *Chlorella sp.* SLA-04 genome was sequenced using Nanopore sequencing. High molecular weight DNA was sequenced using a MinION sequencer and the ligation sequencing kit (SQK-LSK109 by Oxford Nanopore) per manufacturer's instructions. Briefly, 1 µg of genomic DNA was used for library preparation that included end preparation and ligation to Nanopore adapters. 350 ng of the DNA library was used and sequenced on the MinION sequencer. Sequencing generated 1.97 M reads totaling 5.64 gigabases (Gb) of data.

*Chlorella sp.* SLA-04 mRNA sequencing was performed using a MinION sequencer and the direct RNA sequencing kit (SQK-RNA002 by Oxford Nanopore) per manufacturer's instructions. mRNA was isolated and purified from total RNA using magnetic Dynabeads coated with Oligo$(dT)_{25}$ adapters. Purification using magnetic beads yielded 338 ng of mRNA from 22 µg of total RNA. 338 ng of mRNA was used for Nanopore library preparation that included reverse transcription of cDNA scaffold and ligation of Nanopore adapters. 168 ng of mRNA library was used and sequenced on the MinION Nanopore sequencer. Sequencing generated 1.12 M reads totaling 1.35 Gb of data.

Genome Assembly and Annotation Nanopore raw DNA reads (5.64 GB) were basecalled and analyzed for quality using MinKNOW 4.1.22 software in fast basecalling mode. Quality reads were submitted to CANU v1.9, FLYE v2.5-g0c3de5b, and Miniasm v0.3-r179 programs for long-read genome assembly[23–25]. CANU, FLYE, and Miniasm assembly pipelines each generated contig level assemblies with minor gaps. Manual curation of assemblies was

performed by aligning contigs from the three assembly pipelines using minimap2 v2.17-r954-dirty[26]. Each gap within the individual assemblies was completely bridged by contigs from at least one other genome assembly. Assemblies were merged to form a chromosome level genome with no gaps. Chromosomes were considered complete if they had flanking telomeric repeats $(CCCTAAA)_n$ on both termini with no gaps. Coverage depth of DNA reads was calculated using samtools v1.13[27].

RepeatMasker v4.0.9 was used to identify repetitive elements in the *Chlorella sp.* SLA-04 genome using Repbase and Dfam as reference databases[28–30]. LTRharvest v1.5.10 was used to identify and fetch the sequences of long-terminal repeats in the SLA-04 genome by performing a six open reading frame (ORF) translation of the genome and scanning each ORF longer than 75 amino acids for homology to LTR retrotransposon protein domains from Pfam[31–33]. To identify active retrotransposons, we extracted 13 sequences that contained *gag* and *pol* domains flanked by long-terminal repeats and aligned them in MAFFT v7.429[34]. To verify LTR retrotransposon families, we aligned the capsid domains of diverse families of LTR retrotransposons from Repbase and those identified in SLA-04. SLA-04 retrotransposons clustered with retrotransposons from Copia retrotransposons found in plants.

Raw mRNA Nanopore reads were mapped to the SLA-04 genome using minimap2 and a consensus sequence and transcript counts were generated using Stringtie v2.0 and Gffread v0.11.6[35,36]. The mRNA consensus data was used to aid the *de novo* annotation prediction software MAKER v3.01.03 to generate a transcriptome of SLA-04[37,38]. Gene functions were assigned by submitting protein sequence to the Kyoto Encyclopedia for Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) and Eggnog mapper webservers with default settings[39,40].

Genome completeness was assessed by BUSCO v5.4.2 using 1519 single copy orthologs found in the Chlorophyta clade[41,42]. A schematic of the SLA-04 genome (**Fig. 2.1**) was generated using the ggplot2 v3.3.3 package within RStudio v1.3.959[43]. GC content was determined using the RStudio seqinr package v4.2-5[44].

Structural-based protein annotation was performed following the ColabFold-FoldSeek-EggNOG (CoFFE) pipeline[45]. Briefly, amino acid sequences were submitted to the ColabFold Alphafold2-batch notebook through google colab with default settings[46–48]. The highest-ranking predicted protein data base (pdb) structure with predicted local distance difference test on the Cα atoms (pLDDT) > 70 was selected for each analyzed protein. PDB structures were submitted to Foldseek v3.915ef7d for structural comparison against the AlphaFold database of protein structures. For each SLA-04 protein we kept the AlphaFold database hit with the greatest bit score. We enforced a e-value cutoff of 1xe-05. Sequences of AlphaFold database best hits of each SLA-04 protein were annotated using EggNOG mapper webserver with default settings.

Comparative Analyses

Phylogenetic Analysis. Phylogenetic analysis was performed on all chromosome level green algal genomes (n=10) (**Sup Fig. 2.3**) and all Trebouxiophyceae genomes (n=25) (**Fig. 2.2**) within the NCBI database. 18S rDNA sequences were identified and extracted from chromosome level algal genomes (n=10) using SSU-align v0.1.1[49,50] [49, 50]. Ribosomal sequences were trimmed with trimal v1.4.rev22 with the –nogap flag, aligned with MAFFT with the –auto flag, and a maximum likelihood phylogenetic tree of the 18S multiple sequence alignment was generated using IQtree v1.6.1 using default settings[51,52]. Phylogeny based on orthologous proteins was performed on all Trebouxiophyceae genomes (n=25) using OrthologR v0.4.0 in

RStudio[53]. Briefly, OrthologR performed a reciprocal best hit blast for orthology inference between algal species. Orthologous sequences were aligned based on amino acid sequence using the Needleman-Wunsch method and codons were aligned using pal2nal[54]. The Comeron's method was used to predict dN/dS estimations. Orthologous protein sequences were aligned with MAFFT and a maximum likelihood phylogenetic tree was generated using IQtree[51,52]. Trees were rendered using the RStudio ggtree v2.0.4 package[55].

## Results

### General Genome Characteristics

Chlorella sp. SLA-04 Genome Assembly Statistics To determine a telomere-to-telomere genome of *Chlorella sp.* SLA-04, high molecular weight DNA was extracted from SLA-04 cultures for long-read sequencing using a MinION sequencer from Oxford Nanopore (**Supplementary Figure 2.1**). Nanopore sequencing resulted in 5.64 Gb of sequence data in a total of 1.97 M reads (**Supplementary Table 2.1**). Sequencing reads were assembled using a combination of FLYE, Miniasm, and CANU[23–25], resulting in a 38.8 Mb complete genome with an average of 117x coverage. The SLA-04 genome consists of 13 linear chromosomes, ranging from 1.20 to 5.25 Mbp in length with an average GC content of 65.87%, and circular mitochondrion and chloroplast sequences (**Figure 2.1, Supplementary Table 2.1**). This chromosomal architecture is consistent with previously published *Chlorella* genomes, which contain 12-16 nuclear chromosomes[56–58].

Genome Completeness Analysis. Genome completeness was evaluated using several distinct but complementary metrics. First, BUSCO was used to calculate the percentage of orthologous Chlorophyta genes within the *Chlorella sp.* SLA-04 genome[42]. The SLA-04 genome

contains 93.4% of Chlorophyta orthologs (n=1418) which surpasses the 90% BUSCO cutoff considered to represent "good" genome assemblies (**Supplementary Table 2.1**)[59]. Second, repetitive elements were identified and mapped to unique chromosomal locations, indicating that this assembly provides sufficient coverage to unambiguously assign chromosomal location to

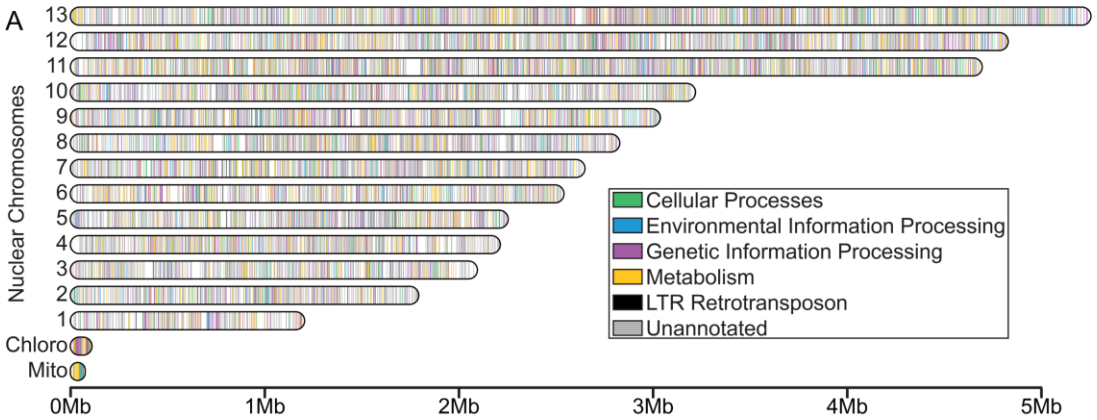



Figure 2.1. Complete genome of Chlorella sp. SLA-04. (A) Chromosome level assembly of Chlorella sp. SLA-04 reveals thirteen nuclear chromosomes, and circular chloroplast and mitochondrial genomes. Genes were annotated with KEGG automatic annotation server (KAAS) and color coded according to six different functional categories.

repetitive elements that have historically frustrated complete genome assemblies (**Figure 2.1, Supplementary Figure 2.2, Supplementary Table 2.2)**[28,60]. Finally, each nuclear chromosome is composed of a single contiguous well-supported sequence flanked by telomeric repeats $(CCCTAAA)_n$ (**Figure 2.1**)[61]. Collectively, these metrics suggest that the SLA-04 genome constitutes a complete telomere-to-telomere algal genome.

Annotation of Chlorella Sp. SLA-04 *Chlorella sp.* SLA-04 mRNA was extracted, and directly sequenced using a MinION sequencer. Sequences were submitted to MAKER v3.01.02 and aligned to the SLA-04 genome to predict protein coding sequences (CDSs) and splice sites[38,62]. The SLA-04 genome contains 10,310 CDSs that span 45.73% of the genome with an

average gene length of 3,475 bp (**Supplementary Table 2.1**). EggNOG mapper was used to assign gene function resulting in 5,757 (55.8%) CDSs with predicted function (**Figure 2.2A**)[40]. Annotated gene functions appear evenly distributed across the nuclear and plastid chromosomes with little functional clustering (**Figure 2.1**). Genes coding for proteins that participate in protein synthesis, amino acid metabolism and chaperones are the most abundant gene categories in the SLA-04 genome (**Figure 2.2**).

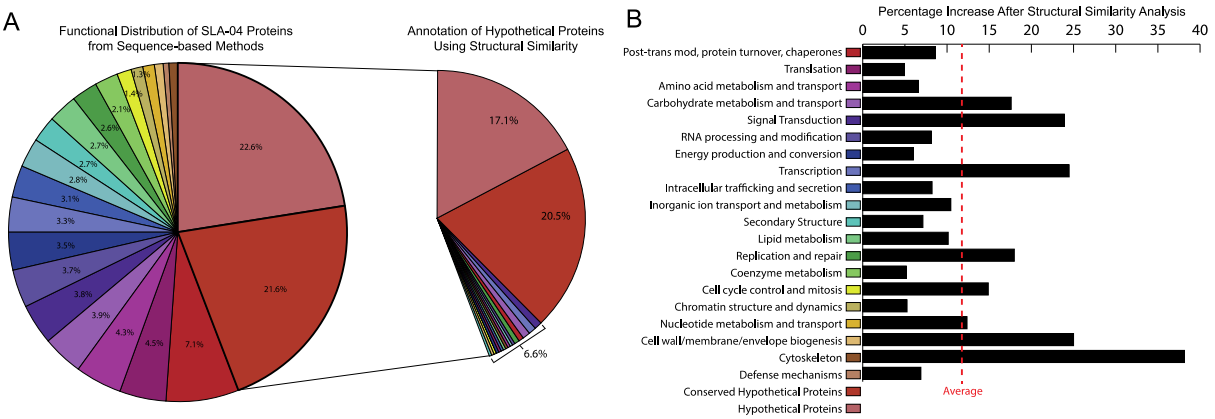



Figure 2.2. Annotation using protein structural similarity improves functional understanding of SLA-04 proteome. A) 10,310 protein sequences from SLA-04 were submitted to Eggnog mapper for annotation using sequence-based methods resulting in 5,757 sequences (55.8%) with predicted function. Remaining sequences (2,226 conserved hypothetical and 2,327 hypothetical proteins) were submitted to Colabfold structural prediction and resulting structures were compared to the Alphafold and Uniprot structural databases using Foldseek to identify structural homologs of hypothetical SLA-04 proteins. Sequences of structural homologs were submitted to Eggnog Mapper resulting in the annotation of an additional 6.6% of the SLA-04 proteome. B) The percent increase in annotation after structural similarity analysis was calculated by dividing the number of proteins annotated using structural prediction methods by the number of original annotated proteins from sequence-based methods. Proteins were grouped based on their Clusters of Orthologous Genes (COG) functionality

Annotation Using Protein 3D Structural Homology The 4,553 (44.2%) CDSs that could not be assigned putative functions using sequence-based annotation, consisted of "hypothetical" and "conserved hypothetical" proteins with unknown functions. Hypothetical proteins are CDSs with no predicted function and no homology to sequences in the EggNOG database, whereas

conserved hypothetical proteins are annotated CDSs that have homology to sequences within the EggNOG database, but the function of these homologous sequences remains unknown. To further improve annotation of the SLA-04 genome, we used the ColabFold-FoldSeek-EggNOG

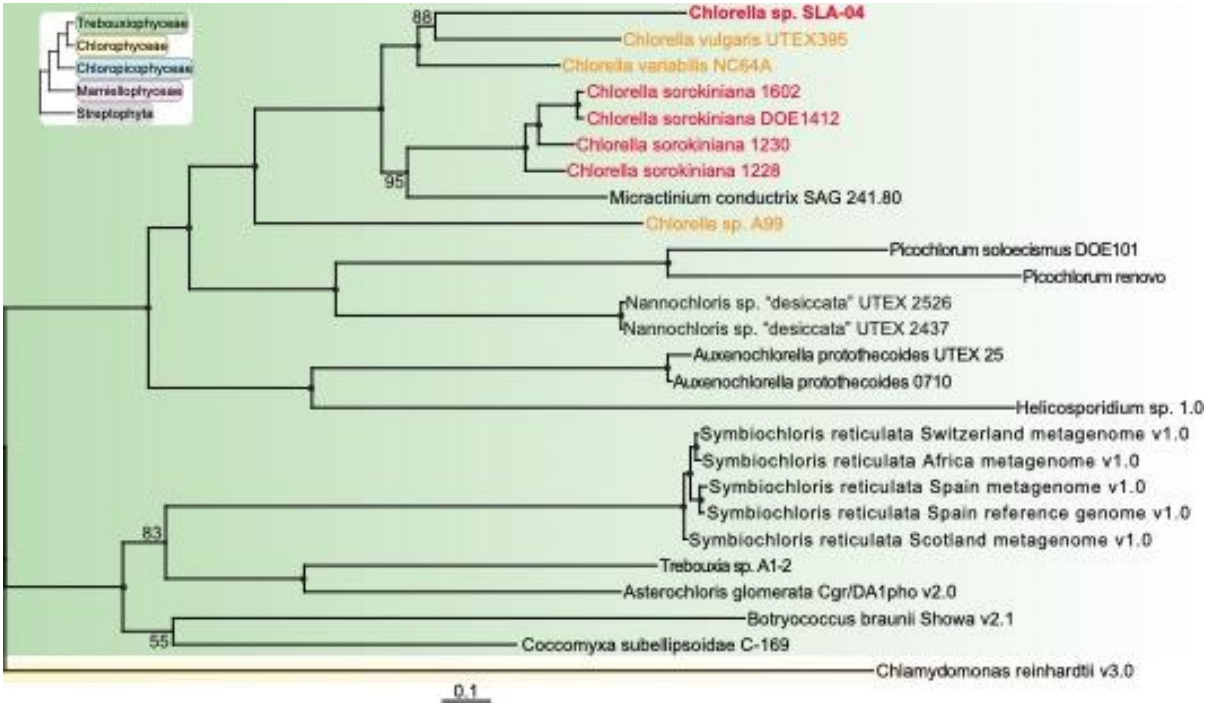



Figure 2.3. Phylogenomic analysis of green algae performed using orthologous genes indicates that SLA-04 is distinct from other Chlorella sorokiniana species. Maximum likelihood phylogenetic analysis based on a concatenation of 215 orthologous proteins sequences of available Trebouxiophyceae genomes from NCBI and PhycoCosm (JGI) databases. Chlamydomonas reinhardtii was used as an outgroup. Tree scale indicates the number of amino acid substitutions per site between samples. Bootstrap values of 100 are represented by black dots on nodes. Bootstrap values less than 100 are provided. Classes of algae are denoted via color on tree branches. Red text denotes Chlorella sorokiniana species. Orange text represents remaining algae classified within the Chlorella genus.

(CoFFE) pipeline to predict the 3D structures of hypothetical and conserved hypothetical proteins to query the protein databank for structural homologs[45]. Consistent with functional importance, 3D structures of proteins are more conserved during evolution than primary amino-acid sequences[63]. Thus, we hypothesized that the integration of structure prediction into genome

annotation would improve our functional understanding of the SLA-04 genome. In fact, structure-based annotation resulted in high-confidence functional annotations for 677 of the CDSs that were previously assigned to the hypothetical and conserved hypothetical categories using Eggnog mapper[40] (**Figure 2.2A**). Structure-based annotation identifies proteins primarily associated with formation of the cytoskeleton and cell wall, transcription, and signal transduction (**Figure 2.2B**).

Phylogenetic Analysis of Chlorella Sp. SLA-04 Initial taxonomic classification of *Chlorella sp.* SLA-04 was based on 5.8S and internal transcribed spacer (ITS) regions of the ribosomal small subunit[18]. ITS regions contain highly variable sequences and can be used to distinguish genera; however, phylogenetic analysis performed using a single gene can limit species level classification and lack statistical support for certain phylogenetic nodes[64–66]. Phylogenomic analysis performed using orthologous proteins or 18S rDNA sequences, reveal that SLA-04 shares a closest common ancestor with *Chlorella vulgaris UTEX 395* (**Figure 2.3, Supplementary Figure 2.3**). Further, the *Chlorell*a genus (n=8) forms a paraphyletic clade interspersed with *Micractinium conductrix SAG 241.80* (**Figure 2.3**). These data suggest that the current taxonomy of Trebouxiophyceae may not capture the true phylogenetic history of these algae and we anticipate that the addition of more complete algal genome sequences will provide new taxonomic perspectives.

Genomic Analysis of Metabolic Pathways in *Chlorella Sp*. SLA-04

To evaluate the metabolic capability encoded within the *Chlorella sp*. SLA-04 genome, we submitted the complete chromosome level genome to the Kegg Automatic Annotation Server (KAAS) and KofamKOALA for metabolic mapping[39,67]. These programs annotate and map

proteins to established metabolic pathways. According to this analysis, the SLA-04 genome contains all genes necessary for core metabolic pathways including photosynthesis, fatty acid synthesis, carbohydrate synthesis, and the citric acid cycle.

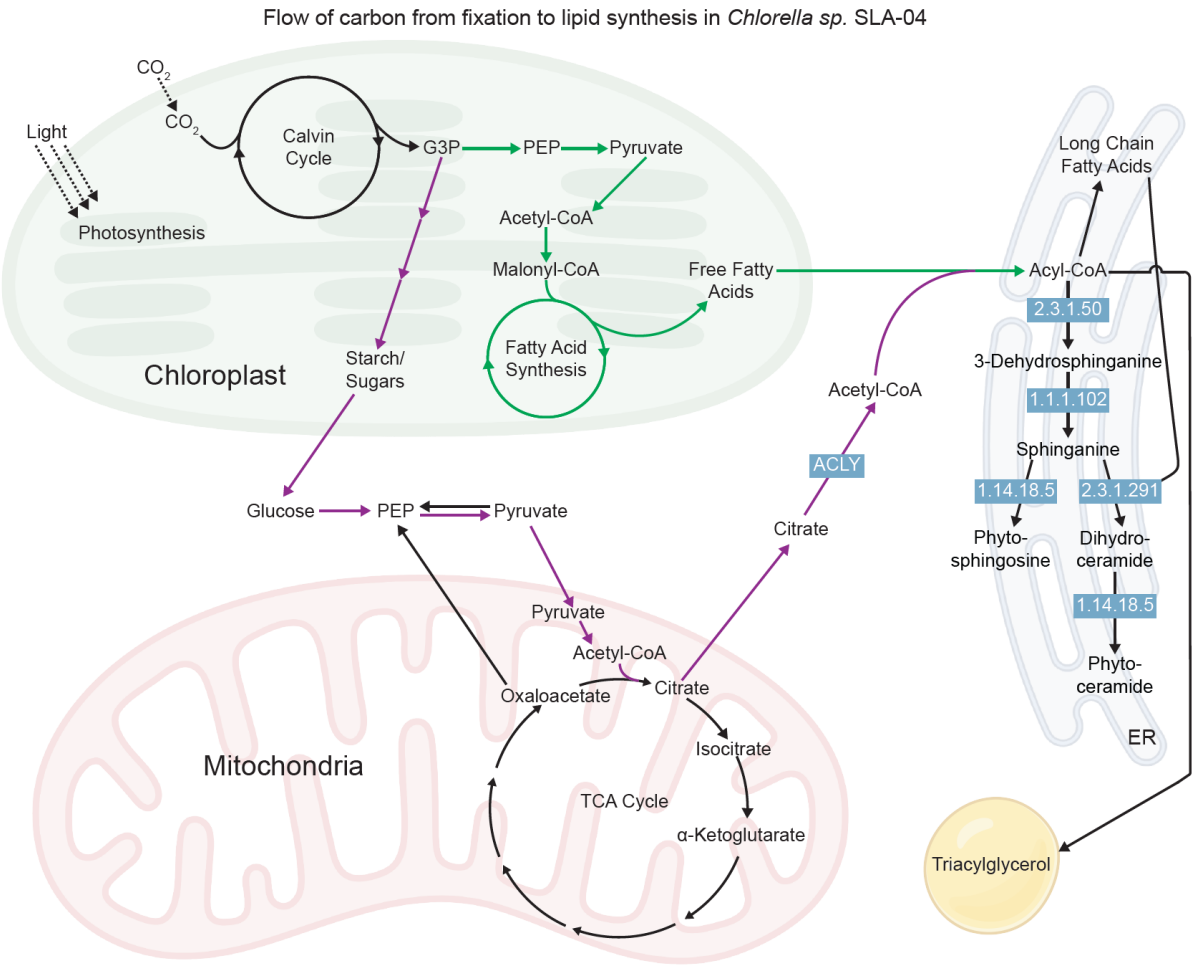



Figure 2.4. Flow of carbon from fixation to lipid synthesis in Chlorella sp. SLA-04. Schematic of carbon flow from fatty acid biosynthesis and TCA cycle to triacylglycerol and sphingolipid biosynthesis in algae. Green arrows represent the flow of carbon from photosynthesis directly to fatty acids within the chloroplast. Purple arrows represent the flow of carbon from photosynthesis to fatty acids through starch synthesis and the TCA cycle. Black arrows represent additional pathways involved in central carbon metabolism. Blue boxes show enzymes of interest. Enzymes are labeled using KEGG and International Union of Biochemistry and Molecular Biology (IUBMB) nomenclature

Genes involved in lipid synthesis are critical for biofuel production and stress response to environmental changes[68–70]. The SLA-04 genome encodes ATP citrate lyase EC 2.3.3.8 (ACLY), involved in replenishing cytosolic acetyl-CoA pools from citrate (**Figure 2.4**). This enzyme is present in some, but not all green algal genomes (**Supplementary Figure 2.4**). Cytosolic acetyl-CoA pools are necessary for fatty acid synthesis within the endoplasmic reticulum. The presence and transcription of ATP citrate lyase ACLY within the SLA-04 genome suggests that this organism can utilize TCA cycle intermediates for fatty acid synthesis in addition to using photosynthate directly (**Supplementary Table 2.3**).

Metabolic network analysis of SLA-04 genes involved in lipid synthesis also reveals the presence and transcription of genes coding for the proteins necessary for sphingolipid biosynthesis (**Figure 2.4, Supplementary Table 2.3**). Similar to ACLY, enzymes involved in sphingolipid synthesis are only present in a subset of green algal genomes (**Supplementary Figure 2.4**). Phytoceramide and phytosphingosine are two sphingolipids synthesized within the endoplasmic reticulum (ER) of land plants and algae that remodel the cell membranes to maintain fluidity in response to temperature and osmotic stress[68,69,71]. The ability to synthesize phytoceramide and phytosphingosine suggest that SLA-04 maintains the ability to synthesize and remodel cell membrane lipids in response to changes in temperature and osmotic stress.

## Discussion

Algae thrive in diverse aquatic ecosystems and identifying the genotypes responsible for growth across diverse environments requires a comprehensive understanding of genomic and metabolic variation across habitats. Using ultra-long-read Nanopore sequencing, we determined a complete chromosome level assembly of the *Chlorella sp.* SLA-04 genome (**Figure 2.1**).

Annotation of the SLA-04 genome using sequence-based methods resulted in an incomplete representation of SLA-04 metabolism with just over half of the proteins having assigned functions. While traditional sequence-based annotation methods identified core metabolic pathways within SLA-04, a large portion of the proteome remained hypothetical with unknown function (**Figure 2.2A**).

The structure of a protein dictates its function[72]. By predicting the structure of hypothetical proteins and comparing them to a structural database we were able to predict the function of proteins that were unrecognizable at the sequence level. While structural annotation identified additional proteins within all functional categories, some categories such as cytoskeleton and cell wall proteins were overrepresented in this analysis (**Figure 2.2**). The overrepresentation of cytoskeleton and cell wall proteins could be a consequence of functional diversification or an underrepresentation of cytoskeleton and cell wall protein sequences within databases used for sequence-based annotation (NCBI or EggNOG). Regardless, we found that using structure-based annotation methods aid in generating a more complete representation of algal metabolism. Interestingly, 37.6% of SLA-04 proteins continue to escape functional predictions following structural annotation, and 20.5% of these sequences are conservation within other algae and plants. Evolutionary conservation and retention of proteins across algae and plants suggests these proteins are functionally important, and this analysis helps focus future efforts on conserved aspect of the proteome that remain functionally uncharacterized.

While annotating the SLA-04 proteome, we identified lipid synthesis pathways that may allow SLA-04 to respond to changing environmental conditions. ATP citrate lyase (ACLY) generates cytosolic acetyl-CoA from citrate and is a critical enzyme for fatty acid (FA) biosynthesis

and modification[73–75]. Algal fatty acids are initially synthesized in the chloroplast and exported into the cytosol where they are modified into triacylglycerols (TAG) or sphingolipids within the ER[69,76–80] (**Figure 2.4**). Modification and synthesis of additional fatty acids within the cytosol and ER requires cytosolic acetyl-CoA. The presence and transcription of genes encoding ACLY suggests that SLA-04 can utilize starch-derived TCA intermediates as a precursor for lipid synthesis. The ability to interconvert between storage molecules enables the metabolic flexibility to synthesize desired biomolecules regardless of the molecular resources available within a cell. Since starches and lipids are important precursors for bioproducts like biodiesel and ethanol, this work demonstrates how genome sequences may improve our ability to link bioproduct production to specific organisms and environmental conditions.

Acetyl-CoA has many cytosolic functions, but we hypothesize that SLA-04 may generate additional cytosolic acetyl-CoA in part to synthesize long term storage lipids and sphingolipids. Sphingolipids are membrane lipids synthesized within the ER that contain long chain FAs. Sphingolipids modify the algal cell membrane to maintain membrane fluidity in response to temperature and osmotic stress[68–70]. The synthesis of sphingolipids such as phytoceramide and phytosphingosine requires cytosolic acetyl-CoA to modify and elongate free FAs that originated from the chloroplast[81](**Figure 2.4**). Similar to *ACLY*, genes encoding enzymes responsible for phytoceramide and phytosphingosine biosynthesis are present and being transcribed in the SLA-04 genome (**Figure 2.4, Supplementary Table 2.3**). Together, the capacity for sphingolipid biosynthesis and cytosolic acetyl-CoA generation by ACLY suggests that SLA-04 may have the metabolic capacity to modulate cell membrane fluidity to rapidly adapt to diurnal and seasonal temperature and osmotic changes (*e.*g., alkalinity and/or salinity) within the environment[82].

## Conclusion

*Chlorella sp.* SLA-04 grows well in a wide range of media conditions such as (i) high (>100 mM)[83] and low (<5 mM)[18] alkalinities, (ii) circumneutral pH[84] and high (>10.2) pH[83], and (iii) in media containing 0 to 35,000 ppm salt[85]. The genome analysis presented here, together with prior experimental results indicate that SLA-04 has robust and adaptable growth characteristics, with potential for use in diverse engineered growth systems for biofuel and bioproduct applications. More broadly, this comprehensive view of metabolism within SLA-04 has demonstrated how complete algal genomes can be leveraged to discover new genotypes and identify algal strains most suitable for production of desired bioproducts.

## Data Access

The raw and processed genomic and transcriptomic data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under BioProject accession number PRJNA766673 and BioSample accession number SAMN21857598.

## Acknowledgements

This work was supported by the Bioenergy Technologies Office (BETO) of the U.S. Department of Energy (DOE) for the Productivity Enhanced Algae and ToolKits (PEAK) program [grant number #DE-EE0008247].

## References

1. Kuypers, M. M. M., Marchant, H. K. & Kartal, B. The microbial nitrogen-cycling network. *Nature Reviews Microbiology* vol. 16 263–276 (2018).

2. Walker, J. C. G. The Oxygen Cycle. 87–104 (1980) doi:10.1007/978-3-662-24940-6_5.

3. Behrenfeld, M. J. Biospheric Primary Production During an ENSO Transition. *Science* vol. 291 2594–2597 (2001).

4. Field, C. B. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* vol. 281 237–240 (1998).

5. Evans, J. H. The Survival of Freshwater Algae During Dry Periods: Part I. An Investigation of the Algae of Five Small Ponds. *The Journal of Ecology* vol. 46 149 (1958).

6. Varshney, P., Mikulic, P., Vonshak, A., Beardall, J. & Wangikar, P. P. Extremophilic micro-algae and their potential contribution in biotechnology. *Bioresour Technol* vol. 184 363–372 (2015).

7. Blanc-Mathieu, R. *et al.* An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* vol. 15 1103 (2014).

8. Merchant, S. S. *et al.* The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. *Science* vol. 318 245–250 (2007).

9. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* vol. 30 693–700 (2012).

10. Kingsford, C., Schatz, M. C. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* vol. 11 21 (2010).

11. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* vol. 21 (2020).

12. Mehrotra, S. & Goyal, V. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics, proteomics & bioinformatics* vol. 12 164–171 (2014).

13. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* vol. 35 2193–2198 (2018).

14. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature Biotechnology* vol. 34 518–524 (2016).

15. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology* vol. 39 1348–1365 (2021).

16. Dohm, J. C., Peters, P., Stralis-Pavese, N. & Himmelbauer, H. Benchmarking of long-read correction methods. *NAR Genomics and Bioinformatics* vol. 2 (2020).

17. Sutton, J. M., Millwood, J. D., Case Mccormack, A. & Fierst, J. L. Optimizing experimental design for genome sequencing and assembly with Oxford Nanopore Technologies. *Gigabyte* vol. 2021 1–26 (2021).

18. Vadlamani, A., Viamajala, S., Pendyala, B. & Varanasi, S. Cultivation of Microalgae at Extreme Alkaline pH Conditions: A Novel Approach for Biofuel Production. *ACS Sustainable Chemistry & Engineering* vol. 5 7284–7294 (2017).

19. Bold, H. C. The Morphology of Chlamydomonas chlamydogama, Sp. Nov. *Bulletin of the Torrey Botanical Club* vol. 76 101 (1949).

20. Mustapa, M., Sallehudin, N., Mohamed, M., Mohammad Noor, N. & Raus, R. Decontamination of Chlorella sp. Culture using antibiotics and antifungal cocktail treatment. *ARPN Journal of Engineering and Applied Sciences* vol. 11 (2016).

21. Goemann, H. M. *et al.* Aboveground and belowground responses to cyanobacterial biofertilizer supplement in a semi-arid, perennial bioenergy cropping system. *Global Change Biology. Bioenergy* vol. 13 1908–1923 (2021).

22. Weyrich, L. S. *et al.* Laboratory contamination over time during low-biomass sample analysis. *Molecular Ecology Resources* vol. 19 982–996 (2019).

23. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* vol. 32 2103–2110 (2016).

24. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. *Genome Research* vol. 27 722–736 (2017).

25. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* vol. 37 540–546 (2019).

26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* vol. 34 3094–3100 (2018).

27. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* vol. 10 (2021).

28. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

29. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).

30. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**, 2 (2021).

31. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427–D432 (2019).

32. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276–277 (2000).

33. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

34. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res* **47**, W5–W10 (2019).

35. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* vol. 9 304 (2020).

36. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* vol. 33 290–295 (2015).

37. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* vol. 18 188–196 (2007).

38. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics* vol. 48 (2014).

39. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* vol. 35 W182-5 (2007).

40. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* vol. 38 5825–5829 (2021).

41. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* vol. 31 3210–3212 (2015).

42. Manni, M. e, Matthew, Seppey, M., Felipe, & Evgeny. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *arXiv pre-print server* (2021) doi:None arxiv:2106.11799.

43. Hadley, W. Ggplot2: Elegrant graphics for data analysis. (2016).

44. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. 207–232 (2007) doi:10.1007/978-3-540-35306-5_10.

45. Ruperti, F., Papadopoulos, N., Musser, J. & Arendt, D. Beyond sequence similarity: cross-phyla protein annotation by structural prediction and alignment. (2022) doi:10.1101/2022.07.05.498892.

46. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

47. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. (2021) doi:10.1101/2021.10.04.463034.

48. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* vol. 19 679–682 (2022).

49. Eddy, S. R. Homology searches for structural RNAs: from proof of principle to practical use. *RNA* **21**, 605–607 (2015).

50. Cannone, J. J. *et al.* The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* vol. 3 2 (2002).

51. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* vol. 30 772–780 (2013).

52. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* vol. 32 268–274 (2015).

53. Drost, H.-G., Gabel, A., Grosse, I. & Quint, M. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. *Molecular Biology and Evolution* vol. 32 1221–1231 (2015).

54. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* vol. 34 W609–W612 (2006).

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

56. Hovde, B. T. *et al.* Genomic characterization reveals significant divergence within Chlorella sorokiniana (Chlorellales, Trebouxiophyceae). *Algal Research* vol. 35 449–461 (2018).

57. Higashiyama, T., Maki, S. & Yamada, T. Molecular organization of Chlorella vulgaris chromosome I: presence of telomeric repeats that are conserved in higher plants. *Molecular and General Genetics MGG* vol. 246 29–36 (1995).

58. Arriola, M. B. *et al.* Genome sequences of Chlorella sorokiniana UTEX 1602 and Micractinium conductrix SAG 241.80: implications to maltose excretion by a green alga. *The Plant Journal* vol. 93 566–586 (2018).

59. Hanschen, E. R., Hovde, B. T. & Starkenburg, S. R. An evaluation of methodology to determine algal genome completeness. *Algal Research* vol. 51 102019 (2020).

60. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* vol. Chapter 4 Unit 4.10 (2009).

61. Fulnečková, J. *et al.* Dynamic Evolution of Telomeric Sequences in the Green Algal Order Chlamydomonadales. *Genome Biology and Evolution* vol. 4 248–264 (2012).

62. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* vol. 12 491 (2011).

63. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* vol. 77 499–508 (2009).

64. Álvarez, I. & Wendel, J. F. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* vol. 29 417–434 (2003).

65. Evans, K. M., Wortley, A. H. & Mann, D. G. An Assessment of Potential Diatom "Barcode" Genes (cox1, rbcL, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in Sellaphora (Bacillariophyta). *Protist* vol. 158 349–364 (2007).

66. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* vol. 6 361–375 (2005).

67. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* vol. 36 2251–2252 (2020).

68. Li, Y. *et al.* Sphingolipids in marine microalgae: Development and application of a mass spectrometric method for global structural characterization of ceramides and glycosphingolipids in three major phyla. *Analytica Chimica Acta* vol. 986 82–94 (2017).

69. De Bigault Du Granrut, A. & Cacas, J.-L. How Very-Long-Chain Fatty Acids Could Signal Stressful Conditions in Plants? *Frontiers in plant science* vol. 7 1490–1490 (2016).

70. Resemann, H. C. *et al.* Convergence of sphingolipid desaturation across over 500 million years of plant evolution. *Nature Plants* vol. 7 219–232 (2021).

71. Kong, J. N. *et al.* Regulation of Chlamydomonas flagella and ependymal cell motile cilia by ceramide-mediated translocation of GSK3. *Molecular Biology of the Cell* vol. 26 4451–4465 (2015).

72. Kendrew, J. C. *et al.* A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* vol. 181 662–666 (1958).

73. Bauer, D. E., Hatzivassiliou, G., Zhao, F., Andreadis, C. & Thompson, C. B. ATP citrate lyase is an important component of cell growth and transformation. *Oncogene* vol. 24 6314–6322 (2005).

74. Hynes, M. J. & Murray, S. L. ATP-Citrate Lyase Is Required for Production of Cytosolic Acetyl Coenzyme A and Development in Aspergillus nidulans. *Eukaryotic Cell* vol. 9 1039–1048 (2010).

75. Liang, M.-H. & Jiang, J.-G. Characterization and nitrogen deficiency response of ATP-citrate lyase from unicellular alga Dunaliella tertiolecta. *Algal Research* vol. 20 77–86 (2016).

76. Fan, J. *et al.* Genomic Foundation of Starch to Lipid Switch in Oleaginous Chlorella. *Plant Physiology* pp.01174.2015 (2015) doi:10.1104/pp.15.01174.

77. Goncalves, E. C., Wilkie, A. C., Kirst, M. & Rathinasabapathi, B. Metabolic regulation of triacylglycerol accumulation in the green algae: identification of potential targets for engineering to improve oil yield. *Plant Biotechnology Journal* vol. 14 1649–1660 (2016).

78. Avidan, O. & Pick, U. Acetyl-CoA synthetase is activated as part of the PDH-bypass in the oleaginous green algaChlorella desiccata. *Journal of Experimental Botany* vol. 66 7287–7298 (2015).

79. Bellou, S. & Aggelis, G. Biochemical activities in Chlorella sp. and Nannochloropsis salina during lipid and sugar synthesis in a lab-scale open pond simulating reactor. *Journal of Biotechnology* vol. 164 318–329 (2013).

80. Tan, K. W. M. & Lee, Y. K. The dilemma for lipid productivity in green microalgae: importance of substrate provision in improving oil yield without sacrificing growth. *Biotechnology for Biofuels* vol. 9 (2016).

81. Aid, F. Plant Lipid Metabolism. (2020) doi:10.5772/intechopen.81355.

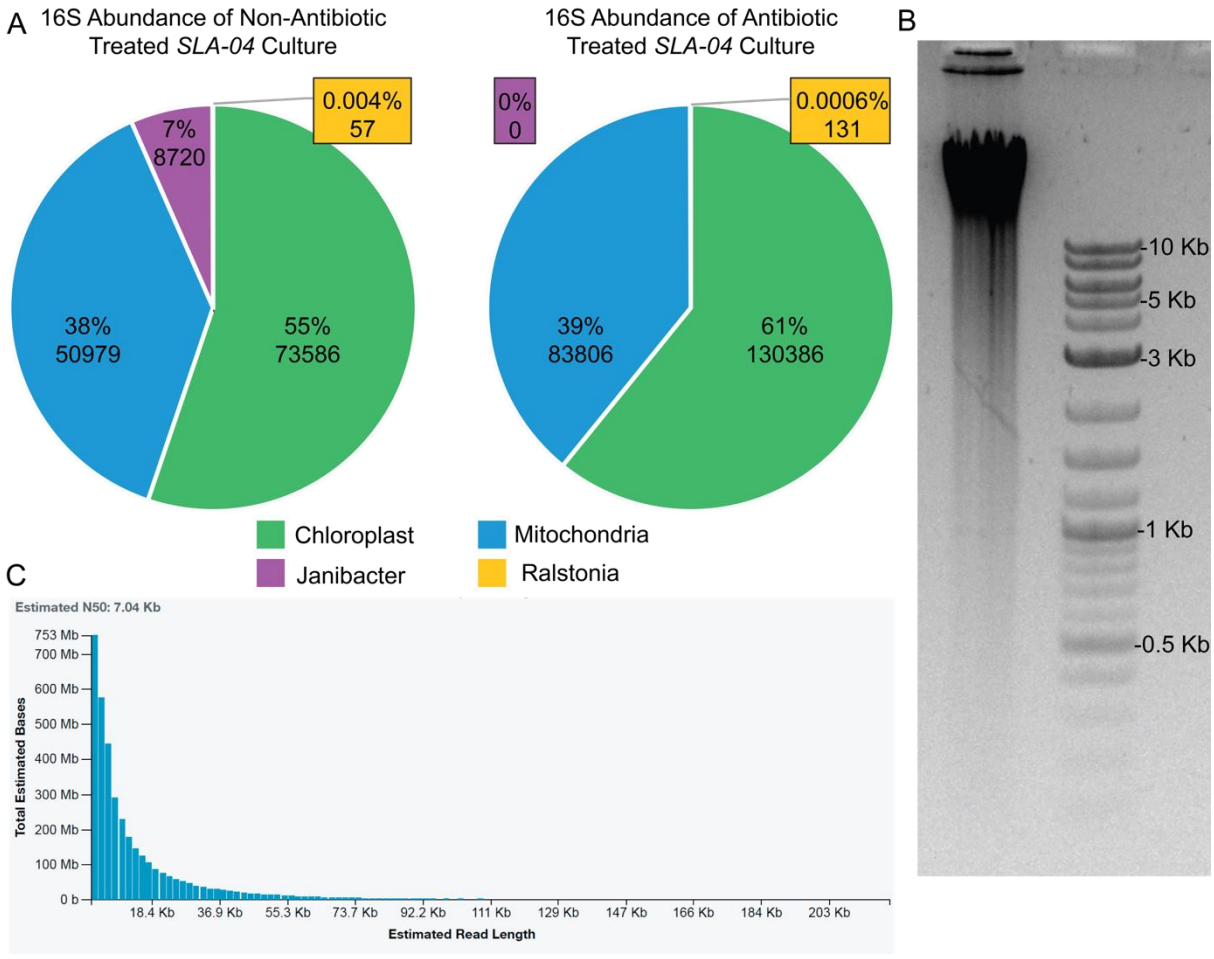
82. Cael, B. B., Heathcote, A. J. & Seekell, D. A. The volume and mean depth of Earth's lakes. *Geophysical Research Letters* vol. 44 209–218 (2017).

83. Vadlamani, A., Pendyala, B., Viamajala, S. & Varanasi, S. High Productivity Cultivation of Microalgae without Concentrated CO2 Input. *ACS Sustainable Chemistry & Engineering* vol. 7 1933–1943 (2019).

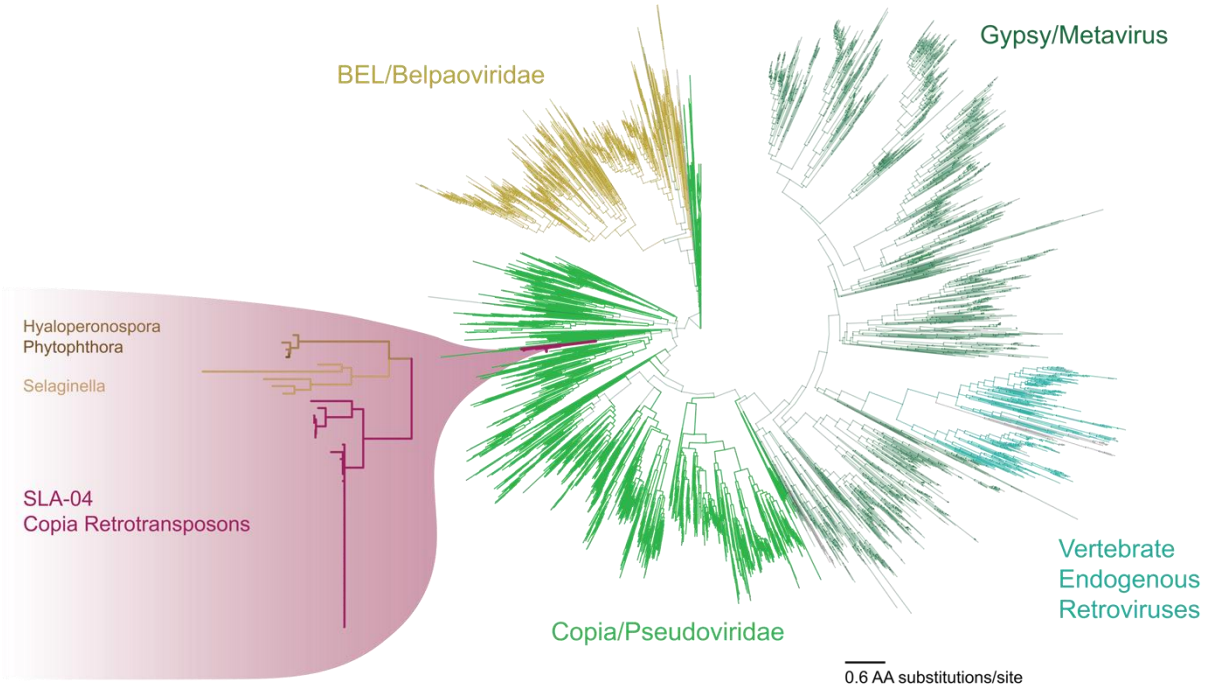
84. Hanifzadeh, M., Garcia, E. C. & Viamajala, S. Production of lipid and carbohydrate from microalgae without compromising biomass productivities: Role of Ca and Mg. *Renewable Energy* vol. 127 989–997 (2018).

85. Hanifzadeh, M.-M. Approaches for Sustainable Production of Microalgae with High Productivity and Flexible Composition. (2018).

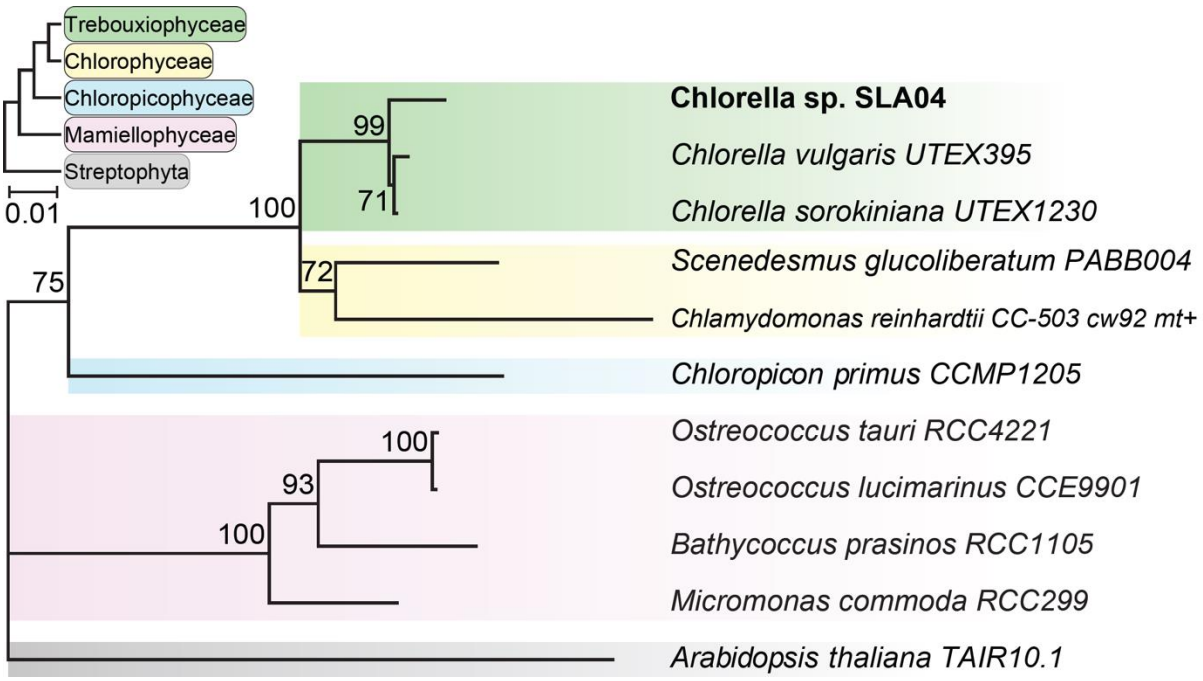
## Supplementary Figures



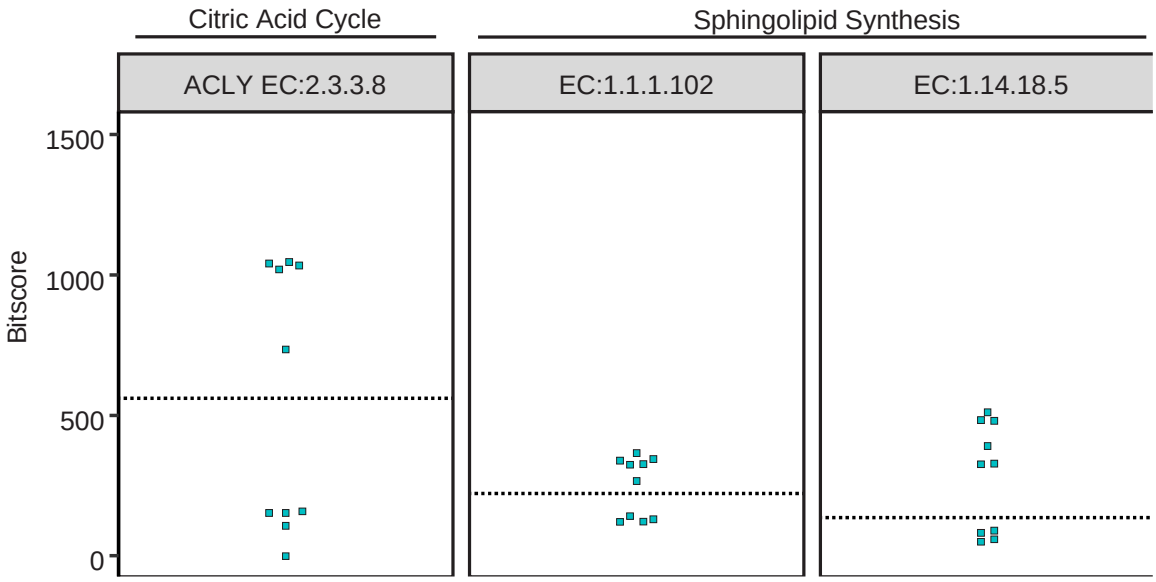

Supplementary Figure 2.1. DNA isolation and sequencing parameters for Chlorella sp. SLA-04. (A) An axenic culture of SLA-04 was confirmed by sequencing the V4 region of the 16S rRNA gene on an Illumina MiSeq platform [21]. Pie charts indicate raw read numbers from an untreated (left) and antibiotic treated culture (right). The antibiotic-treated culture was used for DNA extraction and genomic analysis. Ralstonia sequences are common contaminants of DNA purification columns [22]. (B) Gel of high molecular weight DNA from Chlorella sp. SLA-04 using liquid N2 grinding and gravity column purification. (C) Nanopore output of sequence lengths from SLA-04 sequencing run.




Supplementary Figure 2.2. Phylogenetic tree of LTR retrotransposon capsid domains places SLA-04 retrotransposons among Copia/Pseudoviridae LTR retrotransposons. Hidden Markov Models for LTR retroelement capsid domains were downloaded from Pfam. 752,604 ORFs from NCBI Virus and Repbase were queried to identify LTR retroelements capsid genes in the SLA-04 genome. Redundant sequences were removed with CD-Hit. Sequences were aligned in MAFFT and trimmed using Trimal and manual curation. Final alignment of 75 AA core capsid domain residues from 5520 sequences were used to build a maximum likelihood phylogenetic tree in Fasttree2. Tree was annotated in ggtree.




Supplementary Figure 2.3. Phylogenomic analysis of green algae performed using 18S rDNA. 18S rDNA gene sequences of Chlorella sp. SLA-04 and complete algae genomes from NCBI were used to make a phylogenetic tree using maximum likelihood. Branch lengths represent the number of nucleotide substitutions per site. Arabidopsis thaliana was used as an outgroup. Bootstrap values are labeled next to each node. Classes of algae are denoted via color on tree branches.




Supplementary Figure 2.4. Presence of ATP citrate lyase and sphingolipid enzymes across algal genomes. Algal protein coding sequences from all chromosome level algal genomes (n=10) within the NCBI database were submitted to KofamKOALA to predict function. KofamKOALA assigned KEGG Orthologs (KOs) to protein sequences by homology search against a database of profile hidden Markov models (KOfam) with pre-computed adaptive score thresholds. Each datapoint represents the likelihood of a functional protein being present within a single algal genome. Genes are considered functional if the bitscore is above the adaptive threshold (dotted line). Genes of interest involved in fatty acid and sphingolipid synthesis are reported.

# Supplementary Tables

| Assembly and annotation statistics of *Chlorella sp.* SLA-04 | |
|---|---|
| SLA-04 Genome Assembly: | |
| Assembled genome size | 38.8 Mb |
| Chromosomes | 13 |
| Average contig size | 1,436,022 bp |
| N50 | 2.20 Mb |
| N90 | 762,653 bp |
| GC content | 65.87% |
| Coverage | 117 X |
| | |
| SLA-04 Genome Annotation: | |
| Annotated protein coding genes | 10,310 |
| Annotated tRNA genes | 100 |
| Average gene length | 3,475 bp |
| Average CDS length | 1,745 bp |
| % of genome assembly as genes | 91.07% |
| % of genome assembly as CDS | 45.73% |
| | |
| Busco: Complete genes | 93.4% |
| Busco: Fragmented genes | 2.4% |
| Busco: Missing genes | 4.2% |
| | |
| Genes with KEGG function | 6,434 (62.4%) |
| Cellular processing genes | 770 |
| Environmental info. processing genes | 307 |
| Genetic info. processing genes | 1976 |
| Metabolism genes | 2427 |
| Unannotated genes | 3,876 (37.6%) |

Supplementary Table 2.1. Genome statistics of *Chlorella sp.* SLA-04 genome.

```
total length:   38807484 bp  (38807484 bp excl N/X-runs)
GC level:         65.87 %
bases masked:     403707 bp ( 1.04 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
Retroelements         2312       302824 bp    0.78 %
   SINEs:               21         1111 bp    0.00 %
   Penelope             14          747 bp    0.00 %
   LINEs:             1347       133696 bp    0.34 %
    CRE/SLACS            1           77 bp    0.00 %
     L2/CR1/Rex         30         1469 bp    0.00 %
     R1/LOA/Jockey     336        25654 bp    0.07 %
     R2/R4/NeSL          7          361 bp    0.00 %
     RTE/Bov-B          68         5553 bp    0.01 %
     L1/CIN4            71        22474 bp    0.06 %
   LTR elements:       944       168017 bp    0.43 %
     BEL/Pao            31         3818 bp    0.01 %
     Ty1/Copia         378       128222 bp    0.33 %
     Gypsy/DIRS1       438        29928 bp    0.08 %
       Retroviral       82         4306 bp    0.01 %

DNA transposons        772        52810 bp    0.14 %
   hobo-Activator      124         8667 bp    0.02 %
   Tc1-IS630-Pogo       99         6362 bp    0.02 %
   En-Spm                0            0 bp    0.00 %
   MuDR-IS905            0            0 bp    0.00 %
   PiggyBac              2           62 bp    0.00 %
   Tourist/Harbinger    41         3105 bp    0.01 %
   Other (Mirage,        3          439 bp    0.00 %
    P-element, Transib)

Rolling-circles        116         7169 bp    0.02 %

Unclassified:           12         4049 bp    0.01 %

Total interspersed repeats:      359683 bp    0.93 %


Small RNA:             108        39406 bp    0.10 %

Satellites:             67         7508 bp    0.02 %
Simple repeats:          0            0 bp    0.00 %
Low complexity:          0            0 bp    0.00 %
==================================================
```

Supplementary Table 2.2. Summary of repetitive elements identified by RepeatMasker using Dfam and Repbase as a combined reference database.

| Gene ID | Gene Name | KEGG Ortholog | Coverage | FPKM | TPM |
|---|---|---|---|---|---|
| SLA-04_00282-RA | **ATP citrate (pro-S)-lyase [EC:2.3.3.8]** | K01648 | 97.7 | 136.7 | 171.5 |
| SLA-04_07917-RA | **ATP citrate (pro-S)-lyase [EC:2.3.3.8]** | K01648 | 82.4 | 115.3 | 144.6 |
| SLA-04_04864-RA | **serine palmitoyltransferase [EC:2.3.1.50]** | K00654 | 11.2 | 15.6 | 19.6 |
| SLA-04_04303-RA | **serine palmitoyltransferase [EC:2.3.1.50]** | K00654 | 22.9 | 32.0 | 40.2 |
| SLA-04_04491-RA | **3-dehydrosphinganine reductase [EC:1.1.1.102]** | K04708 | 16.4 | 22.9 | 28.7 |
| SLA-04_04653-RA | **sphingoid base N-palmitoyltransferase [EC:2.3.1.291]** | K23727 | 8.3 | 11.6 | 14.5 |
| SLA-04_00395-RA | **sphingoid base N-palmitoyltransferase [EC:2.3.1.291]** | K23727 | 38.2 | 53.5 | 67.1 |
| SLA-04_04461-RA | **sphingolipid C4-monooxygenase [EC:1.14.18.5]** | K04713 | 32.9 | 46.0 | 57.7 |

Supplementary Table 2.3. Active transcription of lipid synthesis genes within the Chlorella sp. SLA-04 genome. mRNA transcript counts were determined by mapping the mRNA reads to the SLA-04 genome using Integrated Genome Viewer (IGV). Coverage represents the total number of mRNA reads mapped back to each gene. Fragments Per Kilobase Million (FPKM) and Transcripts Per Million (TPM) adjusted coverage values are also reported for each gen

# CHAPTER THREE

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

## Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Murat Buyukyoruk

Contributions: Designed research, wrote the manuscript, performed CRISPRCasFinder, CRISPRDetect, and CRISPRCasTyper searches, analyzed data, edited the manuscript

Co-Author: William S. Henriques

Contributions: Analyzed data and edited the manuscript

Co-Author: Blake Wiedenheft

Contributions: Designed research, wrote the manuscript, analyzed data, edited the manuscript

Manuscript Information

Murat Buyukyoruk, William S. Henriques, Blake Wiedenheft

*The CRISPR Journal*

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

Mary Ann Liebert, Inc.
Vol. 6, No. 3
Published online, 1 June 2023
https://doi.org/10.1089/crispr.2022.0106

## Abstract

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


## Introduction

Repetitive DNA sequences are diverse and abundant in a broad range of species, from bacteria to mammals, where they participate in many different biological functions[1]. CRISPRs belong to a specific family of DNA repeats found in bacteria and archaea that all share a common architecture[2,3]. Each CRISPR locus consists of a series of short repeat sequences (20-50 bps) separated by unique spacer sequences of a similar length (**Figure 3.1A**)[4]. The repeat sequences within a CRISPR locus are generally conserved but repeats between different CRISPR loci can vary in both sequence and length[2]. The diversity of these elements makes them challenging to confidently identify, especially when the number of repeat sequences within a locus is small.

The discovery of CRISPR-based adaptive immunity and the ongoing adaptation of these systems for applications in genome editing has engaged scientists from nearly every discipline

and many perspectives. Thus, the recent paper by van Riet *et al.*, which reported the discovery of more than 12,500 CRISPR loci in the human reference genome (GRCh38.p13; GCF_000001405.39), comes as a surprise[5]. They report that hCRISPRs are similar to prokaryotic CRISPRs in terms of array architecture, repeat length, repeat sequence, and that hCRISPR are flanked by CRISPR-associated (*cas*) genes (**Figure 3.1B**)[5].

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

## Methods

### CRISPR Predictions

The human reference genome (GRCh38.p13; GCF_000001405.39) was downloaded from the NCBI RefSeq Assembly database (193 sequences) (August 22nd, 2022)[6]. Local versions of CRISPRCasFinder[7] (v4.2.20; CasFinder v2.0.3), CRISPRDetect[8] (v3), and CRISPRCasTyper[9] (v1.6.4) were used to identify putative CRISPR arrays in the downloaded human reference genome. We used the following search command for CRISPRCasFinder *"-minSP 21 -maxSP 72 -fast*", and *"-array_quality_score_cutoff 0 -minimum_word_repeatation 2 -minimum_no_of_repeats 2*" for CRISPRDetect. Default parameters were used for the CRISPRCasTyper.

Predicted putative arrays from CRISPRCasFinder, CRISPRDetect, and CRISPRCasTyper were sorted based on the assigned confidence score. Evidence-level scores 3 or 4 by CRISPRCasFinder, array quality score above 4 by CRISPRDetect, and probability score above 0.75 by CRISPRCasTyper were considered as high-confidence arrays. Results from each CRISPR detection tool were compared to identify shared hits. Twenty candidate arrays predicted by CRISPRDetect were removed from the list of high-confidence arrays due to low confidence scores for these same arrays by CRISPRCasFinder. The remaining 72 arrays were then considered candidate arrays. Two of these arrays with the highest confidence scores from each of the three detection methods were manually inspected.

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

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

The mitochondria genomes of *Vicia faba* (X59246.1)[14] and the assembly of *Acanthamoeba polyphaga* mimivirus genome (GCF_000888735.1)[15] were downloaded from the NCBI RefSeq

database (January 12th, 2023). Previously reported CRISPR-like sequences from each genome were extracted and manually inspected.

Identification of Cas Genes in the Human Genome

CRISPRCasFinder and CRISPRCasTyper were used to identify *cas* genes in the human reference genome. CRISPRCasFinder failed to predict any putative *cas* gene in the human reference genome. Putative *cas* genes predicted by CRISPRCasTyper were subjected to a filtering step according to E-value (< 10-5) and sequence coverage cutoff (sequence coverage > 50%). In addition, translated CDS sequences of both the human reference genome (n=60,090) and the T2T assembly of the human genome (n=65,591) were downloaded from the NCBI RefSeq Assembly database[6,10]. One hundred and twenty one *cas* gene HMM profiles from CRISPRCasdb were used to query the CDS sequences using MacsyFinder (v1.0.5; HMMER 3.2.1)[16,17]. Queries were performed using the following parameters: "*macsyfinder --sequence-db <open_reading_frames> --db-type gembase -d <CRISPR_subtype_definitions> -p <HMM_profiles> -w 50 -vv all*". Results from the HMMER analysis were filtered with the same E-value threshold and sequence coverage criteria described above.

Comparison of Repeat Length Distribution of CRISPRs from Prokaryotes and HCRISPRs

The repeat sequences from 19,483 unique putative hCRISPR arrays were extracted using a Python v2.7 script that uses the Bio SeqIO package[18]. In addition, the repeats of bacterial and archaeal CRISPRs were downloaded from CRISPRCasdb (n=28,712). The distribution of repeat lengths was visualized in RStudio by comparing the list of repeats from CRISPRs and hCRISPRs using ggplot[19].

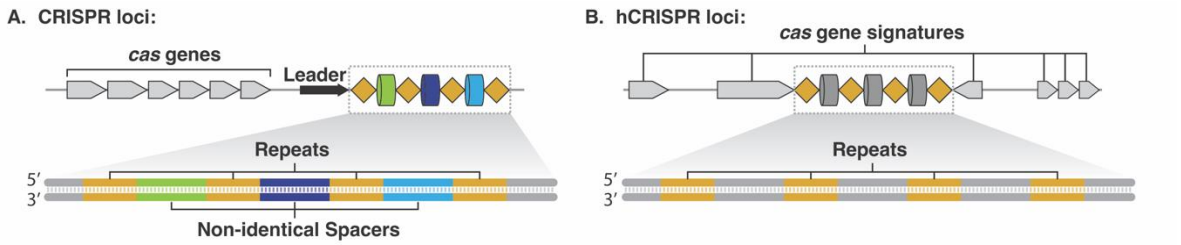



Figure 3.1. Comparison of architectural features in CRISPRs and hCRISPRs. A. Structure of typical CRISPR-Cas locus in bacteria and archaea. B. Structure of a recently reported "hCRISPR".

To understand how the putative CRISPR arrays reported by van Riet *et al.*, may have been missed by the rest of the scientific community and to learn what these new CRISPRs might teach us about the evolution and function of CRISPR systems, we searched the human reference genome (GRCh38.p13; GCF_000001405.39) for CRISPRs using a combination of CRISPRCasFinder[7], CRISPRDetect[8] and CRISPRCasTyper[9] according to the search parameters defined by the authors[5]. Collectively, we identified a total of 19,483 unique putative arrays in the human reference genome (Supplemental Data 1).

Results from CRISPRCasFinder (n=12,616) and CRISPRCasTyper (n=28) were nearly identical to those reported by van Riet *et al.*, while CRISPRDetect (n=10,056) identified 6,403 additional arrays (**Fig. 3.2A**). The exact search parameter used for CRISPRDetect were not provided by the authors (i.e., *-array_quality_score_cutoff 0 -minimum_word_repeatation 2 -minimum_no_of_repeats 2*), so the additional CRISPR loci identified here are anticipated to reflect differences in the search criteria.

CRISPR prediction tools assign confidence scores based on a combination of parameters. These parameters include the number of repeats in an array, conservation of repeat

sequences, and the presence of non-identical spacer sequences[7–9,20]. CRISPRDetect incorporates additional factors, such as the identification of previously described repeat sequences and detection of conserved *cas* genes (i.e., *cas1* or *cas2*), which are used to calculate confidence scores. While confidence scores are not mentioned by van Riet *et al.*[5], we sorted the arrays according to scores calculated by each algorithm. CRISPRCasFinder identified a total of three arrays with high confidence scores, CRISPRDetect identified 61 arrays with high confidence, and CRISPRCasTyper assigned eight arrays with high confidence (**Figure 3.2B**).

To further evaluate these predictions, we manually inspected the two examples presented by van Riet *et al.*[5] (neither of which earned a high confidence score by any of the search methods), as well as two arrays with the highest confidence scores from each of the three detection methods (**Figure 3.2C** and Supplementary Data 2). The two hCRISPR examples presented by van Riet *et al.* include one long and one short locus (**Figure 3.2C**, top)[5]. The long array contains eight "repeat" sequences that are 29% identical to each other and are separated by unique spacers sequences that vary in length (8-75 bps) (**Figure 3.2** and Supplementary Data 1 and 2). The short "hCRISPR" contains two identical 28 bp repeats and a single 39 bp spacer sequence.

We cannot definitively eliminate the possibility that the short array (composed of two repeats) is a CRISPR, but no *cas* genes are found flanking the CRISPR, the repeats do not correspond to any of repeats in CRISPRCasdb[16], and the spacer sequence only matches to itself. In addition to the two examples provided by van Riet *et al.,* we selected two arrays with high confidence scores from each of the three different CRISPR detection algorithms (**Figure 3.2C**, middle). High confidence arrays identified by CRISPRCasFinder contain "spacer" sequences

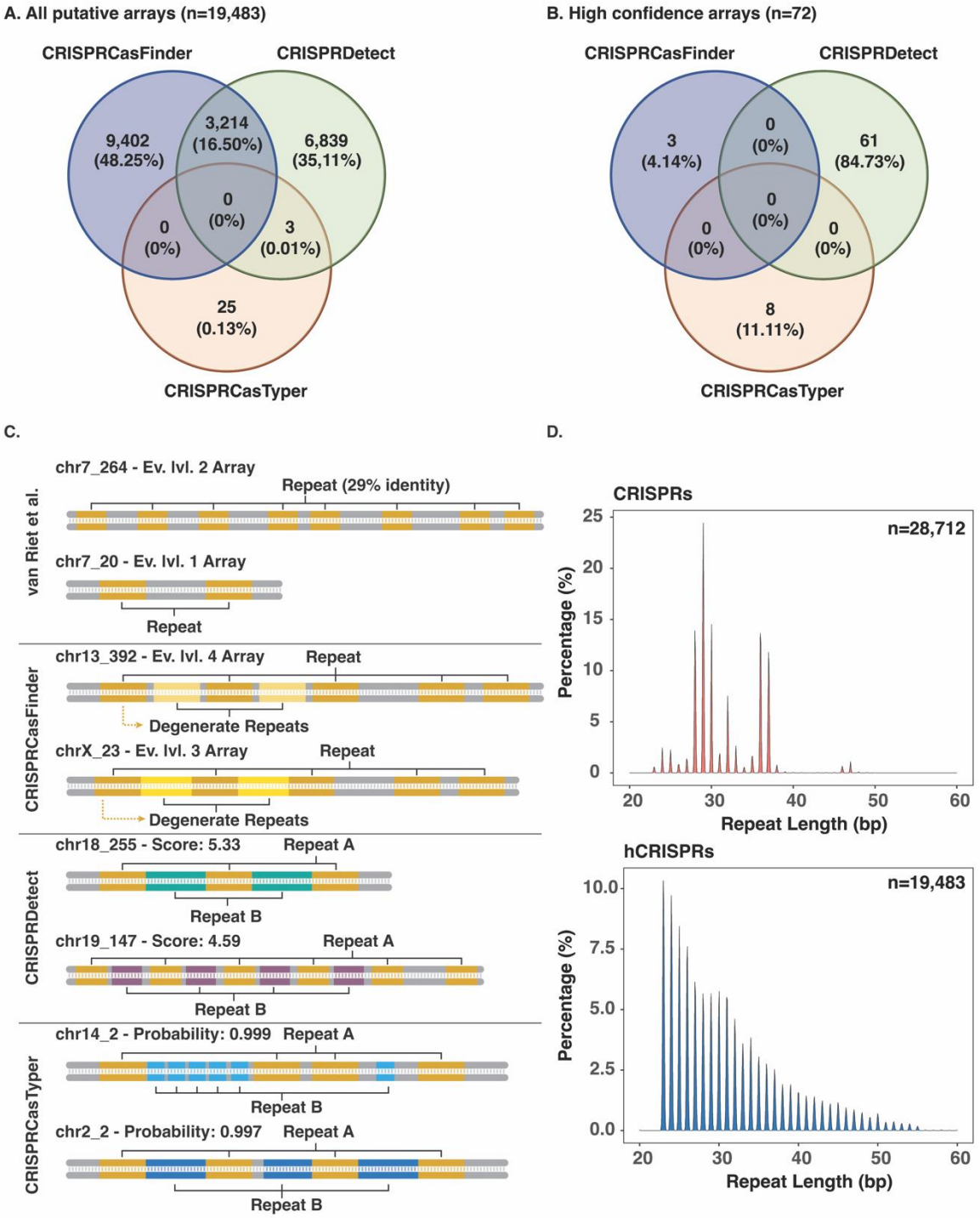



Figure 3.2. Repetitive elements in the human genome do not contain architectural features consistent with CRISPR loci. A) Human reference genome searched with CRISPRCasFinder, CRISPRDetect and CRISPRCasTyper. A total of 19,483 unique putative arrays predicted. 3,214 arrays were detected by both CRISPRCasFinder and CRISPRDetect. Only three arrays predicted by both CRISPRDetect and CRISPRCasTyper and no shared array predictions between

CRISPRCasFinder and CRISPRCasTyper. B) CRISPRCasFinder predicted only three arrays with high confidence (evidence-level 3 or 4), while CRISPRDetect predicted 61 candidate arrays (array quality score > 4) and CRISPRCasTyper assigned eight arrays as high confidence (probability > 75%). None of these arrays share between any of the three CRISPR detection methods. C) Array architecture of two examples presented by van Riet et al. and the top two arrays from each of the three detection methods. Neither of the arrays presented by the authors are high confidence predictions. Two of the predicted arrays failed to follow repeat-structure-repeat structure due to the presence of degenerate repeats found in the spacers. The other five arrays possess spacers that share high similarities between each other, indicating an array with tandem repeats. D) The direct repeats of bacterial and archaeal CRISPRs from CRISPRCasdb were downloaded and used to compare them to the distribution of hCRISPR repeat lengths. All repeat sequences were extracted and the number of each repeat corresponding to a particular length is plotted for bacterial and archaeal genomes (top, n=28,712) and the human reference genome (bottom, n=19,483).

that are degenerate versions of the repeat sequences, whereas putative arrays identified using CRISPRDetect and CRISPRCasTyper all contain "spacer" sequences that are 60% to 72% identical to one another. Thus, these arrays contain an alternating series of two repeats (i.e., Repeat A and Repeat B), rather than repeats separated by unique spacer sequences (**Figure 3.2C**, bottom).

To determine if the repeat sequences in the putative human CRISPR loci are similar in length to those typically found in prokaryotic CRISPRs, we downloaded the direct repeats of bacterial and archaeal CRISPRs from CRISPRCasdb[16] (n=28,712) and compared them to the distribution of repeat lengths from putative CRISPR loci in the human genome (**Figure 3.2D**). CRISPR repeats identified in bacterial and archaeal genomes are primarily 27-28 bp or 36-37 bp in length, while repeats identified in the human reference genome are primarily 23 bp and decrease in frequency as the repeat length increases. While the average length of the repeat sequences is similar (30 bp as identified by van Riet *et al.*), the distributions are very different.

In addition to CRISPR detection, the authors also report the detection of *cas* genes in the flanking regions of 336 hCRISPRs[5]. However, neither the sequences for these ORFs nor

significance values for these *cas* genes are provided. Our efforts to reproduce the identification of *cas* genes in the human reference genome using CRISPRCasFinder and CRISPRCasTyper failed. CRISPRCasFinder failed to predict any putative *cas* gene, while CRISPRCasTyper failed to predict significant hits (E-value < $10^{-5}$ and sequence coverage > 50%). To further verify these results, we downloaded 121 profile Hidden Markov Models (HMMs) for *cas* genes from CRISPRCasdb[16] and queried the human genome using MacSyFinder (v1.0.5)[21]. Consistent with CRISPRCasFinder and CRISPRCasTyper, we find no significant hits (E-value < $10^{-5}$ and sequence coverage > 50%) to *cas* genes in the human genome.

The human reference genome (GRCh38.p13) screened by van Riet *et al.* was originally assembled nearly 20 years ago[6]. While early drafts of the human genome provided critical new insight into human biology, these assemblies preferential excluded many repetitive sequences. However, the recently released Telomere-to-Telomere (T2T) assembly of the human genome, has resolved this problem[10]. Due to the repetitive nature of the human genome, we hypothesized the T2T genome might contain CRISPRs not found in earlier versions. To test this hypothesis, we searched the T2T genome (GCF_009914755.1 T2T-CHM13v2.0) using the methods described above. Collectively, we identified a total of 10,481 putative CRISPR arrays in the T2T genome (Supplementary Data 3). CRISPRDetect identified 78 putative CRISPRs with high confidence (52 arrays were identical to those predicted in human reference genome); however, none of these arrays maintain the architectural characteristics of a CRISPR.

Repetitive sequences make up ~50% of the human genome[10,22]. This is not a unique feature of the human genome, but characteristic of many eukaryotic genomes, where repetitive elements make up 1-80% of genomes[23–25]. Several annotation tools have been developed to

classify eukaryotic repetitive elements and notably, none have found evidence for CRISPRs[26,27]. Using the same methods described here for finding CRISPRs in prokaryotic genomes, we found 3,498 putative CRISPRs in the *Aedes aegypti* genome (310 Mb), 1,419 putative CRISPRs in the *Arabidopsis thaliana* genome (30 Mb), 490 putative CRISPRs in the *Caenorhabditis elegans* genome (15 Mb). In this small, non-representative sample of genomes, the number of putative CRISPRs identified correlates with the size of the genome, which is similar to the distribution of putative CRISPRs identified by van Riet *et al.* ("hCRISPRs are found across all chromosomes and their distribution is directly correlated to the chromosome length").

Apart from the work by van Riet *et al.,* CRISPR loci have been previously reported in the mitochondrial genome of a plant (*Vicia faba*) and the genome of a giant mimivirus that infects single-celled eukaryotes[3,15]. However, neither of these repeats meet the CRISPR criteria described above (**Supplemental Figure 3.1**)[2,28].

Overall, the evidence for functional CRISPR-Cas systems in the human genome and other eukaryotic genomes remains unsupported. While the phylogenetic and functional diversity of CRISPRs continue to expand, (to date) CRISPR-Cas adaptive immune systems remain restricted to prokaryotes.

## References

1. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2011).

2. Jansen, Ruud., Embden, Jan. D. A. van, Gaastra, Wim. & Schouls, Leo. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* **43**, 1565–1575 (2002).

3. Mojica, F. J. M., Díez-Villaseñor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology* **36**, 244–246 (2000).

4. Ishino, Y., Krupovic, M. & Forterre, P. History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *J Bacteriol* **200**, (2018).

5. van Riet, J. *et al.* CRISPRs in the human genome are differentially expressed between malignant and normal adjacent to tumor tissue. *Commun Biol* **5**, 1–13 (2022).

6. Schneider, V. A., Graves-Lindsay, T. & Howe, K. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849–864 (2017).

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

8. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016).

9. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *The CRISPR Journal* **3**, 462–469 (2020).

10. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

11. Matthews, B. J. *et al.* Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature* **563**, 501–507 (2018).

12. Pucker, B. *et al.* A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLOS ONE* **14**, e0216233 (2019).

13. CeS, C. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–8 (1998).

14. Flamand, M. C., Goblet, J. P., Duc, G., Briquet, M. & Boutry, M. Sequence and transcription analysis of mitochondrial plasmids isolated from cytoplasmic male-sterile lines of Vicia faba. *Plant Mol Biol* **19**, 913–923 (1992).

15. Levasseur, A. *et al.* MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. *Nature* **531**, 249–252 (2016).

16. Pourcel, C. *et al.* CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Research* **48**, D535–D544 (2020).

17. Prakash, A., Jeffryes, M., Bateman, A. & Finn, R. D. The HMMER Web Server for Protein Sequence Similarity Search. *Current Protocols in Bioinformatics* **60**, 3.15.1-3.15.23 (2017).

18. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

19. Villanueva, R. A. M. & Chen, Z. J. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives* **17**, 160–167 (2019).

20. Alkhnbashi, O. S., Meier, T., Mitrofanov, A., Backofen, R. & Voß, B. CRISPR-Cas bioinformatics. *Methods* **172**, 3–11 (2020).

21. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLOS ONE* **9**, e110726 (2014).

22. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

23. Goemann, C. L. C. *et al.* Genome sequence, phylogenetic analysis, and structure-based annotation reveal metabolic potential of Chlorella sp. SLA-04. *Algal Research* **69**, 102943 (2023).

24. Hufford, M. B. *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).

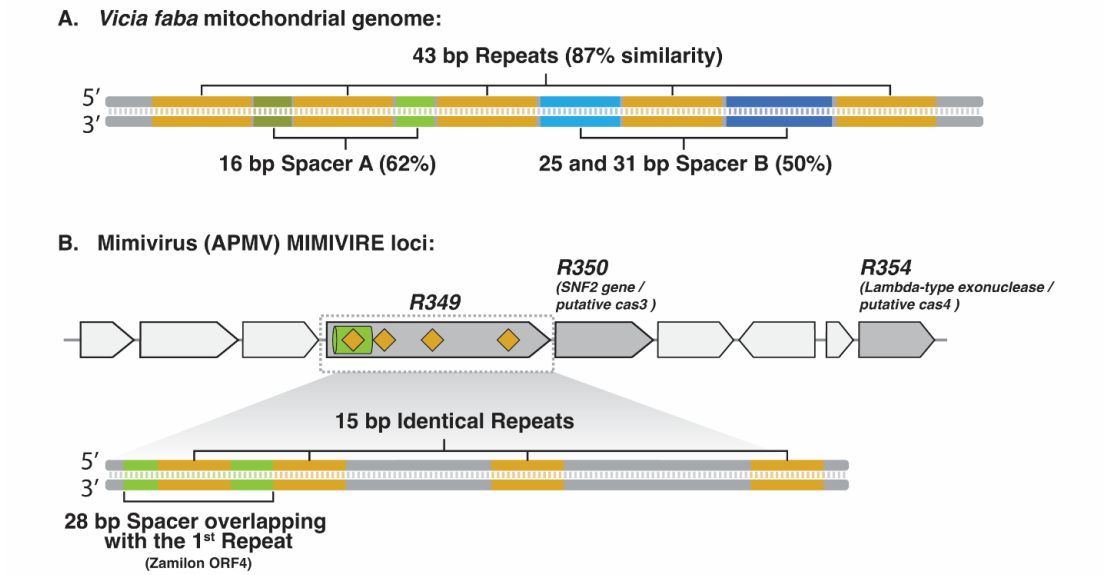
25. Udriște, A.-A., Iordachescu, M., Ciceoi, R. & Bădulescu, L. Next-Generation Sequencing of Local Romanian Tomato Varieties and Bioinformatics Analysis of the Ve Locus. *International Journal of Molecular Sciences* **23**, 9750 (2022).

26. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).

27. Hubley, R., Finn, R. D. & Clements, J. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, (2016).

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

Supplementary Figures




Supplementary Figure 3.1. Previously reported CRISPR-like elements in the mitochondrial genome of a plant (Vicia faba) and the genome of a giant mimivirus (APMV). A) Array architecture of CRISPR-like sequence from V. faba mitochondrion consisted of five similar but non-identical repeats of 43 bp that are interspaced by four small spacers. The two 16 bp spacer sequences are 62 % identical and the two longer spacer sequences are 50% identical. B) MIMIVIRE loci contains four 15 bp repeats with uneven spacing, located within the coding sequence of uncharacterized ORF (R349) of Mimivirus APMV. A 28 bp sequence that includes sequence flanking the first repeat, matches to the ORF4 gene in the Zamilon genome. The MIMIVIRE genome lacks a Cas1 homolog and encodes only distant homologs of Cas3 and Cas4, which are annotated as an SNF2-containing domain and Lambda-type endonuclease, respectively.

Supplemental Information

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

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

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

**GCAGGAGCTGTGCCTGGAGAGGCC**GACTTGAGGAAGTTTCCGGCACGGAGAGCC

TGTCGGGGG**GCAGGAGCTGGGCCTGGAGAGGCC**ACCGTGAGGCGTGAGCTGGGC

CTAATGAGATTAGTGTGAG**GCAGGATCTGGGCCTGTTGAAGCT**GCTGGGAGGCAG

GCAGGACCTTGTTTGGGGAAGGCCTCTGTGAG**GCAAAAGCTGGGCCTGGAGAGGC**

**C**ATTGTCAC**GCAGGAGCTGGGCCTGTCCAGGCC**ATTGGGAGGCAGAAGGTGGGCC

TGGAAAGCTTGACTTGAGGCAGTTTGGGGCCTACAAAGGCCAGTAGGAGCTGG**GCA**

**GGAGCTGAGCCAAAAGAGGTT**GTTCTGAGGCAGGAGTTGGACCTCCAGACACAGT

CAGAAGAAAGAGCACAGCCTGTAGAGGCCACCTGGGG**GCAGCAGCTGGGCCTAGA**

**GATGCC**AACGAAAG**GCAGGAGCTGGGCCTTGAGAGGCT**

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

**TGCTATGGGCCCTTCCAGAGATTCCCCA**CCCAGCTCCCCTCCAATGGCCAATGTCC

ATGTCATGGTT**TGCTATGGGCCCTTCCAGAGATTCCCCA**

- **CRISPRCasFinder Predictions:**

>chr13_392 | NC_000013_392 | Start: 109318667 | Stop: 109318921 | Evidence Level: 4 | Num. of predicted repeat: 5 – Repeat, Degenerate Repeat, **Consensus,** Mismatch

**TTACAGAGAGCTGATTGGCCCATT**TTACAGAGAGCTGATTGGTCCATTTTACAGA**GAGCTGATTGGTCCATT**TTGACAGGTGCTGATTGGTGAGTTTACAATCCCCGAGTTAA**ACACAGAGTGCTGATTGGTGCAG**TTACAATCCTCTAGCTAGACATAACAGTTCTCCAAGTCCCCACCAGATTAGCTAG**ATACAGAGTGCTGATTGGTGCAG**CCACAAACCCCAAGCTAG**ACACAGAGTGCTGATTGGTGCAT**

>chrX_23 | NC_000023_23 | Start: 3170226 | Stop: 3170464 | Evidence Level: 3 | Num. of predicted repeat: 5 – Repeat, Degenerate Repeat, **Consensus,** Mismatch

**CACATCCTGCTGATTGGTCCATTTTACAGAGCACTGATTGGTCCATTTTACAGAGTGCTGATTGGTCTGTTTTACAGAGTGCTAATTGGTCTGTTTCACAGAGTGCTTATTGGTGCGTTT**CCAATCCTTTAGCTAGACACAAAAGTTCTGCAAGTCCCCACCCAATTAGCTAGA**CACAGAGTGCTGATTGGTGCGTTT**ACAAACCTTTAGCTAGA**CACAGAGTGCTGATTGGTGCATTT**

- **CRISPRDetect Predictions:**

>chr18_255 | NC_000018.10 | Start: 78220037 | Stop: 78220212 | Score: 5.33 | Num. of predicted repeat: 3 – Repeat A, Repeat B, **Consensus,** Mismatch

**GTTCTCCTTTGCTCTGTAGAGTCCTATTCCTCTACCTCATCAGGTAGGTCTGTGTGTGTGAGAATGTTTGTTCTCCTTTGCCCTGTAGAGTCCTATTCGTCTACCTTGTCGTGTAGGTTTGTGTGTGGGCGAGTATTTGTTCTCCTTTGCCCTGTAGAGTCCTATTCCTCTACCT**

>chr19_147 | NC_000019.10 | Start: 28434544 | Stop: 28434889 | Score: 4.59 | Num. of predicted repeat: 6 – Repeat A, Repeat B, **Consensus,** Mismatch

**AGGGGTTCTTGCCTCTCCCCCAGAAAAGCGGGACTTGCCTCTAAGGGTGAAGG**
**AGAAGGGGTTGAGAGGTTCTTGCCCCTCTCCCAGAAAAGCGGGACTTGCTGCT**
**GACGGTGAAGGACAAGGGTTTGAGGGGTTCTTGCCCCTCCCCCAGAAAAGTGA**
**GACTTGCTGCTAACAGTGAAGGACAAGATGTTGAGGGGTTCTTGCCCCTCCCCC**
**AGAAAAGTGGGACTTGCCATTAAGGGTGAAGGAGAAGGGGTTGAGGGGTTCTT**
**GCCCCTCCCCCAGAAAAG**CAGAGAAGAGGTAGAGACATGGAGAGAAGGGGT**TGG**
**GGTTCTTGCCCCTCCCCCAGAAAAG**

- **CRISPRCasTyper Predictions:**

> chr14_2_1 | NC_000014.9 | Start: 30570622 | Stop: 30570874 | Probability: 0.999 | Num. of predicted repeat: 4 – Repeat A, Repeat B, **Consensus**

**GTGTGTATATATAGCACATATATAT**AACACAC**ATATAT**GT**ATATAT**ATAC**ATATAT**AACA
TATATGGTGTGT**ATATAT**G**GTGTGTATATATAGCACATATATAT**GGT**GTGTGTATATAT**
**AGCACATATATAT**ATGGTGTGTGT**ATATAT**ATATGGT**GTGTGTATATATAGCACATATA**
**TAT**

>chr2_2_1 | NC_000002.12 | Start: 47310846 | Stop: 47311074 | Probability: 0.997 | Num. of predicted repeat: 4 – Repeat A, Repeat B**, Consensus,** Mismatch

**ATATACACACACACATATATATTTAGAGA**GAGA**ATATTGTGAGTGTGGTGTGTGTG**
**TGTGAGAT**ATATTTAC**ACACATACACACACATATATATAGAGAGAATATTATATGTG**
**CATGTGATGTGTGTCTGAGATATATTTACACACATATATATATAGAGA**GAGA**ATATT**

ATGTGTGTGGTGTGTGTGTGTGAGAGATATTTATATACACACACATATATATATATATC

# CHAPTER FOUR

# THE DIVERSE EVOLUTIONARY HISTORIES OF DOMESTICATED METAVIRAL CAPSID GENES IN MAMMALS

## Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: William S. Henriques

Contributions: Conceptualization, methodology, investigation & data collection, bioinformatics analysis, writing – original draft, writing – review & editing

Co-Author: Janet M. Young

Contributions: Methodology, investigation & data collection, bioinformatics analysis, writing – review & editing

Co-Author: Artem Nemudryi

Contributions: Conceptualization, investigation & data collection, writing – review & editing

Co-Author: Anna Nemudraia

Contributions: Conceptualization, investigation & data collection, writing – review & editing

Co-Author: Blake Wiedenheft

Contributions: Conceptualization, methodology, writing – original draft, writing – review & editing

Co-Author: Harmit S. Malik

Contributions: Conceptualization, methodology, writing – original draft, writing – review & editing

Manuscript Information

William S. Henriques, Janet M. Young, Artem Nemudryi, Anna Nemudraia, Blake Wiedenheft & Harmit S. Malik

*Molecular Biology and Evolution*

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

Oxford University Press
Submitted 25 January 2024
Published 20 March 2024
Volume 41, Issue 4
https://doi.org/10.1093/molbev/msae061

## Abstract

Selfish genetic elements comprise significant fractions of mammalian genomes. In rare instances, host genomes domesticate segments of these elements for function. Using a complete human genome assembly and 25 additional vertebrate genomes, we re-analyzed the evolutionary trajectories and functional potential of capsid genes domesticated from *Metaviridae*, a lineage of retrovirus-like retrotransposons. Our study expands on previous analyses to unearth several new insights about the evolutionary histories of these ancient genes. We find that at least five independent domestication events occurred from diverse *Metaviridae*, giving rise to three universally-retained single-copy genes evolving under purifying selection and two gene families unique to placental mammals with multiple members showing evidence of rapid evolution. In the *SIRH/RTL* family, we find diverse amino-terminal domains, widespread loss of protein-coding capacity in *RTL10* despite its retention in several mammalian lineages, and differential utilization of an ancient programmed ribosomal frameshift in *RTL3* between the domesticated capsid and protease domains. Our analyses also reveal that most members of the *PNMA* family in mammalian genomes encode a conserved putative amino-terminal RNA-binding domain both adjoining and independent from domesticated capsid domains. Our analyses lead to a significant correction of previous annotations of the essential *CCDC8* gene. We show that this putative RNA-binding domain is also present in several extant *Metaviridae,* revealing a novel protein domain configuration in retrotransposons. Collectively, our study reveals the divergent outcomes of multiple domestication events from diverse *Metaviridae* in the common ancestor of placental mammals.

## Introduction

At least half of the human genome is derived from selfish genetic elements [1–4]. Transposons are selfish genetic elements that encode key proteins needed to replicate and spread independent of host control[5–7]. In the human genome, a small number of transposons retain the capacity to 'jump' to new regions of the genome, whereas the majority are found immobilized and in various stages of degradation[2,8]. New transposon insertions can be harmful; they can interrupt host genes, alter the expression patterns of adjacent genes, or lead to ectopic recombination [9]. Although some new insertion events in somatic cells are implicated in disease [10,11], most insertions are functionally inconsequential[12,13].

Although somatic insertions in multicellular organisms are an evolutionary 'dead-end', transposon insertions in the germline are heritable. These insertions become a substrate for natural selection. Insertions that are deleterious to host fitness are rapidly eliminated from the population by purifying selection, whereas insertions that are inconsequential accumulate mutations at rates consistent with genomic mutation rates (reviewed in [14,15]). However, on rare occasions, pieces of mutationally decaying transposons acquire host functions and are subsequently protected from mutational abrasion by a process known as gene domestication, exaptation, or co-option[16,17]. After acquiring host function, such transposon-derived regions are no longer capable of autonomous replication and evolve just as any other host gene subject to purifying selection.

In mammalian genomes, the most common class of selfish elements are retrotransposons, which generate new copies using a 'copy-and-paste' mechanism[18–21]. Retrotransposons are classified into two broad groups: long-terminal-repeat (LTR) retrotransposons and non-long-

terminal-repeat (non-LTR) retrotransposons, which include both autonomous LINE-1 retrotransposons and non-autonomous Alu and SVA elements that rely on LINE-1 machinery. LTR retrotransposons include endogenous retroviruses (ERVs)[22], as well as four other families of LTR retrotransposon from the order of reverse-transcribing viruses, *Ortervirales: Metaviridae* (formerly Ty3/Gypsy)*, Caulimoviridae, Belpaoviridae,* and *Pseudoviridae* (formerly Ty1)[18]. LTR retrotransposons encode structural group-specific antigen (*gag)* genes and enzymatic polymerase (*pol*) genes, flanked by LTRs. Their 'exogenous' (transmissible) counterparts also encode an envelope protein (*env*)[23,24], which mediates membrane fusion needed for infection, and can subvert immune defenses via an immunosuppressive domain[25].

ERV genes have been repeatedly domesticated in mammals for diverse host functions, including reproduction and viral defense. The *syncytin* env-like genes of placental mammals present a spectacular example of domestication, mediating membrane fusion events needed to form multinucleated syncytial trophoblast cells in the placenta[26,27]. *Syncytin* gene domestication is a remarkable example of convergence, with at least seven independent events in different mammalian orders[27,28], and in unusual lineages of lizard[29] and fish[30]. In addition to domestication for placental function, multiple retroviral envelope genes have also been domesticated for retroviral defense in mice[31,32], humans[33], sheep[34], cats[35,36], and chickens[37]. These co-opted envelope genes illustrate two key concepts about gene domestication. First, the same viral protein can be independently repurposed for similar functions in different hosts. Second, domesticated copies of a single viral domain can serve diverse functions even within a single host.

The *gag* genes of reverse-transcribing viruses have been best studied in retroviruses such as HIV-1, and to a lesser extent in *Metaviridae* such as Ty3. Gag genes encode a polyprotein that includes capsid, matrix, and nucleocapsid domains that together package the viral genome during the virion assembly process[21,38]. Like *env*, *gag* has been domesticated both for retroviral restriction in mice (*e.g.,* the *Fv1* gene[39–41]), as well as for other critical host functions[16,42–45]. At least four domesticated capsid genes have been shown to assemble into capsid-like structures[46–50], which perform essential functions in host reproduction[43,44] and neuronal function[46,51]. One of these *gag*-derived genes is vertebrate *Arc (A*ctivity-*R*egulated, *C*ytoskeletal-associated), which originated from an ancient retroelement from the clade *Metaviridae* (formerly known as Ty3/Gypsy). ARC protein forms capsid-like structures and functions in the brain to regulate learning and memory as a signaling hub and messenger RNA shuttle in neurons[46,51]. An independent domestication event from a distinct *Metaviridae* lineage led to *dArc1* in *Drosophila* species, which also forms capsid-like structures and packages mRNA for intercellular neuronal signaling[48,52]. Together, *Arc* and *dArc1* elegantly demonstrate the functional convergence of independently domesticated genes in animal lineages. Although other domesticated *gag* genes also encode proteins capable of assembling into capsid-like structures, most remain functionally uncharacterized[50].

Both independent ancient *Metaviridae* domestication events and post-domestication duplications gave rise to over two dozen domesticated *gag*-like genes in the human genome, including *ARC* [42,45] and two small gene families, often referred to as the *PNMA* (Paraneoplastic Ma antigens) and *SIRH/RTL* (Sushi-Ichi-related Retrotransposon-Homolog/RetroTransposon-Like) families[16,42,53]. Among the best characterized of these genes are *Peg10* (Paternally

Expressed Gene 10, a *SIRH/RTL* family member) and *Rtl1* (retrotransposon-like 1)[54–56], both of which encode proteins that contain capsid domains and are essential for successful embryonic development[43,44,49]. Like ARC, PEG10 viral-like particles (VLPs) can package nucleic acid; PEG10 VLPs have been repurposed as delivery vehicles for custom nucleic acid cargos with potential therapeutic applications[57,47,49]. While roles are emerging for some *PNMA* and *SIRH/RTL* genes[47,58–63], most remain functionally uncharacterized. Similarly, in-depth evolutionary characterization has only been performed for a few of these genes, revealing a mixture of retention and lineage-specific gene loss in some species [16,59,61,62], which is surprising given the essentiality of orthologs of these genes in mice.

Given the important biological functions of domesticated capsid-like human genes, and their therapeutic potential, we sought a more in-depth structural and evolutionary understanding of these genes, focusing on different fates of domesticated genes. Building on previous evolutionary surveys across the family[42,45], and on detailed surveys of individual genes[16,59,61]. We also took advantage of state-of-the-art tools for structure prediction and homology detection - AlphaFold, Foldseek, and DALI[64–66] - and the availability of the first telomere-to-telomere assembly of the human genome and 25 additional vertebrate genomes, to carry out a systematic bioinformatic study of the architecture, evolution, expression, and predicted protein structure of capsid-derived sequences.

We found that most of the approximately two dozen *bona fide* domesticated metaviral capsid-like genes show clear signatures of purifying selection. About half of these have been strictly retained across placental mammals, suggesting functions common to all lineages. Other genes have undergone lineage-specific gene duplication, pseudogenization, or loss, implying

lineage-specific retention or loss of functions. Most striking among these is the *RTL10* gene, whose protein-coding capacity has only been retained in a subset of mammalian lineages. For a small subset of domesticated genes, we find evidence of positive selection, indicating their involvement in ongoing genetic conflicts or acquisition of novel functions. While re-examining the domain architecture, we revise and clarify structural features associated with domesticated genes. We find that the *RTL3* domesticated genes encode either a capsid-protease or separately retained capsid and protease genes in different mammalian lineages. We find no evidence for a canonical retroviral matrix-like domain neighboring the capsid in either the domesticated capsid genes or their active metaviral relatives. Instead, we find that amino-terminal domains (NTD) are widely divergent between domesticated capsid gene families, reflecting their domestication from at least four distinct *Metaviridae* families with distinct domain architectures. Among these divergent NTDs, we identify a novel, putative RNA-binding domain (RBD) encoded by the *PNMA* family, which has undergone dramatic expansion in mammalian genomes and can be retained either with or without an associated capsid domain. Our study reveals that recurrent domestication of the *gag* domain from structurally diverse *Metaviridae* gave rise to genes with distinct evolutionary trajectories and structural features in placental mammals.

## Results

### Divergent Metaviral Capsid-derived Genes in the Human Genome

Previous analyses identified 85 human genes with homology to retroviral- or retrotransposon-encoded *gag* genes[42], demonstrating that *gag* genes have been recurrently repurposed for host function. We updated this analysis using the newly available and complete

Telomere-to-Telomere (T2T) human genome assembly (including the Y chromosome)[67,68]. We focused on the capsid domain of the *gag* gene, which shares homology across reverse transcribing viruses[69]. To ensure capture of remote homologs, we built Hidden Markov Models (HMMs) guided by atomic structures and AlphaFold predictions of retroviral and retrotransposon capsid domains. We generated a separate HMM for each of the three clades of endogenous retroviruses (*Orthoretrovirinae)* as well as *Spumavirinae*[70] and the three major clades of LTR retrotransposons found in vertebrates: *Metaviridae* (previously known as Ty3/gypsy), *Pseudoviridae* (previously Ty1/Copia), and *Belpaoviridae*[18,69] (see Methods, **Supplementary Figure 4.1, Supplementary Table 4.1**).

Querying a six-open reading frame translation (between stop codons) of the T2T assembly with our custom capsid-specific HMMs, we identified 3,140 discrete capsid-like sequences from *Orthoretrovirinae* and *Metaviridae*, but none from *Pseudoviridae*, *Belpaoviridae,* or *Spumavirinae,* consistent with previous reports[71,72]. Our survey found 24 of the 85 previously identified metaviral-derived capsid genes[42]. The major reason for this disparity is that the previous study contained SCAN domains, which have been proposed to originate from ancient retrotransposon *gag* domains[73–76]. However, none of our HMMs for the capsid domain detected the numerous known SCAN-domain-containing proteins in the human genome. To keep our analysis focused, we did not build separate HMMs for the SCAN domain.

Of the 3,140 discrete capsid-like sequences, we found only 467 capsid sequences that contain both a start codon and have the potential to encode a full-length capsid domain, whereas the remaining 2,673 contain premature stop codons or frameshift mutations and are unlikely to encode proteins capable of assembling into functional capsids. To understand the relationship

between these 467 sequences, we aligned them to each other. This alignment allowed us to identify and remove poorly aligning sequences likely due to internal deletions, ultimately yielding 212 well-aligned open-reading frames that encode a full-length capsid domain. We used this alignment to construct an unrooted maximum-likelihood phylogenetic tree (**Figure 4.1**).

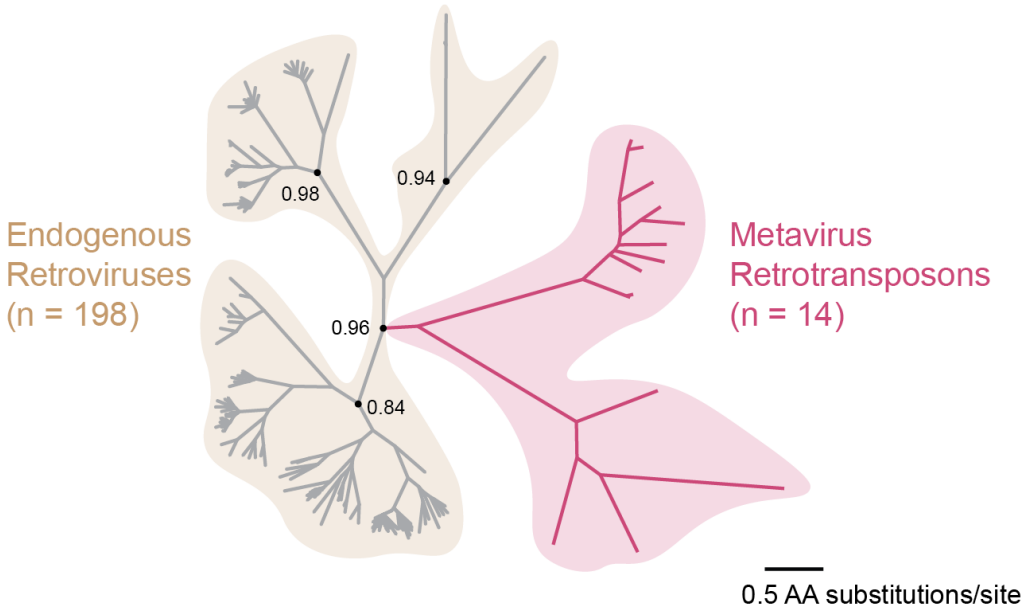



Figure 4.1. Full-length capsid-like ORFs in the human genome. We generated a maximum-likelihood phylogenetic tree of 212 full-length capsid-like ORFs in the human genome from an alignment of the full-length capsid domain (238 positions). Hidden Markov Model profile searches identified capsid sequences from endogenous retroviruses (grey branches), and metavirus retrotransposons (red branches). Maximum likelihood-based support values at selected nodes were calculated in FastTree2 (Price et al. 2010). The multiple sequence alignment and Newick files are available in the Supplementary Data.

We found a notable difference in phylogenetic branching patterns between endogenous retroviruses and *Metaviridae*. The majority of endogenous retroviral ORFs cluster with similar sequences related by short branches. This branching pattern is most consistent with differences that arose during recent autonomous replication[14]. However, since it is difficult to distinguish recent insertions from evolutionarily young domesticated genes, we cannot exclude the

possibility that some ERV-derived capsid sequences have been recently domesticated. In contrast, the single clade representing capsid-like sequences from *Metaviridae* has significantly fewer members, most of which are separated by long branches. This finding is consistent with previous analyses, which showed that *Metaviridae* ceased active transposition in an ancient ancestor of modern mammals[14,16,71]. Thus, any remaining metaviral capsid-like sequences are much more likely to be the result of host domestication[53,77]. Indeed, we found no adjacent long-terminal repeats flanking genomic sequences of metaviral capsids genes using LTRHarvest [78], confirming that these sequences do not represent retrotransposons capable of autonomous transposition. For this reason, we focus the remainder of our study on the 24 intact metaviral-derived capsid-like sequences that remain in the human genome (we also identified two additional pseudogenes, one related to *PNMA6E/PNMA6F* and the other related to *LDOC1*). Our phylogenetic tree (**Figure 4.1**) only contains 14 of these 24 *Metaviral* sequences because we excluded duplicates and partial capsid genes. However, in subsequent analyses, we included both full-length and partial capsid-like sequences because previous studies have shown that even domesticated metaviral genes encoding a truncated capsid domain can still be functional[42].

Our analyses confirm that previous catalogs of intact human domesticated metaviral capsid genes[42,45] were complete. These 24 human metaviral capsid-like sequences correspond to almost all of the previously identified domesticated capsid genes[42,45], including *ARC*, and members of the *PNMA* and *SIRH/RTL* gene families. The only exception is *ASPRV1* (or SASPase). A previous study identified *ASPRV1* as encoding a capsid-like sequence[42]. However, a subsequent study suggested it only encodes protease but not capsid domains[45]. Our analysis (both HMM analyses and AlphaFold predictions) confirm that *ASPRV1* encodes a protease domain but

not a capsid domain. Previous studies of the *PNMA* gene family also identified four additional members of the family that did not appear among the capsid-like sequences we identified: *CCDC8, PNMA8A, PNMA8B and PNMA8C*[79,80]. We independently identified these genes by examining sequence adjacent to capsids (see below) and included them in all subsequent analyses because of their close evolutionary ties to the capsid domain in the *PNMA* family. Multiple gene nomenclatures have been adopted by different studies over the years. To avoid confusion, we list all alternative gene names to help cross-reference previous publications with our results (**Supplementary Table 4.2**).

## Domesticated Metaviral Capsid Genes Experienced Distinct Evolutionary Retention Fates

Previous studies showed that the 24 human metaviral capsid-encoding genes were domesticated in an ancient mammalian ancestor[53] and have subsequently been conserved in at least a few mammalian species[16,59,61,81]. We investigated their evolutionary retention in a deeper sampling of representative placental mammalian genomes to identify any lineage-specific changes and to investigate their evolutionary constraints and origins. For this, we performed searches of 17 additional representative mammalian genomes (14 placental mammals, two marsupials, and one monotreme – **Supplementary Table 4.3**) using sensitive HMM searches as well as iterative blast searches with both nucleotide and protein queries. We assigned the resulting sequences to orthologous groups using phylogenies and/or sequence similarity. We generated in-frame nucleotide alignments for each group (Methods, see Supplementary Materials for alignment files) and looked for inactivating frameshifts and/or premature stop codons relative to the annotated human ORF. In all cases, the full-length human annotation is well-supported by conservation in distantly related placental mammal genomes (Supplementary Datasets).

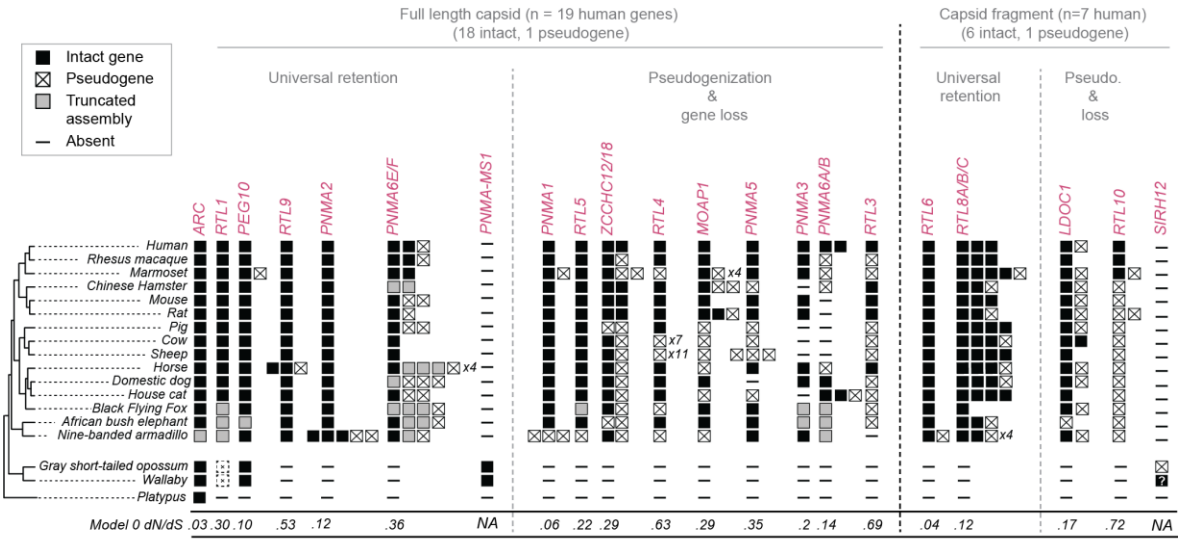



Figure 4.2. Metaviral-derived capsid genes show distinct evolutionary trajectories across placental mammals. Using phylogenetic trees (Figure S2-S7), we assigned each metaviral-derived capsid gene to one of 19 orthologous groups (labeled in red and Table S2) that are restricted to placental mammals, or to two marsupial groups (PNMA-MS1 and SIRH12). Some full-length metaviral-derived capsid genes are universally retained across placental mammals as intact genes (n), whereas others have experienced lineage-specific loss (-), pseudogenization (x) or duplication events (a second n or x within a column). Boxes containing an "X" (i.e., x) represent sequences with obvious inactivating mutations (frameshifts and/or premature stops). Gray boxes represent sequences that are truncated by gaps in the assembly. The two pseudogenes depicted for RTL1 in opossum and wallaby (dashed squares with crosses) were previously reported as gene fragments (Edwards et al. 2008). Our analysis (performed using similar methods) did not reveal convincing RTL1 homology in marsupial genomes. Most sequences are represented by individual boxes, but in cases where pseudogenized duplicates are numerous, the number of pseudogene duplicates is represented as xN (e.g., PNMA6E/F has 4 pseudogenized duplicates in horse). The status of SIRH12 in wallaby is denoted with a "?" to indicate uncertainty. The previously reported wallaby SIRH12 ORF (Ono et al. 2011) (with an exact match to the macEug2 version of the reference genome assembly) encodes a 107 amino acid protein, but contains a frameshifting change in a newer assembly (mMacEug1). SIRH12's ortholog in opossum is a pseudogene. The previously reported PNMA-MS2 pseudogene (Iwasaki et al. 2013) is not shown, because there are no apparently intact representatives of this sequence. The multiple sequence alignments and Newick tree files used to assign orthologous groupings are available in the Supplementary Data

Our phylogenomic analyses (**Figure 4.2, Supplementary Figures 4.2-4.7**) confirm previous estimates of the age of these genes[45,46,53,82,83]. Most domesticated metaviral capsid genes are at least ~100 million years (My) old, with orthologs found in shared syntenic locations in

diverse placental mammals but not in marsupials. *PEG10* is slightly older (at least ~160 My), with additional marsupial paralogs that have been previously described[84,85], whereas *ARC* is much older (at least ~350 My), with orthologs found across all tetrapods including birds, reptiles, and amphibians (**Figure 4.2**, **Supplementary Figure 4.2**). Despite their ancient origins, the repertoire of domesticated genes has not remained static across placental mammals. We found that four orthologous groups in our trees (*PNMA6E/F, ZCCHC12/18, PNMA6A/B*, and *RTL8A/B/C*; **Figure 4.2, Supplementary Figure 4.3-4.5**) contain more than one human gene (with orthologs present in other species), indicating that more recent lineage-specific duplication events continue to shape some branches of the gene family. For example, *PNMA6E* and *PNMA6F* appear to be the result of a duplication in the common ancestor of primates (**Supplementary Figure 4.5**). Primates are not unusual in this regard; many other mammal lineages also appear to have undergone independent duplications in their *PNMA6E/F* family (**Supplementary Figure 4.5**). An alternative possibility is that the independent, lineage-specific *PNMA6E/F* duplications could instead be the result of recurrent, lineage-specific gene conversion events, reminiscent of the evolutionary dynamics previously described in histone variant genes [86] and antiviral IFIT genes [87]. After accounting for lineage-specific duplications, the 24 human domesticated metaviral capsid-like genes group into 19 clades, which likely represents the number of domesticated capsid genes in the common ancestor of all placental mammals.

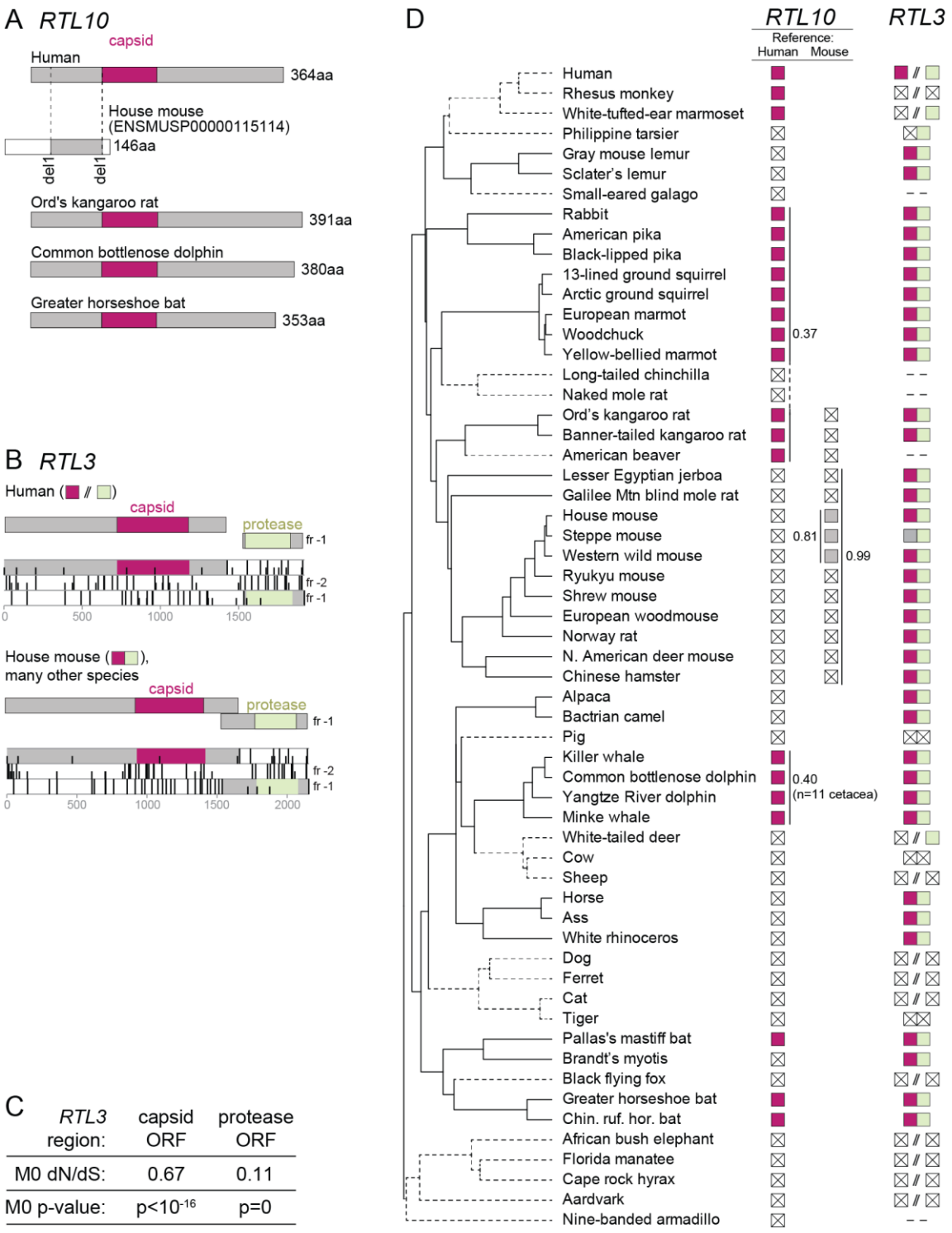



Figure 4.3. RTL3 and RTL10 exhibit domain-specific and lineage-specific patterns of conservation. A. 364aa of the capsid-containing RTL10 protein are conserved in human, some rodents, cetacea, and some bats, while the annotated mouse RTL10 protein is only 146aa and the ORF has two frameshifting 1 bp deletions with respect to human RTL10, resulting in limited

protein-coding homology and the absence of a predicted capsid domain. B. Simplified depictions of RTL3 showing capsid homology (magenta) and protease homology in the -1 reading frame (light green). In mouse and many other species, the two ORFs overlap and therefore likely encode a capsid-protease fusion protein via a programmed ribosomal frameshift. In contrast, in human, the two ORFs do not overlap, and either encode separate proteins, or represent functional loss as seen in other simian primate genomes. Three-frame ORF analysis of RTL3 in human and mouse showing ATG start codons (short vertical bars) and stop codons (tall vertical bars). HMM homology is shown in magenta/green, and stop-free regions containing each domain are shown in gray. C. RTL3 ORFs are subject to purifying selection for protein-coding function, with dN/dS of <1, and maximum likelihood tests ("M0 p-value") that support the dN/dS shown over the null model of neutral evolution. dN/dS values reported here differ from those reported in Figure2/Figure 4 due to the use of additional sequences. D. Summary of RTL10 and RTL3 status in an expanded set of mammalian genomes. The species tree is a trimmed version from Upham et al. 2019. Branches depicted using dashed lines show lineages where the RTL3 capsid-protease fusion protein has been lost. Filled squares at the terminus of each branch represent intact ORFs and squares containing a cross represent sequences with obvious inactivating mutations (frameshifts and/or premature stops). Gray boxes represent sequences that are truncated due to genome assembly gaps, and "-" symbols indicate that we found no matching sequence. For RTL10, human and mouse encode different ORFs due to frame-shifts: the first column ("Human") shows that a capsid-containing ORF is retained in a limited number of diverse mammalian clades, and is lost or pseudogenized in many other genomes. The second column ("Mouse") shows that even closely-related rodent genomes do not preserve the ORF found in mouse. Vertical lines with adjacent numbers show the dN/dS of these ORFs in selected clades. The multiple sequence alignments used for the PAML analyses are available in the Supplementary Data.

Of these 19 clades, eight are universally retained as at least single intact genes across all placental mammals queried, suggesting they confer an important biological function common to all mammals (**Figure 4.2;** *ARC*, *RTL1*, *PEG10*, *RTL9*, *PNMA2* and *PNMA6E/F*, *RTL6* and *RTL8*). Not all these eight universally conserved genes contain full-length capsid-like regions (i.e., *RTL6* and *RTL8*), confirming that a full-length capsid-like domain is not necessary for the new host-related function of some of these genes[88]. In contrast to the eight universally retained genes, we found that 11 of the domesticated full-length and fragmented capsid genes that were in the common ancestor of placental mammals have since undergone gene loss or pseudogenization in one or more mammalian lineages, suggesting that their function is not universally required in

mammals or is redundant with other genes (**Figure 4.2**). For example, *RTL10* was present in the common ancestor of placental mammals but is retained as an intact ORF in only the three out of eighteen mammalian genomes we surveyed. Even though *RTL10* has been repeatedly lost during mammalian evolution, deeper examination in additional species (**Figure 4.3A,D**) reveals that *RTL10* is also retained under purifying selection as an intact gene in at least three diverse mammalian linages: primates, basal glires, and cetaceans (see below).

Our findings are contingent on the caveats associated with reliable genome assembly. Despite significant advances since the last compendium of domesticated genes was created, most of these genome assemblies are still incomplete, leading to cases where an assembly gap results in a truncated sequence (*e.g.,* armadillo *ARC*). This problem is compounded in the case of recent duplicates, which are likely undercounted in *de novo* assemblies that can collapse highly similar sequences (**Figure 4.2**). Finally, genome assemblies contain small numbers of sequencing errors, so a minority of the apparent pseudogenes could represent intact genes with sequencing errors. Given these caveats, where possible, we used additional sequences from sister lineages to gain more confidence that inactivating mutations are not simply genome assembly or sequencing errors.

## Domesticated Metaviral Capsid Genes are Retained under Purifying and Positive Selection

The cadence and locations at which amino acid changes accumulate during evolution can also provide important clues about a protein's cellular function. We investigated the selective pressures on metaviral genes following their domestication (**Figure 4.2, Figure 4.4**), estimating the ratio of non-synonymous (dN, amino acid altering) to synonymous (dS, amino acid preserving) nucleotide substitutions using the codeml algorithm from the PAML suite[89]. Most

protein-coding genes exhibit low dN/dS (<< 1), indicating purifying selection, where non-synonymous mutations are less likely to be tolerated than synonymous mutations. dN/dS close to 1 indicates neutral evolution (no protein-coding constraint), and dN/dS >1 indicates positive selection (adaptive evolution favors amino acid changes).

We first estimated the overall dN/dS of alignments of intact ORFs from across placental mammals (up to n=18 species), using PAML's model 0 that assumes uniform selective pressures across all sites and all lineages. We compared the likelihood of this observed dN/dS to that of a model 0 with fixed dN/dS of 1, allowing us to test whether the observed dN/dS ratio was a better fit than neutral evolution. We find that most *Metaviridae*-origin capsid genes evolve under overall purifying selection, with *ARC* showing the strongest purifying selection (dN/dS=0.03). *RTL6* also shows strong purifying selection (dN/dS=0.04), despite its involvement in immune functions (which is often associated with rapid evolution)[61].

In contrast, for *RTL10*, we were barely able to rule out the null hypothesis of neutral evolution (dN/dS=0.72, close to the neutral ratio of 1) using our initial sampling (n=3 intact primate genes). However, our deeper alignments for *RTL10* provided greater statistical power, revealing evidence of purifying selection in basal glires (rodents/lagomorphs) and in cetaceans (whales/dolphins), despite pseudogenization in other lineages (**Figure 4.3A,D**). It remains possible that in some lineages, *RTL10* could have functions that are unrelated to its protein-coding capacity, for example as a non-coding RNA. Indeed, data from the International Mouse Phenotyping Consortium (see below) indicates that mouse knockouts of *RTL10* have both behavioral and morphological phenotypes, despite apparent frameshifting mutations in the capsid-encoding region of the mouse gene (**Figure 4.3A**). Nevertheless, *RTL10* represents an

| Clade | Gene | Chrom. | M0 dN/dS (Placentals) | M0 vs M0 fixed p-value (Placentals) | Primate Retention | M0 dN/dS (Primates) | M8 vs M8a p-value (Primates) | M8 % selected | dN/dS of selected sites | Sites with BEB >= 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| b | ARC | 8 | 0.03 | 0 | 25 | 0.013 | 1 | NA | NA | NA |
| c | RTL1 | 14* | 0.30 | 0 | 25 | 0.192 | 1 | NA | NA | NA |
| | PEG10 | 7* | 0.10 | 0 | 25 | 0.092 | 0.1 | NA | NA | NA |
| | **RTL3-CA** | **X** | **0.69** | **3.7E-4** | **4** | **1.862** | **6.7e-04** | **25.9** | **5.5** | **9[a]** |
| | RTL3-PR | X | 0.13 | 0 | 8 | 0.785 | 0.08 | NA | NA | NA |
| | **RTL9** | **X** | **0.53** | **0** | **24** | **0.74** | **1.7e-07** | **0.1** | **17.8** | **2[1] (Q25, K489)** |
| d | **RTL4** | **X** | **0.63** | **0** | **17** | **1.06** | **1.3e-04** | **3.6** | **9.6** | **4 (P32, C71, E185, T303)** |
| | **RTL5** | **X** | **0.22** | **0** | **25** | **0.378** | **0.00152** | **0.6** | **7.1** | **2 (R353, R365)** |
| | RTL6 | 22 | 0.04 | 0 | 25 | 0.064 | 1 | NA | NA | NA |
| | LDOC1 | X | 0.17 | 0 | 25 | 0.058 | 1 | NA | NA | NA |
| | RTL10 | 22 | 0.72 | 0.14 | 25 | 0.605 | 0.99 | NA | NA | NA |
| | RTL8A/B | X | 0.12** | 0 | 25 | 0.082 | 0.95 | NA | NA | NA |
| | RTL8C | X | 0.12** | 0 | 23 | 0 | 1 | NA | NA | NA |
| | PNMA1 | 14 | 0.06 | 0 | 25 | 0.148 | 0.23 | NA | NA | NA |
| | MOAP1 | 14 | 0.29 | 0 | 25 | 0.356 | 0.77 | NA | NA | NA |
| | PNMA2 | 8 | 0.12 | 0 | 25 | 0.208 | 1 | NA | NA | NA |
| | PNMA3 | X | 0.2 | 0 | 24 | 0.279 | 1 | NA | NA | NA |
| e | **PNMA5** | **X** | **0.35** | **0** | **25** | **0.552** | **1.7e-04** | **5.5** | **3.3** | **9[b]** |
| | PNMA6A/B | X | 0.14 | 0 | 5 | 0.245 | 1 | NA | NA | NA |
| | **PNMA6E** | **X** | **0.36***** | **0** | **15** | **0.535** | **0.04** | **2.1** | **4.3** | **1[&2] (K118)** |
| | **1-491** | | | **0** | **15** | **0.613** | **0.02** | **23.1** | **2.4** | **1[&2] (K118)** |
| | 492-1431 | | | 0 | 15 | 0.501 | 1 | NA | NA | NA |
| | PNMA6F | X | 0.36*** | 0 | 18 | 0.488 | 0.48 | NA | NA | NA[&] |
| | 1-152 | | | | 18 | 0.388 | 0.33 | NA | NA | NA[&] |
| | 153-1416 | | | | 18 | 0.495 | 0.9 | NA | NA | NA[&] |
| | ZCCHC12 | X | 0.29**** | 0 | 24 | 0.367 | 0.06 | NA | NA | NA[&] |
| | 1-344 | | | | 24 | 0.324 | 1 | NA | NA | NA |
| | **345-1221** | | | | **24** | **0.342** | **0.02** | **7.4** | **2.6** | **1[3] (R283)** |
| | ZCCHC18 | X | 0.29**** | 0 | 10 | 0.927 | 0.04 | 40.9 | 2.4 | None |
| | **CCDC8-ext** | **19** | **0.15** | **0** | **28** | **0.399** | **1.5e-05** | **11.6** | **2.5** | **5[c]** |
| | **PNMA8A** | **19** | **0.32** | **0** | **24** | **0.329** | **0.05** | **1.0** | **6.7** | **2[&4] (P319, P406)** |
| | 1-787 | | | | 24 | 0.251 | 0.06 | NA | NA | NA |
| | **788-1320** | | | | **24** | **0.435** | **0.04** | **1.1** | **8.6** | **2[4] (P56, P143)** |
| | PNMA8B | 19 | 0.15 | 0 | 25 | 0.161 | 0.48 | NA | NA | NA |
| | PNMA8C | 19 | 0.17 | 0 | 26 | 0.142 | 0.57 | NA | NA | NA |

[a](W25, Q34, A147, I149, P171, G212, Y259, W439 ,Q472)
[b](P83, E132, M139, R143, S252, I255, V259, K276, H406)
[c](K141, E390, H442, R462, S494)

*Imprinted gene
**RTL8A/B/C were treated as a single gene across placentals but split in primates
***PNMA6E/F were treated as a single gene across placentals but split in primates
****ZCCHC12/18 were treated as a single gene across placentals but split in primates
[1]Amino acids are from *Pan paniscus* sequence
[2]Amino acids are from *Theropithecus gelada* sequence
[3]Amino acids are from *Callithrix jaccus* sequence
[4]Amino acids are from *Colobus angolensis* sequence
[&] Alignment shows evidence of recombination by GARD. Following two rows show results for GARD-inferred segments, with the start and end coordinates indicated.

Figure 4.4. Evolutionary rates of domesticated metaviral genes across placental mammals and primates. We generated in-frame alignments for each gene across placental mammals, or across primates, and analyzed the evolutionary selective pressures on each using PAML's codeml algorithm. We used the placental mammal and primate alignments to assess 'overall' selective pressure, assuming a single dN/dS ratio across sites and lineages ("M0 = model 0"). Lower dN/dS ratios indicate stronger purifying selection and dN/dS=1 is neutral evolution. To test for positive selection, we analyzed primate alignments using PAML's codeml algorithm (codon model=2, initial dN/dS = 0.4, cleandata=0), but this time we compared the log likelihoods of an evolutionary model that allows for a subset of residues under positive selection (model 8) with a paired model that only allows purifying and neutral selection (models 8a). For genes where the

maximum likelihood test indicates positive selection, we report the proportion of sites estimated to be under positive selection, the dN/dS of this class of sites, and the identity of sites with a >90% posterior probability of being members of the positively selected class (codeml's "BEB" Bayes Empirical Bayes method). The multiple sequence alignments used for the PAML analysis are available in the Supplementary Data.

unusual case; for most domesticated genes, there is unambiguous evidence of their retention under purifying selection. Overall, our evolutionary analyses indicate that most domesticated metaviral capsid genes have been retained under strong selective pressures during the last 100 million years of placental mammal evolution, suggesting that they each play important, non-redundant roles in mammalian biology.

Despite an overall signature of purifying selection, domesticated capsid genes might still be subject to positive selection (dN/dS >1) at a subset of codons as seen for host immunity related genes such as *Fv1*[40,41]. To test this possibility, we collected sequences from the simian primate lineage, because this taxonomic group provides good statistical power to detect site-specific positive selection[90]. We analyzed these alignments using PAML's maximum likelihood tests that ask whether an evolutionary model that allows a subset of sites with dN/dS>1 (NSsites model 8) is a better fit to the observed sequence data than a matched model that only allows sites under neutral and purifying selection (dN/dS of 0-1) (NSsites model 8a). These tests reveal that ten of these nineteen ancient metaviral genes contain a subset of sites evolving under positive selection in primates (**Figure 4.4**), suggesting ongoing involvement in evolutionary arms races. Our findings of positive selection are intriguing given previous studies that implicate three of the positively selected genes (*CCDC8, RTL5, RTL9*) in immune-related functions[61,63,91], suggesting that host-pathogen genetic conflict may be driving the rapid evolution of at least some of these genes.

Domesticated Capsid Genes are Structurally Diverse

Next, we compared the domain architecture of metaviral-derived human genes to those of their metaviral progenitors, using contemporary methods for protein structure prediction and algorithms for detecting structural homology [64–66]. To identify metaviral relatives, we queried both the Repbase database, as well as six-frame translations of non-mammalian vertebrate lineages where *Metaviridae* might be still actively retrotransposing: chicken (*Gallus gallus),* alligator (*Alligator mississippiensis),* painted turtle (*Chrysemys picta bellii)*, anole lizard (*Anolis carolensis*), African clawed frog (*Xenopus laevis*), and coelacanth (*Latimeria chalumnae*) genomes. We aligned these metaviral capsid sequences with the mammalian capsid-like genes and generated a maximum-likelihood phylogenetic tree of the resulting 949-sequence alignment (**Figure 4.5A**).

Consistent with previous studies[42,45], our phylogeny broadly separates the domesticated metaviral genes into three groups: the *PNMA* family and the *ARC* gene (each represented by monophyletic clades), as well as a looser group containing the *SIRH/RTL* family members (including PEG10). Closer examination of the *SIRH/RTL* group subdivides that family into three monophyletic clades: *RTL1*, *PEG10*, and the remaining family members. Some previous publications have considered *RTL1* and the other *SIRH/RTL* genes as a single family[16,92]. However, this conclusion was based on analysis of fewer metaviral outgroup sequences. Our inclusion of a denser sampling of metaviruses reveals the independent ancestry of *RTL1, PEG10* and the remaining *SIRH/RTL* genes (**Figure 4.5A**, **Supplementary Figure 4.7**) and is consistent with other published in-depth analyses[42,45]. These five monophyletic clades indicate that the mammalian metaviral-derived genes derived from at least five independent domestication events. Three of these 5 clades are represented by single mammalian genes (*ARC, PEG10* and *RTL1*),

whereas two other clades (*PNMA* and *SIRH/RTL*) have expanded to contain many mammalian members, demonstrating that gene duplications (rather than independent retrotransposition events) probably led to the expansion of each gene family[42,45].

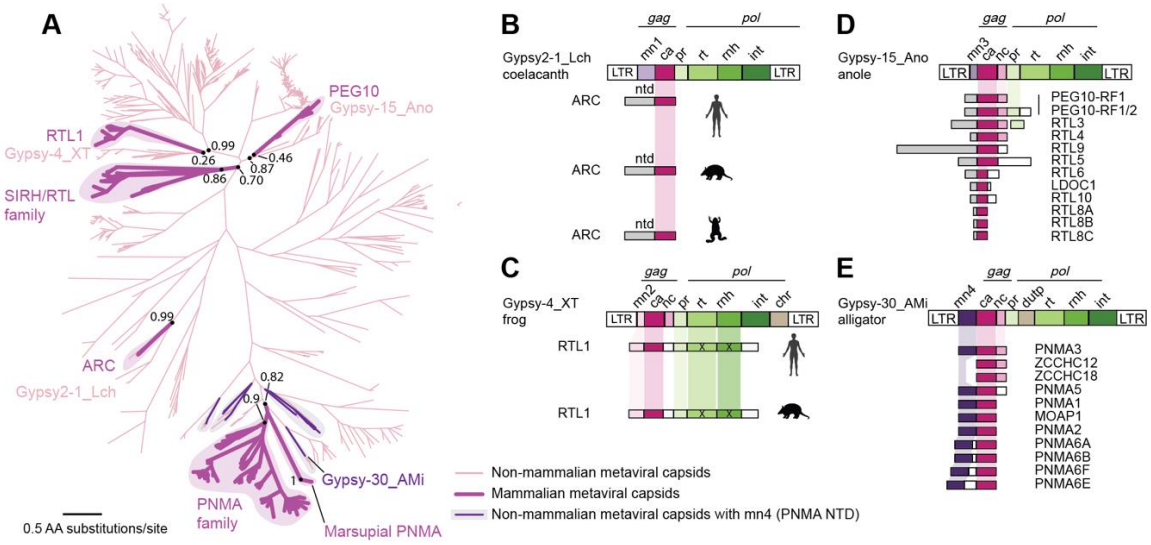



Figure 4.5. Four independent metavirus domestication events include structurally distinct N-terminal domains. A. A maximum likelihood phylogenetic tree of 949 capsid sequences from 24 vertebrate genomes and Repbase. 119 domesticated metaviral capsid genes in mammals (dark purple, highlighted) and 830 metaviral capsid-like ORFs (light pink) from selected non-mammalian vertebrate genomes: chicken (n=1), alligator (n=13), painted turtle (n=32), anole lizard (n=310), African clawed frog (n=291) and coelacanth (n=15) and consensus vertebrate metaviral elements from the database Repbase (n=175), aligned across 172 positions in the capsid domain. Maximum likelihood-based support values calculated in FastTree2. Metaviral gag genes containing the PNMA N-terminal domain are phylogenetic neighbors to mammalian PMNA family genes (grey highlight, deep purple lines). The closest consensus Repbase sequence for each domestication is indicated (XT – Xenopus tropicalis, African clawed Frog, Ano – Anolis carolensis, anole lizard, Lch – Latimeria chalumnae, coelacanth, AMi – Alligator mississippiensis, American alligator). B-E. Domain architecture (not to scale) of human Metaviridae-derived capsid genes, organized according to major clades in the tree shown in panel A. Colored boxes indicate domains within each open reading frame identified by HMM profile searches, structural prediction, and structural homology searches (mn - Metaviral N-terminus, numbered 1-4 to indicate four unique N-terminal domains in the visualized metaviruses, ntd - N-terminal domain, ca - capsid, nc - nucleocapsid, pr - protease, dutp - dUTPase, rt - reverse transcriptase, rnh – RNAaseH, int - integrase, chr – chromodomain, LTR – Long Terminal Repeat). The multiple sequence alignment and Newick files are available in the Supplementary Data.

To explore the domestication events giving rise to each clade, we selected a representative consensus Repbase sequence from the nearest metaviral clades in the tree (**Figure 4.5A**), considering these as 'proxies' for the (presumed ancestral) active metaviruses that were originally domesticated. In our tree, *PEG10* and the other non-*RTL1 SIRH/RTL* family members branch very closely and share the same ancestral proxy (Gypsy-15_Ano). As previously indicated, *PEG10* and the other non-*RTL1 SIRH/RTL* genes likely derive from distinct domestication events since they do not form a monophyletic clade on the tree, appear to have different ages, and have slightly different distribution, with *PEG10* also present in marsupials whereas the others are present in only placental mammals. Nevertheless, the phylogeny indicates they derived from similar, likely closely-related ancestral metaviral sequences. Because of these close relationships, we treat the *PEG10* and the other non-*RTL1 SIRH/RTL* genes as a single group despite their independent domestication (e.g., **Figure 4.5D**). We emphasize that these extant metaviruses are imperfect proxies for the true ancestral viruses. For example, since PEG10 contains a classic programmed ribosomal frameshift, its metaviral ancestor almost certainly did too; however, the ancestral proxy we selected for the PEG10 clade (Gypsy-15_Ano) does not contain a frameshift.

We closely examined the domain architectures of the four 'ancestral proxy' metaviral sequences and the domesticated genes that associate with each group (**Figure 4.5B-E**). Many of the domesticated metaviral capsid genes include additional domains, which sometimes include additional metaviral-derived regions, as well as protein segments with no recognizable viral homology[42,45]. We first focused on annotating non-capsid domains of 'ancestral proxy' metaviruses found in our genome scan and consensus sequences from the database Repbase[93]

using LTRHarvest, HMMER, PFAM (now part of Interpro), AlphaFold, Foldseek, and DALI[64–66,78,94,95]. We then compared the resulting patterns to those observed in the domesticated mammalian metaviral genes. For the domesticated genes, we also used these additional non-capsid HMMs to search six-frame translations of 16kb of flanking genomic sequence on each side of the capsid region, to rule out the possibility of any gene mis-annotation (especially of genes that might have unannotated programmed frameshifts) or subsequent insertions having split the original metaviral domestication. Although several of our annotations agree with previously published observations[42,45], our analyses also reveal new insights that significantly revise the proposed domain architecture for some domesticated genes. For completeness, we both summarize previous observations as well as discuss our new findings below.

First, while many metaviral gag-derived genes encode a complete capsid domain, several in the *SIRH/RTL* family encode a truncated capsid domain that has only the N-terminal lobe (*e.g., LDOC1*) (**Figure 4.5D**, **Supplementary Figure 4.6**). These truncations are seen across mammalian orthologs of each gene, indicating that they occurred soon after gene birth prior to the separation of the different mammalian orders. Notably, the AlphaFold predictions indicate that proteins encoded by these 'half-capsid' genes have absolutely conserved the structural fold of the N-lobe from the ancestral metavirus, whereas adjacent domains have diversified, and are typically predicted to be a mix of single alpha helices and disordered regions of varying lengths (schematized in **Figure 4.5**, full predictions in **Supplementary Figure 4.8**). A recent report suggests the N-lobe-encoding *RTL8* regulates the function of the full-length capsid-encoding *PEG10*[88]**.** Other truncated capsids may perform analogous regulatory functions of full-length capsid proteins.

Second, several *SIRH/RTL* family genes, as well as *RTL1*, encode full-length metaviral capsid domains (and in two cases, additional domains such as nucleocapsid, protease, reverse transcriptase, and RNaseH) fused to long N- or C-terminal regions that have low confidence structural predictions (gray boxes, **Figure 4.5C-D**). Our in-depth structural analysis finds no additional recognizable domains in these regions and no evidence that these regions are metaviral-derived. Although these regions are likely unstructured, they are nevertheless well conserved across placental mammalian orthologs, indicating their functional importance.

Third, *PEG10* encodes a protease domain after a -1 ribosomal frameshift that is nearly universally conserved across placental mammals; ~30% of translating ribosomes undergo the frameshift, resulting in two different protein products: a shorter majority product containing only the *gag*-related region, and a longer minority product that also contains a protease domain[57,96,97]. RTL1 proteins contain protease, and (likely catalytically inactive but structurally conserved) reverse transcriptase and ribonuclease H (RNase H) domains[98].

Like *PEG10*, a programmed frameshift and downstream protease domain have also been suggested for the *RTL3* gene[16,77], although the functional relevance or evolutionarily history has not been well-defined (*e.g.,* it is not noted in the RefSeq annotations). We used numerous additional placental mammal genomes to investigate *RTL3*'s putative programmed frameshift (**Figure 4.3B,D**). Although *RTL3* is clearly a pseudogene in several major mammalian clades (*Carnivora*, *Bovidae*), both protease and capsid-containing ORFs remain intact in many other lineages. In these instances, the protease domain is very well conserved and shows much stronger evidence of amino acid constraint (overall dN/dS=0.11) than the capsid-containing ORF (overall dN/dS=0.67; dN/dS of capsid domain alone=0.55) (**Figure 4.3C**). In the mouse *Rtl3*

ortholog, a simple -1 frameshift would result in a fusion protein that contains both capsid and protease domains. However, in human *RTL3*, a -1 frameshift would not be sufficient to produce a fusion protein, because the stop-free regions containing the two domains do not overlap and the intervening region contains stop codons in all three reading frames. Furthermore, in many other simian primate species, one or both domains of *RTL3* clearly acquired inactivating mutations in ancestral species (**Figure 4.3B,D**). The most likely explanation is that *RTL3* encodes a capsid-protease fusion protein via a programmed frameshift in mouse and many other mammalian lineages. In contrast, in human and other simian primates. *RTL3* encodes independent capsid and protease proteins. Similar putative programmed frameshifts have also been reported for *PNMA3* and *PNMA5*[99]. However, the sequences following the putative frameshifts are poorly conserved (unlike *PEG10* or *RTL3*) and we found no evidence for downstream functional domains; thus, their functional relevance remains unclear.

### Diverse N-terminal Structures in Domesticated Genes Derived from Metaviral Ancestors

The most surprising finding from our reannotation emerged from the analysis of the four 'ancestral proxy' metaviruses, each of which has a distinct domain architecture (**Figure 4.5B-E**). Of the four ancestral proxies, only the ARC-like metavirus lacks a nucleocapsid (NC) domain immediately downstream of the capsid domain. In contrast, a recognizable NC is present in the other three ancestral proxy metaviruses and has been maintained in some but not all domesticated genes in these three clades. NC domains bind viral RNAs to help package them into capsid[100]. Their presence in some domesticated genes suggests encoded proteins that are more likely to have RNA-binding capabilities, as has been demonstrated for PEG10[47,49]. However, ARC packages RNA even though it lacks a canonical nucleocapsid domain[46]. The four

‘ancestral proxy’ metaviruses also differ in other idiosyncratic ways; a chromodomain is found only in the *RTL1* ancestor, and dUTPase only in the *PNMA* family ancestor. However, neither the chromodomain nor the dUTPase was retained in any of the domesticated genes, so we do not study these domains further.

Previous analyses have described the N-terminal region of metaviral ORFs as matrix domains, by analogy to retroviral matrix domains that function to facilitate membrane association and virion formation[42,45,46]. However, this domain assignment has been based more on analogy than actual homology. Typically, retroviral matrix domains fold into a globular core composed of four alpha helices with a conserved basic surface patch that facilitates membrane interaction[101,102]. This helical core is also capable of binding tRNA, which regulates membrane association during virion formation[103]. We explored the domain architecture of both the four ‘ancestral proxy’ metaviruses and derived domesticated genes using a combination of BLAST, AlphaFold, DALI and Foldseek. We were surprised to find no evidence of a canonical retroviral matrix domain in any of the four metaviruses.

We found that the N-terminal domains (*i.e.,* upstream of the capsid) encoded by the *gag* gene of each of these four ‘ancestral proxy’ metaviruses are completely dissimilar to each other and surprisingly variable in length (**Supplementary Table 4.4**). According to AlphaFold, these domains are often predicted to be a single long alpha helix flanked by disordered regions (**Figure 4.5, Supplementary Table 4.4,** data not shown). Our analyses reveal a previously underappreciated degree of domain complexity in a poorly studied region of the *gag* gene of metaviruses and retrotransposons in general.

Given their distinct evolutionary origins and domain architecture, we examined each of the four clades of domesticated genes in more detail to infer functions. We first examined the ~350 My-old *ARC* gene, analyzing the mammalian orthologs described above (**Figure 4.2, Supplementary Figure 4.2**), as well as a more distant frog ortholog. Notably, ARC's entire gene architecture is conserved across species, not only the capsid domain. However, upon comparison of ARC with 'ancestral proxy' metaviral sequences from the coelacanth, we found homology between capsid domains using reciprocal HMM searches. However, N-terminal HMMs built from 11 coelacanth metaviral sequences were not homologous to N-terminal HMMs built from vertebrate ARC orthologs, which predict a ~200 amino acids long coiled-coil domain (schematicized in **Figure 4.5B)**. Furthermore, the ARC N-terminal region HMM did not reveal matches to any metaviral retrotransposons in the Repbase database. Additional efforts using HHpred[104], DALI[64], and Foldseek[66] also failed to clarify the evolutionary origin of this 200 amino acid N-terminal domain of ARC. We cannot rule out the possibility that this unusual N-terminal domain derived from a metaviral lineage that does not yet have a sequenced representative, especially since the domestication event is so old. Alternatively, it remains possible that the domain was acquired from non-viral sequences during or following domestication. However, we note that no homologous matches were found in any animal genomes outside *ARC* itself. Thus, while the capsid domain is unambiguously of metaviral origin, the origin of the ARC N-terminus remains mysterious and should be a focus of future analyses given its importance in ARC function[105].

The N-terminal regions of RTL1 and the *SIRH/RTL* family (and their closest metaviral relatives) are predicted to encode single alpha helices and disordered regions ranging from 25

residues (RTL8) to 1169 residues (RTL9) with no clear domain architecture. The N-termini of each gene are conserved across placental orthologs but are not similar between different genes in this family, suggesting they are functionally important domains, despite our inability to predict their function through structural homology (**Supplementary Figure 4.9**). This reveals that the N-terminus of the *SIRH/RTL* family has diverged widely since domestication, whereas the capsid core has retained nearly perfect structural homology to a metaviral capsid domain. This contrasts starkly with the fate of the N-terminus in the *PNMA* family, which contains a highly conserved N-terminus that we discuss in more detail in the next section.

## A Novel Putative RNA-binding Domain in the PNMA Family

Our structural modeling and similarity searches of the ~175 amino acid N-terminal region in the *PNMA* clade and its closest metaviral relative revealed a predicted ~100 amino acid RNA-binding domain (RBD) with mixed alpha-helical and beta sheet topology that was unrecognized in any previous analyses of either metaviral sequences or the domesticated *PNMA* gene family (**Figure 4.5E, Figure 4.6**). Using AlphaFold's structural prediction for the N-terminal domain from human PNMA1 as a query in DALI and Foldseek structural homology searches, we found high-scoring matches to RNA-binding domains found in many proteins, including an RNA recognition motif of the p65 subunit of the *Tetrahymena thermophila* telomerase complex (PDB: 7LMA)[106] and human MARF1 (meiosis regulator and mRNA stability factor 1, PDB: 2DGX, no structure-associated publication) (**Figure 4.6B**). Due to several missing residues in the Gypsy-30_Ami consensus sequence, we instead used a closely related genomic sequence (also from *Alligator mississippiensis*) for structural predictions (**Supplementary Figure 4.10A**). The

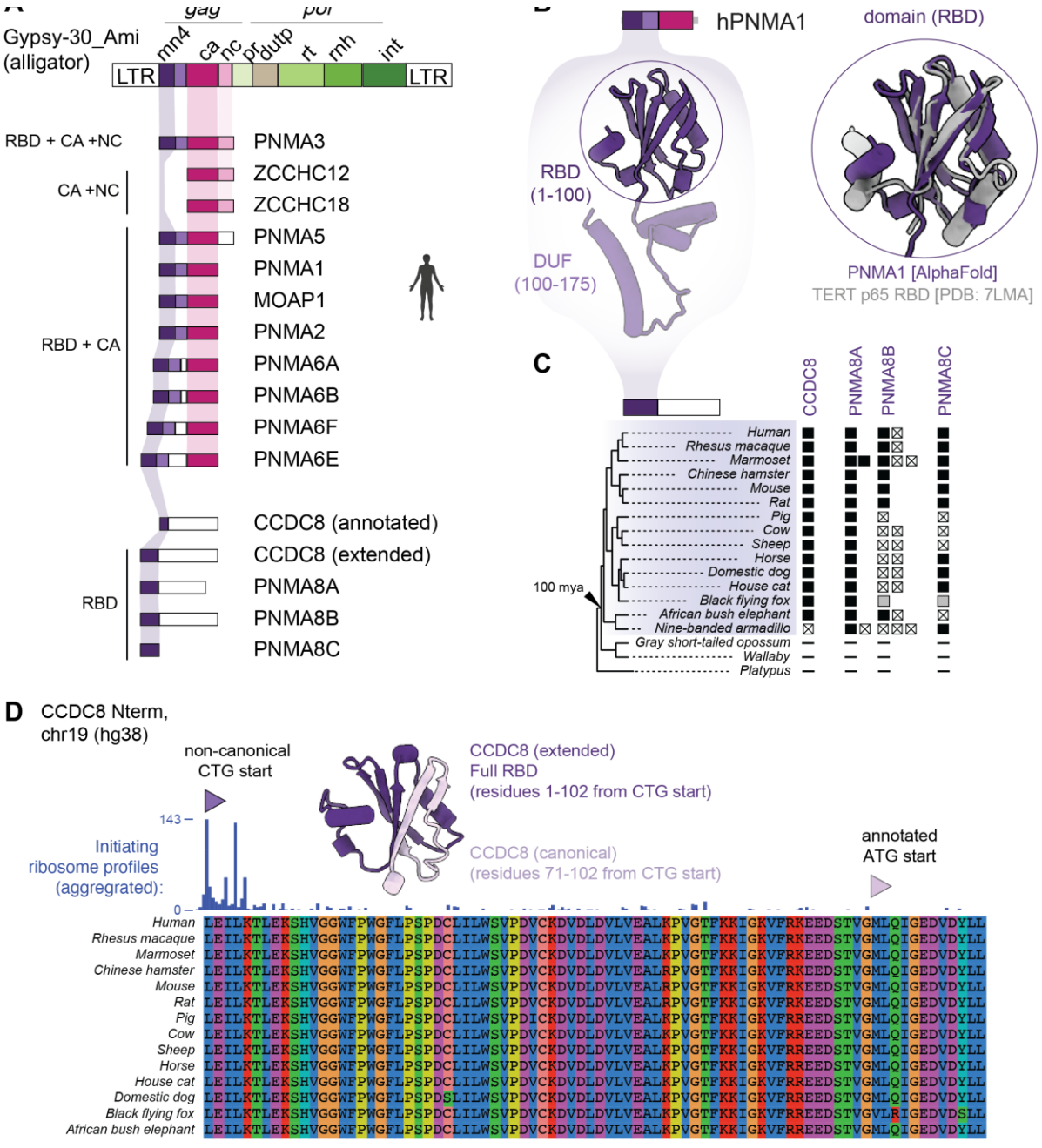



Figure 4.6. Predicted RNA-binding domain in the PNMA family and related metaviruses. A. Illustration of a metavirus from the alligator genome colored by domain. The first ~100 amino acids of related capsids in the human genome are predicted to form an RNA Binding Domain (RBD). This domain is also found independent of the capsid. B. AlphaFold structural prediction of PNMA1 (purple), shown alone (left) as well as superimposed on an experimentally determined structure (gray, right) of the RNA binding domain (PDB: 7LMA) from telomerase p65 (RMSD between 41 pruned C$\alpha$ atoms). C. Retention of putative RBD-only genes in

placental mammals. Filled squares represent intact genes and squares containing a cross represent sequences with obvious inactivating mutations (frameshifts and/or premature stops). Gray boxes represent sequences that are truncated due to genome assembly gaps, and '-' symbols represent cases where we find no matching sequence. A known species tree is shown on the left and was obtained by pruning a whole-genome tree available via the UCSC genome browser. D. CCDC8 translation initiates at an upstream non-canonical CTG start codon (dark purple). Aggregated data from many ribosomal profiling studies displayed via the GWIPS-viz genome browser (Michel et al. 2014) show an accumulation of initiating ribosomes at the non-canonical CTG start site (dark purple) rather than the canonical start site (light purple). The additional 70 N-terminal amino acids are highly conserved across mammals. The extended N-terminus is predicted (Alphafold) to encode a full-length RBD. The Alphafold predictions, multiple sequence alignments, and Newick files used to assign orthologous groups are available in the Supplementary data.

predicted Gypsy-30 structure has clear structural homology with human PNMA1, marsupial PNMA, and many modern metaviruses (**Supplementary Figure 4.10A**), demonstrating that this domain was present in the ancestral metavirus before domestication in mammals.

To identify all instances of human proteins that contain this novel putative RBD, we generated an HMM and performed several sensitive database searches, using a gathering threshold bit-score of 25 and a length threshold of 100 amino acids. Searching a six-frame translation of the human genome, we found that almost all *PNMA* family members, except *ZCCHC12* and *ZCCHC18,* contain a match to this RBD HMM. Nine of the human RBD-containing genes also contain a capsid domain, but four do not. These four non-capsid genes (*CCDC8, PNMA8A, PNMA8B* and *PNMA8C*) represent the four metaviral-derived genes that we had originally missed based on our capsid-focused HMM search strategy (**Figure 4.6C**). Thus, the only domain that links these four genes to the rest of the members of the domesticated PNMA gene family is the putative RBD, highlighting its likely functional significance. Three of these non-capsid genes encode proteins in which the putative RBD is fused to unstructured C-terminal regions. We also performed the same cross-species analysis as before to investigate the

evolutionary retention of this domain. We found that all *PNMA* family members, including capsid-encoding and capsid-lacking genes, show significant conservation of the putative RBD across placental mammals **(Figure 4.4, Figure 4.6C, Supplementary Figure 4.3**).

Our identification of the RBD also allowed to us to correct a major mis-annotation of one of the *PNMA* gene family members. While most of the HMM matches spanned the full length of the putative RBD, our initial survey found only a partial match to the RBD HMM in the annotated *CCDC8* protein, suggesting it contained an apparently truncated form of the domain. This truncation seemed surprising, so we explored *CCDC8* genomic sequences more closely. Using both evolutionary conservation and ribosome profiling studies that identified translation start sites, we found strong evidence that the correct initiation codon for *CCDC8* translation is a non-canonical upstream CTG start codon, encoding a 608 amino acid protein containing a full-length RBD HMM match (**Figure 4.6D**) instead of the currently annotated 538-residue protein. Thus, identifying the novel putative RBD allowed us to correct the annotation of the essential and rapidly evolving *CCDC8* gene[107].

The *PNMA* gene architecture demonstrates that a metavirus containing an RBD-capsid-NC domain architecture was active in an ancient mammalian ancestor. To determine how prevalent this potential RBD is in modern metaviruses, we queried Repbase using our *PNMA* family N-terminal HMM. We identified 12 consensus metaviral sequences containing this RBD domain, representing reptile, fish, and amphibian elements. We also searched six-frame translations of genomes in which metaviruses are still active: alligator (*Alligator mississippiensis),* painted turtle (*Chrysemys picta bellii)*, anole lizard (*Anolis carolensis*), African clawed frog (*Xenopus laevis*), coelacanth (*Latimeria chalumnae*), zebrafish (*Danio rerio*), and

fugu (*Takifugu rubripes*) (**Supplementary Table 4.3**). Our search yielded 217 RBD-like ORFs, including representatives from all genomes searched, consistent with our observations from the Repbase database. These results indicate that metaviruses with an RBD-capsid-NC architecture are still actively retrotransposing in reptile, amphibian, and fish lineages. By aligning these metaviral sequences with mammalian RBD sequences and generating a maximum likelihood phylogeny, we conclude that the domestication of RBD-like *PNMA* genes most likely occurred once in each of the common ancestors of placental mammals and marsupials, and was spread within those lineages by subsequent duplication (**Supplementary Figure 4.3, Supplementary Figure 4.10**), in agreement with inferences based on the capsid domain (**Figure 4.3**) [45]. Strikingly, structural predictions for this domain share a high degree of structural homology across placental mammals, marsupials, and non-mammalian vertebrates (**Supplementary Figure 4.10B**) suggesting that this protein fold is still critical to modern domesticated function. The persistence of the putative RBD in active metaviruses and its widespread conservation in domesticated PNMA-family genes highlight the importance of this domain, though its function for RNA-binding or some other function remains to be demonstrated in domesticated genes or in metaviral transposition. Thus, our detailed reanalysis of domesticated genes reveals not only an unanticipated diversity of domain architectures in metaviruses, but also an idiosyncratic lineage-specific evolutionary history of mammalian genes derived from these elements.

## Discussion

Domestication of genes derived from viruses and transposons can significantly expand the coding potential of host genomes. Our analysis of *Metaviridae*-derived capsid-like genes in human and placental mammal genomes reaffirms the broad conclusions of previous studies

[42,45,53,75,80] but extends these studies with several novel conclusions. Like previous analyses, we find 24 intact human capsid-containing genes derived from five independent germline integrations of ancient, diverse Metaviruses. However, we show that the evolutionary fate of these domesticated genes varies widely following domestication. Three ancient germline integrations of metaviruses created universally conserved single-copy genes (*ARC, RTL1, PEG10*), whereas the others serially duplicated to become the *SIRH/RTL* and *PNMA* gene families[42,45,53,75]. Our analysis reveals that these 24 intact human capsid-containing genes and their mammalian orthologs fall into 19 orthologous groupings, likely reflecting a count of 19 domesticated capsid genes in the common ancestor of placental mammals. However, deeper examination reveals that only 8 of these 19 ancestral genes are universally conserved among placental mammals, suggesting these 8 genes perform important functions that have remained largely unchanged in ~100 My of mammalian evolution. In contrast, other domesticated genes have experienced losses or duplications, suggesting lineage-specific function. The most extreme example of this lineage-specific retention is *RTL10*, which is only retained as a protein-coding gene in primates, basal glires, and cetaceans but lost in other mammalian lineages (**Figure 4.3**). Several of the lineage-specific domesticated genes are known to function in placenta or the brain, both organs that exhibit distinct anatomic and functional diversity among mammals[61,108,109]. Almost all metaviral genes reveal signatures of having been preserved via purifying selection in some mammalian lineages. However, we show that some of these genes evolve under positive selection, predicting their involvement in evolutionary arms races and consistent with recent reports that some of them may carry out immune-related functions[61,63,110].

Domesticated metaviral genes are expressed in diverse tissues. Systematic analysis of gene expression patterns across 54 adult human tissues (Genotype-Tissue Expression Consortium, GTEx), together with placenta from the Human Protein Atlas and an independent study[111–113] enabled us to compare expression patterns for different domesticated genes (**Figure S11**). Some domesticated genes are widely expressed in all or most human tissues (*e.g., RTL8C, PNMA1*), whereas others are tissue-restricted (*e.g., PNMA6F* is expressed principally in brain, *RTL9* principally in testis, and *RTL1* principally in placenta). We found robust evidence of expression for all but one domesticated gene in at least one human tissue. Only *RTL4* does not exhibit robust expression in any of the tissues sampled, which could be a limitation of tissues sampled by GTEx, or because bulk RNA-seq is not sufficiently sensitive.

What are the functions of these domesticated genes? Although the organismal and cellular functions of most domesticated metaviral genes remain functionally uncharacterized, important inroads into functional characterization have been made for several metaviral genes. *ARC* has been extensively studied for its role in learning and memory[51,114–118]. *ARC* knockout mice exhibit deficits in learning, memory and sleep[119–121]. However, the exact role of the capsid domain remains unclear. In addition, Kaneko-Ishino, Ishino, and colleagues have studied the function of metaviral capsid genes in mice by generating knockouts of *PEG10*[43], *RTL1*[44,122], *LDOC1*[81], and *RTL4*[59], *RTL9*[63], *RTL8*[62], and fluorescent reporter knock-in alleles of *RTL5* and *RTL6*[61]. These in-depth studies reveal that many domesticated genes have profound functional consequences on organismal fitness. For instance, loss of *PEG10* leads to complete early embryonic lethality[43], *RTL1* loss leads to partial lethality at late fetal/early neonatal stages and abnormal behavioral and musculoskeletal phenotypes in surviving animals[44,122–125], whereas loss

of *LDOC1* leads to abnormal placental morphology[81]. Additional studies suggest other metaviral-derived genes function in cognition or in innate immunity in the brain[59,61,62,109,125]. Complementing these in-depth studies focused on individual genes, eight of nine knockouts of domesticated genes generated and phenotyped by the International Mouse Phenotyping Consortium reveal significant behavioral and/or physiological phenotypes (**Supplementary Table 4.5**)[107]. However, so far there is little overlap between the genes knocked out by the IMPC KO and other studies – only *RTL4* and *RTL5* have been knocked out by both the IMPC and in an independent study[59,61], and in these cases, results between the studies did not match. For example, Irie and colleagues were able to identify differences in noradrenaline levels in *RTL4* knockout mice that they associated with increased impulsivity, reduced attention, and memory deficits[59]. However, these defects were not observed in the high throughput analyses by IMPC. Likewise, Irie and colleagues identified a role for *RTL5* in innate immunity in the mouse brain[61], but this function was not tested in the IMPC's high throughput knockout analysis.

In addition to their native endogenous functions, aberrant expression of some domesticated metaviral genes can lead to autoimmune disease. For example, *PNMA1* encodes a protein associated with the autoimmune disorder paraneoplastic syndrome (PNMA1: paraneoplastic Ma antigen 1). Other *PNMA* genes may have similar phenotypes, although it is currently unclear whether their autoimmune consequences are related to their endogenous function or the result of aberrant expression[50,79,126,127]. Little is known about the endogenous cellular function of the *PNMA* family genes. Previous studies have revealed that *MOAP1* (*PNMA4)* likely plays a role in regulating apoptosis, but beyond these studies in cancer cell lines, functional studies remain sparse[58,128–132]. In summary, mouse knockout studies indicate that

domesticated metaviral genes perform important functions; their loss leads to profoundly deleterious outcomes from embryonic lethality to cognitive deficits to organ abnormalities. These studies suggest that the domestication and subsequent proliferation of metavirus-derived genes led to novel non-redundant and essential functions[59,61,81,125]. However, the link between their molecular function and their metaviral origins (*e.g.,* capsid-mediated transport of RNA by ARC and PEG10) remains an active area of investigation. Our findings of lineage-specific retention of some domesticated genes further suggest that some of the functions derived from these domestication events may themselves be lineage-specific.

One of the most important contributions of our study is its novel insights into the domain architecture of *Metaviridae*-derived domesticated genes. Previous sequence homology studies showed that adjacent metaviral domains (*e.g.,* protease RT, RNase H) have been occasionally retained alongside the capsid-like region, that unstructured regions with no recognizable homology have been added to metaviral-derived sequences of some genes, and that some domesticated capsid-like genes retain only one of the two capsid lobes[16,42,45]. Using recently developed structural prediction and homology search tools, our analysis reveals that each of the five ancient domestications has a distinct N-terminal region, with no recognizable homology to what has previously been defined as a retroviral matrix domain. While the capsid domain has preserved structural homology across a wide-range of reverse-transcribing viruses [69], the N-terminus of the *gag* domain of reverse transcribing viruses displays considerably more variability (**Figure 4.5, Supplementary Table 4**). This finding suggests that the N-terminus has been an unappreciated hotspot of evolution in the structural genes of reverse-transcribing viruses. Furthermore, our analyses reveal that one of the four ancestral viruses and most of its descendant

domesticated genes (the *PNMA* family) have an N-terminal region with clear structural homology to a previously-unrecognized RNA-binding domain (RBDs). The putative RBD is structurally district from the matrix domain typically associated with the N-terminus of the retroviral capsids, but these domains may perform a functionally analogous role, where interactions with RNA promote capsid assembly and membrane association. While most *PNMA* family members retain the capsid- and RBD-like domains, several have lost the capsid but still contain the putative RBD. In these instances, the RBD is often fused to predicted disordered regions. Although the previous focus has been on *Metaviridae*-derived capsid domains, our findings highlight a putative RNA-binding domain of unknown function within the *Metaviridae gag* domain that has been captured in essential protein-coding genes in placental mammals and even amplified via gene duplication.

Our discovery of a novel RBD is especially relevant to significantly amend the annotation of *CCDC8* by identifying a non-canonical start codon that extends the N-terminal protein-coding region by 70 amino acids to now comprise an intact RBD. *CCDC8* is completely conserved in mammals, including the 70 N-terminal residues. Moreover, a knockout mouse strain from IMPC reveals an essential role for *CCDC8* at least in mice[107]. This is important because previous experiments on *CCDC8* inadvertently relied on overexpression of a truncated *CCDC8*, which we predict would not be sufficient to recapitulate its essential function. This highlights the utility of our evolutionary and structural insights for future studies to reveal the function of one of the few essential domesticated capsid genes in mammals. RBD retention in domesticated *PNMA* family genes like *CCDC8* further suggests that RNA interactions may be an important function, independent of whether they contain an adjacent capsid-like region.

The *PNMA*-encoded putative RNA-binding domain (RBD) is present in some modern *Metaviridae* that continue to actively circulate in amphibians, reptiles, and fish. Thus, our analysis of domesticated genes not only reveals an unprecedented diversity of domain architecture in *Metaviridae*, but also identifies a previously undescribed putative RNA-binding domain in vertebrate metaviral retrotransposons that has an unknown role in the metaviral lifecycle. Viral capsid domains are already known to function to package viral RNA genomes, and the C-terminal nucleocapsid is already predicted to bind RNAs. The prevalence of a tripartite RBD-capsid-NC architecture in ancient and active metaviruses indicates that each of these domains encodes a non-redundant function, which may further facilitate metaviral capsid-RNA interactions or encode other functions to potentially defend metaviral RNAs from host defenses. Thus, by capturing snapshots in time based on when they were domesticated, metaviral-derived host genes provide important archeological insights into metaviral protein domains and retrotransposition strategies. The mysterious N-terminal non-capsid domain of the ARC protein may represent just such an archeological clue about an ancient domain with no extant homologs that was acquired either from *Metaviridae* or host genomes ~350 million years ago.

## Methods

### Building Capsid HMMs

We queried a library of diverse capsid-like sequences from *Retroviridae, Metaviridae, Pseudoviridae,* and *Belpaoviridae*[18,70,133] using five previously generated PFAM HMMs (5 HMMs from Viral_Gag CL0148 and 10 from Gag-polyprotein CL0523[95]). However, these PFAM clans do not precisely match the capsid domain. Therefore, we built new HMMs that precisely identified the capsid domain in each major clade of retroviruses (n=4 clades) and LTR

retrotransposons (n=4 clades). To build custom capsid HMMs for each group of reverse transcribing viruses, we queried NCBI's non-redundant protein database [134] with PFAM seed alignments (in cases where the HMM covered the full-length capsid domain) or with single sequences using a single iteration of PSI-BLAST[135]. Sequences identified by PSI-BLAST were aligned in MAFFT [136] and the alignment was submitted to HHpred to identify the capsid domain using structural homology prediction. Alignments were trimmed to the full-length capsid domain, and pressed into profile HMMs using hmmbuild from the HMMER3 package [94] to generate eight retrovirus or retrotransposon-specific HMMs (**Supplementary Figure 4.1, Table S1**). Each clade of LTR retroelement is represented by a single capsid HMM, except for *Metaviridae*. When our initial analysis did not identify a strong hit for ARC, we built a second metaviral HMM using previously generated ARC structures[137].

To verify the specificity of capsid HMMs, a single sequence from each class of LTR retroelement was queried with each of the eight HMMs constructed. Each HMM identified a single domain from only one of these control sequences with a high bit score and significant E-value, while returning low-scoring hits or no hit at all for capsid sequences from other LTR retroelement clades (*e.g.,* the ERV2 HMM identified the MLV capsid with a bit-score of 16.1 and an e-value of 3.1e-06, while the ARC HMM identified the metaviral control capsid with a bit score of 42.3 and an e-value of 3.7e-14). In sequences associated with previously determined structures[138–141], HMM matches overlap exactly with the bi-lobed alpha-helical capsid domains. In sequences with no known molecular structure, AlphaFold [65] was used to predict tertiary structure. In these unstudied sequences, HMMs identified domains for which AlphaFold predicted bi-lobed alpha-helical architecture (**Supplementary Figure 4.1C-D**).

Identifying Capsid Genes in Vertebrate Genomes

To generate a comprehensive catalog of capsid-like sequences in the human genome, we generated a six-open reading frame translation of the recently complete T2T genome (GCF_009914755)[67,68] using the EMBOSS[142] tool getorf with the following parameters: -minsize 300 (minimum ORF size 100 codons); -find 1 (return regions bounded by a start and stop codon); -table 1 (use the standard genetic code, but allowing non-canonical start codons, due to our finding that CCDC8 uses a previously unannotated CTG start codon, **Figure 4.6D**). The resulting dataset contains 2,094,180 peptide sequences that we queried for capsid homology using the previously described HMMs, using hmmscan with default parameters and with a post-scan gathering threshold bit-score of 25 and a length threshold of 125 amino acids. We repeated our search on 25 additional vertebrate genomes downloaded from the UCSC genome browser site or NCBI to identify as many metaviral capsid domains as possible for our subsequent phylogeny of metaviral capsid domains. We filtered matching sequences to ensure that if we had required a canonical start codon (ATG), the ORF would still be >= 300bp in size. This approach could be less sensitive to capsid domains interrupted by large insertions or introns; however, a similar search of predicted spliced transcripts found in RefSeq yielded no additional hits. Furthermore, previously annotated capsid-containing domesticated genes have no introns within the capsid region[42,45]. Additionally, we downloaded Repbase 25.03[93] and generated a six-frame translation of all sequences to identify the capsid domains in previously characterized consensus sequences of vertebrate LTR retroelements.

Matching Human Capsid-like ORF Sequences to Previously Annotated Genes

To determine if capsid-genes identified using the strategy above were previously annotated, or uncharacterized sequences, we used each sequence to retrieve query NCBI's RefSeq database using blastn, considering only exact matches. We obtained the full-length open reading frames of the matching annotated genes for use in our subsequent analyses.

Analysis of Domain Architecture and Genetic Context

To identify other metaviral fragments (protein domains or LTRs) in or adjacent to each human domesticated metaviral capsid gene, we extracted flanking nucleic acid sequences 16 kb upstream of the start codon and 16 kb downstream of the stop codon from the T2T genome assembly. These ~33 kb genomic sequences were queried for long terminal repeats (LTRs) using LTRHarvest[78]. We generated new six-frame translations of the ~33 kb extracted sequences using getorf (see above parameters). We queried translated sequences with the entire PFAM 34.0 database to identify any recognizable protein domains nearby as well as our capsid and RBD HMMs. These HMM search results allowed us to identify the conserved non-canonical start site of *CCDC8*. This non-canonical start-site was verified by visualizing Ribo-seq data for initiating ribosomes using GWIPS-vis[143].

Phylogenetic Analysis of Capsid-like Sequences

To understand the evolutionary relationships between capsid-like sequences in the human and 26 other vertebrate genomes, we built two initial maximum likelihood phylogenetic trees (**Figure 4.1, Figure 4.5A, Supplementary Table 4.3**). Full-length capsid sequences were defined as ORFs longer than 150 amino acids, containing HMM alignments longer than 125

amino acids (**Supplementary Figure 4.1**). Duplicate sequences were removed using CD-Hit (sequence identity threshold flag, -c 1)[144]. Unique sequences were aligned using MAFFT[136] (using the more accurate L-ins-I method). Alignments were trimmed and filtered using trimAl (gt = 0.5)[145] and visually inspected, removing any sequences containing large internal insertions and deletions. A phylogenetic tree was then constructed in Fasttree2[146] using the JTT evolutionary model with gamma-shape rate variation based on ProtTest model selection[147]. Phylogenetic trees were visualized with the ggtree package in R[148]. Tree topologies were verified by building phylogenetic trees in IQTREE[149] using both UFBOOT and traditional bootstrap parameters. We found that metaviral sequences consistently separated from endogenous retroviral sequences in the human phylogeny built from 238 aligned positions (**Figure 4.11**). Mammalian metaviral orthologs consistently separated into 4-5 clades with distinct groups of non-mammalian metaviral sequences as the closest-branching outgroups in a phylogenetic tree built from 172 aligned positions (**Figure 4.5A**).

Identification of Mammalian Orthologs

To identify potential orthologs of the human domesticated metaviral-derived capsid-like genes, we performed thorough searches of 17 additional representative mammalian genomes (14 placental mammals, two marsupials, and one monotreme – **Supplementary Table 4.3**) using several complementary methods. Our process was iterative, combining multiple rounds of database searches followed by sequence analysis (alignment building and orthology checking via phylogenies).

We initiated the first search using the full-length human ORF nucleotide sequences and their translations as initial queries for tblastn searches (protein query versus nucleotide database)

of predicted transcripts in the NCBI non-redundant nucleotide database, specifying the species of interest for each search. In subsequent iterations, we used our initial alignments and phylogenies (see below) to identify species/gene combinations where more attention was needed, either because a gene seemed to be missing, or because the gene prediction appeared to be truncated. To search for these 'missing' genes, we employed additional strategies that are better suited for detection of degenerate pseudogenes. We added blastn searches (nucleotide query versus nucleotide database) and we searched genome assemblies rather than predicted gene sets. In some cases, we added queries from non-human species, choosing closely related species to the one with the missing gene. For some of the apparent pseudogenes, we added sequences from sister taxa (outside the original 18 mammalian species selected) to investigate whether inactivating mutations are real rather than genome assembly errors; seeing inactivating mutations shared across related species provides confidence they are real mutations. We complemented these blast search results by searching six-frame genome translations using our metaviral capsid and putative RBD HMMs (as we did for human), ensuring that all matches were included in our final sequence collection.

We generated alignments and phylogenies to help determine orthologous relationships (depicted in **Figure 4.2**) as well as to understand which sequences are intact genes versus pseudogenes. Blastn searches against a database of all human transcripts were also useful tools in this step, often immediately clarifying that a query sequence obviously grouped with a single human gene to the exclusion of all others. For example, using each human metaviral-derived sequence as a blast query against all human genes revealed the very high similarity between genes within each of the *PNMA6E/F*, *PNMA6A/B* and *RTL8A/B/C* groups, indicating that these

human genes fall into only three clades, rather than one for each individual human gene. Using our initial blastn-based group assignments, we generated ‘master’ in-frame nucleotide alignments, using the MACSE frame-aware alignment tool [150], alignments that we added to and refined during subsequent search/analysis iterations.

All except four of the HMM matches we found among whole genome six frame translations fit very clearly into a single master in-frame alignment. The four exceptions (bosTau9 chr3:77464558-77464911; pteAle1 NW_006435837.1:195339-195764; mMonDom1 NC_077233.1:146410251-146410589 and mMacEug1 CM051816.1:211971879-211972229) are RBD-only matches, have only remote HMM homology, and blastn searches of the RefSeq representative genomes shows that neither is preserved as an ORF outside of very closely related taxa. These sequences are therefore likely to be degenerated metaviral sequences; while they meet the criteria we defined for HMM matches, they do not appear to be true domesticated genes. One additional HMM match (mMonDom1 NC_077230.1:27205883-27206248) is clearly a recent retrotransposon insertion, as RepeatMasker recognizes it as an ERV9_MD element (ERV1 family) and finds LTRs in the flanking regions, and a blastn search reveals many almost-identical copies in the opossum genome assembly. We did not include these five sequences in our final datasets.

Analyzing our in-frame nucleotide alignments using a custom script allowed us to classify each sequence as intact (complete ORF), incomplete due to genome assembly gaps (*i.e.,* we observed one or more N bases and/or the start/end of an assembly contig near the homologous region) or a clear pseudogene (*i.e.,* contains frameshift and/or premature stop codons with respect to the full-length annotated human ORF, or is significantly truncated without

evidence of assembly gaps). Our script makes an exception for the *PEG10* gene because it has been shown to contain a programmed ribosomal frameshift; thus, we permitted frameshifts observed within 10 codons of the expected location.

During our analyses, we noticed that NCBI's predicted transcripts are sometimes annotated to indicate that potential frameshifts and/or premature stop codons that were observed in genome assemblies have been 'corrected', with the idea they most likely represent genome sequencing errors. While this approach makes sense for genes that are rarely pseudogenized, we found that for some of these domesticated capsid genes, it likely underestimates the number of true pseudogenes. For this reason, we ensure that we always used corresponding sequences extracted directly from genome assemblies, rather than the predicted transcripts in the NR database. For most apparent pseudogenes, we were able to gain additional confidence that inactivating mutations are real by adding sequences from closely related taxa, where we observed that identical inactivating mutations are often shared between taxa.

In order to more formally confirm orthology relationships, we selected full-length intact genes from all species, generated amino acid translations, and realigned them using MAFFT [136]. Because sequence divergence is very high between the major groups shown in **Figure 4.5A**, we aligned sequences only within each group. We manually trimmed each alignment to contain only well-aligning regions, used ProtTest to select the best-fitting evolutionary model, and used PHYML[151] to generate phylogenies, with 100 bootstrap replicates (JTT or LG substitution model, --pinv e --alpha e -f e). Visual inspection of the resulting trees (**Supplementary Figures 4.2-4.7**) allowed confident assignment of orthology relationships. To confirm orthology assignments for pseudogenes and recent duplicates, we also generated nucleotide-based

phylogenies of alignments for individual groups, again using PHYML (parameters: GTR substitution model, flags --pinv e --alpha e -f e). We also checked that species relationships roughly followed the expected species tree (**Figure 4.2**). Finally, we obtained gene age estimates in My (millions of years) using the TimeTree database[152].

Analysis of Evolutionary Selective Constraints

We performed two complementary analyses of the evolutionary selective pressures acting on these genes. First, we used sequences from diverse placental mammals to estimate the 'overall' selective pressure on each gene. From our master in-frame nucleotide alignments for each gene, we selected only the intact genes, from only the 18 placental mammal species shown in **Figure 4.2**. We removed any in-frame gaps from the resulting alignments, and manually trimmed each to contain only well-aligning regions. We generated a phylogeny from each trimmed alignment (using PHYML with the GTR substitution model, estimating the proportion of invariable sites, shape of the gamma distribution and nucleotide frequencies). We then used the alignment and tree as inputs to the codeml algorithm from the PAML package [89] to estimate overall dN/dS, using "model 0", which makes the simplifying assumption that all lineages and all amino acid sites evolve under uniform selective pressure (additional parameters: codon model=2, initial dN/dS = 0.4, cleandata=0). We also ran codeml under model 0 but fixing dN/dS=1 to examine the null hypothesis that sequences evolve neutrally. We determined the level of statistical support for rejecting the null (neutral) hypothesis in favor of the alternative model where dN/dS has the value estimated by model 0 by comparing twice the difference in log-likelihoods of these two model 0 runs to a chi-squared distribution with 1 degree of freedom (**Figure 4.4**, "M0 (Placentals)" columns).

Second, while the overall dN/dS estimates indicate that purifying selection is the prevailing evolutionary pressure operating on these genes, we wanted to explore the hypothesis that positive selection might be acting at a subset of amino acid positions in some of these genes. For this position-specific analysis we used primate genomes, which provide a powerful clade in which to look for selection (the placental mammal alignments contain much greater diversity, such that synonymous substitutions begin to show saturation). We therefore used blastn searches of NCBI's non-redundant database to collect primate orthologs for each gene, again using reciprocal blastn searches against a local database of all human genes to ensure one-to-one orthology and selecting only intact genes for our analysis. We generated in-frame alignments using MACSE, and adjusted and trimmed alignments manually. We ran the GARD algorithm (parameters: general discrete model of site-to-site rate variation, 3 rate classes) on each alignment [153] to check alignments for recombination, which can give rise false-positive findings of positive selection [154]. Four alignments showed evidence of recombination (*PNMA6E*, *PNMA6F*, *PNMA8A* and *ZCCHC12*), each with a single recombination breakpoint; in these cases we split each alignment into segments at the breakpoints indicated by GARD, and ran the following analysis on each segment separately. Again, we estimated phylogenies using PHYML (parameters as for model 0 analysis) and used PAML's codeml algorithm (codon model=2, initial dN/dS = 0.4, cleandata=0), but this time we compared the log likelihoods of an evolutionary model that allows for a subset of residues under positive selection (model 8) with a paired model that only allows purifying and neutral selection (models 8a). For genes where the maximum likelihood test indicates positive selection, we report in **Figure 4.4** the proportion of sites estimated to be under positive selection, the dN/dS of this class of sites, and the identity of sites

with a >90% posterior probability of being members of the positively selected class (codeml's "BEB" Bayes Empirical Bayes method). For these genes, we also checked the robustness of evidence favoring positive selection by re-running analyses with alternative codeml parameters (codon model=3, starting dN/dS=3); all positive selection findings remained consistent under alternative parameter combinations.

Structure Predictions and Homology Modelling

To gain structural insight into the regions of human metaviral sequences not annotated by either our capsid HMMs or Pfam, we downloaded all human metaviral sequences from AlphaFold's Protein Structure Database[155] and submitted unannotated portions of each protein to DALI and Foldseek to identify structural homologues[64,66]. This strategy consistently identified an RNA binding domain within the N-terminus of PNMA-family proteins with hundreds of hits in DALI. For example, the N-terminus of PNMA1 identified 939 significant hits (Z-score > 2) in the PDB, the majority of which were RNA-binding domains. The top ten hits all contained Z-scores above 6.6 for RNA-binding proteins (**Figure 6,** data not shown). AlphaFold was also used to predict the structures of the closest consensus metaviral sequences from Repbase, and structural homology between metaviral sequences was determined using structural superposition (mmaker command in the ChimeraX Daily build)[156]. This strategy confirmed the near perfect retention of capsid domains in RTL, ARC, and in the PNMA family, while in the *SIRH/RTL* family only the N-terminal lobe of the capsid domain is retained. This approach also revealed that most metaviral-derived genes have N-terminal regions with predicted disordered regions and no clear globular domains (**Figure 4.5B-E, Supplementary Figure 4.9**).

Phylogenetic Analysis of Putative RBD Sequences

To understand the evolutionary history of the putative RBD of the PNMA family, we built two maximum likelihood phylogenetic trees (**Supplementary Figure 4.3, Supplementary Figure 4.10**). The first was built using only a subset of putative metaviral RBD sequences and all mammalian sequences to confirm orthology relationships of the mammalian PNMA genes (**Supplementary Figure 4.3 –**see details in Methods section, "Identification of mammalian orthologs"). To understand the evolutionary history of the PNMA putative RBD in the context of modern metaviruses, we queried 22 vertebrate genomes (**Supplementary Table 4.3**) and vertebrate metaviruses from Repbase with a single HMM built from an alignment of the first 200 amino acids of the PNMA family genes (except for ZCCHC12/18, which do not contain this N-terminus) (hmmsearch – default parameters used). This search yielded 408 full-length putative RBDs, defined as ORFs longer than 100 amino acids with bit-score > 25. Duplicate sequences were removed using cd-hit (sequence identity threshold, -c 1)[144]. Unique sequences were aligned using MAFFT[136] (using the more accurate L-ins-I method). Alignments were visually inspected and any sequences containing large insertions and deletions were removed. The remaining 250 sequences were used to construct a phylogenetic in Fasttree2[146] using the JTT evolutionary model with gamma-shape rate variation based on ProtTest model selection[147]. Phylogenetic trees were visualized with the ggtree package in R[148]. Alphafold predictions were made from the sequence at the indicated position on the tree, and all Alphafold predictions were subsequently aligned using the mmaker command in ChimeraX (**Supplementary Figure 4.10**).

Expression Analysis

Tissue-specific expression data was downloaded from the Genotype-Tissue Expression (GTEx) Project, supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS[111]. The median gene level TPM dataset was obtained from the GTEx Portal on October 13, 2021. A publicly available placental RNA-seq dataset was downloaded from the Human Protein Atlas on October 12, 2021 (https://www.proteinatlas.org/about/download)[157]. A third placental RNA-seq data set was downloaded from https://www.obgyn.cam.ac.uk/placentome/[113], and fragments per kilobase per million (FPKM) values were converted to transcript per million (TPM) values. A heatmap of expression was made in R using the pheatmap package with k-means clustering turned on for columns (tissue type), but not for rows (gene)[158] (with flags: color = viridis(50, option = "G", direction = -1), cutree_cols = 6, gaps_row = c(1,2, 13, 23), cluster_rows = FALSE) (**Supplementary Figure 4.11**).

International Mouse Phenotyping Consortium

HGNC gene names were submitted as queries to the IMPC database [107]; results are summarized in **Supplementary Table 4.5**.

## Data and Resource Availability

All sequences used in this study are from publicly available genomic sequences (except for Repbase, which was downloaded in 2020 through a Montana State University institutional subscription). Supplementary Table 4.3 includes a list of the primary vertebrate genomes (including version) used in this study, with their respective source (either UCSC or NCBI)

indicated. Coordinates from these genomes, and additional sequences used outside of these primary genomes can be found in the metadata directory of the Supplementary Data.

## References


1. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**, 657–663 (1999).

2. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

3. Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409**, 847–849 (2001).

4. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics* **7**, e1002384 (2011).

5. Dawkins, R. *The Selfish Gene*. (Oxford University Press, 1976).

6. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).

7. Orgel, L. E. & Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).

8. Deininger, P. L. & Batzer, M. A. Mammalian Retroelements. *Genome Res.* **12**, 1455–1465 (2002).

9. Boissinot, S., Davis, J., Entezam, A., Petrov, D. & Furano, A. V. Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences* **103**, 9590–9594 (2006).

10. Burns, K. H. Transposable elements in cancer. *Nat Rev Cancer* **17**, 415–424 (2017).

11. Burns, K. H. Our Conflict with Transposable Elements and Its Implications for Human Disease. *Annu Rev Pathol* **15**, 51–70 (2020).

12. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).

13. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol* **23**, 481–497 (2022).

14. Mager, D. L. & Stoye, J. P. Mammalian Endogenous Retroviruses. *Microbiology Spectrum* **3**, 3.1.09 (2015).

15. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539–561 (2020).

16. Brandt, J., Veith, A.-M. & Volff, J.-N. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. *Cytogenet. Genome Res.* **110**, 307–317 (2005b).

17. Modzelewski, A. J., Gan Chong, J., Wang, T. & He, L. Mammalian genome innovation through transposon domestication. *Nat Cell Biol* **24**, 1332–1340 (2022).

18. Krupovic, M. *et al.* Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. *Journal of Virology* **92**, e00515-18 (2018).

19. Finnegan, D. J. Retrotransposons. *Curr. Biol.* **22**, R432-437 (2012).

20. Burns, K. H. & Boeke, J. D. Human Transposon Tectonics. *Cell* **149**, 740–752 (2012).

21. Dodonova, S. O., Prinz, S., Bilanchone, V., Sandmeyer, S. & Briggs, J. A. G. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *PNAS* **116**, 10048–10057 (2019).

22. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691–703 (2009).

23. Doolittle, R. F., Johnson, M. S. & McClure, M. A. Origins and Evolutionary Relationships of Retroviruses. *The Quarterly Review of Biology* **64**, 1–30 (1989).

24. Hayward, A. Origin of the retroviruses: when, where, and how? *Current Opinion in Virology* **25**, 23–27 (2017).

25. Ashkenazi, A., Faingold, O. & Shai, Y. HIV-1 fusion protein exerts complex immunosuppressive effects. *Trends in Biochemical Sciences* **38**, 345–349 (2013).

26. Kim, F. J., Battini, J.-L., Manel, N. & Sitbon, M. Emergence of vertebrate retroviruses and envelope capture. *Virology* **318**, 183–191 (2004).

27. Dupressoir, A., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* **33**, 663–671 (2012).

28. Lavialle, C. *et al.* Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120507 (2013).

29. Cornelis, G. *et al.* An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proceedings of the National Academy of Sciences* **114**, E10991–E11000 (2017).

30. Henzy, J. E., Gifford, R. J., Kenaley, C. P. & Johnson, W. E. An Intact Retroviral Gene Conserved in Spiny-Rayed Fishes for over 100 My. *Molecular Biology and Evolution* **34**, 634–639 (2017).

31. Ikeda, H., Laigret, F., Martin, M. A. & Repaske, R. Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. *J Virol* **55**, 768–777 (1985).

32. Taylor, G. M., Gao, Y. & Sanders, D. A. Fv-4: identification of the defect in Env and the mechanism of resistance to ecotropic murine leukemia virus. *J Virol* **75**, 11244–11248 (2001).

33. Frank, J. A. *et al.* Evolution and antiviral activity of a human protein of retroviral origin. *Science* **378**, 422–428 (2022).

34. Varela, M., Spencer, T. E., Palmarini, M. & Arnaud, F. Friendly Viruses. *Annals of the New York Academy of Sciences* **1178**, 157–172 (2009).

35. Ito, J. *et al.* Refrex-1, a Soluble Restriction Factor against Feline Endogenous and Exogenous Retroviruses. *Journal of Virology* **87**, 12029–12040 (2013).

36. Ito, J., Baba, T., Kawasaki, J. & Nishigaki, K. Ancestral Mutations Acquired in Refrex-1, a Restriction Factor against Feline Retroviruses, during its Cooption and Domestication. *Journal of Virology* **90**, 1470–1485 (2016).

37. Robinson, H. L., Astrin, S. M., Senior, A. M. & Salazar, F. H. Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. *Journal of Virology* **40**, 745–751 (1981).

38. Olson, E. D. & Musier-Forsyth, K. Retroviral Gag protein - RNA interactions: Implications for specific genomic RNA packaging and virion assembly. *Semin Cell Dev Biol* **86**, 129–139 (2019).

39. Bénit, L. *et al.* Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J Virol* **71**, 5652–5657 (1997).

40. Yap, M. W., Colbeck, E., Ellis, S. A. & Stoye, J. P. Evolution of the Retroviral Restriction Gene Fv1: Inhibition of Non-MLV Retroviruses. *PLOS Pathogens* **10**, e1003968 (2014).

41. Young, G. R., Yap, M. W., Michaux, J. R., Steppan, S. J. & Stoye, J. P. Evolutionary journey of the retroviral restriction gene Fv1. *PNAS* **115**, 10130–10135 (2018).

42. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends in Genetics* **22**, 585–589 (2006).

43. Ono, R. *et al.* Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet* **38**, 101–106 (2006).

44. Sekita, Y. *et al.* Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nat Genet* **40**, 243–248 (2008).

45. Kokošar, J. & Kordiš, D. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol Biol Evol* **30**, 1015–1031 (2013).

46. Pastuzyn, E. D. *et al.* The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **172**, 275-288.e18 (2018).

47. Abed, M. *et al.* The Gag protein PEG10 binds to RNA and regulates trophoblast stem cell lineage specification. *PLOS ONE* **14**, e0214110 (2019).

48. Erlendsson, S. *et al.* Structures of virus-like capsids formed by the Drosophila neuronal Arc proteins. *Nat Neurosci* 1–4 (2020) doi:10.1038/s41593-019-0569-y.

49. Segel, M. *et al.* Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science* **373**, 882–889 (2021).

50. Xu, J. *et al.* PNMA2 forms immunogenic non-enveloped virus-like capsids associated with paraneoplastic neurological syndrome. *Cell* **187**, 831-845.e19 (2024).

51. Nikolaienko, O., Patil, S., Eriksen, M. S. & Bramham, C. R. Arc protein: a flexible hub for synaptic plasticity and cognition. *Seminars in Cell & Developmental Biology* **77**, 33–42 (2018).

52. Ashley, J. *et al.* Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172**, 262-274.e11 (2018).

53. Kaneko-Ishino, T. & Ishino, F. Mammalian-specific genomic functions: Newly acquired traits generated by genomic imprinting and LTR retrotransposon-derived genes in mammals. *Proc Jpn Acad Ser B Phys Biol Sci* **91**, 511–538 (2015).

54. Butler, M., Goodwin, T., Simpson, M., Singh, M. & Poulter, R. Vertebrate LTR Retrotransposons of the Tf1/Sushi Group. *J Mol Evol* **52**, 260–274 (2001).

55. Ono, R. *et al.* A Retrotransposon-Derived Gene, PEG10, Is a Novel Imprinted Gene Located on Human Chromosome 7q21. *Genomics* **73**, 232–237 (2001).

56. Charlier, C. *et al.* Human–Ovine Comparative Sequencing of a 250-kb Imprinted Domain Encompassing the Callipyge (clpg) Locus and Identification of Six Imprinted Transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. *Genome Res* **11**, 850–862 (2001).

57. Clark, M. B. *et al.* Mammalian gene PEG10 expresses two reading frames by high efficiency -1 frameshifting in embryonic-associated tissues. *J. Biol. Chem.* **282**, 37359–37369 (2007).

58. Foley, C. J. *et al.* Dynamics of RASSF1A/MOAP-1 association with death receptors. *Mol Cell Biol* **28**, 4520–4535 (2008).

59. Irie, M. *et al.* Cognitive Function Related to the Sirh11/Zcchc16 Gene Acquired from an LTR Retrotransposon in Eutherians. *PLOS Genetics* **11**, e1005521 (2015).

60. Lee, Y. H., Pang, S. W., Poh, C. L. & Tan, K. O. Distinct functional domains of PNMA5 mediate protein-protein interaction, nuclear localization, and apoptosis signaling in human cancer cells. *J. Cancer Res. Clin. Oncol.* **142**, 1967–1977 (2016).

61. Irie, M. *et al.* Retrovirus-derived RTL5 and RTL6 genes are novel constituents of the innate immune system in the eutherian brain. *Development* **149**, dev200976 (2022).

62. Fujioka, Y. *et al.* Targeting retrovirus-derived Rtl8a and 8b causes late onset obesity and neurodevelopmental defects. 2023.05.28.542606 Preprint at https://doi.org/10.1101/2023.05.28.542606 (2023).

63. Ishino, F. *et al.* Retrovirus-Derived RTL9 Plays an Important Role in Innate Antifungal Immunity in the Eutherian Brain. *International Journal of Molecular Sciences* **24**, 14884 (2023).

64. Holm, L. DALI and the persistence of protein shape. *Protein Science* **29**, 128–140 (2020).

65. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

66. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 1–4 (2023) doi:10.1038/s41587-023-01773-0.

67. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

68. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).

69. Krupovic, M. & Koonin, E. V. Homologous Capsid Proteins Testify to the Common Ancestry of Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses. *Journal of Virology* **91**, e00210-17 (2017).

70. Gifford, R. J. *et al.* Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* **15**, 59 (2018).

71. Blikstad, V., Benachenhou, F., Sperber, G. O. & Blomberg, J. Endogenous retroviruses. *Cell. Mol. Life Sci.* **65**, 3348–3365 (2008).

72. Sperber, G. O., Airola, T., Jern, P. & Blomberg, J. Automated recognition of retroviral sequences in genomic data—RetroTector©. *Nucleic Acids Research* **35**, 4964–4976 (2007).

73. Ivanov, D., Stone, J. R., Maki, J. L., Collins, T. & Wagner, G. Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain. *Mol. Cell* **17**, 137–143 (2005).

74. Emerson, R. O. & Thomas, J. H. Gypsy and the birth of the SCAN domain. *J Virol* **85**, 12043–12052 (2011).

75. Kaneko-Ishino, T. & Ishino, F. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front Microbiol* **3**, 262 (2012).

76. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *PNAS* **114**, E2401–E2410 (2017).

77. Brandt, J. *et al.* Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* **345**, 101–111 (2005a).

78. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

79. Schüller, M., Jenne, D. & Voltz, R. The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. *J. Neuroimmunol.* **169**, 172–176 (2005).

80. Pang, S. W., Lahiri, C., Poh, C. L. & Tan, K. O. PNMA family: Protein interaction network and cell signalling pathways implicated in cancer and apoptosis. *Cell Signal* **45**, 54–62 (2018).

81. Naruse, M. *et al.* Sirh7/Ldoc1 knockout mice exhibit placental P4 overproduction and delayed parturition. *Development* **141**, 4763–4771 (2014).

82. Iwasaki, S. *et al.* Identification of a novel PNMA-MS1 gene in marsupials suggests the LTR retrotransposon-derived PNMA genes evolved differently in marsupials and eutherians. *DNA Res.* **20**, 425–436 (2013).

83. Edwards, C. A. *et al.* The evolution of the DLK1-DIO3 imprinted domain in mammals. *PLoS Biol* **6**, e135 (2008).

84. Suzuki, S. *et al.* Retrotransposon Silencing by DNA Methylation Can Drive Mammalian Genomic Imprinting. *PLOS Genetics* **3**, e55 (2007).

85. Ono, R. *et al.* Identification of tammar wallaby SIRH12, derived from a marsupial-specific retrotransposition event. *DNA Res* **18**, 211–219 (2011).

86. Molaro, A., Young, J. M. & Malik, H. S. Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals. *Genome Res* **28**, 460–473 (2018).

87. Daugherty, M. D., Schaller, A. M., Geballe, A. P. & Malik, H. S. Evolution-guided functional analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. *eLife* **5**, e14228 (2016).

88. Campodonico, W., Black, H. H., Lau, C. I. & Whiteley, A. M. The gag-like gene RTL8 antagonizes PEG10-mediated virus like particles in humans. 2023.02.03.527044 Preprint at https://doi.org/10.1101/2023.02.03.527044 (2023).

89. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).

90. McBee, R. M., Rozmiarek, S. A., Meyerson, N. R., Rowley, P. A. & Sawyer, S. L. The effect of species representation on the detection of positive selection in primate gene data sets. *Mol Biol Evol* **32**, 1091–1096 (2015).

91. Wei, M. *et al.* Inhibition of HIV-1 assembly by coiled-coil domain containing protein 8 in human cells. *Sci Rep* **5**, 14724 (2015).

92. Youngson, N. A., Kocialkowski, S., Peel, N. & Ferguson-Smith, A. C. A Small Family of Sushi-Class Retrotransposon-Derived Genes in Mammals and Their Relation to Genomic Imprinting. *J Mol Evol* **61**, 481–490 (2005).

93. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).

94. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011).

95. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427–D432 (2019).

96. Shigemoto, K. *et al.* Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting. *Nucleic Acids Res* **29**, 4079–4088 (2001).

97. Shiura, H. *et al.* PEG10 viral aspartic protease domain is essential for the maintenance of fetal capillary structure in the mouse placenta. *Development* **148**, dev199564 (2021).

98. Lynch, C. & Tristem, M. A Co-opted gypsy-type LTR-Retrotransposon Is Conserved in the Genomes of Humans, Sheep, Mice, and Rats. *Current Biology* **13**, 1518–1523 (2003).

99. Wills, N. M., Moore, B., Hammer, A., Gesteland, R. F. & Atkins, J. F. A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. *J Biol Chem* **281**, 7082–7088 (2006).

100. Muriaux, D. & Darlix, J.-L. Properties and functions of the nucleocapsid protein in virus assembly. *RNA Biol* **7**, 744–753 (2010).

101. Murray, P. S. *et al.* Retroviral Matrix Domains Share Electrostatic Homology: Models for Membrane Binding Function throughout the Viral Life Cycle. *Structure* **13**, 1521–1531 (2005).

102. Hamard-Peron, E. & Muriaux, D. Retroviral matrix and lipids, the intimate interaction. *Retrovirology* **8**, 15 (2011).

103. Bou-Nader, C. *et al.* HIV-1 matrix-tRNA complex structure reveals basis for host control of Gag localization. *Cell Host & Microbe* **29**, 1421-1436.e7 (2021).

104. Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**, 2237–2243 (2018).

105. Eriksen, M. S. *et al.* Arc self-association and formation of virus-like capsids are mediated by an N-terminal helical coil motif. *The FEBS Journal* **288**, 2930–2955 (2021).

106. He, Y. *et al.* Structures of telomerase at several steps of telomere repeat synthesis. *Nature* **593**, 454–459 (2021).

107. Koscielny, G. *et al.* The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Research* **42**, D802–D809 (2014).

108. Cho, G., Lim, Y., Zand, D. & Golden, J. A. Sizn1 Is a Novel Protein That Functions as a Transcriptional Coactivator of Bone Morphogenic Protein Signaling. *Molecular and Cellular Biology* **28**, 1565–1572 (2008).

109. Takaji, M., Komatsu, Y., Watakabe, A., Hashikawa, T. & Yamamori, T. Paraneoplastic Antigen-Like 5 Gene (PNMA5) Is Preferentially Expressed in the Association Areas in a Primate Specific Manner. *Cerebral Cortex* **19**, 2865–2879 (2009).

110. Jiang, X. *et al.* Overexpressed coiled-coil domain containing protein 8 (CCDC8) mediates newly synthesized HIV-1 Gag lysosomal degradation. *Scientific Reports* **10**, 11416 (2020).

111. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

112. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).

113. Gong, S. *et al.* The RNA landscape of the human placenta in health and disease. *Nat Commun* **12**, 2639 (2021).

114. Carmichael, R. E. & Henley, J. M. Transcriptional and post-translational regulation of Arc in synaptic plasticity. *Semin Cell Dev Biol* **77**, 3–9 (2018).

115. Epstein, I. & Finkbeiner, S. The Arc of cognition: Signaling cascades regulating Arc and implications for cognitive function and disease. *Semin Cell Dev Biol* **77**, 63–72 (2018).

116. Mabb, A. M. & Ehlers, M. D. Arc ubiquitination in synaptic plasticity. *Semin Cell Dev Biol* **77**, 10–16 (2018).

117. Newpher, T. M., Harris, S., Pringle, J., Hamilton, C. & Soderling, S. Regulation of spine structural plasticity by Arc/Arg3.1. *Semin Cell Dev Biol* **77**, 25–32 (2018).

118. Okuno, H., Minatohara, K. & Bito, H. Inverse synaptic tagging: An inactive synapse-specific mechanism to capture activity-induced Arc/arg3.1 and to locally regulate spatial distribution of synaptic weights. *Semin Cell Dev Biol* **77**, 43–50 (2018).

119. Plath, N. *et al.* Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. *Neuron* **52**, 437–444 (2006).

120. Manago, F. *et al.* Genetic Disruption of Arc/Arg3.1 in Mice Causes Alterations in Dopamine and Neurobehavioral Phenotypes Related to Schizophrenia. *Cell Rep* **16**, 2116–2128 (2016).

121. Suzuki, A., Yanagisawa, M. & Greene, R. W. Loss of Arc attenuates the behavioral and molecular responses for sleep homeostasis in mice. *Proceedings of the National Academy of Sciences* **117**, 10547–10553 (2020).

122. Kitazawa, M., Tamura, M., Kaneko-Ishino, T. & Ishino, F. Severe damage to the placental fetal capillary network causes mid- to late fetal lethality and reduction in placental size in Peg11/Rtl1 KO mice. *Genes Cells* **22**, 174–188 (2017).

123. Kitazawa, M. *et al.* Deficiency and overexpression of Rtl1 in the mouse cause distinct muscle abnormalities related to Temple and Kagami-Ogata syndromes. *Development* **147**, dev185918 (2020).

124. Kitazawa, M., Sutani, A., Kaneko-Ishino, T. & Ishino, F. The role of eutherian-specific RTL1 in the nervous system and its implications for the Kagami-Ogata and Temple syndromes. *Genes Cells* **26**, 165–179 (2021).

125. Chou, M.-Y. *et al.* RTL1/PEG11 imprinted in human and mouse brain mediates anxiety-like and social behaviors and regulates neuronal excitability in the locus coeruleus. *Human Molecular Genetics* **31**, 3161–3180 (2022).

126. Dalmau, J. *et al.* Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. *Brain* **122**, 27–39 (1999).

127. Henry, K. Paraneoplastic syndromes: Definitions, classification, pathophysiology and principles of treatment. *Seminars in Diagnostic Pathology* **36**, 204–210 (2019).

128. Tan, K. O. *et al.* MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. *J Biol Chem* **276**, 2802–2807 (2001).

129. Tan, K. O. *et al.* MAP-1 is a mitochondrial effector of Bax. *PNAS* **102**, 14623–14628 (2005).

130. Fu, N. Y., Sukumaran, S. K. & Yu, V. C. Inhibition of ubiquitin-mediated degradation of MOAP-1 by apoptotic stimuli promotes Bax function in mitochondria. *PNAS* **104**, 10051–10056 (2007).

131. Huang, N.-J. *et al.* The Trim39 ubiquitin ligase inhibits APC/CCdh1-mediated degradation of the Bax activator MOAP-1. *Journal of Cell Biology* **197**, 361–367 (2012).

132. Law, J. *et al.* Modulator of apoptosis 1 (MOAP-1) is a tumor suppressor protein linked to the RASSF1A protein. *J Biol Chem* **290**, 24100–24118 (2015).

133. Vargiu, L. *et al.* Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **13**, 7 (2016).

134. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**, D61–D65 (2007).

135. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

136. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res* **47**, W5–W10 (2019).

137. Hallin, E. I. *et al.* Structure of monomeric full-length ARC sheds light on molecular flexibility, protein interactions, and functional modalities. *J. Neurochem.* **147**, 323–343 (2018).

138. Qu, K. *et al.* Structure and architecture of immature and mature murine leukemia virus capsids. *PNAS* **115**, E11751–E11760 (2018).

139. Acton, O. *et al.* Structural basis for Fullerene geometry in a human endogenous retrovirus capsid. *Nature Communications* **10**, 5822 (2019).

140. Ball, N. J. *et al.* Structure of a Spumaretrovirus Gag Central Domain Reveals an Ancient Retroviral Capsid. *PLOS Pathogens* **12**, e1005981 (2016).

141. Nielsen, L. D., Pedersen, C. P., Erlendsson, S. & Teilum, K. The Capsid Domain of Arc Changes Its Oligomerization Propensity through Direct Interaction with the NMDA Receptor. *Structure* **27**, 1071-1081.e5 (2019).

142. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276–277 (2000).

143. Michel, A. M. *et al.* GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Research* **42**, D859–D864 (2014).

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

145. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

146. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

147. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).

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

149. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

150. Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N. & Delsuc, F. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Molecular Biology and Evolution* **35**, 2582–2584 (2018).

151. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321 (2010).

152. Kumar, S. *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* **39**, msac174 (2022).

153. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006).

154. Anisimova, M., Nielsen, R. & Yang, Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**, 1229–1236 (2003).

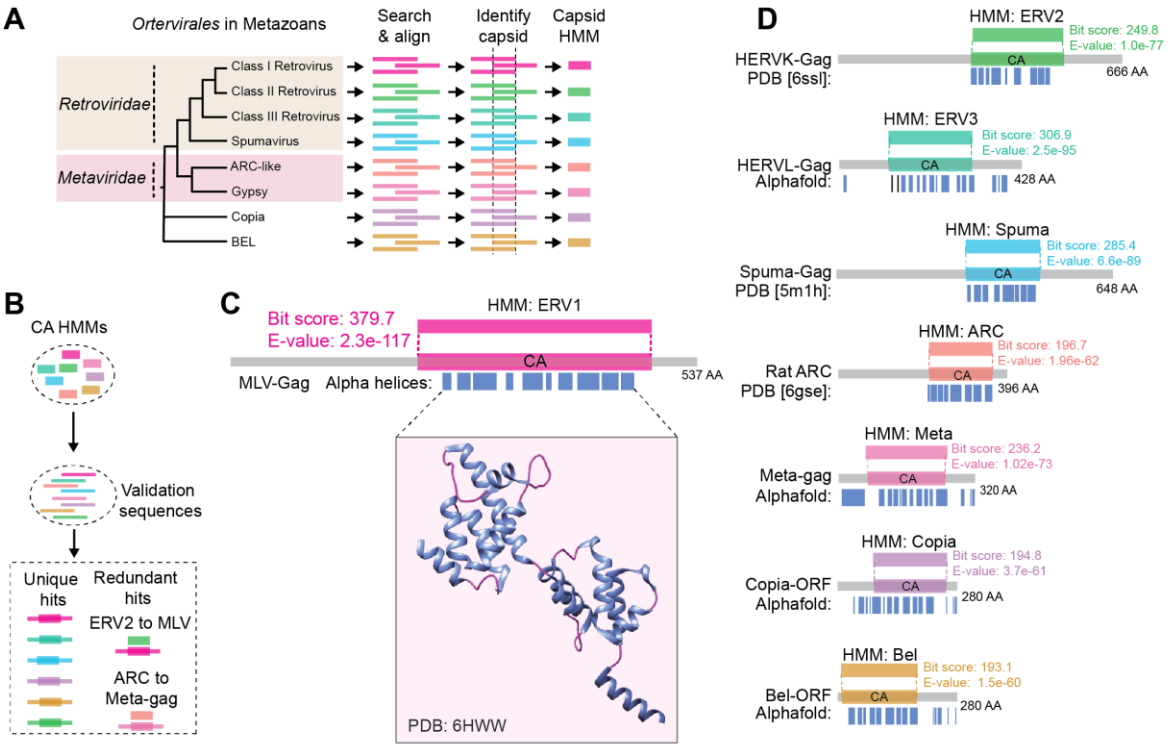
155. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).

156. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).

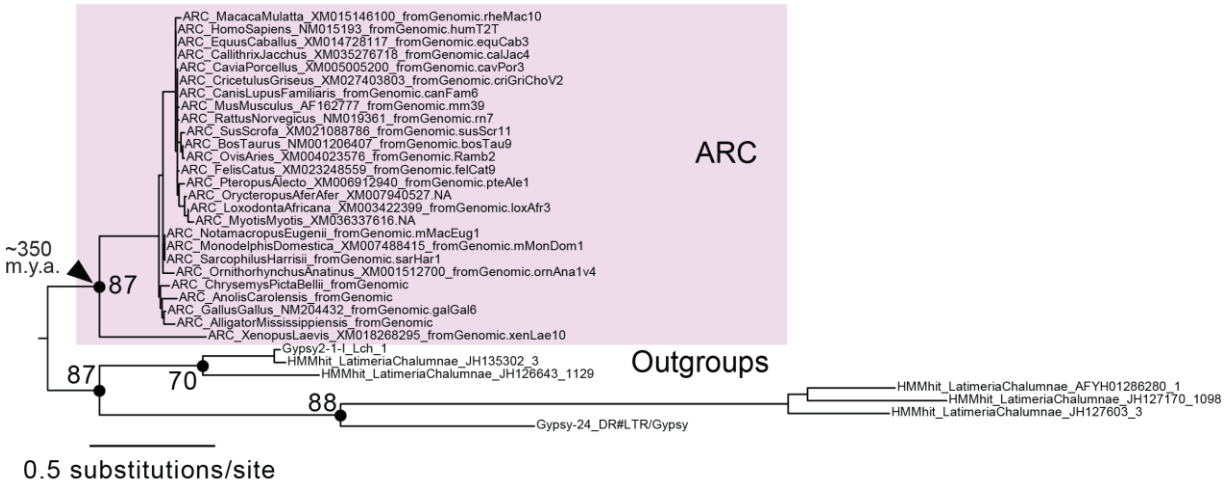
157. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

158. Kolde, R. pheatmap: Pretty Heatmaps. (2019).

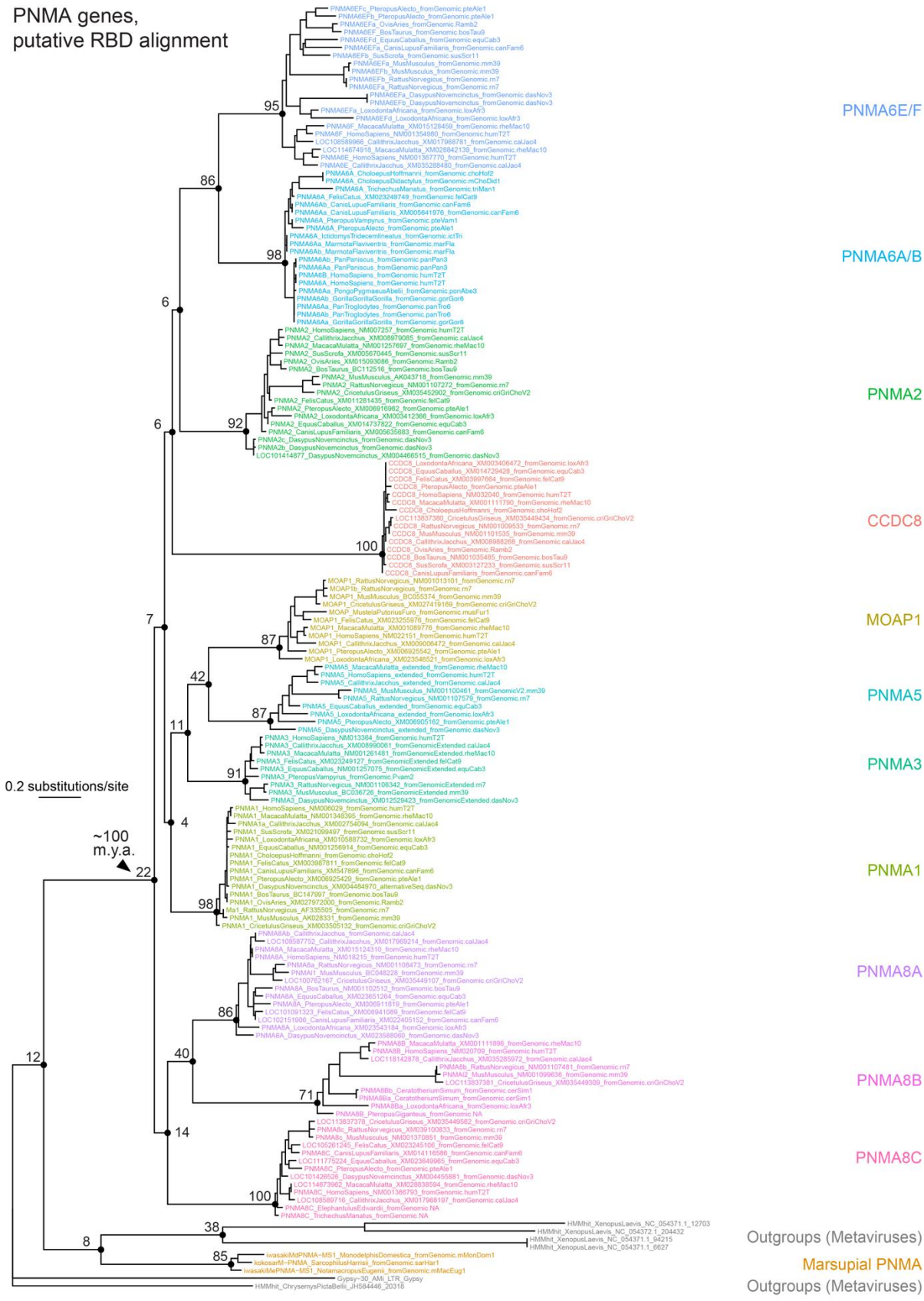
Supplementary Figures



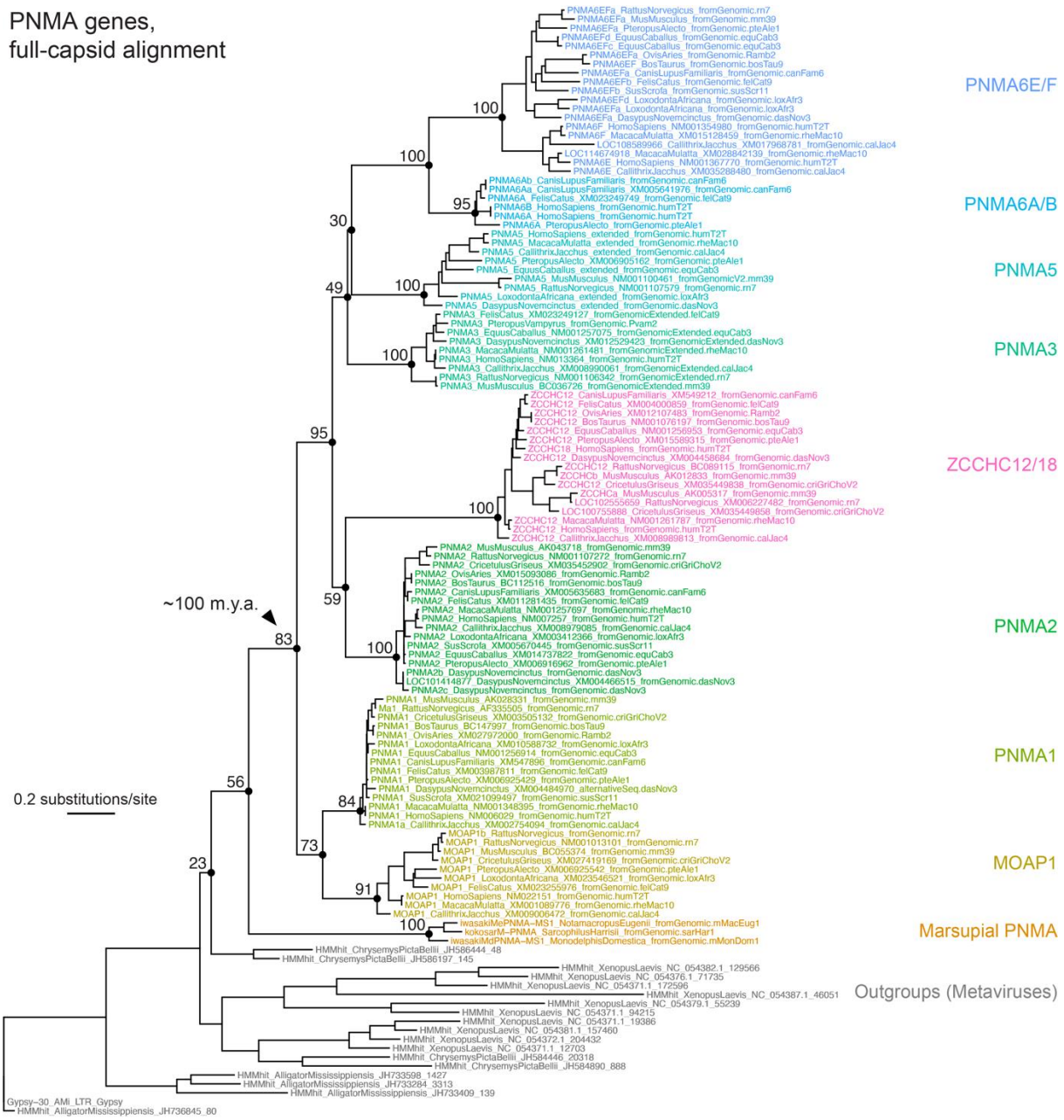

Supplementary Figure 4.1. Constructing Hidden Markov Models for the capsid domain of LTR retrotransposons. A. Seed sequences from four groups from Retroviridae, two groups of Metaviridae (Gypsy/Ty3), and one group each of Pseudoviridae (Copia/Ty1) and Belpaoviridae (BEL) were used to generate clade-specific full-length capsid HMMs (schematic phylogeny derived from (Gifford et al. 2018; Krupovic et al. 2018). Colors used in (A) are consistent in every other panel. B. Capsid HMMs were validated by querying a pool of sequences containing a single sequence from each of the clades in (A). Only two redundant hits were observed (see Methods) C. The capsid HMMs build for Class 1 retroviruses (HMM: ERV1 – colored magenta) identifies the full-length capsid domain (PDB: 6HWW) of the murine-leukemia virus (MLV) gag open reading frame (ORF) (colored grey). Alpha helices are schematicized in 2D representation below the ORF (vertical blue bars). D. Summary views of the same analysis conducted in (C) for all custom capsid HMMs shows that these HMMs cover the full-length capsid domain and identify positive control sequences with high bit-scores and low e-values. Off-target hits for ERV2 and ARC HMMs had bit scores an order of magnitude lower and E-values many orders of magnitude higher (not shown – see Methods). Gag-like open reading frames containing capsid domains are represented as grey lines with length and name indicated. The name and length of the capsid HMM are shown above the ORF, with the capsid domain identified highlighted on the ORF in the same color as in (A). Some capsids begin with a short beta-hairpin. In the case of HERVL-Gag, this beta-hairpin is indicated with two vertical black lines underneath the open reading frame




Supplementary Figure 4.2. Phylogeny of ARC proteins. ARC-like sequences and outgroup metaviruses from coelacanth (Latimeria chalumnae - Lch) and zebra fish (Danio rerio – DR) were translated and aligned using MAFFT. The alignment was manually trimmed to a well-aligned region, and a phylogenetic tree was built using the LG amino acid substitution model with 100 bootstrap replicates in PHYML. Bootstrap values are shown only at key nodes in the tree. Because orthologs can be found from frog to mammals, the domestication event must have occurred before the divergence of these lineages ~350 million years ago (Kumar et al. 2022), as reported previously by Pastuzyn et al. (2018).

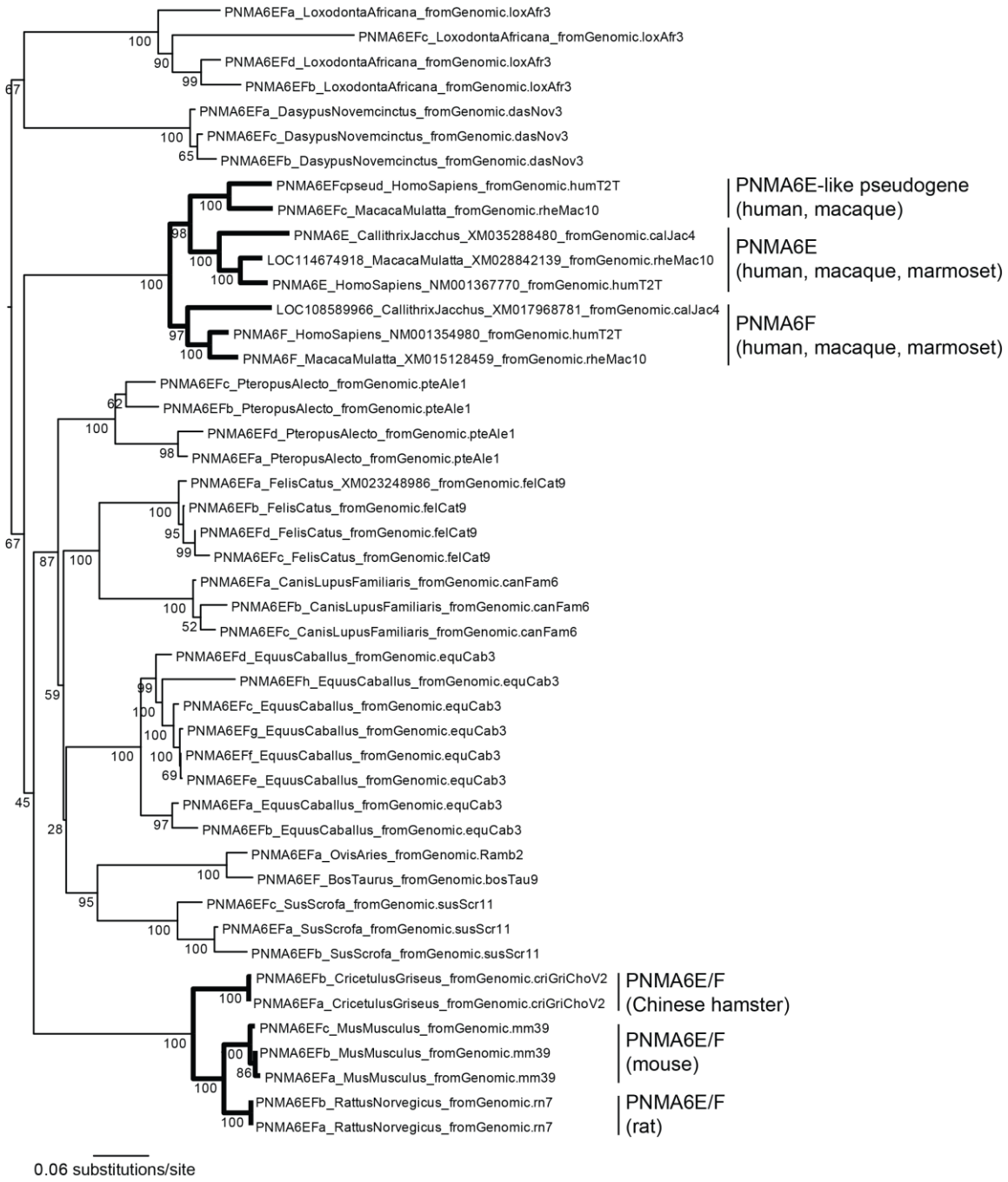

PNMA genes,
putative RBD alignment
PNMA6E/F
PNMA6A/B
PNMA2
CCDC8
MOAP1
PNMA5
PNMA3
PNMA1
PNMA8A
PNMA8B
PNMA8C
Outgroups (Metaviruses)
Marsupial PNMA
Outgroups (Metaviruses)
0.2 substitutions/site
~100
m.y.a.
95
86
98
6
6
92
100
7
87
42
87
11
91
4
22
98
86
40
12
71
14
100
38
8
85

Supplementary Figure 4.3. Phylogeny of the putative RNA-binding domain of PNMA proteins. Intact PNMA genes were translated and aligned using MAFFT along with capsid HMM matches (i.e. presumed active metaviral sequences) from the African clawed frog (Xenopus laevis), the painted turtle (Chrysemys picta bellii) and the American alligator (Alligator mississippiensis), as well as the Gypsy-30_Ami sequence from Repbase (the 'ancestral proxy' we selected for this gene group). The alignment was manually trimmed to a well-aligned region: because there is no single region shared by all PNMA genes, we chose for this tree to focus on the putative RBD (Figure 4.5), as it is shared by a larger number of the domesticated genes than the capsid region. This tree is therefore missing the ZCCHC12 and ZCCHC18 sequences: they are included in the corresponding capsid-region phylogeny (Supplementary Figure 4.4). We used the trimmed alignment to generated a phylogeny in PHYML using the JTT amino acid substitution model with 100 bootstrap replicates. Numbers show bootstrap support of the nearby nodes with black dots. PNMA6E and PNMA6F form a single mixed clade that was not well-resolved in this analysis, which we explored further with a more focused PNMA6E/F-only tree (Supplementary Figure 4.5).

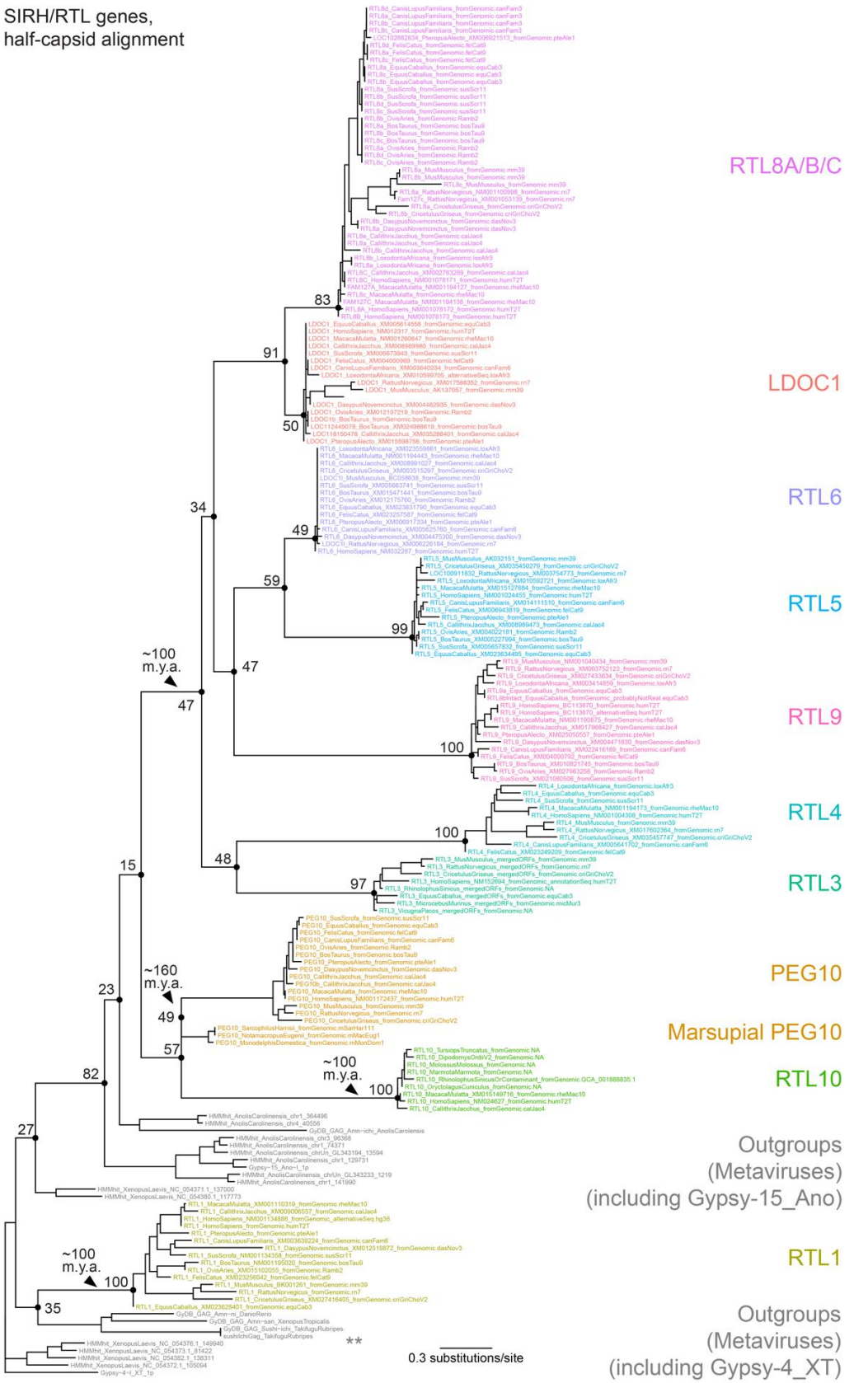



Supplementary Figure 4.4. Phylogeny of PNMA protein capsid domains. PNMA genes were translated and aligned using MAFFT along with capsid HMM matches (i.e. presumed active metaviral sequences) from the African clawed frog (Xenopus laevis), the painted turtle (Chrysemys picta belli) and the American alligator (Alligator mississippiensis), as well as the Gypsy-30_Ami sequence from Repbase (the 'ancestral proxy' we selected for this gene group). The alignment was manually trimmed to a well-aligned region: because there is no single region shared by all PNMA genes, and to confirm the gene relationships observed in the PNMA family N-terminal putative RNA-binding domain, we chose for this tree to focus on the capsid region. This tree is therefore missing the CCDC8 and PNMA8A/B/C sequences: these genes are included in the RBD region tree (Supplementary Figure 4.3). We used the trimmed alignment to

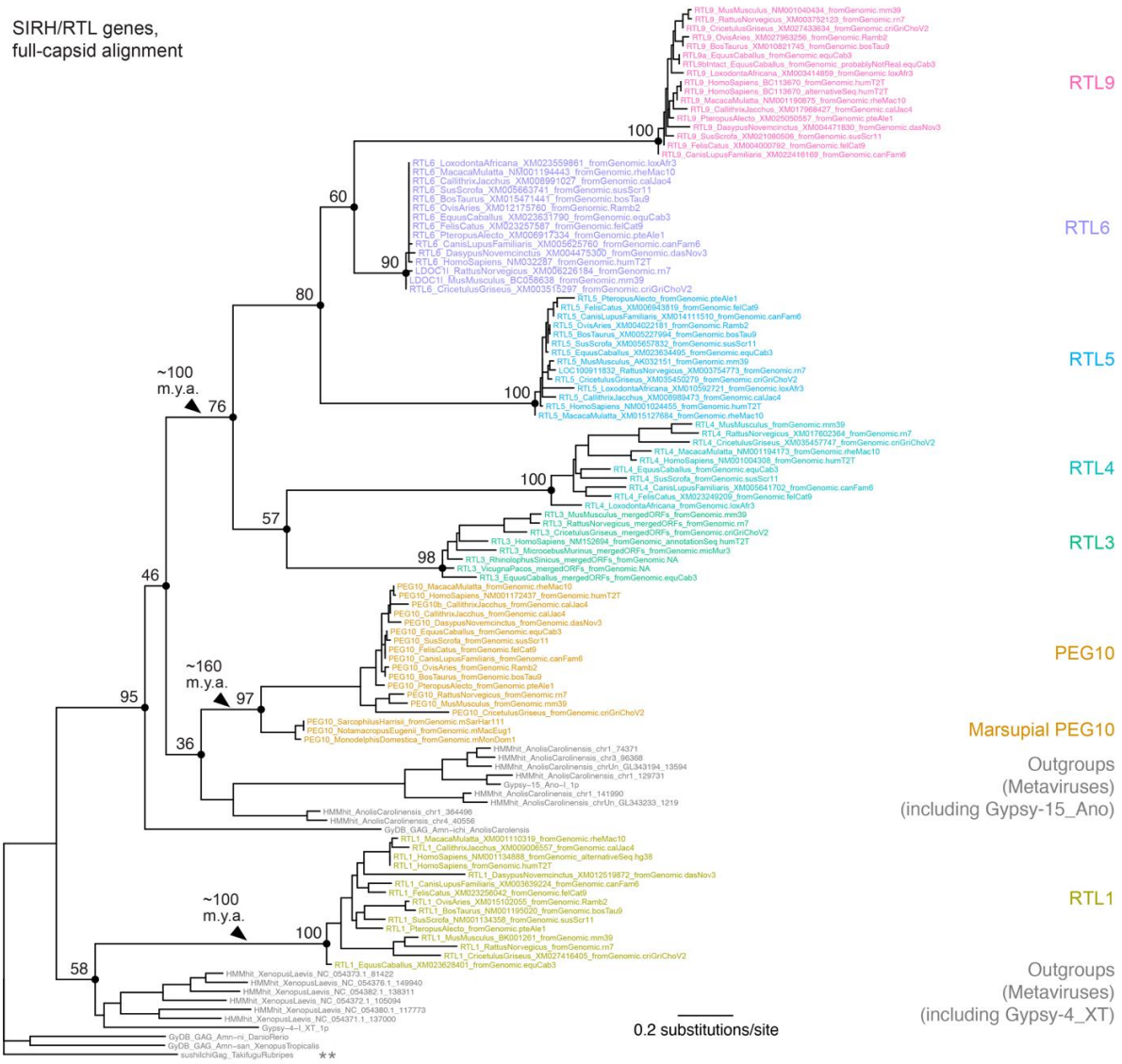
generated a phylogeny in PHYML using the JTT amino acid substitution model with 100 bootstrap replicates. Numbers show bootstrap support of the nearby nodes with black dots. PNMA6E and PNMA6F form a single well-supported clade.



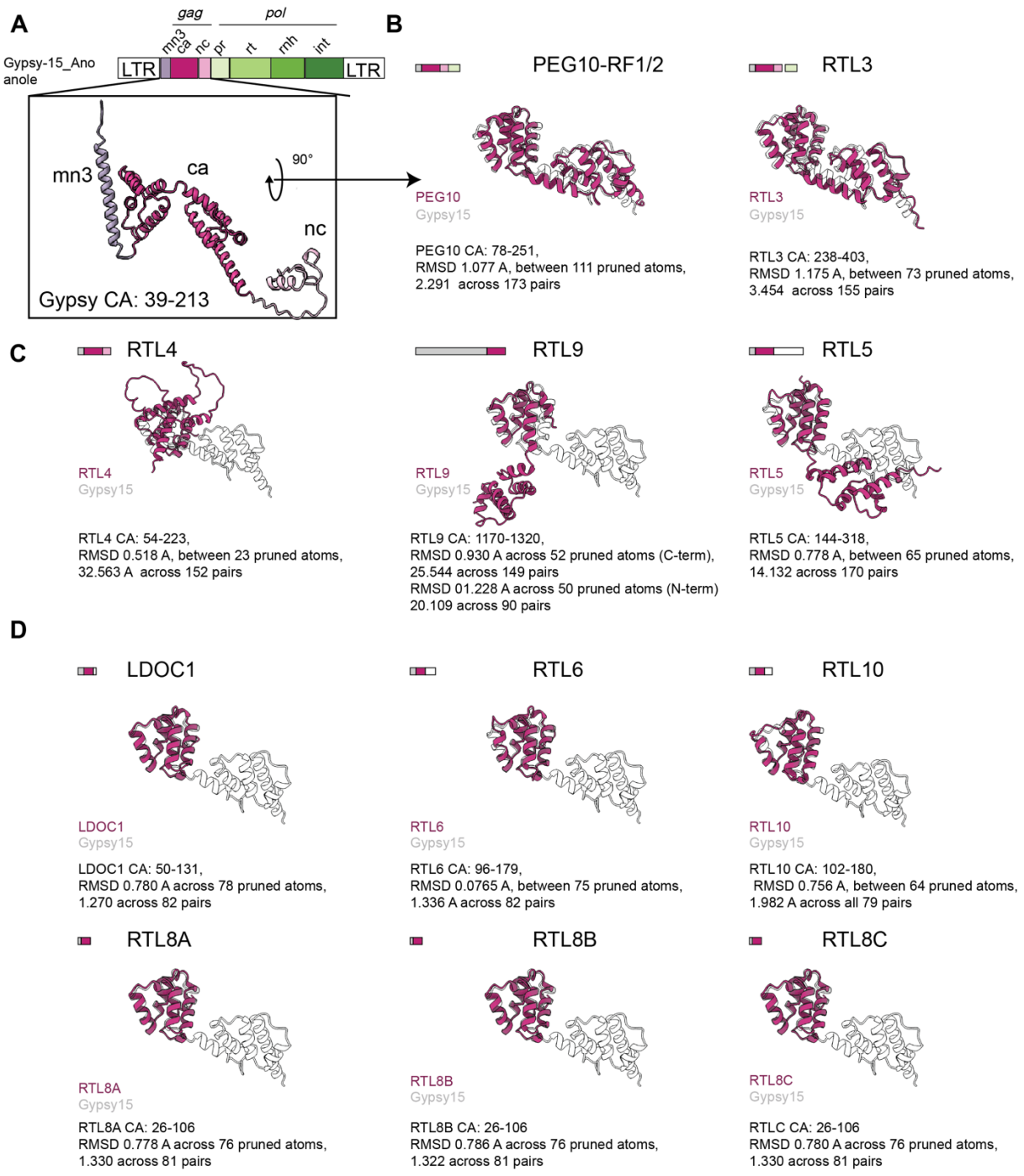

Supplementary Figure 4.5. PNMA6E/F has duplicated multiple times in distinct lineages. PNMA6E/F nucleotide sequences (including intact genes and pseudogenes) were translated and aligned in MACSE, with manual adjustment. We then manually trimmed the alignment to remove a poorly-aligned internal repetitive region and removed alignment positions where >30% of sequences contained a gap. A phylogeny was generated using PHYML with 100 bootstrap replicates (node labels represent bootstrap support). In several instances (indicated with bold branches in the tree, and labels to the right) PNMA6E/F sequences group with paralogs from the same species, rather than with presumed orthologs from other species.


SIRH/RTL genes,
half-capsid alignment
RTL8A/B/C
LDOC1
RTL6
RTL5
RTL9
RTL4
RTL3
PEG10
Marsupial PEG10
RTL10
Outgroups
(Metaviruses)
(including Gypsy-15_Ano)
RTL1
Outgroups
(Metaviruses)
(including Gypsy-4_XT)
~100
m.y.a.
~160
m.y.a.
~100
m.y.a.
~100
m.y.a.
0.3 substitutions/site

Supplementary Figure 4.6. Phylogenetic tree for SIRH/RTL family (half capsid alignment). Intact genes from the SIRH/RTL family were translated and aligned using MAFFT together with outgroup metaviral sequences from genomes of the anole lizard (Anolis carolensis), Western clawed frog (Xenopus tropicalis), and African clawed frog (Xenopus laevis), as well as selected Repbase sequences ('ancestral proxies'), the sushi-ichi element (NCBI accession AAC33525.1) and selected metaviral sequences from the GyDB database of metaviral elements (Llorens et al. 2011). The alignment was manually trimmed to well-aligned regions, which in this case comprises the half-capsid region shared by all family members. The trimmed alignment was degapped and used to generated a phylogeny in PHYML with the JTT amino acid substitution model and 100 bootstrap replicates. Numbers represent bootstrap support of selected key nodes, highlighted with black dots. **Indicates the position of the sushi-ichi metaviral capsid protein, used in some previous publications as an outgroup sequence. In this and in other phylogenetic analyses (e.g. Figure 3A, as well as Figure 2 of Kokosar & Kordis 2013, and Figure S5.1 of Campillos et al, 2006), RTL1 consistently groups distinctly from the other SIRH/RTL proteins, suggesting independent domestication.

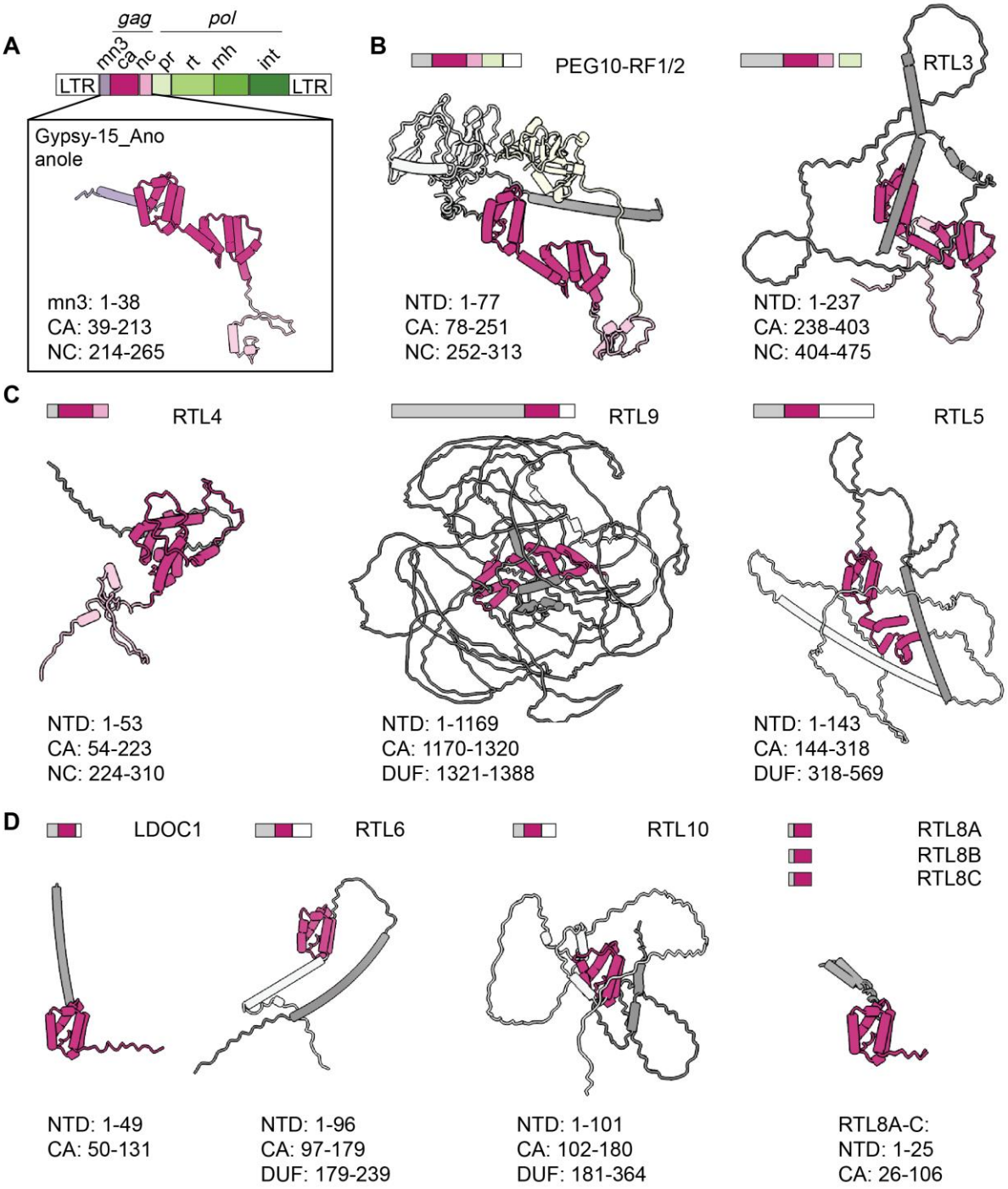



Supplementary Figure 4.7. Phylogenetic tree for SIRH/RTL family (full capsid alignment). Intact genes from the SIRH/RTL family were translated and aligned using MAFFT together with outgroup metaviral sequences from genomes of the anole lizard (Anolis carolensis) and African clawed frog (Xenopus laevis), as well as selected Repbase sequences ('ancestral proxies'), the sushi-ichi element (NCBI accession AAC33525.1) and selected metaviral sequences from the GyDB database of metaviral elements (Llorens et al. 2011). The alignment was manually trimmed to well-aligned regions, which in this case comprises the full-length capsid region. The trimmed alignment was degapped and used to generated a phylogeny in PHYML with the JTT amino acid substitution model and 100 bootstrap replicates. Numbers represent bootstrap support of selected key nodes, highlighted with black dots. **Indicates the position of the sushi-ichi metaviral capsid protein, used in some previous publications as an outgroup sequence. In this and in other phylogenetic analyses (e.g. Figure 3A, as well as Figure 2 of Kokosar & Kordis 2013, and Figure S5.1 of Campillos et al, 2006), RTL1 consistently groups distinctly from the other SIRH/RTL proteins, suggesting independent domestication.



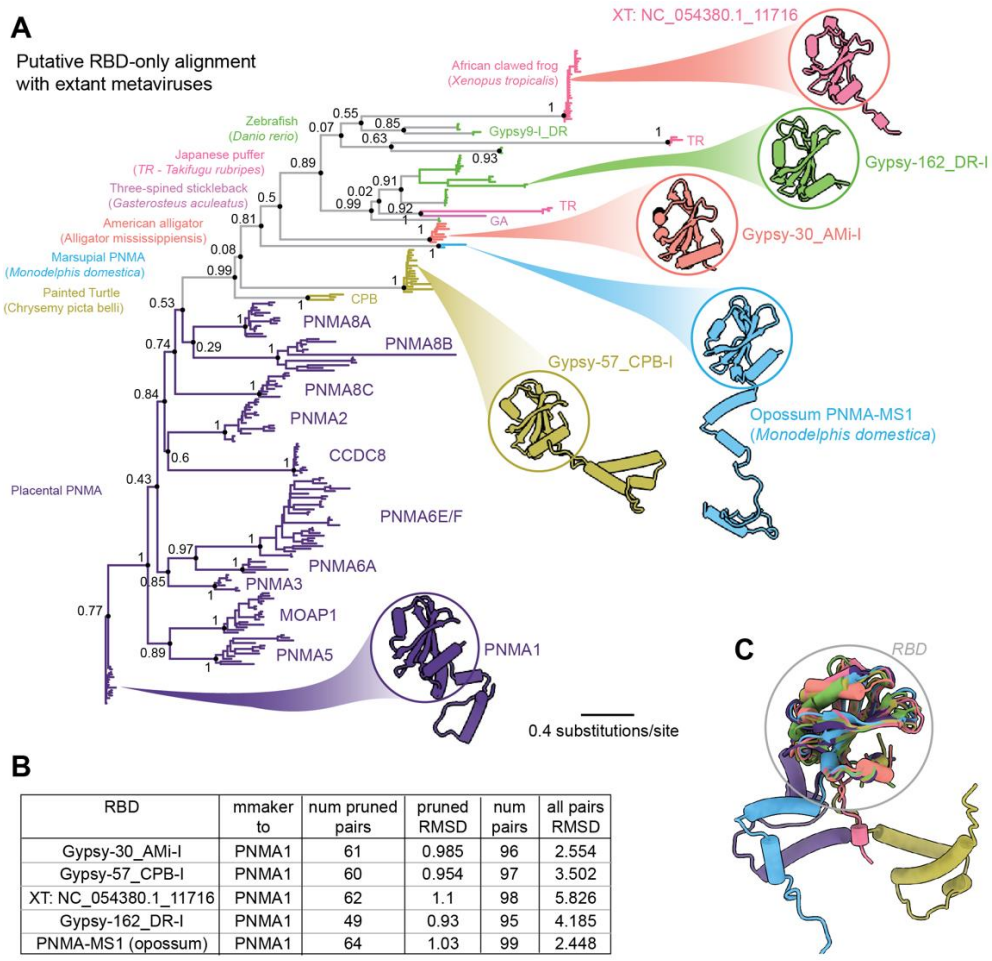

Supplementary Figure 4.8. Conservation and divergence of the capsid domain in the SIRH/RTL family. A. The AlphaFold-predicted structure of the gag domain of the 'ancestral proxy' metavirus of the SIRH/RTL family, Gypsy-15_Ano, colored by N-terminus (mn3 - lilac), capsid (ca -magenta), and nucleocapsid (nc - light pink) domains. Additional pol domains are not shown (pr – protease, rt – reverse transcriptase, rnh – RNAseH, int – integrase). We compared the AlphaFold predicted structure of the capsid domain of each human protein in the SIRH/RTL

family to this ‘ancestral proxy’ metavirus for the full-length capsid domains (B-C) and N-terminal lobe only of the capsid domain (D) (rotated by 90o compared to (A) to clearly show the two capsid lobes). Capsid domain regions identified by HMMER were aligned to the Gypsy-15_Ano capsid domain in ChimeraX with the mmaker command. The coordinates of the capsid domain and the root mean square deviation (RMSD) for alignment of the structure predictions is shown below each overlay. B. PEG10 and RTL3 are highly similar to the full-length capsid domain of Gypsy-15_Ano. C. RTL4, RTL9, and RTL5 have structurally similar N-terminal lobes, but diverge either in predicted orientation or structure of the C-terminal lobe. D. Truncated capsid genes that have lost the C-terminal lobe are structurally identical to the N-terminal lobe of Gypsy-15_Ano. Overall, the N-terminal lobe of the capsid domain is structurally identical across all proteins encoded by the SIRH/RTL family of genes.




Supplementary Figure 4.9. Lack of structural conservation in the N-terminal domain of the SIRH/RTL family. We compared the length and architecture of AlphaFold-predicted structures for the N-terminal domain (NTD) upstream of the capsid (CA) domain (magenta) of the 'ancestral proxy' metavirus Gypsy-15_Ano (A) to the SIRH/RTL family of genes (B-D). The N-terminal domain of the SIRH/RTL family lacks conservation and is made up of predicted alpha helices and disordered regions from 25 amino acids long (RTL8A-C) to 1169 amino acids (light

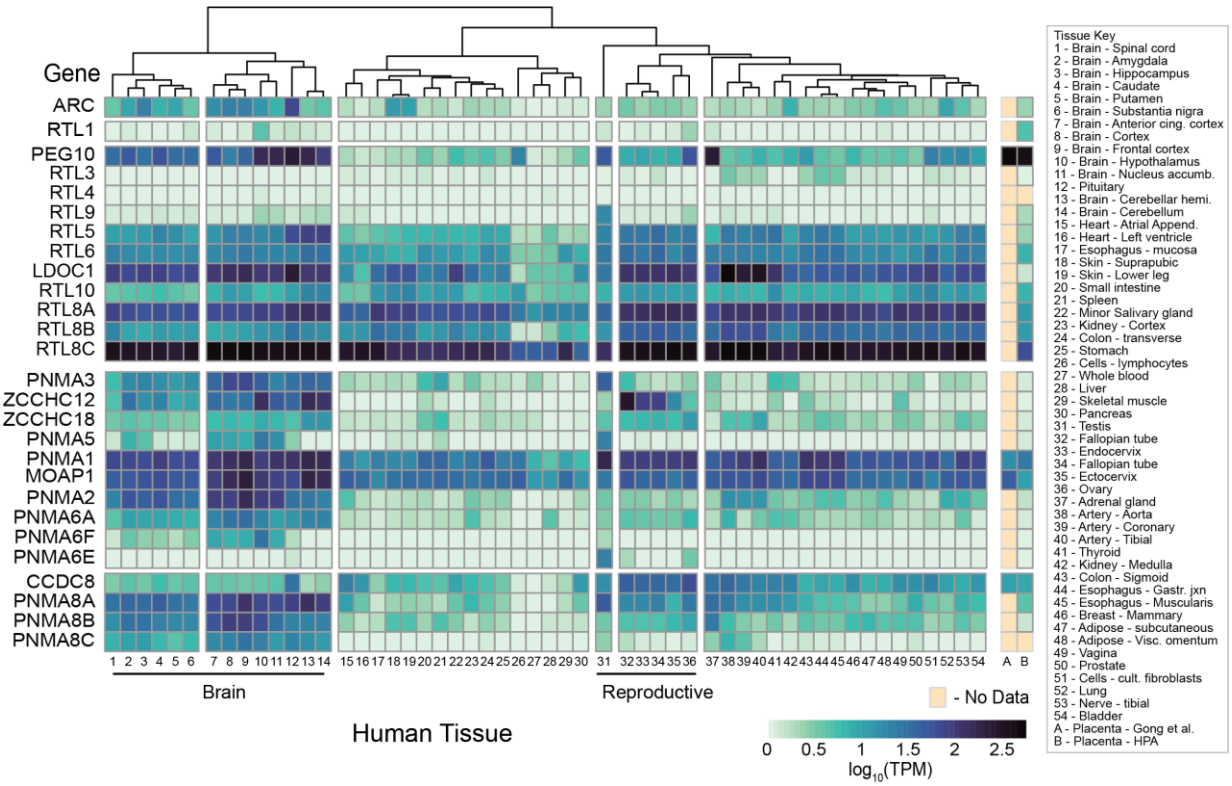
purple). The amino-acid coordinates for each domain are listed below each structure. A. The ancestral proxy Gypsy-15_Ano lacks a canonical matrix domain and has instead a 53 amino acid alpha helical NTD (mn3 – N-terminus of Gypsy15-Ano, ca – capsid, nc – nucleocapsid, pr – protease, rt – reverse transcriptase, rnh – RNAseH, int – integrase). B. PEG10 and RTL3 have preserved canonical CA and NC domains, but have divergent N-termini. C. The other full-length capsid genes have neither NC domains nor conserved N-terminal domains. D. Truncated capsid genes in the SIRH/RTL family have divergent N-termini in both length and predicted structure.




| RBD | mmaker to | num pruned pairs | pruned RMSD | num pairs | all pairs RMSD |
|---|---|---|---|---|---|
| Gypsy-30_AMi-I | PNMA1 | 61 | 0.985 | 96 | 2.554 |
| Gypsy-57_CPB-I | PNMA1 | 60 | 0.954 | 97 | 3.502 |
| XT: NC_054380.1_11716 | PNMA1 | 62 | 1.1 | 98 | 5.826 |
| Gypsy-162_DR-I | PNMA1 | 49 | 0.93 | 95 | 4.185 |
| PNMA-MS1 (opossum) | PNMA1 | 64 | 1.03 | 99 | 2.448 |

Supplementary Figure 4.10. The evolutionary origins of the putative RNA-binding domain of the PNMA family. The putative RNA-binding domain (RBD) of the PNMA family originated from metaviruses in two separate domestication events, one in marsupials and one in placental mammals. A. Metaviral sequences containing putative RBDs are found in fish, reptiles and amphibians. Putative RNA-binding domains were found by querying either six-ORF translations of 22 genomes or a six-ORF translation of the Repbase database with a putative PNMA RBD hidden Markov model (HMM). Sequences with significant hits (bit-score > 25) longer than 100 amino acids were aligned and the alignment used to construct an unrooted maximum likelihood phylogenetic tree in FastTree2 using the JTT model of amino acid evolution and gamma rate parameters. Support values are maximum-likelihood estimates of bootstrap support. Popouts show AlphaFold predictions of the identified putative RNA-binding domains for each major clade, visualized in ChimeraX. (Abbreviations: “DR” Danio rerio/zebrafish; “GA” Gasterosteus aculeatus/three-spined stickleback; “XT” Xenopus tropicalis/Western clawed frog; “AMi” Alligator mississippiensis/American alligator; “CPB” Chrysemys picta bellii/painted turtle). Gypsy-30_AMi, the ‘ancestral proxy’ used for the PNMA genes in Figure 4, is shown near the top of this tree. B-C. Structural overlay of the first ~100 residues of AlphaFold predictions of PNMA1 with representative metaviral sequences from amphibians, fish, and reptiles indicates a

striking degree of structural conservation in the putative RBD (RMSD of ~1 Å across ~60 pruned alpha carbons, calculated from mmaker in ChimeraX – table of statistics in B).




Supplementary Figure 4.11. Human expression of domesticated metaviral capsid genes. Bulk RNA-sequencing data for 54 human tissues was downloaded from the GTEX portal, and median transcript per million (TPM) value plotted for each gene in each tissue. Data from the placenta were downloaded from the Human Protein Atlas and from Gong et al. and converted to median TPM values. Expression data reveals that most domesticated metaviral genes are widely expressed across tissues, while some are only expressed in specific tissues (e.g. RTL1 in the brain and placenta). Numbers indicate specific tissues from GTEX (Lonsdale et al. 2013), placental data collated from A (Gong et al. 2021) and B (Sjöstedt et al. 2020).

Supplementary Tables

| LTR Retroelement | MSA (num seqs) | MSA (length) | HMM length | Bit-score | E-value |
|---|---|---|---|---|---|
| ***Retroviridae* - Class I ERV** | 149 | 294 | 249 | 379.7 | 2.30E-117 |
| ***Retroviridae* - Class II ERV** | 173 | 270 | 213 | 249.8 | 1.00E-77 |
| ***Retroviridae* - Class III ERV** | 17 | 203 | 197 | 306.9 | 2.50E-95 |
| ***Retroviridae* - Spumavirus** | 14 | 175 | 175 | 285.4 | 6.60E-89 |
| ***Metaviridae* - Gypsy** | 96 | 220 | 192 | 236.2 | 1.02E-73 |
| ***Metaviridae* – ARC** | 55 | 197 | 149 | 196.7 | 1.96E-62 |
| ***Pseudoviridae* - Copia** | 212 | 207 | 174 | 194.8 | 3.70E-61 |
| ***Belpaoviridae* - BEL** | 221 | 204 | 180 | 193.1 | 1.50E-60 |

Supplementary Table 4.1. Model statistics for custom capsid Hidden Markov Models (HMMs) build from multiple sequence alignments (MSA) of capsid amino acid sequences. Bit-scores and E-values are for the control sequence from each clade used to test the HMMs (**Supplementary Figure 4.1**).

| **Official human gene (HGNC) symbol** | ***SIRH* name ("sushi-ichi-related retrotransposon homolog")** | ***RTL* name ("retrotransposon like")** | ***MART* name ("mammalian retrotransposon-derived")** | ***PNMA* name ("paraneoplastic Ma antigen")** | **Gene name in Kokosar et al, 2013** | **Additional synonyms** |
|---|---|---|---|---|---|---|
| ***ARC*** | | | | | *ARC* | *Arg3.1, hArc, KIAA0278* |
| ***RTL1*** | *SIRH2* | *RTL1* | *Mart1* | | *RTL1* | *HUR1, Mar1, PEG11* |
| ***LDOC1*** | *SIRH7* | *RTL7* | *Mart7* | | *LDOC1* | *BCUR1, Mar7, RTL7* |
| ***PEG10*** | *SIRH1* | *RTL2* | *Mart2* | | *PEG10* | *EDR, HB-1, Mar2, MEF3L, RGAG3, KIAA1051* |
| ***RTL10*** | | *RTL10* | | | *C22orf29* | *BOP, FLJ21125* |
| ***RTL3*** | *SIRH9* | *RTL3* | *Mart3* | | *ZCCHC5* | *Mar3, ZHC5, FLJ38865* |
| ***RTL4*** | *SIRH11* | *RTL4* | *Mart4* | | *ZCCHC16* | *Mar4, FLJ46608* |
| ***RTL5*** | *SIRH8* | *RTL5* | *Mart5* | | *RGAG4* | *6430402L03Rik, MAR5, KIAA2001* |
| ***RTL6*** | *SIRH3* | *RTL6* | *Mart6* | | *LDOC1L* | *dJ1033E15.2, Mar6, DKFZp761O17121* |
| ***RTL8A*** | *SIRH6* | *RTL8A* | *Mart8A* | | *FAM127B* | *CXX1b, DKFZP564B147* |
| ***RTL8B*** | *SIRH4* | *RTL8B* | *Mart8B* | | *FAM127C* | *CXX1c, MAR8B* |
| ***RTL8C*** | *SIRH5* | *RTL8C* | *Mart8* | | *FAM127A* | *CXX1, Mar8, MAR8C, MART8C* |
| ***RTL9*** | *SIRH10* | *RTL9* | *Mart9* | | *RGAG1* | *Mar9, KIAA1318* |
| ***CCDC8*** | | | | | *CCDC8* | *3M3, p90, PPP1R20,* |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | *DKFZp564K032 2* |
| ***MOAP1*** | | | | *PNMA4* | *MOAP1* | *MAP-1* |
| ***PNMA1*** | | | | *PNMA1* | *PNMA1* | *MA1* |
| ***PNMA2*** | | | | *PNMA2* | *PNMA2* | *MA2, MM2, RGAG2* |
| ***PNMA3*** | | | | *PNMA3* | *PNMA3* | *MA3, MA5, MGC132756, MGC132758* |
| ***PNMA5*** | | | | *PNMA5* | *PNMA5* | *KIAA1934* |
| ***PNMA6A*** | | | | *PNMA6A* | *PNMA6A* | *MA6, PNMA6, PNMA6C, MGC15827* |
| ***PNMA6B*** | | | | | *PNMA6B* | *PNMA6D* |
| ***PNMA6E*** | | | | *PNMA6E* | | |
| ***PNMA6F*** | | | | *PNMA6F or PNMA6BL* | *PNMA6BL* | |
| ***PNMA8A*** | | | | *PNMA8A* | *PNMAL1* | *KIAA1183L, FLJ10781* |
| ***PNMA8B*** | | | | *PNMA8B* | *PNMAL2* | *KIAA1183* |
| ***PNMA8C*** | | | | | | |
| ***ZCCHC12*** | | | | *PNMA7A* | *ZCCHC12* | *SIZN, SIZN1, FLJ16123* |
| ***ZCCHC18*** | | | | *PNMA7B* | *ZCCHC18* | *SIZN2* |

Supplementary Table 4.2. Gene names table. Compilation of all synonyms for the gene names used in this and previous studies

| Common name | Scientific name | Source | Genome |
|---|---|---|---|
| African bush elephant | Loxodonta africana | UCSC | loxAfr3.fa |
| African clawed frog | Xenopus laevis | NCBI | GCF_017654675.1_Xenopus_laevis_v10.1_genomic.fna |
| American alligator | Alligator mississippiensis | UCSC | allMis1.fa |
| Black flying fox | Pteropus alecto | NCBI | GCF_000325575.1_ASM32557v1_genomic.fna |
| Chicken | Gallus gallus | UCSC | galGal6.fa |
| Chinese Hamster | Cricetulus griseus | UCSC | criGriChoV2.fa |
| Cow | Bos taurus | UCSC | bosTau9.fa |
| Domestic dog | Canis lupus familiaris | UCSC | canFam6.fa |
| Gray short-tailed opossum | Monodelphis domestica | NCBI | GCF_027887165.1_mMonDom1.pri_genomic.fna |
| Green anole | Anolis carolinensis | UCSC | anoCar2.fa |
| Horse | Equus caballus | UCSC | equCab3.fa |
| House cat | Felis catus | UCSC | felCat9.fa |
| Human | Homo sapiens | NCBI | GCF_009914755.1_T2T-CHM13v2.0_genomic.fna |
| Japanese pufferfish | Takifugu rupripes | UCSC | fr3.fa |
| Marmoset | Callithrix jacchus | UCSC | calJac4.fa |
| Mouse | Mus musculus | UCSC | mm39.fa |
| Nine-banded armadillo | Dasypus novemcinctus | UCSC | dasNov3.fa |
| Painted turtle | Chrysemys picta bellii | UCSC | chrPic1.fa |
| Pig | Sus scrofa | UCSC | susScr11.fa |
| Platypus | Ornithorhynchus anatinus | NCBI | GCF_004115215.2_mOrnAna1.pri.v4_genomic.fna |
| Rat | Rattus norvegicus | UCSC | rn7.fa |
| Rhesus macaque | Macaca mulatta | UCSC | rheMac10.fa |
| Sheep | Ovis aeries | NCBI | GCF_016772045.1_ARS-UI_Ramb_v2.0_genomic.fna |
| Wallaby | Notamacropus eugenii | NCBI | GCA_028372415.1_mMacEug1.pri_genomic.fna |
| West Indian Ocean coelacanth | Latimeria chalumnae | UCSC | latCha1.fa |
| Zebra fish | Danio rerio | UCSC | danRer11.fa |

Supplementary Table 4.3. Table of genomes used in this study

| | Gypsy-4 | Gypsy-15 | Gypsy 2-1 | |
|---|---|---|---|---|
| Gypsy-4<br>1-68 | | | | |
| Gypsy-15<br>1-28 | -33<br>8%<br>21% | | | Pairwise BLAST score<br>Percent ID<br>Percent similarity |
| Gypsy 2-1<br>1-196 | -123<br>8%<br>17% | -155<br>6%<br>7% | | Pairwise BLAST score<br>Percent ID<br>Percent similarity |
| Gypsy-30<br>1-100 | -47<br>15%<br>28% | -70<br>8%<br>12% | -100<br>12%<br>23% | Pairwise BLAST score<br>Percent ID<br>Percent similarity |

Supplementary Table 4.4. BLAST statistics for consensus Repbase metaviral sequences used as proxies for ancestral metaviruses. Translated N-terminal domains from the start codon to the start of the capsid domain were aligned and scored using pBLAST.

| Fig 3 Clade | | Gene | Chr | KO | Zygosity | n | Tissue | Phenotype | p-value |
|---|---|---|---|---|---|---|---|---|---|
| b | | ARC | 8 | Plath et al. 2006 | -/- | 5-25 | brain | Disrupted memory consolidation | Sig |
| | | ARC | 8 | Manago et al. 2016 | -/- | 6-22 | brain | Disrupted social behavior, object recognition | Sig |
| | | ARC | 8 | Suzuki et al. 2020 | -/- | 4-14 | brain | Disrupted sleep regulation | Sig |
| c | | RTL1 | 14* | Sekita et al. 2008 | +/- (Pat. KO) | 13-35 | placenta | Disruption of placental development, lethal | Sig |
| | | RTL1 | 14* | Kitazawa et al. 2017 | +/- (Pat. KO) | 16-97 | placenta | Disruption of placental development, lethal | Sig |
| d | Full CA | PEG10 | 7* | Ono et al. 2006 | +/- (Pat. KO) | 2-53 | placenta | Embryonic lethal | Sig |
| | | RTL3 | X | IMPC | -/Y | 8 (M) | brain | Increased freezing behavior | Sig |
| | | | | | | | brain | Increased aggression | Sig |
| | | RTL9 | X | Ishino et al. 2023 | -/- | 2-8 | brain | Loss of antifungal immunity | NP |
| | | RTL4 | X | IMPC | -/Y | 7 (F) | NA | NA | NA |
| | | RTL4 | X | Irie et al. 2015 | -/Y | 7 (F) | brain | Increased impulsivity, memory deficits | Sig |
| | | RTL5 | X | IMPC | -/Y | 15 (M) | eye | abnormal eye morphology | Sig |
| | | | | | -/Y | 15 (M) | kidney | enlarged kidney | Sig |
| | | | | | -/Y | 15 (M) | heart | enlarged heart | Sig |
| | | | | | -/Y | 15 (M) | heart | abnormal heart morphology | Sig |
| | | | | | -/- | 8 (F) | blood | increased circulating total protein | Sig |
| | | | | | -/Y | 15 (M) | eye | microphthalmia | Sig |
| | | | | | -/Y | 15 (M) | kidney | abnormal kidney morphology | Sig |
| | | | | | -/Y | 8(M) | brain | Increased freezing behavior | Sig |
| | | RTL5 | X | Irie et. al. 2022 | -/- | 2-4 | brain | microglial innate immunity | NP |
| | Truncated CA | RTL6 | 22 | Irie et al. 2022 | -- | 2-4 | brain | microglial innate immunity | NP |
| | | LDOC1 | X | Naruse et al. 2014 | -/- // -/+ | 3-37 | placenta | Abnormal placenta, elevated progesterone | Sig |
| | | RTL10 | 22 | IMPC | -/- | 9 (F) | brain | hyperactivity | Sig |
| | | | | | -/- | 13 (M/F) | brain | abnormal vocalization | Sig |
| | | | | | -/- | 13 (M/F) | skin | abnormal coat/hair pigmentation | Sig |
| | | RTL8A/B/C | X/X/X | Fujioka et al. 2023[1] | DKO | 5-14 | fat tissue | Increased body weight | Sig |
| | | | | | | 5-14 | brain | Disrupted social behavior | Sig |
| e | | PNMA1 | 14 | IMPC | NP | NP | NA | None | NA |
| | | PNMA4 | 14 | No | NA | NA | NA | NA | NA |
| | RBD + CA | PNMA2 | 8 | IMPC | NP | NP | NA | None | NA |
| | | PNMA3 | X | IMPC | -/- | 8 (F) | blood | abnormal blood uric acid levels | Sig |
| | | PNMA5 | X | No | NA | NA | NA | NA | NA |
| | | PNMA6A | X | No | NA | NA | NA | NA | NA |
| | | PNMA6E/F | X | No | NA | NA | NA | NA | NA |
| | CA | ZCCHC12 | X | IMPC | NP | NP | NA | None | NA |
| | | ZCCHC18 | X | No | NA | NA | NA | NA | NA |
| | | CCDC8-ext | 19 | No | NA | NA | NA | NA | NA |
| | | CCDC8-original | 19 | IMPC | -/- | NA | NA | preweaning lethality, complete penetrance | Sig |
| | | | | | -/+ | 8 (M) | fat | increased total body fat amount | Sig |
| | RBD | PNMA8A | 19 | No | NA | NA | NA | NA | NA |
| | | PNMA8B | 19 | No | NA | NA | NA | NA | NA |
| | | PNMA8C | 19 | No | NA | NA | NA | NA | NA |

*Imprinted gene
[1]bioRxiv preprint

Abbreviations:
Sig = Significant
Pat = Paternal
NP = Not provided
NA = Not available
DKO = Double knockout
IMPC = International Mouse Phenotyping Consortium

Supplementary Table 4.5. Phenotypes associated with knockouts of domesticated metaviral genes in mice. P-values are reported as significant ($p<0.05$), not provided by the source (NP), or not applicable (NA) if no knockout has been generated.

# CHAPTER FIVE

# THE RETROTRANSPOSON-DERIVED CAPSID GENES *PNMA1* AND *PNMA4* MAINTAIN REPRODUCTIVE CAPACITY

## Contribution of Authors and Co-Authors

Manuscript in Chapter 5

Author: Thomas W.P. Wood

Contributions: Conceptualization, Methodology, Formal Analysis, Investigation, Original Draft, Review and Editing, Visualization

Co-Author: William S. Henriques

Contributions: Methodology, Formal Analysis, Investigation, Review and Editing, Visualization,

Co-Author: Harrison B. Cullen

Contributions: Methodology, Investigation, Review and Editing,

Co-Author: Mayra Romero

Contributions: Formal Analysis, Investigation, Review and Editing

Co-Author: Cecilia S. Blengini

Contributions: Investigation, Review and Editing

Co-Author: Shreya Sarathy

Contributions: Formal Analysis, Investigation, Review and Editing, Visualization

Co-Author: Julia Sorkin

Contributions: Formal Analysis, Investigation, Review and Editing, Visualization,

Co-Author: Hilina Bekele

Contributions: Investigation, Review and Editing, Visualization

Co-Author: Chen Jin

Contributions: Formal Analysis, Investigation, Review and Editing, Visualization

Co-Author: Seungsoo Kim

Contributions: Formal Analysis, Review and Editing

Co-Author: Alexei Chemiakine

Contributions: Investigation, Review and Editing

Co-Author: Rishad C. Khondker

Contributions: Investigation, Review and Editing,

Co-Author: José V.V. Isola

Contributions: Formal Analysis, Investigation, Review and Editing, Visualization

Co-Author: Michael B. Stout

Contributions: Methodology, Review and Editing, Supervision and Funding Acquisition

Co-Author: Vincenzo A. Gennarino

Contributions: Methodology, Review and Editing, Supervision and Funding Acquisition

Co-Author: Binyam Mogessie

Contributions: Methodology, Investigation, Review and Editing, Visualization, Supervision and Funding Acquisition

Co-Author: Devanshi Jain

Contributions: Methodology, Review and Editing, Visualization, Supervision and Funding Acquisition

Co-Author: Karen Schindler

Contributions: Methodology, Review and Editing, Visualization, Supervision and Funding Acquisition

Co-Author: Yousin Suh

Contributions: Methodology Review and Editing, Supervision and Funding Acquisition

Co-Author: Blake Wiedenheft

Contributions: Methodology, Review and Editing, Supervision and Funding Acquisition

Co-Author: Luke E. Berchowitz

Contributions: Conceptualization, Methodology, Investigation, Original Draft, Review and Editing, Visualization, Supervision and Funding Acquisition

Manuscript Information

Thomas W.P. Wood[1], William S. Henriques[2], Harrison B. Cullen[1], Mayra Romero[3], Cecilia S. Blengini[3], Shreya Sarathy[3], Julia Sorkin[3], Hilina Bekele[4], Chen Jin[5], Seungsoo Kim[5], Alexei Chemiakine[1], Rishad C. Khondker[1], José V.V. Isola[6], Michael B. Stout[6], Vincenzo A. Gennarino[1,7,8], Binyam Mogessie[4], Devanshi Jain[3], Karen Schindler[3], Yousin Suh[1,5], Blake Wiedenheft[2], and Luke E. Berchowitz[1,9*]

*Nature Aging*

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

Springer Nature Limited
Submitted 10 June 2024

## Abstract

The human genome contains 24 *gag*-like capsid genes derived from deactivated retrotransposons conserved among eutherians. Although some of their encoded proteins retain the ability to form capsids and even transfer cargo, their fitness benefit has remained elusive. Here we show that the *gag*-like genes *PNMA1* and *PNMA4* support reproductive capacity during aging. Analysis of donated human ovaries shows that expression of both genes declines normally with age, while several *PNMA1* and *PNMA4* variants identified in genome-wide association studies are causally associated with low testosterone, altered puberty onset, or obesity. Six-week-old mice lacking either *Pnma1* or *Pnma4* are indistinguishable from wild-type littermates, but by six months the mutant mice become prematurely subfertile, with precipitous drops in sex hormone levels, gonadal atrophy, and abdominal obesity; overall they produce markedly fewer offspring than controls. These findings expand our understanding of factors that maintain human reproductive health and lend insight into the domestication of retrotransposon-derived genes.


## Introduction

Almost half of the human genome consists of retrotransposons, ‘parasitic’ sequences that insert themselves into the host genome via an RNA intermediate[1]. While most of these sequences are silenced or mutationally deactivated, they can present opportunities for evolutionary innovation: mutation of a deteriorating retrotransposon can, in rare circumstances, result in a gene that provides a selective advantage to the host in a process termed domestication[2–5].In order for a domesticated gene to be beneficial, however, it needs to acquire a useful expression pattern. To maintain vertical inheritance and avoid extinction, a retrotransposon must express and

proliferate in germ cells or their precursors[6–8].This is no trivial feat, as host cells have numerous strategies to silence retrotransposon sequences including DNA methylation, chromatin modification, and transcriptional interference[7,9]. It is possible to evade these maneuvers, however, as the yeast *Metaviridae* (formerly *Ty3*) long-terminal repeat (LTR) retrotransposons do, by integrating downstream of an essential meiotic transcription factor binding site during germ cell development[10]. Could this evolutionary contest between host and retrotransposon shape the activities of the domesticated genes?

The *PNMA* family of *gag*-like genes, which were domesticated from an ancient vertebrate retrotransposon of the *Metaviridae* clade at least 100 million years ago[11,12] may provide an affirmative answer to that question. There are six full-length *PNMA* genes in humans and five in mice[13]. PNMA1-3 were originally identified as targets of the autoantibodies that cause paraneoplastic neurological syndrome (PNS) in lung and gynecologic cancers. *PNMA4-6* were identified based on homology with *PNMA1-3* but are not implicated in PNS[14–16]. We reasoned that the conservation of *PNMA* genes is unlikely to be based on their autoantigenic properties after ectopic expression in neoplasms. While *Pnma5* is implicated in mouse oocyte development via an unknown mechanism[17], little is known about the non-pathological roles of the other genes other than that some seem to have retained the ability to form capsids when expressed in human cells or bacteria[18,19]. The autoantigenicity of PNMA2, in fact, requires capsid formation[20]. *PNMA1* and *PNMA4* (also called *MOAP-1* or *MAP-1*) harbor capsid-like, linker, and N-terminal RNA-binding domains (RBDs)[12,13] though whether they form capsids *in situ* is unknown. They are, however, positively regulated by the master germ cell transcription factors MYBL1 and STRA8, and their transcripts are bound by the translational regulator DAZL

during gametogenesis[10]. This developmental regulation of *PNMA1* and *PNMA4* expression in male and female gonadal tissue suggested to us that they might serve a reproductive function. Further support for this notion comes from the fact that these two genes co-reside within a syntenic block that has persisted throughout the evolution of placental mammals (~20 mb apart on chromosome 14 in humans and ~20 mb apart on chromosome 12 in mice). Given that marsupials also harbor a single *PNMA* (*mPNMA*) that is highly expressed in pachytene spermatocytes[11,21], the germ cell expression of *PNMA* family genes was either present in the common therian ancestor or convergently evolved in marsupials and eutherians.

We therefore set out to determine whether either *PNMA1* or *PNMA4* affect reproductive function. Through the analysis of donated human ovaries, genome-wide association studies (GWAS), and genetic mouse models, we find that PNMA proteins are necessary for the maintenance of a normal reproductive lifespan. PNMA1 and PNMA4 self-assemble into capsids that exit human cells, and we observed large PNMA4 particles in male mouse gonadal tissue that are consistent with capsid formation.

## Results

### Expression of *PNMA1* and *PNMA4* Declines with Age in Human Ovaries

While previous studies of mouse male gonads indicated that both *PNMA1* and *PNMA4* are expressed at their highest levels in spermatocytes[22], we sought to determine which and to what degree human female gonadal cells express *PNMA*-family genes. We analyzed *PNMA1-5* expression by single-nuclei RNAseq in human ovaries donated from four young individuals in their 20s, four older individuals in their late 40s to early 50s, and in published human oocyte

expression data[23,24]. Among *PNMA1-5*, *PNMA1* and *PNMA4* showed the highest expression levels throughout the ovary (**Figures 5.1A-B, Supplmentary Figure 5.1**). Oocytes expressed the highest levels of *PNMA1* and *PNMA4*. The somatic ovarian cell type that expressed the highest levels of either gene was granulosa cells: in young adult ovaries, 8.5% of granulosa cells expressed *PNMA1* and 5.1% expressed *PNMA4*, possibly indicating specific granulosa subpopulations (**Figure 5.1B**). Comparing younger and older ovaries, *PNMA1* expression was

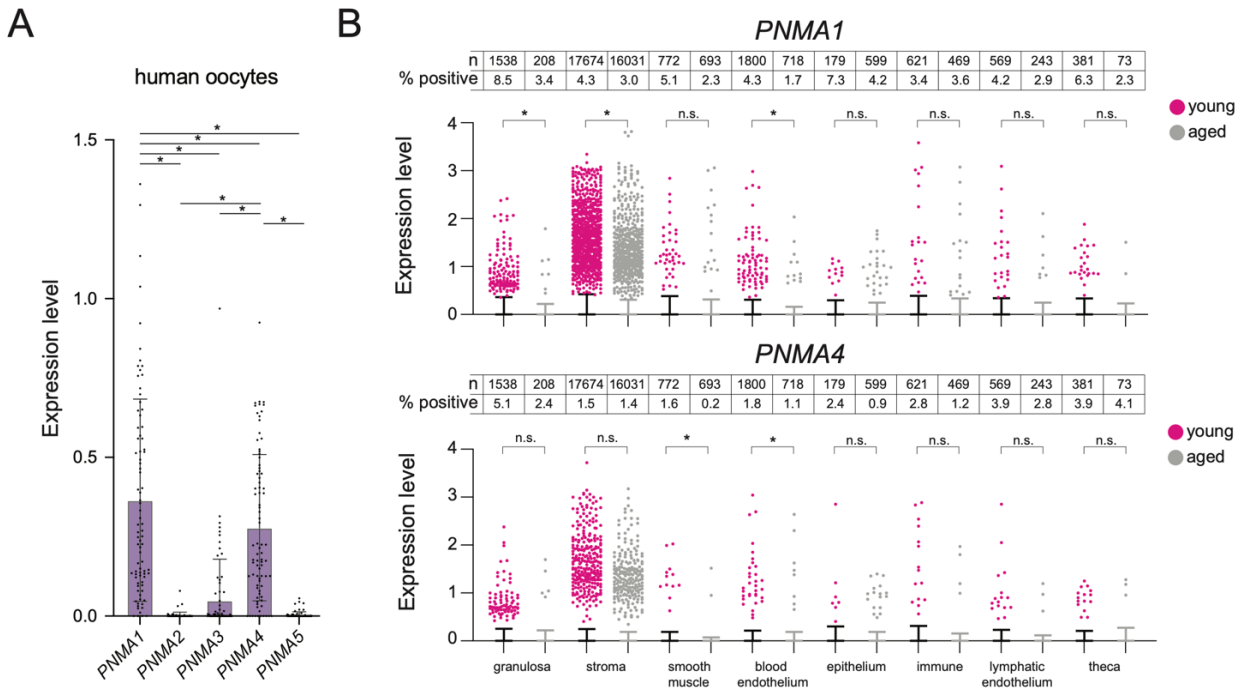



Figure 5.1. PNMA1 and PNMA4 are expressed in human ovaries. (A) Analysis of uniquely mapping single-cell RNAseq reads for PNMA1-5 loci expressed in MII human oocytes (data from Yuan et al., 2021)38. (B) Human ovaries from young (23-29 years of age, N = 4) and older (49-54 years of age, N = 4) donors were analyzed by single-nuclei RNAseq. Uniquely mapping reads for PNMA1 and PNMA4 loci were assigned to ovarian tissue types based on clustering analysis. Note that the majority of values = 0 and lie underneath the x axis (error bars represent SEM). Total n of each cell type and percentage of positive cells is indicated above the plots. Statistical significance (*p < 0.05) was determined by Mann-Whitney test with Welch's correction.

significantly higher in young granulosa, stroma, and blood endothelium cells, while *PNMA4* was significantly higher in young smooth muscle and blood endothelium cells signifying age-related decline in expression of these genes.

In mouse spermatocytes and human fetal ovaries, the 3' untranslated regions (UTRs) of *PNMA1* and *PNMA4* transcripts are bound by DAZL [25,26], which can activate or repress mRNA translation[27].To test whether the PNMA-DAZL interaction could be of regulatory significance, we expressed human *PNMA1*, *PNMA4*, or *SMC1b* (control DAZL target) with or without DAZL in human cells (HEK293T, **Supplementary Figure 5.2A**). DAZL augmented *PNMA1* protein levels and repressed *PNMA4* without altering mRNA levels (**Supplementary Figure 5.2B-C**) supporting the idea that DAZL can regulate the translation of PNMA transcripts.

*PNMA1* and *PNMA4* loci are present in most (if not all) eutherians and if they serve important functions, we expect them to exhibit some degree of conservation. *PNMA1* is universally conserved among eutherians and exhibits a purifying selection signature (**Figure 5.2A**, dN/dS = 0.06)[12]. The *PNMA4* locus is also universal to eutherians but has undergone more diversification: some lineages have acquired a premature stop within the coding sequence leading to pseudogenization and, in primates, an Alu element has integrated in the 3' UTR. Despite its pseudogenization in some mammals, *PNMA4* also shows evidence of purifying selection (**Figure 5.2B**, dN/dS = 0.29)[12]. Consequently, two primary lines of evidence led us to investigate the possibility that these genes could play important roles in reproductive aging: a) their expression in gonadal tissue appears to decline with age and is developmentally regulated at multiple levels and b) both show robust conservation signatures among eutherians.

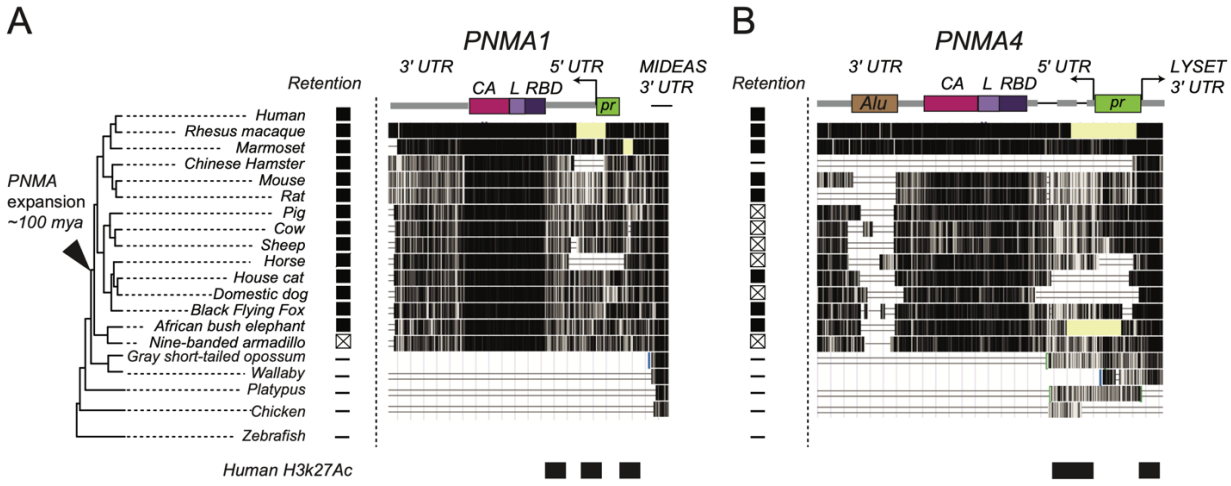



Figure 5.2. PNMA1 and PNMA4 are conserved among eutherians (A, B) Architecture of the human PNMA1 and PNMA4. Loci are colored according to domains: capsid domain (CA, magenta), linker (L, lilac), and RNA-binding domain (RBD, purple). The promoter (green) and transcription start sites are shown using arrows. A black triangle on the phylogenetic tree (left) indicates the point of the first expansion of the ancestral PNMA locus leading to PNMA1-5. PNMA1 is universally retained across placental mammals as an intact gene, whereas PNMA4 has experienced lineage-specific pseudogenization (boxed x's). Conservation at each amino acid is shown with a vertical black line for fifteen eutherian mammals, three marsupials, and two vertebrate outgroups. The histone modification for active transcription (H3K27Ac) in humans is shown along the bottom.

## Male and Female Mice Lacking *Pnma1* or *Pnma4* Show Premature Loss of Fertility

If *PNMA1* and *PNMA4* have been retained in mammals for reproductive functions, then we expect that animals lacking one or more of these genes should show some sort of diminution in reproductive capacity. We therefore used CRISPR-Cas9 with flanking cut sites to generate *Pnma1*$^{-/-}$ and *Pnma4*$^{-/-}$ mouse knockouts (KOs, **Supplementary Figure 5.3A**), and we verified that both null alleles exhibited Mendelian segregation. To validate the knockouts and to assess whether the two genes compensate for one another, we analyzed *Pnma1* and *Pnma4* mRNA and protein in wild type and knockouts. Rather than seeing loss of one gene leading to compensatory upregulation of the other, we observed post-transcriptional anticompensation in both knockouts:

*Pnma1*$^{-/-}$ testes showed an >80% reduction in PNMA4 protein levels (but unaltered mRNA), and *vice versa* for *Pnma4*$^{-/-}$ (**Supplementary Figure 5.3B-C**).

The male knockout mice did not differ from wild-type littermates at two months of age, but very soon their fertility declined, as evidenced by litter size (**Figure 5.3A**): at three months of age, the mutants fathered litters ~70% of the size of wild-type mice, and by six months of age, their litters were only 25-50% as large as those sired by wild-type mice (**Figure 5.3A, Supplementary Figure 3D**). There was a concomitant age-dependent testicular atrophy and decline in epididymal sperm count in single and double mutants compared to controls (**Figure 5.3B-D**). Serum testosterone levels in *Pnma1*$^{-/-}$ were statistically indistinguishable from controls at 1.5 and 3 months but fell precipitously by 6 months of age (**Figure 5.3E**). *Pnma4*$^{-/-}$ and double knockout mice showed less than 10% of the serum testosterone levels found in control animals at any time point. Even though peritubular myoid cells (which help transport sperm) and Sertoli cells (which support the development of immature sperm) appeared normal, from 3 months onward, some seminiferous tubules in single and double mutants were devoid of undifferentiated spermatogonia, spermatocytes, and spermatids (**Figures 5.3F-G, Supplementary Figure 5.4A**). Because of the convoluted organization of testicular tubules, depleted areas could represent loss of developing sperm in either distinct tubules or segments of individual tubules.

To gain insight into spermatogenesis problems that ultimately lead to devoid tubules, we analyzed whether histologically normal tubules contained fewer germ cells or more apoptotic cells at 12 months. Undifferentiated spermatogonia counts, indicated by PLZF staining, did not differ from controls (**Supplementary Figure 5A-B**), but histologically normal tubules in *Pnma* mutants contained more apoptotic cells, as indicated by the presence of fragmented dsDNA using

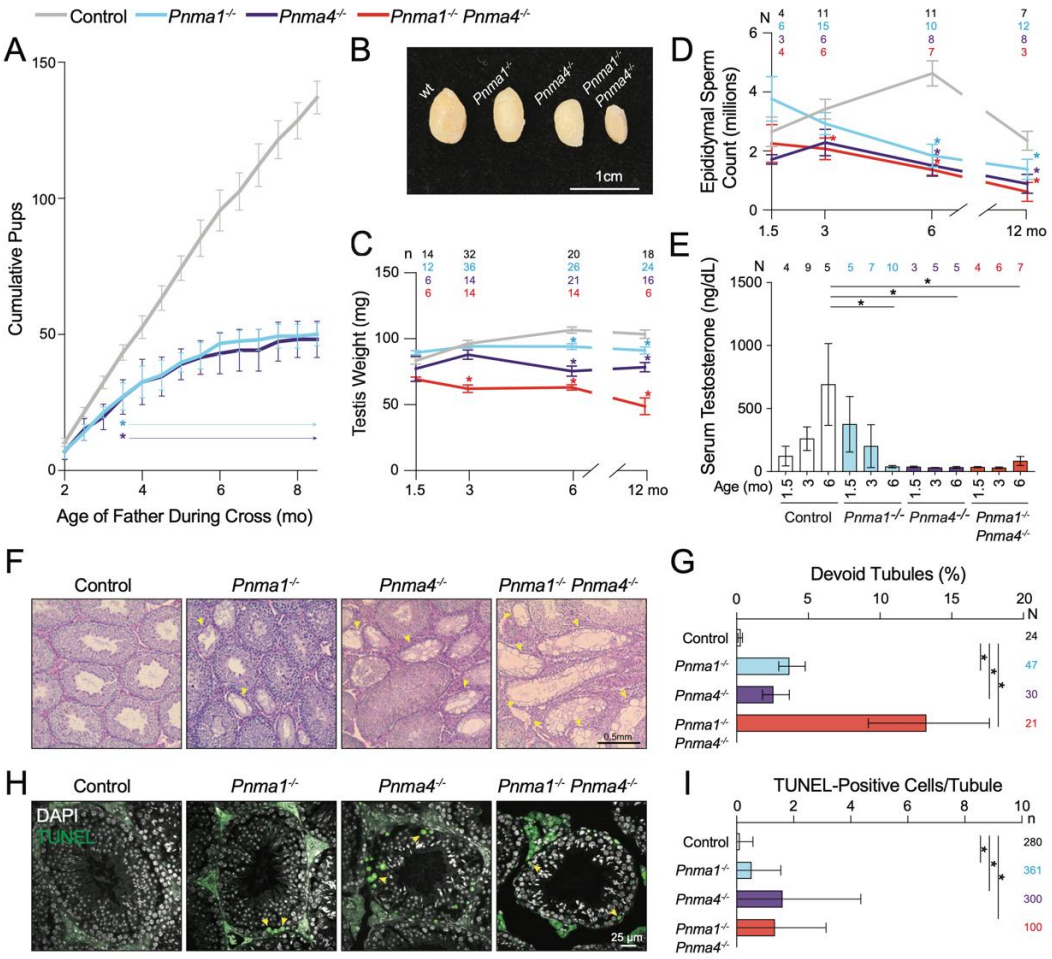



Figure 5.3. Male mice lacking Pnma1 or Pnma4 prematurely lose reproductive capacity (A) Wild type male control mice (gray), Pnma1-/- (cyan), or Pnma4-/- (purple) were crossed biweekly to CF-1 female fertility tester mice (N = 5 pairs for each genotype-timepoint combination). Pup numbers for each cross were recorded. (B, C) Testes (n above graph) from control (pooled wild type and heterozygous, gray), Pnma1-/- (cyan), Pnma4-/- (purple), or Pnma1-/- Pnma4-/- double mutant (red) mice were weighed. Representative images and quantifications are shown. (D) Sperm counts from dissected cauda epididymides. The number of individuals (N) for each genotype and age are shown above. (E) Serum testosterone (from N individual males) at the indicated ages. (F) PAS-stained tubule sections from 6-month-old mouse testes; devoid tubules are indicated by yellow arrowheads.**(G)** Devoid tubules (from n testes) as percentage of total.**(H)** Sections from 12-month-old testes were analyzed by TUNEL and DAPI staining. TUNEL-positive cells are indicated by yellow arrowheads.**(I)** Percent TUNEL-positive cells per tubule (n indicated by genotype to the right). Statistical significance (*$p < 0.05$) was determined by one-way ANOVA with correction for multiple comparisons and student's t-test. Error bars indicate SEM.

terminal deoxynucleotidyl transferase dUTP nick end labelling (TUNEL, **Figures 5.3H-I**). The depleted tubules and germ cell death support the idea that *Pnma1* and *Pnma4* are important for male germ cell development and likely execute their functions prior to spermatid formation.

Like the male mice, female mice lacking *Pnma1* or *Pnma4* were similar to wild type in their reproductive capacity at two months of age, but they birthed smaller litters over time (**Figure 5.4A, Supplementary Figure 5.3E**). This premature loss of fertility correlated with ovarian atrophy, diminished antral follicle counts, and a greater tendency to develop follicular cysts than wild type (**Figures 5.4B-F, Supplementary Figure 5.4B**). Some mutant follicles had lost cumulus granulosa cells, which synthesize steroid hormones and promote oocyte growth, and resembled follicles in mice lacking estrogen receptors (ER) or follicle stimulating hormone (FSH) receptors[28–30]. Other ovarian features did not differ significantly different between mutants and controls with some exceptions (**Supplementary Figure 5.6A-E**). Serum progesterone, FSH, and anti-Müllerian hormone (AMH) levels were statistically indistinguishable from controls at 1.5, 3, and 6 months with two exceptions: progesterone was down in *Pnma4*$^{-/-}$ at 6 months and AMH was elevated in double mutants at 1.5 months. Oocytes isolated from single and double mutants completed meiosis II at 16 hours at similar frequencies compared to controls, indicating that the reduced fertility in mutant animals is unlikely to be caused by severe meiotic defects (**Figure 5.4G**).

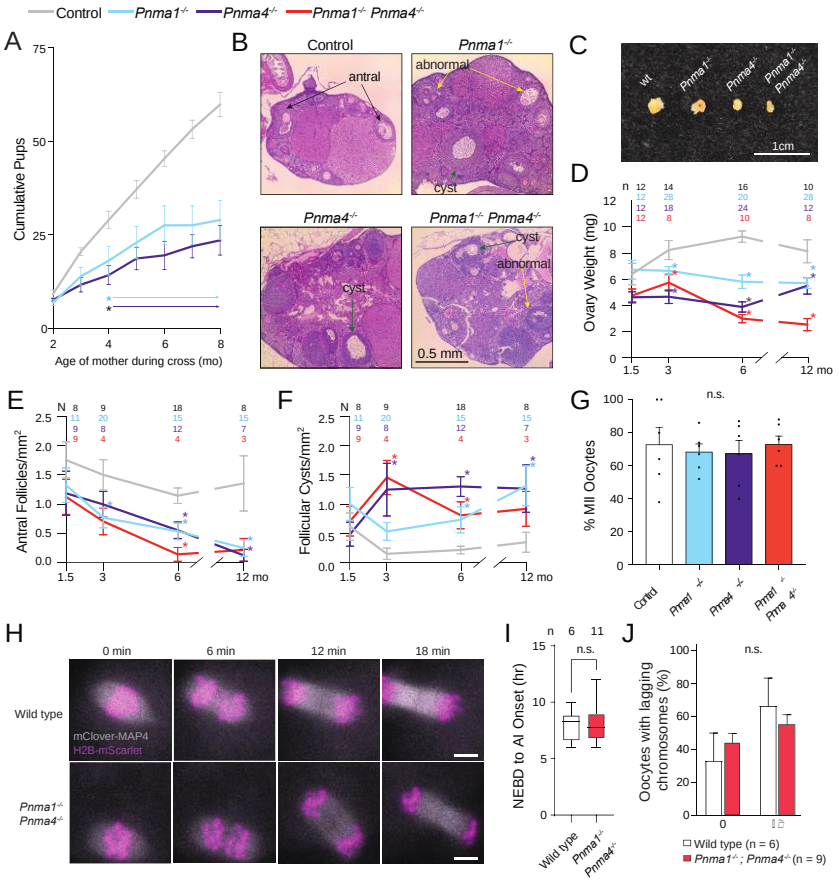



Figure 5.4. Female mice lacking Pnma1 or Pnma4 have age-dependent reproductive defects. (A) Wild type female wild type mice (gray), *Pnma1*$^{-/-}$ (cyan), or *Pnma4*$^{-/-}$ (purple) were crossed to B6D21/J male fertility tester mice (N = 5 pairs for each genotype-timepoint combination). Pup numbers for each cross are plotted over time. (B) PAS-stained ovarian sections from 3-month-old mouse ovaries. Black arrows denote antral follicles, yellow arrows denote abnormal follicles, and green arrows denote follicular cysts. (C) Representative images of ovaries from the four genotypes. (D) Ovaries (n above graph) from control (pooled wild type and heterozygous, gray), *Pnma1*$^{-/-}$ (cyan), *Pnma4*$^{-/-}$ (purple), or *Pnma1*$^{-/-}$ *Pnma4*$^{-/-}$ double mutant (red) mice were weighed. (E) The number of antral follicles at the indicated ages per mm$^2$. The number of individuals (N) for each genotype and age are shown. **(F)** The number of follicular cysts per mm$^2$ by age and genotype (N's in graph).**(G)** Germinal vesicle (GV) oocytes were collected from six-month-old females and meiotically induced. Meiotic progression was analyzed by tubulin immunofluorescence. Percentage of meiosis II oocytes was recorded. **(H)** GV oocytes were collected from seven-month-old wild-type and double mutant mouse ovaries. We injected mRNA encoding *mClover-MAP4* (microtubule-binding protein) and *H2B-mScarlet* (histone) and imaged oocytes live for ~18 hours, recording **(I)** the time from nuclear envelope breakdown (NEBD) to anaphase I and **(J)** lagging chromosome percentages. Statistical significance (*p < 0.05) was determined by determined by one-way ANOVA with correction for multiple comparisons and student's t-test. Error bars indicate SEM.

To determine whether mutant oocytes that do complete both meiotic divisions have subtler meiotic defects, we used high-resolution live imaging to analyze chromosome segregation during oocyte maturation. Oocytes from double mutants at two and seven months did not differ from controls in meiotic chromosome segregation, alignment on the metaphase I and metaphase II spindles, or progression timing (**Figures 5.4H-J, Supplementary Figures 5.7A-G**), suggesting that mutant oocytes that do progress are meiotically competent. To assess the mouse ovarian cell types where *Pnma* genes may act, we analyzed the expression of *Pnma1-5* in single cells at young (4.5 months), peri-estropause (10.5 months), and post-estropause (15.5 months) stages (**Supplementary Figure 5.8**). *Pnma1* and *Pnma4* showed the highest expression levels throughout the ovary compared to other *PNMA* genes (except *Pnma5* specifically in peri-estropause oocytes, **Supplementary Figure 5.8**, orange dots). *Pnma1* and *Pnma4* are primarily co-expressed (by % positive cells) in granulosa, stroma, and theca cells. Orthogonal analysis of *Pnma1* and *Pnma4* expression from a published data set[31] corroborated that these genes are expressed in subpopulations of granulosa and stroma cells (**Supplementary Figure 5.9A-B**). In granulosa, stroma, and theca, we observed a ~10-fold decline in *Pnma1*-positive cells at post-estropause (**Supplementary Figure 5.8**, gray dots). *Pnma4* expression did not change significantly in the same time frame. Together, these data support the idea that these genes are expressed in subpopulations of ovarian cells that support oocyte maturation and viability.

## *Pnma1* and *Pnma4* Mutants Acquire Abdominal Obesity but Appear Behaviorally Normal

In humans and mice, reproductive decline that occurs naturally with age commonly correlates with changes in body composition[32,33]. As male and female *Pnma* mutants aged, we noticed that they rapidly gained abdominal subcutaneous and visceral fat, corresponding to the

time course of their dwindling fertility. At 1.5 months, the knockouts were statistically indistinguishable from controls, but by 3 months, *Pnma1*$^{-/-}$ and double mutants were 20% heavier than controls (**Figure 5.5-B**). By 6 months, all mutants were significantly heavier than controls. Mutant mice did not consume more food compared to controls (**Figure 5.5C**), so the weight gain likely reflects altered metabolism. A similar phenotype is observed post-ovariectomy[34], in mice lacking sex hormone receptors[28–30], and in castrated males[35]. In human and rodent males, increased adiposity arises due to testosterone deficiency[33].

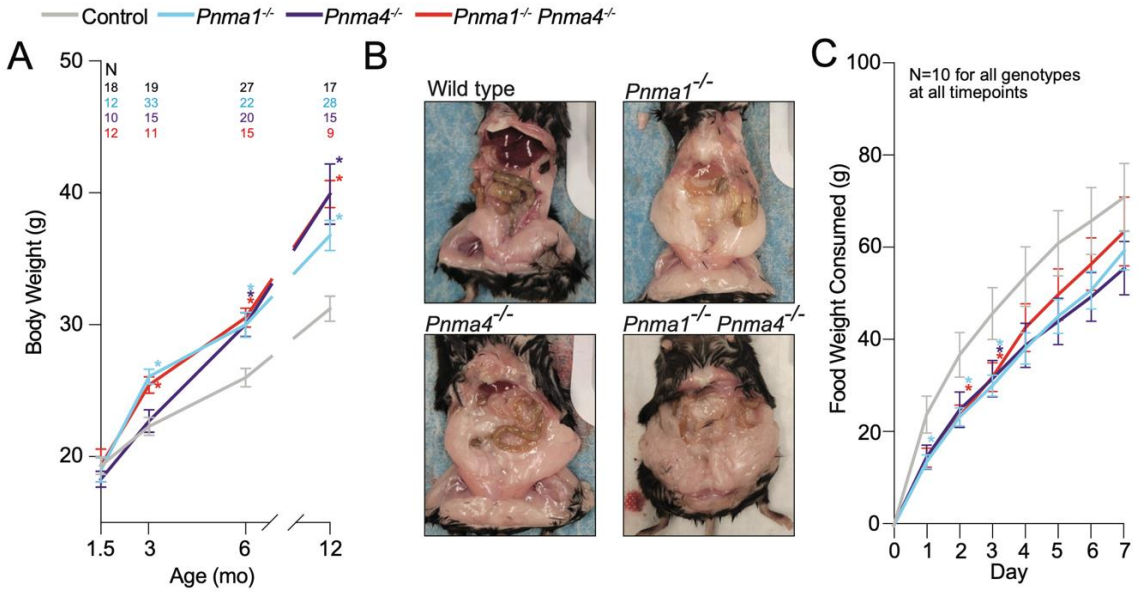



Figure 5.5. Pnma1 and Pnma4 mutants gain abdominal fat with age. (A) Body weights (N above graph) of control (pooled wild type and heterozygote, gray), *Pnma1*$^{-/-}$ (cyan), *Pnma4*$^{-/-}$ (purple), or *Pnma1*$^{-/-}$ *Pnma4*$^{-/-}$ double mutant (red) at the indicated ages. (B) Representative images of abdominal fat at six months in each genotype. (C) Food intake of 3-month wild type, *Pnma1*$^{-/-}$, *Pnma4*$^{-/-}$, or *Pnma1*$^{-/-}$ *Pnma4*$^{-/-}$ double mutant mice (N = 10 mice) was monitored daily for 7 days. Statistical significance (*$p < 0.05$) was determined by one-way ANOVA with correction for multiple comparisons. Error bars indicate SEM.

Like other *PNMA* genes, both *PNMA1* and *PNMA4* are expressed in the brain[36]. We therefore tested basic neurological functions in six-month-old male *Pnma1*$^{-/-}$, *Pnma4*$^{-/-}$, and

double mutants compared to wild type. The Y-maze spontaneous alternation assay revealed no deficits in short-term or spatial memory (**Supplementary Figure 5.10A**). Mutant mice were largely indistinct from controls regarding long-term memory, as assessed by the amount of post-cue freezing behavior after fear conditioning (**Supplementary Figure 5.10B**). In the open field test, mutants did not differ significantly from controls in time spent in the center, distance traveled, and vertical counts (**Supplementary Figure 5.10C-E**). We observed no difference in muscular strength or coordination between single and double mutants in the inverted hang or grip strength tests (**Supplementary Figure 10F-G**). Although these results revealed no obvious behavioral abnormalities in *Pnma1*$^{-/-}$ and *Pnma4*$^{-/-}$ male mice at 6 months of age, it is possible that female mutants or mice at other ages could display differences in cognitive function.

## PNMA1 and PNMA4 Form Capsid-like Structures in Gonadal Tissue

Gag family proteins self-assemble into capsid-like structures that are critical for packaging and delivering molecular cargo. One prominent example is ARC, which has been independently domesticated in mammals and *Drosophila* to encapsulate and traffic mRNA across neuronal synapses[37,38]. To determine whether PNMA1 and/or PNMA4 proteins have retained the ability to self-assemble into capsids, we expressed and purified His-tagged recombinant (mouse) PNMA1 and PNMA4 from *E. coli* which we analyzed by size exclusion chromatography (**Supplementary Figure 11A-B).** *Metaviridae* capsids are amenable to isolation by size-based fractionation because they are large (yeast Ty3 capsids contain 540 gag protein subunits and are ~ 20 megadaltons[39]). We visualized the high molecular weight fractions by negative stain transmission electron microscopy (TEM). Non-uniform capsid-like assemblies are evident in micrographs of both PNMA1 and PNMA4 (16-21 nm diameter for PNMA1 and 36-51 nm for

PNMA4, **Figures 5.6A-B**). Based on their size, these results suggest that PNMA1 forms capsids similar to those formed by PNMA2[19,20], while PNMA4 capsids are similar to those formed by LTR retroelements[39].

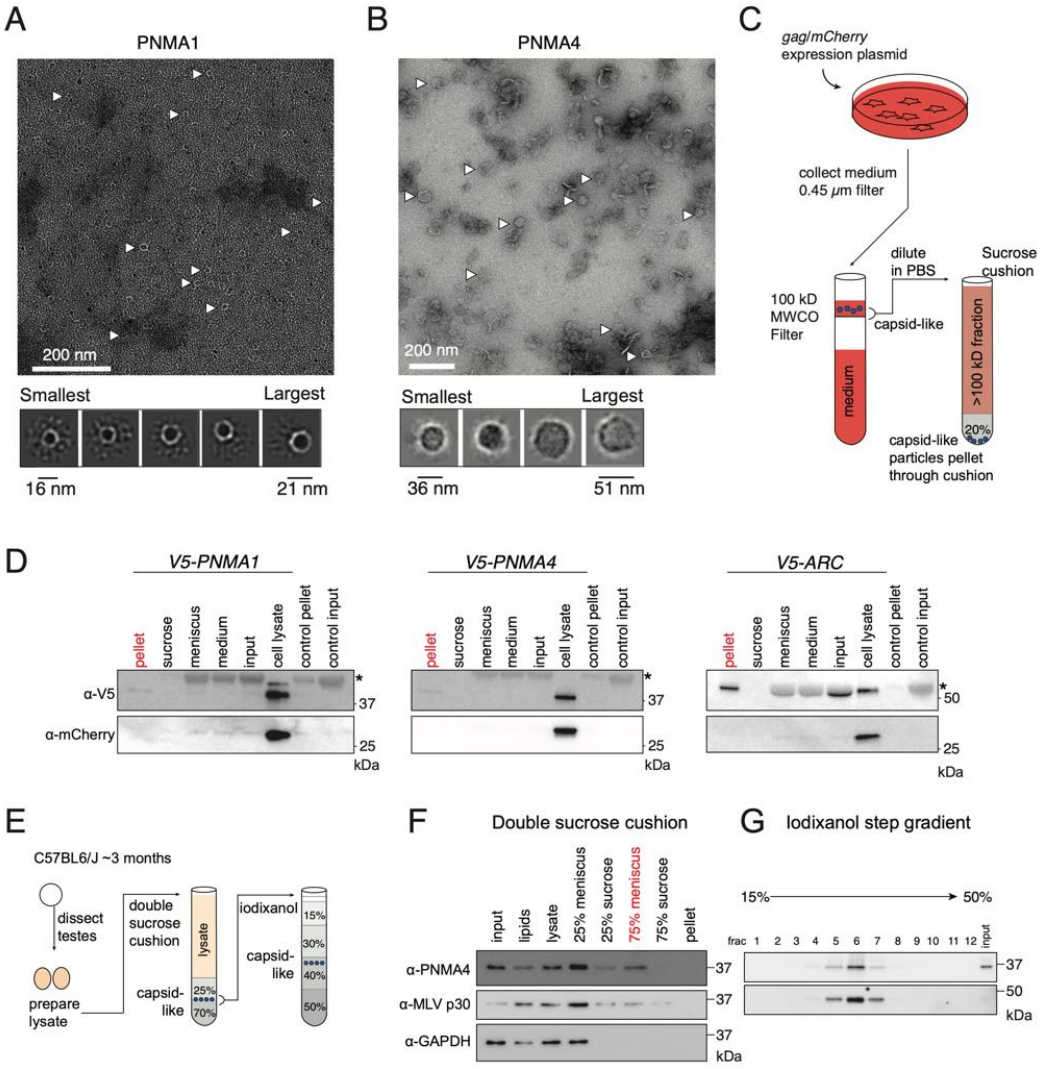



Figure 5.6. PNMA proteins form capsid-like structures that can exit human cells (A, B) Top panels depict TEM micrographs of negatively stained recombinant PNMA1 (A) and PNMA4 (B) protein. Bottom panels depict 2D class averages from cryoSPARC. (C) Experimental setup: 3V5-tagged PNMA1, PNMA4, and ARC (control exported capsid) expression plasmids were transfected into HEK-293T cells. mCherry (control non-capsid) was co-expressed from the transfected plasmid. Culture medium was collected, spun, and filtered to remove cells and debris. Filtrate was fractionated over a 20% sucrose cushion to enrich for capsids, which fractionate as

large particles at the bottom of the cushion (pellet). Sucrose, meniscus, spun media, and pellet fractions were collected. Unfractionated media input and cell lysate samples were also collected.(D) PNMA, PNMA4, ARC (anti-V5), and mCherry protein levels by immunoblot. Untransfected controls (control input) are at the far right of each blot. Asterisks (*) mark the cross-reacting BSA band (abundant in growth medium). Note the presence of BSA in the untransfected controls and absence of BSA in pellet/lysate.(E) Experimental setup: Lysate was prepared from ten testes from wild type C57BL6/J mice were collected at 3 months. Lysate was fractionated by velocity step gradient (38,000 rpm, 3 hours) over a double sucrose cushion (25% and 70%). Capsid-like structures migrate to the interface between 25%-70% sucrose (70% meniscus). The 70% meniscus was further fractionated by isopycnic (equilibrium) centrifugation on an iodixanol step gradient; the photograph shows the centrifuge tube after this step. (F-G) PNMA4, MLV p30 gag, and GAPDH protein levels were determined by immunoblot in each fraction at the end of the double sucrose fractionation and (G) iodixanol fractionation.

To assess capsid formation in mammalian cells, we expressed human *PNMA1*, *PNMA4* and *ARC* (positive control) in HEK293T cells with *mCherry* (non-capsid control, driven by a separate promoter). After 24 hours, we harvested cells and fractionated cell lysate by velocity gradient ultracentrifugation. Because of their size and density, capsids migrate through the gradient, away from the major protein content of the cell. Some PNMA1 and PNMA4 appeared in denser factions (**Supplementary Figure 12**). To determine if these large assemblies have the capacity to exit cells, we expressed *PNMA1*, *PNMA4* and *ARC* for 24 hours in human cells, then collected the growth medium and removed the cells and any debris by filtration. We concentrated the medium using a 100 kDa molecular weight cutoff (MWCO) filter, discarded the < 100 kDa flow-through, fractionated the concentrate by sedimentation over a 20% sucrose cushion, and analyzed the fractions by immunoblot (**Figures 5.6C-D**). Like ARC, some PNMA1 and PNMA4 appeared in the pellet fraction (**Figure 5.6D**), supporting the idea that at least some PNMA proteins exit human cells in a capsid-like form.

To assess whether PNMA proteins form capsid structures *in situ*, we purified PNMA4 capsids from wild-type mouse testes using a procedure that was developed for the purification of

intact VLPs from plant tissue[40]. We focused on PNMA4 because we found its detectability in testes by (polyclonal) immunoblot more reliable than PNMA1. To separate PNMA4 capsids from monomers and oligomers, we fractionated testis lysate over two sucrose cushions of 25% and 70% by ultracentrifugation (**Figure 5.6E**). Because of their size and density, capsid particles migrate to the interface between the 25% and 70% sucrose layers and separate from monomeric proteins, which accumulate at the interface between the lysate and 25% cushion[40]. We observed enrichment of PNMA4 at the 25%-70% interface compared to a control monomeric protein (GAPDH) and co-enrichment with p30 gag from murine leukemia virus (MLV), an enveloped capsid positive control (**Figure 5.6F**). To separate capsids from other large structures such as ribosomes, we collected and re-fractionated the 25%-70% meniscus fraction by isopycnic (equilibrium) ultracentrifugation on an iodixanol (Optiprep) density step gradient. PNMA4 protein isolated from testis tissue co-fractionated (as large particles) with MLV p30 capsid protein (**Figure 5.6G**). The *in situ* biochemical properties of PNMA4 are consistent with capsid formation, which may mediate the protein's pro-fertility function.

## Human Variation at the *PNMA1* and *PNMA4* Loci is Associated with Reproductive Defects

Based on the reduction of fertility in mice lacking *Pnma1*, *Pnma4*, or both, we asked whether subfertility or other reproductive phenotypes in humans are causally linked to genetic variation in the *PNMA1* and/or *PNMA4* loci (**Supplementary Table 1**). We used the variant-to-gene (V2G) and locus-to-gene (L2G) pipelines available from Open Targets Genetics, with a standard p-value cutoff of $5x10^{-8}$, to identify the causal gene from a trait-associated variant or locus identified by GWAS[41,42]. Briefly, from a trait-associated variant (*i.e.*, the polymorphism corresponding to a GWAS peak), these pipelines establish the likely causal gene by weighted

analysis of independent published quantitative trait loci experiments (eQTLs, pQTLs and sQTLs), chromatin interaction experiments, *in silico* functional predictions, and distances between a variant and transcription start sites. We found six variants attributed to *PNMA1* (from nine studies and three publications[43–45]). associated with altered testosterone levels and levels of sex hormone binding globulin (SHBG, levels correlate to testosterone). Two variants attributed to *PNMA4* are associated with altered age of menarche[46] and male puberty timing[47]. We also identified two additional variants associated with elevated body fat (thirteen studies, three publications[48–50]) that are attributed to *PNMA4*. These results support the notion that the alterations in reproductive capacity we observed in mouse *Pnma1* and *Pnma4* knockouts are relevant to humans.

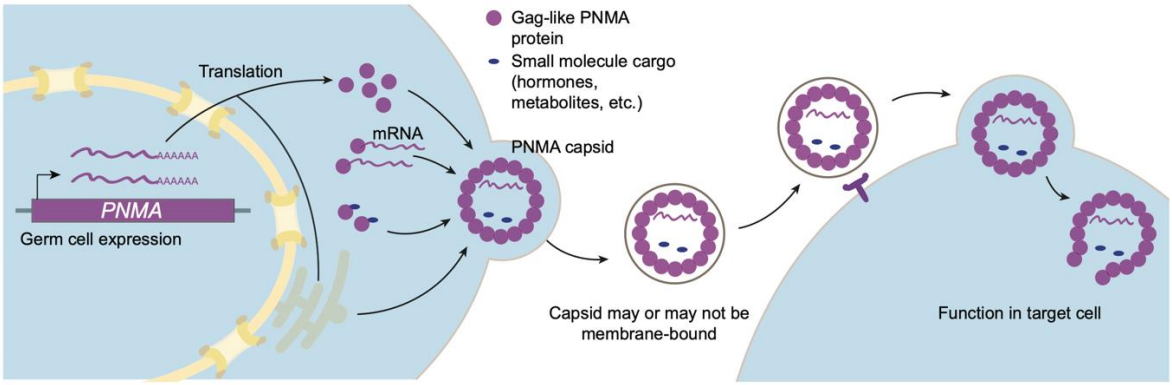



Figure 5.7. Model for PNMA1 and PNMA4 function.

## Discussion

In general, retroelement gag capsids facilitate endogenous and parasitic intercellular communication. Like many other tissues, gonads rely on endocrine, paracrine, and autocrine signaling to support gamete development, and their cells employ a variety of means to communicate with each other. For example, testosterone, produced largely in Leydig cells,

diffuses into the testicular tubules where it binds to androgen receptors on Sertoli cells which then drive the production of factors that support spermatogenesis[51,52]. Progesterone, produced by follicular granulosa cells and corpora lutea, is critical for successful oocyte maturation and ovulation[53,54]. Communication among germ cells also takes place through intercellular bridges that connect the cytoplasm of adjacent cells into a syncytium; males that cannot form these bridges (such as *Tex14* knockouts) are sterile[55] (these structures are important, but not essential, for female fertility[56,57]). These germ cell bridges/syncytia have been proposed to foster dosage compensation in haploid cells, coordination of meiotic entry, and sharing of signals for synchronous cell divisions within seminiferous tubules[55,56,58]. In the ovary, oocytes are surrounded by cumulus/coronal granulosa cells that extend long processes termed transzonal projections through the zona pellucida[59–61]. These projections release vesicles that are thought to contain RNA and have been proposed to enable sharing of RNA between transcriptionally active granulosa cells and the silent oocyte[62,63]. Conversely, oocytes release growth factors that regulate proliferation of granulosa cells. Emerging evidence supports the notion that communication between follicular cells and oocytes is important for maintaining germ cell fitness: aged oocytes transplanted into young follicles become rejuvenated, improving in mitochondrial function and the fidelity of meiotic chromosome segregation, among other measures[64]. Although PNMA1 and PNMA4 are not *essential* for intercellular communication (the double mutant would be sterile), we propose that PNMA1 and PNMA4 capsids could promote the transmission of cytoplasmic signals, whether through bridges, transzonal projections, and/or by transport across plasma membranes (**Figure 5.7**).

The word capsid derives from the Latin word for “case,” and capsids form an elegant means to carry molecular messages. The question we still need to answer is, do the capsids formed by PNMA1 and PNMA4 carry cargo, and if so, what is it? This cargo could conceivably be nucleic acids such as mRNA or tRNA, hormones such as testosterone or FSH, or a metabolite—but whatever the cargo, it likely has important functions for fertility.

The literature provides a few clues as to the likely purpose of the possible cargo. *PNMA4* has been implicated in apoptosis in the context of mis-expression in cancer cells[65–67]. PNMA4 interacts with the pro-apoptotic protein Bax and the pro-survival proteins Bcl-2 and Bcl-X(L) and can mediate caspase-dependent apoptosis when overexpressed[68]. Given that 12-month-old *Pnma4*$^{-/-}$ testes exhibit greater numbers of TUNEL-positive cells, *PNMA4* likely promotes survival in gonadal tissue. In mutant males, we observed an age-dependent intra-organ heterogeneity where some testicular tubule segments were devoid of sperm and sperm precursor cells while others appeared wild type. Other than increased cell death in aged tubules, we did not observe any clear defects in the ‘normal’-appearing seminiferous tubules of mutant animals. Because mutant tubule segments are either devoid or full, without an obvious intermediate state, we speculate that devoid tubule segments have experienced a failure to initiate a wave of spermatogenesis and, once a failure has occurred, that tubule cannot recover.

Female mice lacking *Pnma1* and/or *Pnma4* exhibit a wide range of ovarian abnormalities. *Bmp6*$^{-/-}$, *Bmp15*$^{-/-}$, *Smad2/3*$^{cko}$, and *Smad3*$^{cko}$ female mice exhibit age-dependent subfertility with ovarian histology reminiscent of *Pnma* knockouts[69–71], suggesting that the *Pnma1* and *Pnma4* mutant phenotypes could, in part, be due to downregulated TGF-beta signaling. In mutant females, the oocytes that are produced appear healthy and undergo meiosis. Similarly, while

sperm count declines with age in mutant males, the sperm within the epididymis appear normal. These data are consistent with a model in which the function(s) of PNMA1 and PNMA4 are enacted prior to gamete release from the ovary/testis though we cannot rule out defects in later processes such as implantation. Elucidating their functions will likely require identifying the cargo that is being transported.

Developing yeast gametes thwart proliferation of *Ty3*, a parasitic ancestor of *PNMA* genes, by using amyloid-like condensates of the RNA-binding protein Rim4, which binds retrotransposon mRNA and inhibits its translation[10]. Mammalian analogs of Rim4 include DAZ, DAZL (DAZ-like), and PUM1/2 (Pumilio). Although higher-order assemblies of RNA-binding proteins behave somewhat differently in yeast and mammalian germ cells, DAZ, DAZL, and PUM1/2 are like Rim4 in that they regulate translation in developing gametes, affect gene expression post-transcriptionally by binding the UTRs of their target genes, and form condensates *in vivo*[72–75]. Our analysis of DAZL-mediated regulation of *PNMA* genes supports the idea that, from yeast to humans, RNA-binding proteins with the propensity to form condensates can regulate the translation of retroelement-derived mRNA. We propose that as *Metaviridae* elements gradually evolved beneficial functions and lost proliferative capacity in host meiotic cells, the analogs of condensates acting as a defense mechanism (Rim4) were repurposed in mammals as regulators (DAZL). Could DAZL have maintained a role to protect sperm and eggs from the assault of expressed *Metaviridae* retrotransposons? Several human endogenous retroviruses (hERVs) are robustly expressed in gonads[76,77] and these elements do not appear to be rapidly proliferating, suggesting that the hERV lifecycle is somehow blocked after mRNA transcription. The possibility for condensate-forming RNA-binding proteins to post-

transcriptionally regulate hERVs has precedent: in neurons, the neurodegeneration-associated protein TDP-43 binds to hERV-K mRNA and regulates its protein synthesis[78].

It has been noted that PNMA capsids should be suitable as delivery vehicles for therapeutic agents such as RNA vaccine delivery and gene therapy[18,19,37,79]; such therapies involve bioengineering capsid proteins for the delivery of designer cargo, such as mRNA. The beauty of PNMA1 and PNMA4 is that they would require no engineering to function as therapeutic agents for subfertility because they could naturally harbor pro-fertility cargo. Therapeutic development would focus on PNMA4 because, unlike PNMA1, it is not associated with autoimmunity[80]. Before a therapy for subfertility can be developed, however, we must optimize methods to purify PNMA4 capsids from gonadal tissue, understand precisely what cargos they carry, and assess their therapeutic efficacy in mouse models of subfertility.

## Methods

### Generation of Mutant Mouse Models

*Pnma1*$^{-/-}$ and *Pnma4*$^{-/-}$ mice were generated at the Columbia University Irving Medical Center (CUIMC) Genetically Modified Mouse Model Shared Resource. Two single guide RNAs (sgRNA) were assembled with the Alt-R™ S.p. Cas9 Nuclease V3 (IDT) in an injection buffer containing 5 mM Tris-HCl [pH 8] and 0.2 mM EDTA. This assembly process aimed to form the CRISPR RNP (ribonucleoproteins) with a final concentration of 0.1 μM. The assembly was conducted at room temperature for 10 minutes. Subsequently, the assembled RNP was microinjected into the pronuclei of fertilized B6CBAF1 eggs. Eggs that survived the injection were then transferred into the oviducts of E0.5 pseudo-pregnant surrogates. Pups resulting from this injection were genotyped at the age of 15 days using PCR.

Along with these knockout lines, the following mouse strains were used in this work: C57BL/6J (wild type Jackson Labs: #000664), B6D2F1/J (Jackson Labs: #100006), and CF-1 (Inotiv/Envigo: 033-US).

Mouse Husbandry

Animal studies were conducted in accordance with the Institutional Animal Care and Use Committee (IACUC) policies of CUIMC and the respective NIH policies. All studies were also conducted in accordance with the provisions of the Animal Welfare Act (NIH/DHHS) and the guidelines of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC). All mice were bred at CUIMC, the Jackson Laboratory, or Inotiv/Envigo Animal facilities. Mice were housed at CUIMC. All mouse procedures were conducted at CUIMC except for oocyte dissection, which was conducted at Rutgers University and Yale University. Mice were checked daily, housed in ventilated temperature-controlled cages that were regularly cleaned, containing no more than 5 individuals, and with *ad libitum* access to food and water. Mice were given ample time to equilibrate to housing facilities/experimental rooms as necessary. The animal protocols for Columbia University were: AABO0557 (L.E.B.), AABM7575 (V.A.G.). For Rutgers University: PROTO201702497 (K.S.). For Yale University: 2021-20408 (B.M.). Along with knockout mouse lines created at CUIMC mentioned above, the following mouse strains were used in this work: C57BL/6J (wild type Jackson Labs: #000664), B6D2F1/J (Jackson Labs: #100006), and CF-1 (Inotiv/Envigo: 033-US). The ages of mice are detailed in the respective sections of the manuscript.

Mouse Genotyping

Mouse tail clips were added to 200 µl of 50 mM NaOH in a 1.5 ml tube and placed in a thermomixer at 350 rpm at 55°C overnight. The next day, 25 µl of 1 M Tris-HCL [pH 7.5] was added. Samples were then vortexed and spun at 16,100*g* to pellet debris. Supernatant was transferred to a new tube and frozen at -20°C until genotyping. For PCR we used the following primers:

| | |
|---|---|
| *Pnma1* forward primer | 5'-GAAATGCGCCGTCATAGAATAG-3' |
| *Pnma1* reverse primer | 5'-GGATGGGGGAGAAACTCG-3' |
| *Pnma1*$^{-/-}$ forward primer | 5'-GAAATGCGCCGTCATAGAATAG-3' |
| *Pnma1*$^{-/-}$ reverse primer | 5'-TTCTGTGCCGAAACACTGTC-3' |
| *Pnma4* forward primer | 5'-TCCTCGGAAAGCACTATTGG-3' |
| *Pnma4* reverse primer | 5'-GCTAACATGGGTGCTCGAAT-3' |
| *Pnma4*$^{-/-}$ forward primer | 5'- GTCTCACCTTCCTGTCCTGAGTG-3' |
| *Pnma4*$^{-/-}$ reverse primer | 5'-TAGGCTGTTGAGCTCACTTTGT-3' |

Standard thermocycler amplification conditions were used with Platinum SuperFi II DNA Polymerase and corresponding buffer (Thermo Fisher, REF-12361050), an annealing temperature of 60°C, and 35 cycles. In early rounds of genotyping, following amplification, bands were cut from the gel and extracted using column purification (Macherey-Nagel, REF-740609.250). Samples were submitted for sequencing at Azenta/Genewiz to confirm gene deletions or fragment insertions.

Fertility Assay

To assess mouse fertility, *Pnma1*$^{-/-}$, *Pnma4*$^{-/-}$, or C57BL/6J wild type control mice were crossed to fertility tester mouse strains for 7 (females) or 14 (males) breeding cycles beginning at ~2 months of age. One male was paired with one female in a cage at a time. Female *Pnma1*$^{-/-}$, *Pnma4*$^{-/-}$, or C57BL/6J wild type mice were crossed to B6D2F1/J male tester mice. Mice were paired for a week and then separated. Males were then reintroduced three weeks after being separated (four weeks after the beginning of the prior cross). Pups were counted, removed, and sacrificed before reintroducing mice. Male *Pnma1*$^{-/-}$, *Pnma4*$^{-/-}$, or C57BL/6J wild type mice were crossed to CF-1 female tester mice. Mice were paired for a week and then males separated from female group 'A', with males being paired with another female 'B' after a week of separation from female 'A'. Males were paired with female 'B' for a week, separated for a week and then returning to female 'A' and repeated. This was done to provide more data for male fertility studies. Pup counts from female and male fertility trials were recorded and plotted with age of parents at the time of the cross. Statistical significance ($*p < 0.05$) was determined by student's t-test.

Body and Gonad Weight

Mouse body weight was measured prior to dissection using an analytical balance (Mettler Toledo, MS104TS/00). Ovaries and testis were weighed with as much fat and connective tissue removed as possible without damaging the morphological integrity of the organs.

Immunofluorescence of Testis Sections

Testes were dissected and placed into 3 ml of 4% PFA (16% stock, Electron Microscopy Sciences, REF-15710) in a 15 ml falcon tube. Testes were incubated in fixative at 4°C with

rocking overnight. The following morning, samples were washed 3x (20 minutes) in 1x PBS (10x stock, Fisher Bioreagents, REF BP3994) at 4°C with rocking. Samples were washed ~10 x in 70% EtOH at 4°C with rocking. Samples were submitted in 70% EtOH to the Columbia University Medical Center Molecular Pathology Core for paraffin embedding and slides containing 5 μm sections were prepared by the Rutgers Research Pathology Services Core Facility. Slides were deparaffinized by 3 x 5-minute washes in xylene, followed by rehydration by 3-minute washes in 100% EtOH, 95% EtOH, 70% EtOH and DI water.

TUNEL assays were performed using the Click-iT Plus TUNEL Assay kit (Invitrogen, C10617) according to manufacturers' instructions with the following modifications: after the TdT reaction, detection was done using $1/10^{th}$ of the recommended Click-iT Plus TUNEL reaction cocktail. For immunofluorescence staining of testis sections, antigen retrieval was performed by treating slides for 20 minutes in Sodium Citrate buffer (10 mM Sodium Citrate, 0.05% Tween 20 in Milli-Q water, pH 6.0) at 95-100°C. Slides were blocked for 20 minutes in blocking buffer (0.2% BSA, 0.2% gelatin, 0.05% Tween 20 in PBS), incubated with goat anti-PLZF primary antibody at 1:500 dilution (Human PLZF Antibody, R&D Systems AF2944) at 4°C overnight. Stained slides were blocked by 3 x 10-minute washes in blocking buffer, incubated with donkey anti-goat Alexa Fluor 488 at 1:250 dilution (Invitrogen A-11055) for 60 minutes at 37°C, and blocked by 3 x 10-minute washes in blocking buffer. Slides were mounted using mounting medium containing DAPI (Vectashield). Slides were digitized using EVOS M700 Imaging System (Invitrogen) with a 40× objective and analyzed using QuPath software[81]. Images in figures were captured using a Leica TCS SP8 STED confocal microscope. Images of PLZF-staining were subjected to wavelet-based noise reduction using Leica LAS X software.

Epididymal Sperm Counts

To assess epididymal sperm count in mice, the epididymis was separated from the testis. The cauda epididymis and deferent duct were then dissected and washed in 1 ml of 1x PBS (10x stock, Fisher Bioreagents, REF-BP3994) in a plastic petri dish. The cauda epididymis and deferent duct were then placed in 0.5 ml of 10 mM sodium citrate [pH 6.0] in a plastic petri dish and shredded with needles to release sperm. Epididymal tissue was allowed settle onto the dish prior to the removal liquid supernatant to a 1.5 ml tube. Additional sperm was collected from debris by washing with 0.5 ml of 10 mM sodium citrate [pH 6.0] for 1 ml total suspension. Sperm suspension was then diluted 1: 5 with 10 mM sodium citrate [pH 6.0]. Samples were stored at 4°C before counting. 10 µl of diluted sperm suspension was loaded onto a hemocytometer and five regions were counted. This counting process was repeated 3 times with a fresh diluted sperm sample. To calculate the number of sperm in one sample, the average of the counts was multiplied by 10,000 to account for the dilutions.

Periodic Acid-Schiff (PAS) Staining

Organs were dissected into 3 ml of Bouins solution (Sigma, REF- HT10132-1L) in a 15 ml falcon tube and incubated at 4°C with rocking overnight. The following morning, samples were washed 3 x 20 in MilliQ water at 4°C with rocking. Samples were then washed ~10 x in 70% EtOH at 4°C with rocking, or until the EtOH was clear. Samples were submitted in 70% EtOH to the CUIMC Molecular Pathology Core for paraffin embedding, sectioning, and PAS staining.

### Blood Collection, Sample Preparation, and Hormone Analysis

Mice were anesthetized via isoflurane and sacrificed by cervical dislocation followed by decapitation. 0.5 ml-1 ml of blood was collected into a 1.5 ml Eppendorf tube. Blood was allowed to clot at room temperature for 90 minutes and then the clot was disrupted from adhesion to the tube wall with a wooden applicator stick. Samples were centrifuged at 2000*g* for 15 minutes at room temperature to separate serum from blood. Serum was pipetted into a fresh 1.5 ml Eppendorf tube and frozen at -20°C until shipment for analysis. Samples were shipped overnight on dry ice to the University of Virginia Ligand Core for hormone analysis.

### Ovarian Histology

PAS-stained ovary sections were imaged using an EVOS XL core microscope at 10 x magnification. Images were analyzed using FIJI/ImageJ software. Total area was measured via the freehand selections tool and numbers of ovarian features were quantified including antral follicles, secondary follicles, primary follicles, primordial follicles, corpora lutea, post ovulation follicles, follicular cysts, and instances of follicular atresia. Variability of section size was accounted for by calculating the prevalence of each feature per $mm^2$.

### Oocyte Collection and Maturation

Oocytes arrested in meiotic prophase I were collected from ovaries of 6-month-old females as described previously. Briefly, ovaries were placed in minimal essential medium (MEM) containing 2.5 μM milrinone (Sigma-Aldrich #M4659) to prevent meiotic resumption. Oocytes were isolated by piercing the ovaries with needles. To induce meiotic resumption, oocytes were cultured in milrinone-free Chatot, Ziomek, and Bavister (CZB) medium in an

atmosphere of 5% $CO_2$ in air at 37°C. Oocytes were matured for 16 hours to reach metaphase II, fixed in phosphate buffer saline (PBS) containing paraformaldehyde (PFA) 2% for 20 minutes at room temperature, and permeabilized in PBS containing 0.1% (vol/vol) Triton X-100 and 0.3% (wt/vol) BSA for 20 minutes, and blocked for 10 minutes in blocking buffer (0.3% BSA containing 0.01% Tween in PBS). Immunostaining was performed by incubating cells in Alexa-fluor 488 conjugated primary α-tubulin antibody (1:100; Life Technologies #322588) for 1 hour in a dark, humidified chamber at room temperature followed by 3 consecutive 10-minute incubations in a blocking buffer. After washing, the cells were mounted in 10 µL VectaShield (Vector Laboratories, #H-1000) with 4′, 6- Diamidino-2-Phenylindole, Dihydrochloride (DAPI; Life Technologies #D1306; 1:170).

Oocytes were imaged using a Leica SP8 confocal microscope equipped with a 40X, 1.30 N.A. oil immersion objective. For each image, optical z-slices were obtained using a 1.0 µm step with a zoom setting of 2.5. Numbers of metaphase II eggs were recorded.

Mouse Oocyte Isolation, Culturing, Microinjection and
High-Resolution Confocal Microscopy

All animal work for visualizing meiosis in live oocytes was conducted in accordance with a protocol approved by the Yale University Institutional Animal Care and Use Committee (IACUC number 2021-20408). Wild-type C57BL/6 female mice were purchased from The Jackson Laboratory. Oocytes were isolated from ovaries of wild-type and *Pnma* knockout mice and cultured in M2 medium containing dbcAMP (Sigma, D0627-100MG) to maintain them in prophase arrest. mRNA transcription constructs for synthesis of MAP4 (to mark microtubules) and H2B (to mark chromosomes) were generated as described previously[82]. Oocytes were microinjected with 6 to 8 pl of *in vitro* transcribed mRNAs as described in detail recently[83].

For live imaging of meiotic spindle assembly and chromosome segregation, oocytes were released from prophase arrest by transferring them into M2 medium that did not contain dbcAMP. Confocal timelapse images were acquired using Zeiss LSM 800 and Zeiss LSM 900 microscopes equipped with 40× C-Apochromat 1.2 numerical aperture (NA) water-immersion objectives and environmental chambers maintained at 37°C. Image acquisition using ZEN blue software (Zeiss) was performed at a temporal resolution of 6 min and with a Z-stack thickness of 40.5 μm at 1.5 μm confocal sections.

Plasmid Creation

PCDNA4TO (Thermo Fisher, REF-V102020) was provided by Dr. Yossi Sabo. *DAZL* (human) gene fragments were synthesized by Twist Biosciences and cloned into the otherwise unmodified PCDNA4TO. For *PNMA1, PNMA4* (both human and mouse), *SMC1B* and *ARC* plasmids we modified PCDNA4TO by inserting an *mCherry* reporter driven by the secondary SV40 promoter to act as a transfection efficiency reporter and a monomeric protein control. 3V5 tagged *Pnma1* and *Pnma4* (mouse) and *PNMA1, PNMA4*, *SMC1B* and *ARC* (human) gene fragments were synthesized by Twist Biosciences and cloned into the modified PCDNA4TO backbone in the Berchowitz lab using standard restriction cloning procedures. Plasmid sequences were analyzed using Genewiz/Azenta.

HEK293T Cell Culture

HEK293T cells were grown on 6-well flat bottom with lid culture plates (tissue culture treated, non-pyrogenic, polystyrene, Corning, REF-3516). Cells were maintained in a grown medium consisting of Dulbecco's Modified Eagle Medium (DMEM) +4.5 g/L D-Glucose, -sodium Pyruvate (Gibco, REF-11960-044), supplemented with 10% Heat Inactivated Fetal

Bovine Serum (HI-FBS) (Sigma, REF-F4135), 1% L Glutamine 200 mM (Gibco, REF-25030-081), and 1% penicillin-streptomycin (Sigma, REF-P4458). Cells were incubated at 37°C, with 5% $CO_2$ in a humidified environment (HERAcell 150i, ThermoFisher Scientific). Cells were passaged when they reached ~80% confluency after ~2 days and reseeded at 5 x $10^5$ cells per well of a 6-well plate. During passaging, cells were dissociated using 0.25% Trypsin-EDTA, phenol red (ThermoFisher Scientific, REF-25200056) for 2 minutes at 37°C and inactivated via dilution with un-supplemented DMEM. Cells were centrifuged at 300 RCF for 5 minutes and resuspended in supplemented medium and then reseeded in the necessary dishes.

HEK293T Transfection

On experimental day one, cells were passaged into a 24-well plate (2.5x$10^5$ cells per well) with 450 µl of supplemented DMEM and incubated for 24 hours as described above. On experimental day 2, old media was aspirated and replaced with premixed 250 µl supplemented DMEM and 200 µl OPTI-MEM (Gibco, REF-31985-070) and then placed back in the incubator. To prepare transfection mix tubes: tube 1 (general tube) contained 24 µl per well of OPTI-MEM, and 1ul per well of Lipofectamine 3000 (Thermo Fisher, REF-L3000008), tube 2 (well specific tube, 2A, 2B, etc.) contained 1 µl of P3000 enhancer (Thermo Fisher, REF-L3000008), 1 µg/µl of plasmid DNA, and 23 µl of OPTI-MEM. We added 25 µl of tube 1 to each tube 2 and incubated them in the tissue culture hood for 10-15 minutes at room temperature. The 50 µl of transfection mix was added to the wells using gradual dropwise pipetting, gently rocked to distribute transfection mixture, and cells were placed back in the incubator for 24 hours. On experimental day 3, cells were checked for transfection efficiency using fluorescent reporters

(Evos M5000). If transfection was successful, cells were dissociated as described above and prepared for western/northern blotting or density gradient ultracentrifugation analysis.

Cloning, Expression, and Purification of Recombinant PNMA1 and PNMA4

Full-length mouse *Pnma1* and *Pnma4* (DNA fragments synthesized by Twist Biosciences) were cloned into the first cassette of a pET-Duet plasmid using inverse PCR. The pET-Duet plasmid was previously modified to encode a C-terminal 3C protease cut site) followed by a His tag. For protein expression, the plasmids carrying C-terminally tagged *Pnma1* and *Pnma4* were transformed into the LOBSTR *E. coli* expression strain containing a RIL plasmid (Kerafast)[84]. To obtain purified PNMA1 and PNMA4, 1L of LOBSTR cells containing the *Pnma1* or *Pnma4* plasmids were grown in LB (with 100 μg/mL of ampicillin and 35 μg/mL of chloramphenicol) to an $OD_{600}$ of 0.7 at 37°C, shifted to 18°C for 30 min, and induced with 200 μM IPTG for 18 hrs. The cells were collected by centrifugation and resuspended in 400 mL of cold lysis buffer (50 mM Tris/HCl pH 8.0, 500 mM NaCl, 40 mM imidazole, and 5 mM β-mercaptoethanol). The cells were lysed on ice by sonication (Misonix). Cell debris was removed by centrifugation and the supernatant was incubated with 1 mL of equilibrated Ni-Sepharose 6 Fast Flow resin (GE Life Sciences) and gently mixed in batch for 1 hr at 4° C. The Ni-resin was collected by centrifugation and placed into a 10 mL polypropylene column (Thermo Scientific) and washed with 100 column volumes of wash buffer (10 mM Tris/HCl pH 8.0, 150 mM NaCl, 40 mM imidazole, and 5 mM β-mercaptoethanol). Pnma1 and Pnma4 were eluted with 10 mL of elution buffer (10 mM Tris/HCl pH 8.0, 150 mM NaCl, 300 mM imidazole, and 5 mM β-mercaptoethanol) and analyzed by SDS-PAGE. Samples were then concentrated over a 30 kD

molecular weight cutoff (MWCO) filter to ~500 µl, and run over a Superose 6 Increase HiScale (80 ml volume, Cytiva) size exclusion column. Fractions were analyzed by SDS-PAGE (**Supplementary Figure 5.9**) Finally, the void fractions corresponding to large forms of PNMA1 and PNMA4 were collected for further analysis.

Negative Stain Transmission Electron Microscopy

Aliquots of PNMA1 and PNMA4 were thawed on ice and centrifuged for 5 minutes at 4° C at 16,000*g* to pellet any insoluble protein. Samples for negative stain TEM were prepared as described[85]. Formvar/Carbon 300 Mesh, Cu grids (Electron Microscopy Sciences, FCF300-Cu-50) were glow discharged using 15 mA for 45 s, with a 10 s hold (easiGlow, PELCO). 3 µL (PNMA4) or 5 µL (PNMA1) of sample was spotted on glow-discharged grids at room temperature. After 1 minute, sample was blotted away, washed twice with water, and stained with 2% uranyl acetate (made from SPI supplies, CAS#6159-44-0). Samples were dried for at least ten minutes prior to imaging. Images were acquired using DigitalMicrograph from Gatan at either 15000 or 23000 nominal magnification and converted to MRC files in DigitalMicrograph. Micrographs were imported into cryoSPARC[86] (v4.4.1+240110) (Output constant CTF: true). Capsid-like particles were manually picked and classified using the 2D classification job (Re-center 2D-classes: false, Do CTF correction: false). For PNMA4, the partial staining observed prevented use of the template picker. However, for PNMA1, the complete negative staining allowed us to use the template picker followed by several rounds of 2D classification to identify capsid-like structures.

Velocity Gradient Ultracentrifugation

HEK293T cells were transfected with 3*V5*-tagged *PNMA1*, *PNMA4*, or *ARC* expression plasmids as described above. Approximately ~4 x $10^6$ cells were lysed in polysome lysis buffer (1% Triton-X, 20 mM Tris-HCl [pH 7.5], 10 mM magnesium chloride, 50 mM potassium chloride, 10 µg/mL cycloheximide, 1 mM PMSF, 1x Halt protease and phosphatase inhibitor cocktail (Thermo Fisher Scientific, 78442)). Lysate was cleared by centrifugation at 4˚ C at 20,000*g* for 10 minutes. Lysate was loaded on a 10% to 50% sucrose gradient in polysome lysis buffer which was centrifuged at 4˚ C for 3 hrs at 30,000 rpm in a Beckman SW41Ti rotor. Fractions were collected using a BioComp gradient station and a BioComp Triax flow cell monitoring continuous absorbance at 260 nm. For Western blot analysis, each fraction received 100% TCA to a final concentration of 5% TCA. Precipitate was pelleted by centrifugation and washed once in acetone. We added 99 µl TE, 1 µl 1M DTT, and 50 µl Loading Buffer (9% SDS, 0.75 mM Bromophenol blue, 187.5 mM Tris-HCl [pH 6.8], 30% glycerol, and 810 mM beta-mercaptoethanol); samples were heated at 100°C for 5 min and centrifuged 5 min at 20,000 g prior to analysis by SDS-PAGE.

Western Blotting

To prepare testes lysate (for protein and RNA extraction), one snap-frozen testis was ground by mortar and pestle into a fine powder under liquid nitrogen. 0.75 ml lysis buffer (150 mM NaCl, 50 mM Tris [pH 7.5], 0.25% NP-40, 1 mM PMSF, 1x Halt Protease Inhibitor Cocktail (Pierce)) was added. The lysate was clarified by centrifugation at 20,000 g for 20 minutes, collected to a new tube, and re-clarified by centrifugation at 20,000 g for 20 minutes. 100 µl lysate was combined with 50 µl 3x SDS loading buffer (9% SDS, 0.75 mM Bromophenol blue,

187.5 mM Tris-HCl [pH 6.8], 30% glycerol, and 810 mM β-mercaptoethanol) and samples were boiled at 100°C for 5 minutes, and then centrifuged at 20,000 x g for 5 minutes. To prepare HEK293T lysate, $5x10^5$ cells were resuspended in 100 µl lysis buffer (85 µl RIPA buffer (Sigma, REF-R0278), 10µl TURBO DNase buffer (ThermoFisher Scientific, REF-AM2238), 2 µl TURBO DNase (ThermoFisher, REF-AM2238), 1 µl RNaseA at 5 µg/µl (Lucigen, REF-MRNA092), 1 µl HALT protease inhibitor (ThermoFisher Scientific, REF-1861282), 1 µl PMSF (ThermoFisher, REF-36978)) and incubated at room temperature for 1 minute. 50 µl 3x SDS loading buffer was added, samples were boiled at 100°C for 5 minutes, and centrifuged at 20,000*g* for 5 minutes.

Pre-cast polyacrylamide gels (4%-15% Mini-Protean TGX Precast Gels, Bio-Rad, 10 well REF-4561083, 15 well REF-4561086) were loaded and run on a mini gel system (Mini-PROTEAN Tetra Vertical Electrophoresis Cell, Bio-Rad REF-1658004) at 200V for 35-45 minutes. Gels were run with precision plus protein dual color standards (Bio-Rad, REF-1610374) and with SDS Running Buffer (190 mM glycine, 25 mM Trizma base, 3.5 mM, 1% SDS). Transfers to nitrocellulose (Bio-Rad, REF-1620145) were performed using semi-dry transfer on a trans turbo blot transfer apparatus (Bio-Rad, REF-1704150), with turbo transfer filter paper for nitrocellulose (Bio-Rad 1620215) and stained in Ponceau S to visualize total protein load. Membranes were blocked with 3% milk powder in TBST for 1 hour. Primary antibody solutions were prepared in 50 ml of 1x TBST, 2% Sodium azide, 1% milk 1% BSA plus antibody at necessary concentration. Membranes were incubated with primary antibodies at 4°C with rocking overnight. The next day, primary antibody solution was removed, and the membrane given 2x 10-minute 1x TBST washes. Secondary antibody solutions were prepared in 50 ml of

1x TBST, 1% milk 1% BSA plus antibody at necessary concentration. After ~2 hours (or longer if necessary) secondary antibody solutions were removed from membranes, which were washed 2x (10-minute 1x TBST) before visualization.

Primary antibodies were used at the following concentrations: α-PNMA1 1: 2,000 (Proteintech 13631-1-AP), α-PNMA1/MOAP1 1: 2,000 (Proteintech 15325-1-AP), α-α-tubulin 1: 1,000 (Invitrogen), α-V5 1: 1,000 (Invitrogen), α-DAZL 1: 5,000 (Abcam), α-mCherry 1: 2,000 (Abcam), α-GAPDH 1:5,000 (Thermo Fisher), and α-MLV p30 1: 1,000 (Abcam). α-mouse HRP-conjugated, α-rabbit HRP-conjugated (Cytiva), or α-rat FITC-conjugated (for α-α-tubulin, Thermo Fisher) secondary antibodies were used.

Quantitative PCR

Total RNA was prepared from 400 µl testis lysate prepared as in previous section. SDS was added to 0.5%, 400 µL acid phenol: chloroform 5: 1 (Ambion) was added, and samples were incubated at 1,400 rpm for 30 min at 65°C in a Thermomixer (Eppendorf). Samples were centrifuged for 5 min at 13,000*g*, extracted to 1 mL 100% ethanol and 40 µL sodium acetate [pH 5.5], and precipitation was conducted at 20°C overnight. RT and qPCR was conducted using the Primetime one-step RT-qPCR master mix (IDT) with 80 ng RNA template, 500 nM forward/reverse primer, and 250 nM probe. Reactions were conducted in triplicate (with no RT and no RNA controls) in a CFX96 thermocycler (BioRad) using the following parameters: 50°C 15:00 (1x), 95°C 3:00 (1x), 95°C 0:05, 60°C 0:30 (40x).

***Pnma1:*** F primer 5'- GAATGGCAGGTGTCGGATATAG-3'

R primer 5'- GCGTTGTTGGTCTTGAGGAT-3'

probe 5'- /56-FAM/CGGCGGTTG/ZEN/ATGGAGAGTTTGAGA/3IABkFQ/-3'

***Pnma4:*** F primer 5'- GGACCAGGTGAGGAAGAATTT-3'

R primer 5'- TGGACCTCTAAGGCTCTCTATC-3'

probe 5'-/56-FAM/ACATGGCAG/ZEN/GTGTCAGATGTAGAGA/3IABkFQ/-3'

***PPIA:*** F primer 5'- GGCTATAAGGGTTCCTCCTTTC-3'

R primer 5'- TTTCTCTCCGTAGATGGACCT-3'

probe 5'- /56-FAM/ATGTGCCAG/ZEN/GGTGGTGACTTTACA/3IABkFQ/3'

Extracellular Protein Analysis

HEK293T cells were transfected with 3*V5-PNMA1*, 3*V5-PNMA4*, or 3*V5-ARC* expression plasmids as described above. 40 ml extracellular medium was collected and filtered (0.45 μm) to remove cells and large debris. Cells were collected and lysed in 85 μl RIPA buffer (cell lysate control, analyzed as above). Growth medium was concentrated over a 100 kD MWCO filter to ~500 μL which was diluted to 8 ml in 1x PBS and layered on top of a sucrose cushion of 2 ml 25% sucrose in an SW41 tube (Seton). The tubes were spun for 3 hr at 38,000 rpm at 4˚ C in an SW41Ti rotor. After centrifugation, the following fractions were taken: lipids, medium (from middle of 8 ml), meniscus of 25% sucrose, 25% sucrose, and pellet. A sample of each fraction was taken for analysis by SDS-PAGE/immunoblot.

Analysis of Testicular Capsid-Like PNMA4 Protein

Lysate Preparation 10 snap-frozen testes from wild type (C57BL/6J) mice were ground by mortar and pestle into a fine powder under liquid nitrogen. 1 ml lysis buffer (150 mM NaCl, 50 mM Tris [pH 7.5], 0.25% NP-40, 1 mM PMSF, 1x Halt Protease Inhibitor Cocktail (Pierce)) was added. The lysate was clarified by centrifugation at 20,000 g for 20 minutes, collected to a

new tube, and re-clarified by centrifugation at 20,000 g for 20 minutes and diluted to 8 ml in lysis buffer.

Double Sucrose Gradient 8 ml testis lysate was layered on top of a double sucrose cushion of 2 ml 25% (upper) and 2 ml 70% (lower) in an SW41 tube (Seton). The tubes were spun for 3 hr at 38,000 rpm 4˚ C in an SW41Ti rotor. After centrifugation, the following fractions were taken: lipids, lysate (from middle of 8 ml), meniscus of 25% sucrose, 25% sucrose, meniscus of 70% sucrose, 70% sucrose, and pellet. A sample of each fraction was taken for analysis by SDS-PAGE/immunoblot.

Iodixanol Step Gradient An iodixanol (Optiprep density gradient medium, supplied at 60% iodixanol) step gradient was prepared in an SW41 tube with the following steps: (top) 2 ml 15% (in 1 M NaCl), 3 ml 30%, 3 ml 40%, 3 ml 50% (bottom). The 70% meniscus fraction (~500 µL) was layered on top and the tube was spun at 41,000 rpm for 48 hours at 4˚ C in an SW41Ti rotor. Twelve 1 ml fractions were taken using a pipette. ~40 µL samples (20 µL 3x SDS loading buffer was added) from each fraction were taken for analysis by SDS-PAGE/immunoblot (30 µL load).

Northern Blot Analysis

1 x $10^7$ HEK293T cells were harvested by centrifugation. RNA was extracted using the RNeasy kit (Qiagen, REF-74104) using the manufacturer's protocol. We added 22 mL denaturing mix (15 mL formamide, 5.5 mL formaldehyde, and 1.5 mL 10x MOPS) to 8 mg total RNA in 8 mL and heated at 55°C for 15 min 20 mL of sample (approximately 5 mg) was resolved on a denaturing agarose gel (1.9% agarose, 3.7% formaldehyde, 1x MOPS buffer) for 2.5 h at 80 V. The gel was blotted to a Hybond membrane (GE Healthcare) by capillary transfer

in 10x SSC (1.5 M NaCl, 0.15 M trisodium citrate dihydrate, [pH 7]). The membrane was incubated in a hybridization buffer (0.25 M Na-phosphate [pH 7.2], 0.25 M NaCl, 1 mM EDTA, 7% SDS, and 5% dextran sulfate) at 65°C probed with 32P-labeled V5 DNA probes prepared via Amersham Megaprime DNA labeling kit (GE Healthcare) and Illustra ProbeQuant columns (GE Healthcare), transferred to a phosphor screen, and imaged on a Typhoon imager (GE Healthcare).

*PNMA1-5* Expression Analysis in Human and Mouse Ovary

To assess the expression of *PNMA1-5* in the human ovary, we analyzed single-nuclei RNA-seq datasets obtained from eight individuals. These datasets included samples from four younger women aged 23, 27, 28, and 29, and four older women aged 49, 51, 52, and 54. To assess the expression of *PNMA1-5* in the mouse ovary, we performed single cell RNA sequencing of mouse ovaries from reproductive young (Y, 4.5-month, n=5), peri-estropause (M, 10.5-month, n=7), and post-estropause (O, 15.5-month, n=5) stage. The gene expression data and metadata for these samples were downloaded from the GEO database (accession number: GSE202601). Additionally, human oocyte expression data were sourced from two published datasets available under GEO accession numbers GSE107746 and GSE155179. All gene expression data were normalized using the 'NormalizeData' function from Seurat (v4.4.0)[87] Previously published data from single cell RNA-seq of adult murine ovarian cells was also used (C57BL/6J mice; n=8). To identify cell-types that express *Pnma1* and *Pnma1,* cells positive for expression of these genes were displayed over a dimplot of previously identified clusters of distinct ovarian cell-types using the featureplot command from the Seurat package. To compare expression of these genes across clusters, the average expression levels from each sample was obtained and used for statistical comparison in the GraphPad Prism software. Strip plots are

presented with individual points shown and means ± SEM indicated. For comparisons of means between clusters, one-way ANOVA was used followed by Tukey's post-hoc test. Significant differences were defined at $P < 0.05$.

Analysis of Traits Associated with Human *PNMA1* and *PNMA4* Variation

We gathered information on the traits associated with the *PNMA1* and *PNMA4* genes using the Open Targets Genetics database containing GWAS summary statistics[42] This database predicts the causal genes of traits using the variant-to-gene and locus-to-gene (L2G) pipeline[41]. In the database, we queried the *PNMA1* and *PNMA4* genes to collect the causally-associated traits from the GWAS studies. We then filtered traits by *P*-value < $5x10^{-8}$ and merged them into a table (**Table S1**) labeled with the queried genes.

Statistical Analysis

All statistical analyses were conducted using Prism (V10, GraphPad). Unless indicated otherwise, statistical significance (*$p < 0.05$) was determined by student's t-test (with Welch's correction for single nuclei RNAseq data), multiple t-tests with correction for multiple comparisons, one-way ANOVA with correction for multiple comparisons, or Mann-Whitney test as indicated. For all figures, the number of individuals is denoted by capital N and number of measurements is denoted by lowercase n.

Neurobehavioral Tests

All mouse behavioral tests were conducted within the CUIMC Neurobehavior Core facility. ~6-month-old male *Pnma1*$^{-/-}$, *Pnma4*$^{-/-}$, *Pnma1*$^{-/-}$*: Pnma4*$^{-/-}$ double knockouts or C57BL/6J wild type mice were used for all tests. Mice were given a minimum of 30 minutes to

equilibrate to testing rooms before experiments were conducted. To ensure minimal stress for the animals, the tests progressed from least to most stressful (open field, grip strength, inverted hang, Y maze, fear conditioning).

Open Field Test Individual mice were placed into the center of a clear Plexiglass box (27.31 x 27.31 x 20.32 cm), (Med Associates, ENV-510) in a dim room (~5 lux). Mice were allowed to explore freely for 30 minutes. Mouse behavior, including exploration of different areas of the box, horizontal and vertical locomotion speed and distance, and time spent in the center vs edge were tracked. Tracking was conducted via infrared beams that were embedded along the X, Y, and Z axes of the box and processed by Activity Monitor Version 7 tracking software (Med Associates Inc.). At the end of the test, mice were returned to their home cages and the box was thoroughly cleaned between trials using 70% EtOH.

Grip Strength Test Forelimb grip strength was tested in mice by utilizing a grip strength test meter (BIOSEB). A small metal grid acted as a grasping surface and was connected to the strength meter. The metal grid was positioned so mice would grasp the bottom bar of the grid, with this being parallel to the horizontal plane of the strength meter. The mouse was held by the tail and moved towards the grid until it reached out and grasped the bottom bar of the grid with its forepaws. The mouse's tail was gradually pulled backwards until the mouse's torso was horizontal. When the mouse lost its grip, that force recording was taken as the grip strength reading at peak tension. This test was repeated five times for each mouse. The mean of these five trials was recorded.

Inverted Hang Test Mice were placed on a grid, which was inverted and placed over a (30 x 30 x 30 cm) plastic box with padding at the bottom. Time was started upon inversion of the

grid. Latency to fall was measured for each mouse or after 3 minutes when the trial was ended. At the end of the test, mice were returned to their home cages and the grid was thoroughly cleaned between trials using 70% EtOH.

Y Maze Spontaneous Alternation Mice were placed in a plastic Y-maze comprising 3 arms of equal size (30 x 5 x 30 cm) extending from a central area. The test was conducted in a quiet, dim (~5 lux) room. The rate of spontaneous alternation between arms of the maze was tracked using Ethovision XT software (Noldus) over an 8-minute trial. At the end of the test, mice were returned to their home cages and the maze was thoroughly cleaned between trials using 70% EtOH.

Fear Conditioning Fear conditioning studies were performed as previously described[88]. Training and conditioning trials were performed in two identical shock chambers (30 x 24 x 21 cm), which were inside sound-mitigating chambers to limit outside stimuli (Med Associates). Chambers consist of a clear polycarbonate front wall, a white opaque back wall, two steel side walls and a grid floor with a waste collection pan underneath. Chambers were calibrated to deliver shocks of 0.5 milliamps and 80 decibels. A camera on the door of the attenuating chamber was used to record the experimental session. Sessions were analyzed, and % of time freezing was calculated using VideoFreeze software (Med Associates). For the training session, each chamber was illuminated with white house light. An olfactory cue was presented using a of drop vanilla extract on the back wall of the shock chamber. On experimental day 1, mice were placed into the chamber and allowed to explore freely for 2 minutes. After 2 minutes a pure tone lasting 30 seconds was played (5kHZ, 80dB) to serve as the controlled stimulus (CS). In the last 2 seconds a foot shock (0.5 mA) was administered through the floor grid, which served as the

unconditioned stimulus (US). Over the course of the day 1 trial, each mouse received three CS-US pairings with 90 second intervals, with the mouse being left in the chamber after the final CS-US pairing for an additional 2 minutes. Percentage of time staying frozen still (freezing behavior) was recorded during this 2-minute interval. The mouse was then returned to their home cage. On experimental day 2 contextual conditioning was tested in the same shock chamber. The same chamber illumination and olfactory cue was present but with no tone or foot shock. Mice were placed in the chamber for 5 minutes with no CS or US delivered. Freezing behavior was scored during this 5-minute period and the mice was returned to its home cage. On experimental day three cued conditioning was tested. Contextual cues of the cage were changed by covering the grid floor with smooth white plastic and inserting a piece of curved white plastic to cover the side walls. Near infrared light was used in place of white light, and the olfactory cue was changed from vanilla extract to lemon extract. Mice were placed in the chamber and allowed to freely explore for 3 minutes, followed by 3 minutes of the CS tone (5 kHz, 80 dB). During both 3-minute periods, the percentage of time spent freezing was recorded and scored, with the mouse being returned to its home cage. The chamber was thoroughly cleaned between trials using 70% EtOH, with fresh olfactory cue being applied for each new trial.

## References

1. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691–703 (2009).

2. Kaneko-Ishino, T. & Ishino, F. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front Microbiol* **3**, 262 (2012).

3. Jangam, D., Feschotte, C. & Betrán, E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics* **33**, 817–831 (2017).

4. Volff, J.-N. Cellular genes derived from Gypsy/Ty3 retrotransposons in mammalian genomes. *Annals of the New York Academy of Sciences* **1178**, 233–243 (2009).

5. Brosius, J. Retroposons—Seeds of Evolution. *Science* **251**, 753–753 (1991).

6. Haig, D. Transposable elements: Self-seekers of the germline, team-players of the soma. *BioEssays* **38**, 1158–1166 (2016).

7. Crichton, J. H., Dunican, D. S., MacLennan, M., Meehan, R. R. & Adams, I. R. Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. *Cell Mol Life Sci* **71**, 1581–1605 (2013).

8. McLaughlin, R. N. & Malik, H. S. Genetic conflicts: the usual suspects and beyond. *J Exp Biol* **220**, 6–17 (2017).

9. Yang, F. & Wang, P. J. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. *Seminars in Cell & Developmental Biology* **59**, 118–125 (2016).

10. Laureau, R. *et al.* Meiotic Cells Counteract Programmed Retrotransposon Activation via RNA-Binding Translational Repressor Assemblies. *Developmental Cell* **56**, 22-35.e7 (2021).

11. Kokošar, J. & Kordiš, D. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol Biol Evol* **30**, 1015–1031 (2013).

12. Henriques, W. S. *et al.* The Diverse Evolutionary Histories of Domesticated Metaviral Capsid Genes in Mammals. *Molecular Biology and Evolution* **41**, msae061 (2024).

13. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends in Genetics* **22**, 585–589 (2006).

14. Pang, S. W., Lahiri, C., Poh, C. L. & Tan, K. O. PNMA family: Protein interaction network and cell signalling pathways implicated in cancer and apoptosis. *Cell Signal* **45**, 54–62 (2018).

15. Devine, M. F., Kothapalli, N., Elkhooly, M. & Dubey, D. Paraneoplastic neurological syndromes: clinical presentations and management. *Ther Adv Neurol Disord* **14**, 1756286420985323 (2021).

16. Naville, M. *et al.* Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clin. Microbiol. Infect* **22**, 312–323 (2016).

17. Zhang, X.-L. *et al.* Pnma5 is essential to the progression of meiosis in mouse oocytes through a chain of phosphorylation. *Oncotarget* **8**, 96809–96825 (2017).

18. Segel, M. *et al.* Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science* **373**, 882–889 (2021).

19. Madigan, V. *et al.* Human paraneoplastic antigen Ma2 (PNMA2) forms icosahedral capsids that can be engineered for mRNA delivery. *Proceedings of the National Academy of Sciences* **121**, e2307812120 (2024).

20. Xu, J. *et al.* PNMA2 forms immunogenic non-enveloped virus-like capsids associated with paraneoplastic neurological syndrome. *Cell* **187**, 831-845.e19 (2024).

21. Lesch, B. J., Silber, S. J., McCarrey, J. R. & Page, D. C. Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat Genet* **48**, 888–894 (2016).

22. Soumillon, M. *et al.* Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports* **3**, 2179–2190 (2013).

23. Yuan, L. *et al.* Single-cell transcriptome analysis of human oocyte ageing. *Journal of Cellular and Molecular Medicine* **25**, 6289–6303 (2021).

24. Zhang, Y. *et al.* Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. *Molecular Cell* **72**, 1021-1034.e4 (2018).

25. Zagore, L. L. *et al.* DAZL Regulates Germ Cell Survival through a Network of PolyA-Proximal mRNA Interactions. *Cell Reports* **25**, 1225-1240.e6 (2018).

26. Rosario, R., Smith, R. W. P., Adams, I. R. & Anderson, R. A. RNA immunoprecipitation identifies novel targets of DAZL in human foetal ovary. *Molecular Human Reproduction* **23**, 177–186 (2017).

27. Yang, C.-R. *et al.* The RNA-binding protein DAZL functions as repressor and activator of mRNA translation during oocyte maturation. *Nat Commun* **11**, 1399 (2020).

28. Danilovich, N. *et al.* Estrogen Deficiency, Obesity, and Skeletal Abnormalities in Follicle-Stimulating Hormone Receptor Knockout (FORKO) Female Mice**This investigation was supported in part by the Canadian Institutes of Health Research. *Endocrinology* **141**, 4295–4308 (2000).

29. Heine, P. A., Taylor, J. A., Iwamoto, G. A., Lubahn, D. B. & Cooke, P. S. Increased adipose tissue in male and female estrogen receptor-α knockout mice. *Proceedings of the National Academy of Sciences* **97**, 12729–12734 (2000).

30. Schomberg, D. W. *et al.* Targeted Disruption of the Estrogen Receptor-α Gene in Female Mice: Characterization of Ovarian Responses and Phenotype in the Adult*. *Endocrinology* **140**, 2733–2744 (1999).

31. Isola, J. V. V. *et al.* A single-cell atlas of the aging mouse ovary. *Nat Aging* **4**, 145–162 (2024).

32. Razmjou, S. *et al.* Body composition, cardiometabolic risk factors, physical activity, and inflammatory markers in premenopausal women after a 10-year follow-up: a MONET study. *Menopause* **25**, 89 (2018).

33. Kelly, D. M. & Jones, T. H. Testosterone and obesity. *Obesity Reviews* **16**, 581–606 (2015).

34. Ke, H. Z. *et al.* Effects of CP-336,156, a New, Nonsteroidal Estrogen Agonist/Antagonist, on Bone, Serum Cholesterol, Uterus, and Body Composition in Rat Models. *Endocrinology* **139**, 2068–2076 (1998).

35. Baik, M. *et al.* Testosterone deficiency caused by castration increases adiposity in male rats in a tissue-specific and diet-dependent manner. *Genes & Nutrition* **15**, 14 (2020).

36. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

37. Ashley, J. *et al.* Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172**, 262-274.e11 (2018).

38. Pastuzyn, E. D. *et al.* The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **172**, 275-288.e18 (2018).

39. Dodonova, S. O., Prinz, S., Bilanchone, V., Sandmeyer, S. & Briggs, J. A. G. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. *PNAS* **116**, 10048–10057 (2019).

40. Peyret, H. A protocol for the gentle purification of virus-like particles produced in plants. *Journal of Virological Methods* **225**, 59–63 (2015).

41. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021).

42. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* **49**, D1311–D1320 (2021).

43. Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med* **26**, 252–258 (2020).

44. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* **53**, 185–194 (2021).

45. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet* **53**, 1260–1269 (2021).

46. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *The American Journal of Human Genetics* **104**, 65–75 (2019).

47. Hollis, B. *et al.* Genomic analysis of male puberty timing highlights shared genetic basis with hair colour and lifespan. *Nat Commun* **11**, 1536 (2020).

48. Martin, S. *et al.* Genetic Evidence for Different Adiposity Phenotypes and Their Opposing Influences on Ectopic Fat and Risk of Cardiometabolic Disease. *Diabetes* **70**, 1843–1856 (2021).

49. Karlsson, T. *et al.* Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. *Nat Med* **25**, 1390–1395 (2019).

50. Hübel, C. *et al.* Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **180**, 428–438 (2019).

51. Zirkin, B. R. & Papadopoulos, V. Leydig cells: formation, function, and regulation†. *Biology of Reproduction* **99**, 101–111 (2018).

52. de Mattos, K., Pierre, K. J. & Tremblay, J. J. Hormones and Signaling Pathways Involved in the Stimulation of Leydig Cell Steroidogenesis. *Endocrines* **4**, 573–594 (2023).

53. Bulletti, C., Bulletti, F. M., Sciorio, R. & Guido, M. Progesterone: The Key Factor of the Beginning of Life. *International Journal of Molecular Sciences* **23**, 14138 (2022).

54. Salehnia, M. & Zavareh, S. The Effects of Progesterone on Oocyte Maturation and Embryo Development. *Int J Fertil Steril* **7**, 74–81 (2013).

55. Greenbaum, M. P. *et al.* TEX14 is essential for intercellular bridges and fertility in male mice. *Proceedings of the National Academy of Sciences* **103**, 4982–4987 (2006).

56. Greenbaum, M. P., Iwamori, N., Agno, J. E. & Matzuk, M. M. Mouse TEX14 Is Required for Embryonic Germ Cell Intercellular Bridges but Not Female Fertility1. *Biology of Reproduction* **80**, 449–457 (2009).

57. Nuzhat, N. *et al.* Altered germline cyst and oocyte differentiation in Tex14 mutant mice reveal a new mechanism underlying female reproductive life-span. 814384 Preprint at https://doi.org/10.1101/814384 (2019).

58. Greenbaum, M. P., Iwamori, T., Buchold, G. M. & Matzuk, M. M. Germ Cell Intercellular Bridges. *Cold Spring Harb Perspect Biol* **3**, a005850 (2011).

59. Conine, C. C. & Rando, O. J. Soma-to-germline RNA communication. *Nat Rev Genet* **23**, 73–88 (2022).

60. Coticchio, G. *et al.* Oocyte maturation: gamete-somatic cells interactions, meiotic resumption, cytoskeletal dynamics and cytoplasmic reorganization. *Human Reproduction Update* **21**, 427–454 (2015).

61. Clarke, H. J. Regulation of germ cell development by intercellular signaling in the mammalian ovarian follicle. *WIREs Developmental Biology* **7**, e294 (2018).

62. Macaulay, A. D. *et al.* Cumulus Cell Transcripts Transit to the Bovine Oocyte in Preparation for Maturation1. *Biology of Reproduction* **94**, 16, 1–11 (2016).

63. Macaulay, A. D. *et al.* The Gametic Synapse: RNA Transfer to the Bovine Oocyte1. *Biology of Reproduction* **91**, 90, 1–12 (2014).

64. Wang, H. *et al.* Rejuvenation of aged oocyte through exposure to young follicular microenvironment. 2024.03.11.584343 Preprint at https://doi.org/10.1101/2024.03.11.584343 (2024).

65. Tan, C. T. *et al.* MOAP-1 Mediates Fas-Induced Apoptosis in Liver by Facilitating tBid Recruitment to Mitochondria. *Cell Reports* **16**, 174–185 (2016).

66. Vos, M. D. *et al.* The RASSF1A tumor suppressor activates Bax via MOAP-1. *J Biol Chem* **281**, 4557–4563 (2006).

67. Fu, N. Y., Sukumaran, S. K. & Yu, V. C. Inhibition of ubiquitin-mediated degradation of MOAP-1 by apoptotic stimuli promotes Bax function in mitochondria. *Proceedings of the National Academy of Sciences* **104**, 10051–10056 (2007).

68. Tan, K. O. *et al.* MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. *J Biol Chem* **276**, 2802–2807 (2001).

69. Sugiura, K., Su, Y.-Q. & Eppig, J. J. Does Bone Morphogenetic Protein 6 (BMP6) Affect Female Fertility in the Mouse?1. *Biology of Reproduction* **83**, 997–1004 (2010).

70. Yan, C. *et al.* Synergistic Roles of Bone Morphogenetic Protein 15 and Growth Differentiation Factor 9 in Ovarian Function. *Molecular Endocrinology* **15**, 854–866 (2001).

71. Li, Q. *et al.* Redundant Roles of SMAD2 and SMAD3 in Ovarian Granulosa Cells In Vivo. *Molecular and Cellular Biology* **28**, 7001–7011 (2008).

72. Berchowitz, L. E. *et al.* Regulated Formation of an Amyloid-like Translational Repressor Governs Gametogenesis. *Cell* **163**, 406–418 (2015).

73. Elguindy, M. M. & Mendell, J. T. NORAD-induced Pumilio phase separation is required for genome stability. *Nature* **595**, 303–308 (2021).

74. Sousa Martins, J. P. *et al.* DAZL and CPEB1 regulate mRNA translation synergistically during oocyte maturation. *J Cell Sci* **129**, 1271–1282 (2016).

75. Cao, Q., Padmanabhan, K. & Richter, J. D. Pumilio 2 controls translation by competing with eIF4E for 7-methyl guanosine cap recognition. *RNA* (2009) doi:10.1261/rna.1884610.

76. Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Science Translational Medicine* **7**, 307ra153-307ra153 (2015).

77. Dopkins, N. & Nixon, D. F. Activation of human endogenous retroviruses and its physiological consequences. *Nat Rev Mol Cell Biol* **25**, 212–222 (2024).

78. Durnaoglu, S., Lee, S.-K. & Ahnn, J. Human Endogenous Retroviruses as Gene Expression Regulators: Insights from Animal Models into Human Diseases. *Mol Cells* **44**, 861–878 (2021).

79. Deo, V. K., Kato, T. & Park, E. Y. Chimeric Virus-Like Particles Made Using GAG and M1 Capsid Proteins Providing Dual Drug Delivery and Vaccination Platform. *Mol. Pharmaceutics* **12**, 839–845 (2015).

80. Schüller, M., Jenne, D. & Voltz, R. The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. *J. Neuroimmunol.* **169**, 172–176 (2005).

81. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci Rep* **7**, 16878 (2017).

82. Scheffler, K., Giannini, F., Lemonnier, T. & Mogessie, B. The prophase oocyte nucleus is a homeostatic G-actin buffer. *J. Cell Sci* **135**, 259807 (2022).

83. Scheffler, K., Giannini, F., Lemonnier, T. & Mogessie, B. The prophase oocyte nucleus is a homeostatic G-actin buffer. *J. Cell Sci* **135**, 259807 (2022).

84. Andersen, K. R., Leksa, N. C. & Schwartz, T. U. Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. *Proteins: Structure, Function, and Bioinformatics* **81**, 1857–1861 (2013).

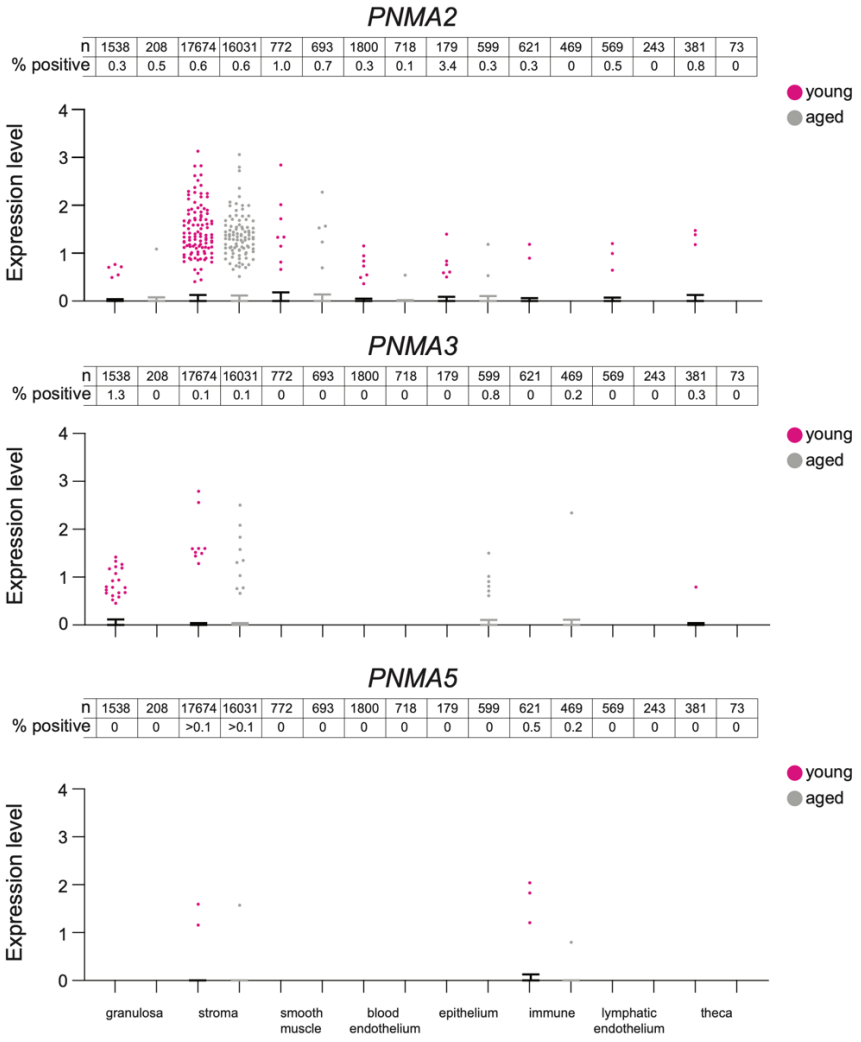
85. Gallagher, J. R., Kim, A. J., Gulati, N. M. & Harris, A. K. Negative-Stain Transmission Electron Microscopy of Molecular Complexes for Image Analysis by 2D Class Averaging. *Curr Protoc Microbiol* **54**, e90 (2019).

86. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290–296 (2017).

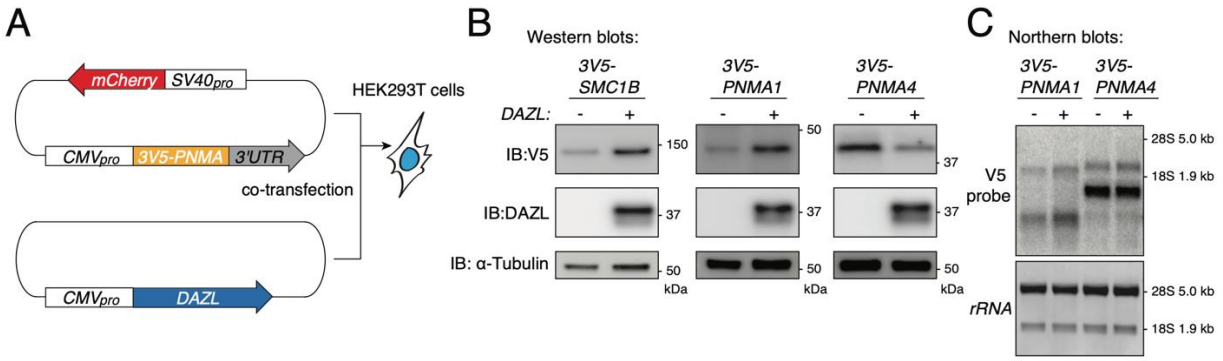
87. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

88. Yang, M. *et al.* Reduced Excitatory Neurotransmission and Mild Autism-Relevant Phenotypes in Adolescent Shank3 Null Mutant Mice. *J. Neurosci.* **32**, 6525–6541 (2012).

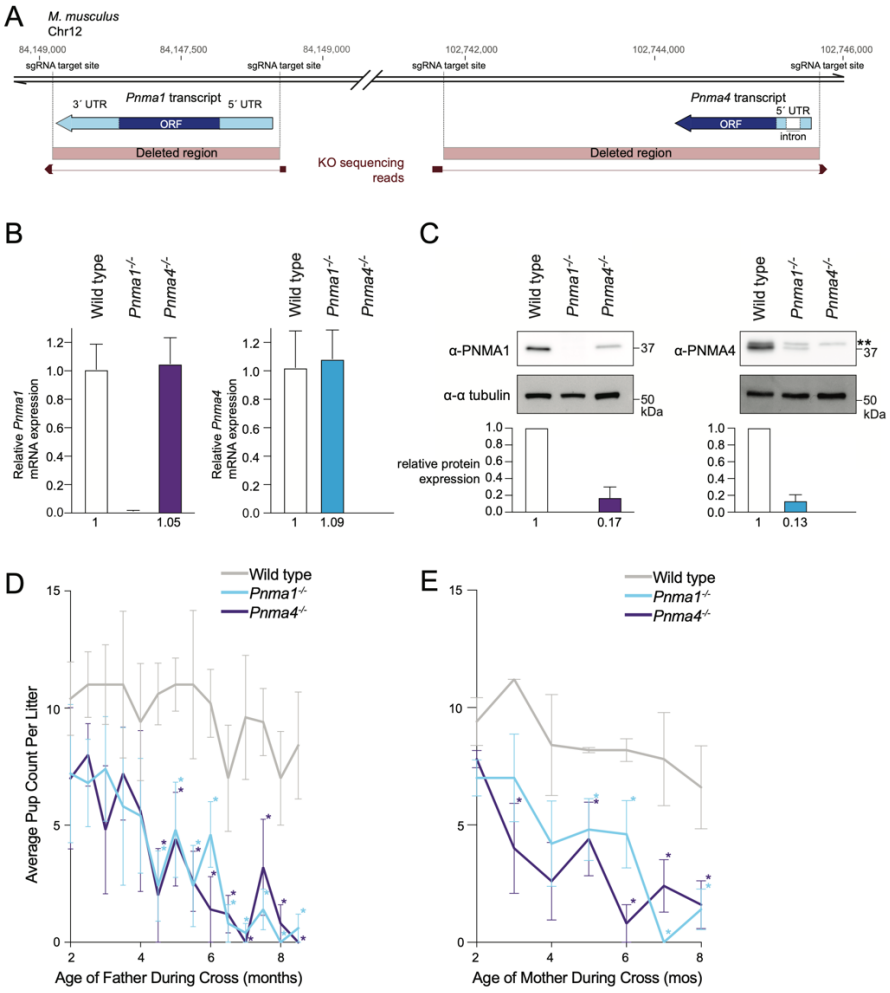
Supplementary Figures




Supplementary Figure 5.1. PNMA2, PNMA3, and PNMA5 expression analysis in human ovaries. Ovaries from reproductively young (23-29 years, magenta, N = 4) and reproductively old (49-54 years, gray, N = 4) donors were analyzed by single-nuclei RNAseq. Uniquely mapping reads for PNMA2, PNMA3, and PNMA5 loci were assigned to ovarian tissue types based on clustering analysis. Statistical significance (*p < 0.05) was determined by Mann-Whitney test.

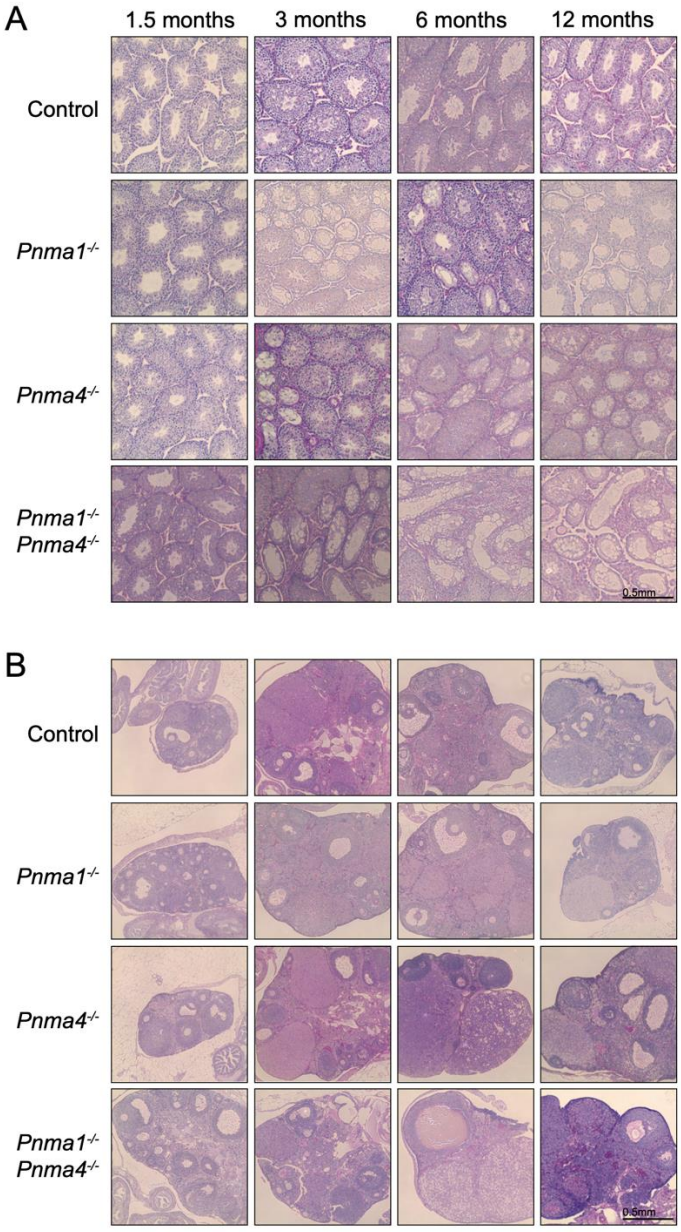



Supplementary Figure 5.2. DAZL has the capacity to post-transcriptionally regulate PNMA1 and PNMA4. (A) 3V5-tagged *PNMA1*, *PNMA4*, or *SMC1B* (positive control DAZL target) expression plasmids were transfected into HEK-293T cells with and without *DAZL* co-transfection. (B) Protein levels of PNMA1, PNMA4, SMC1B, DAZL, and α-tubulin (loading) were analyzed by immunoblot. **(C)** *PNMA1* and *PNMA4* mRNA levels were analyzed by Northern blot.

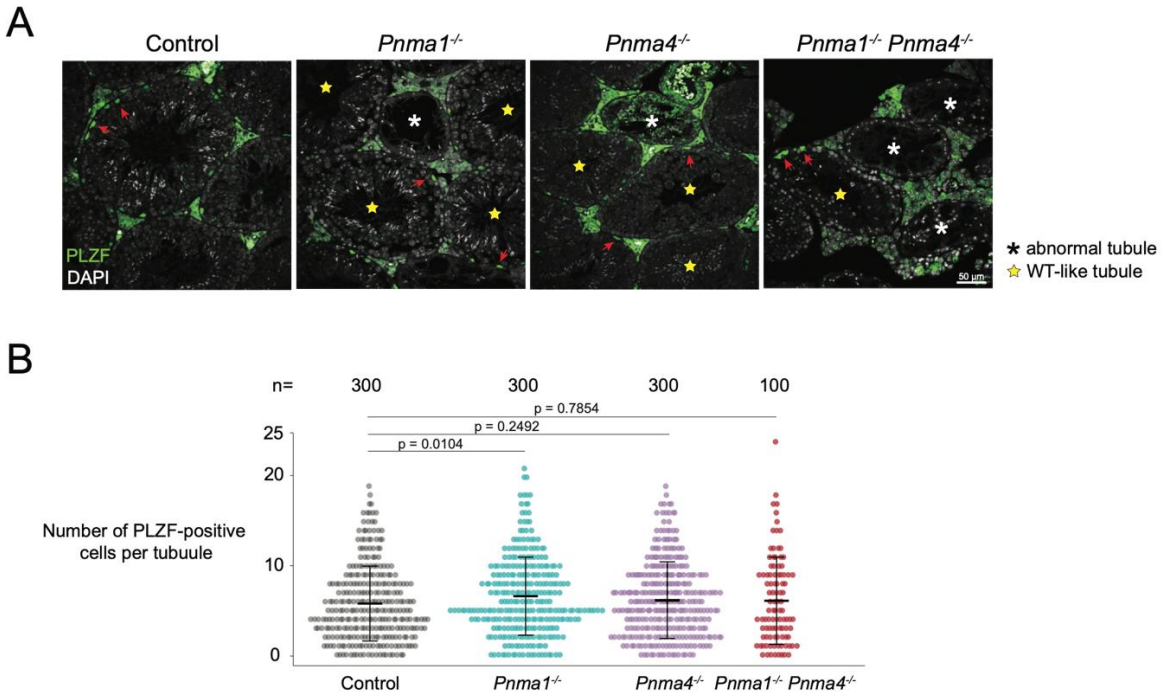



Supplementary Figure 5.3. Diagram of Pnma1 and Pnma4 loci on M. musculus chromosome 12. (A) Pnma1 and Pnma4 deletions were generated by CRISPR/Cas9 genome editing (sgRNA target sites shown in dashed lines). Successful deletion was assessed by PCR/sequencing. Shown are chromosomal coordinates, annotated transcripts (CDS in navy, UTRs in cyan, and introns in white), deleted regions, and sequencing reads (below, maroon) confirming deletion of Pnma1 and Pnma4.(B, C) Analysis of mRNA and protein produced in the Pnma1-/- and Pnma4-/- mutants. A testis was dissected from 3-month wild type (C57BL/6J), Pnma1-/-, and Pnma4-/- mice. Lysate was prepared and split for mRNA and protein analysis. (B) Analysis of Pnma1 and Pnma4 mRNA levels by qPCR (n = 3 replicates). (C) Analysis of PNMA1, PNMA4, and α-tubulin (loading) protein levels by immunoblot. Shown below is a quantification of the ratio of PNMA1 and PNMA4 in knockouts vs. wild type (set to 1) from n = 3 biological replicates. ** denotes presence of an unfortunate cross-reacting band present in all lanes of the α-PNMA4 immunoblot. Error bars indicate SEM. (D) Male wild type mice (gray), Pnma1-/- (cyan), or Pnma4-/- (purple) were crossed biweekly to CF-1 female fertility tester mice (N = 5 pairs for each genotype-timepoint combination). (E) Female wild type mice (gray), Pnma1-/- (cyan), or Pnm-/- (purple) were crossed to B6D21/J male fertility tester mice (N = 5 pairs for each genotype-timepont combination).Pup numbers for each cross were recorded. This is the same

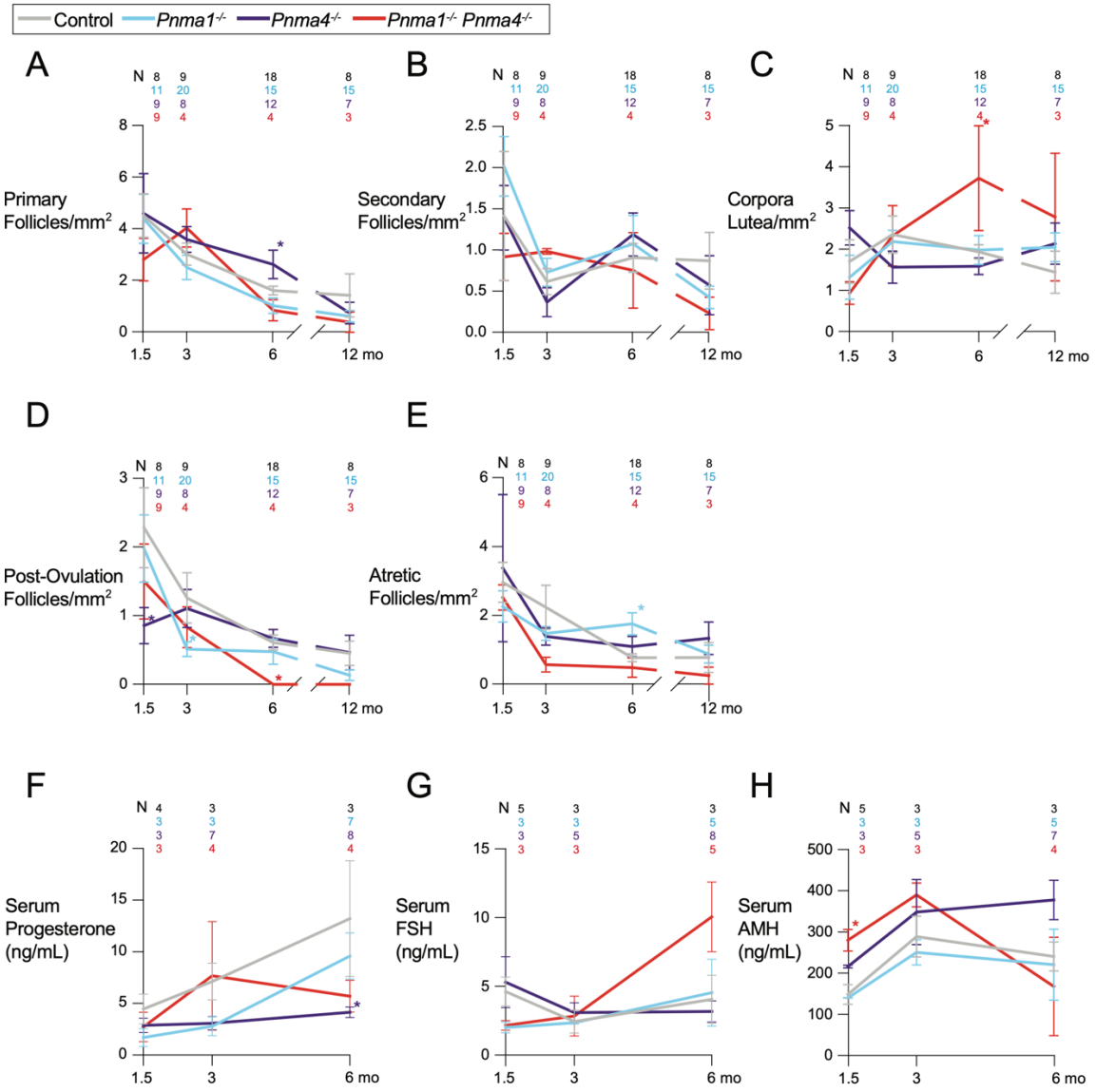
experiment used to generate the data in Figures 2A and 3A, but here average pup count per litter is plotted. Statistical significance (*$p < 0.05$) was determined by student's t-test. Error bars indicate SEM



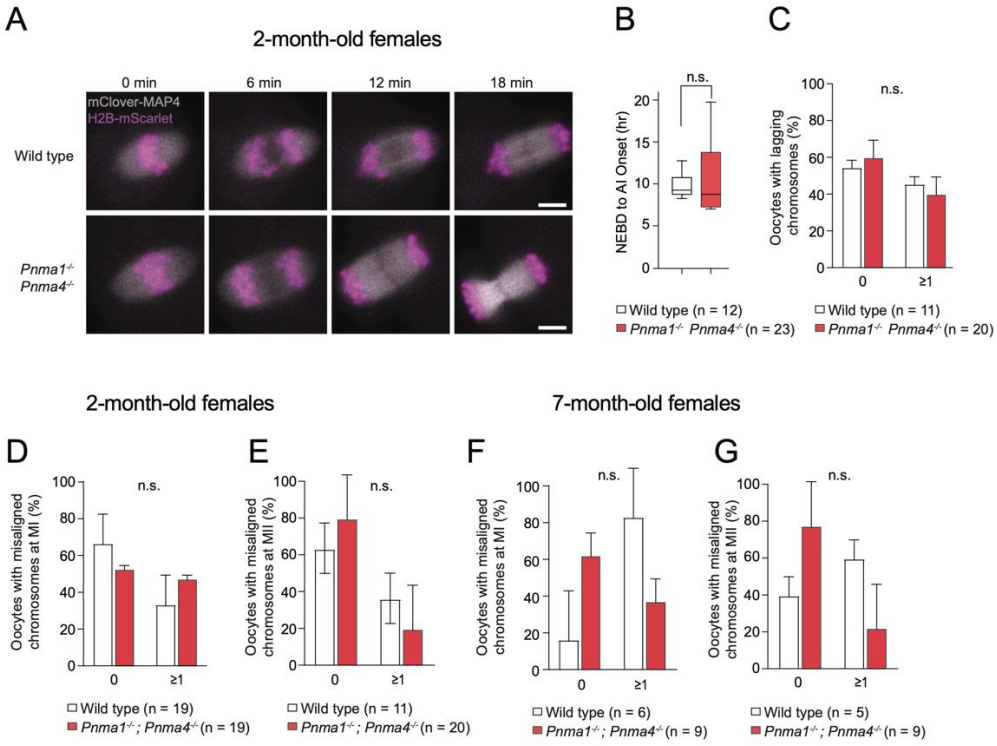

Supplementary Figure 5.4. Defective testicular and ovarian morphology in Pnma1 and Pnma4 mutants. (A) Representative PAS-stained testis sections by age. Quantified in Figure 2. (B) Representative PAS-stained ovary sections by age. Quantified in Figure 3 and S4.



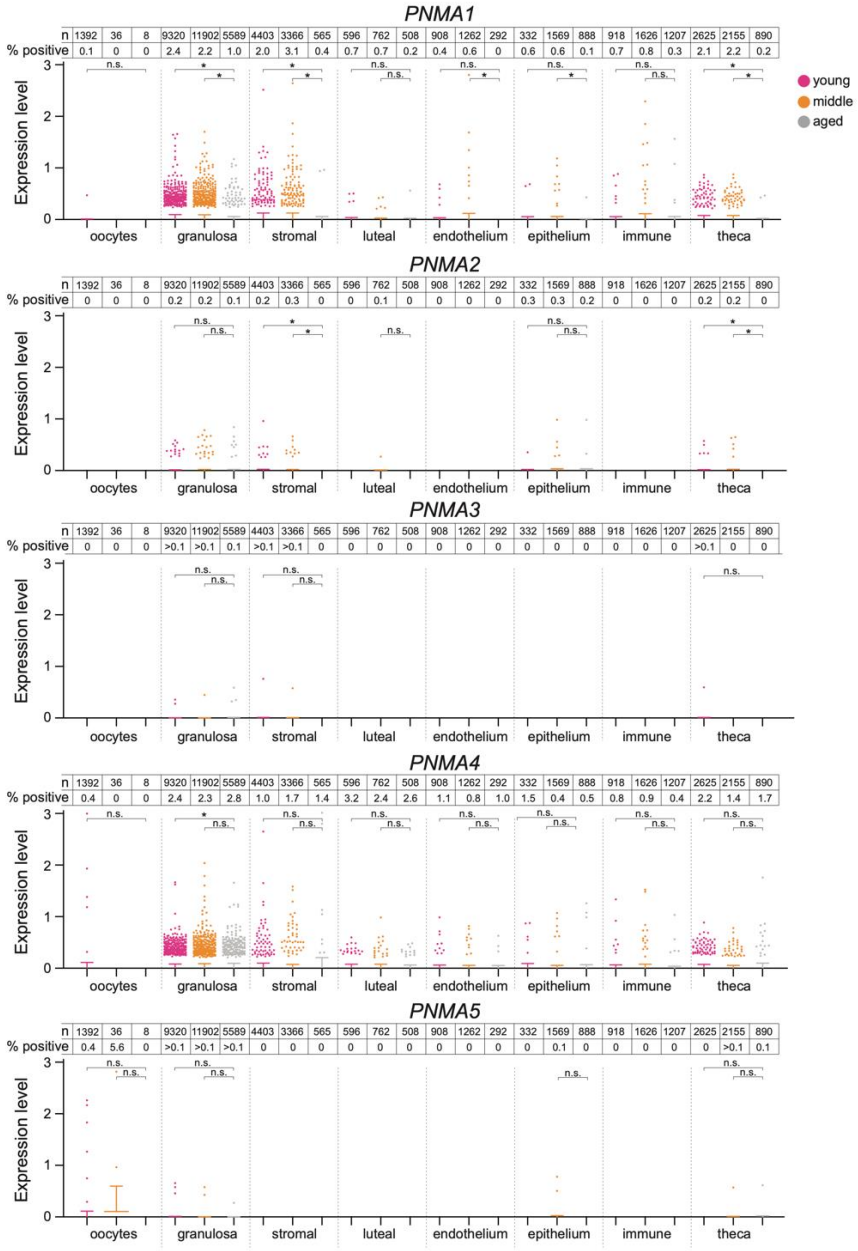

Supplementary Figure 5.5. Analysis of undifferentiated spermatogonial cells by PLZF staining. (A, B) Testes were fixed, paraffin embedded, and sectioned. Sections from 12-month samples were analyzed by TUNEL and DAPI staining. (A) Images and (B) quantifications (from n wild type-like tubules) of PLZF-positive cells (red arrows) are shown. Statistical significance (*p < 0.05) was determined by Mann-Whitney test. Error bars indicate SD.



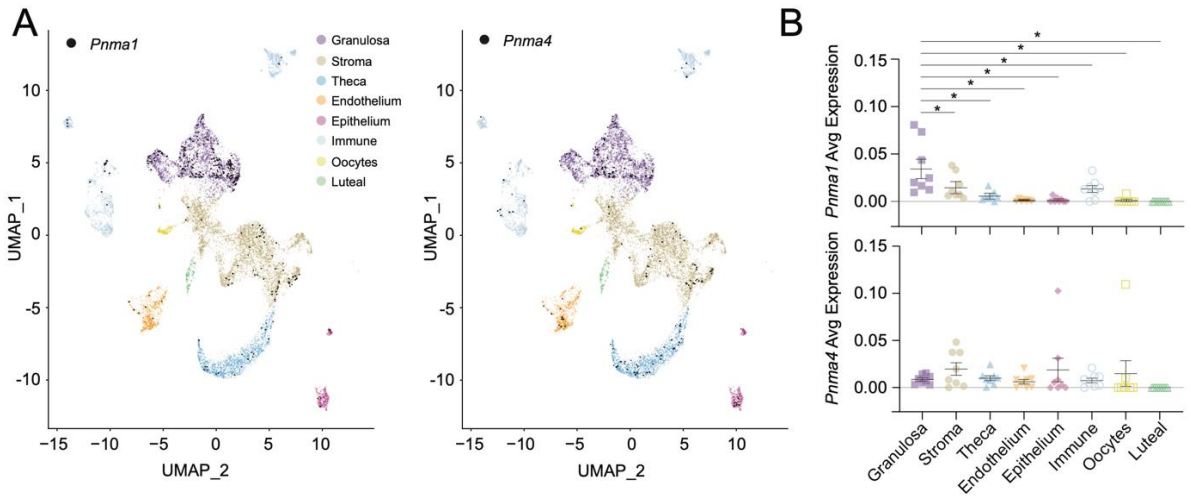

Supplementary Figure 5.6. Several ovarian features are not dramatically altered in Pnma1 and Pnma4 mutants. (A-E) Ovaries from control (pooled wild type and heterozygote, gray), Pnma1-/- (cyan), Pnma4-/- (purple), or Pnma1-/-; Pnma4-/- double mutant (red) mice were fixed, embedded, and PAS-stained. The following features were quantified per unit area (mm2) at the indicated times: (A) primary follicles, (B) secondary follicles, (C) corpora lutea, (D) post-ovulation follicles, and (E) atretic follicles. (F-H) Serum levels of (F) progesterone, (G) FSH, and (H) AMH from females was measured at the indicated ages. Statistical significance (*p < 0.05) was determined by determined by one-way ANOVA with correction for multiple comparisons. Error bars indicate SEM.



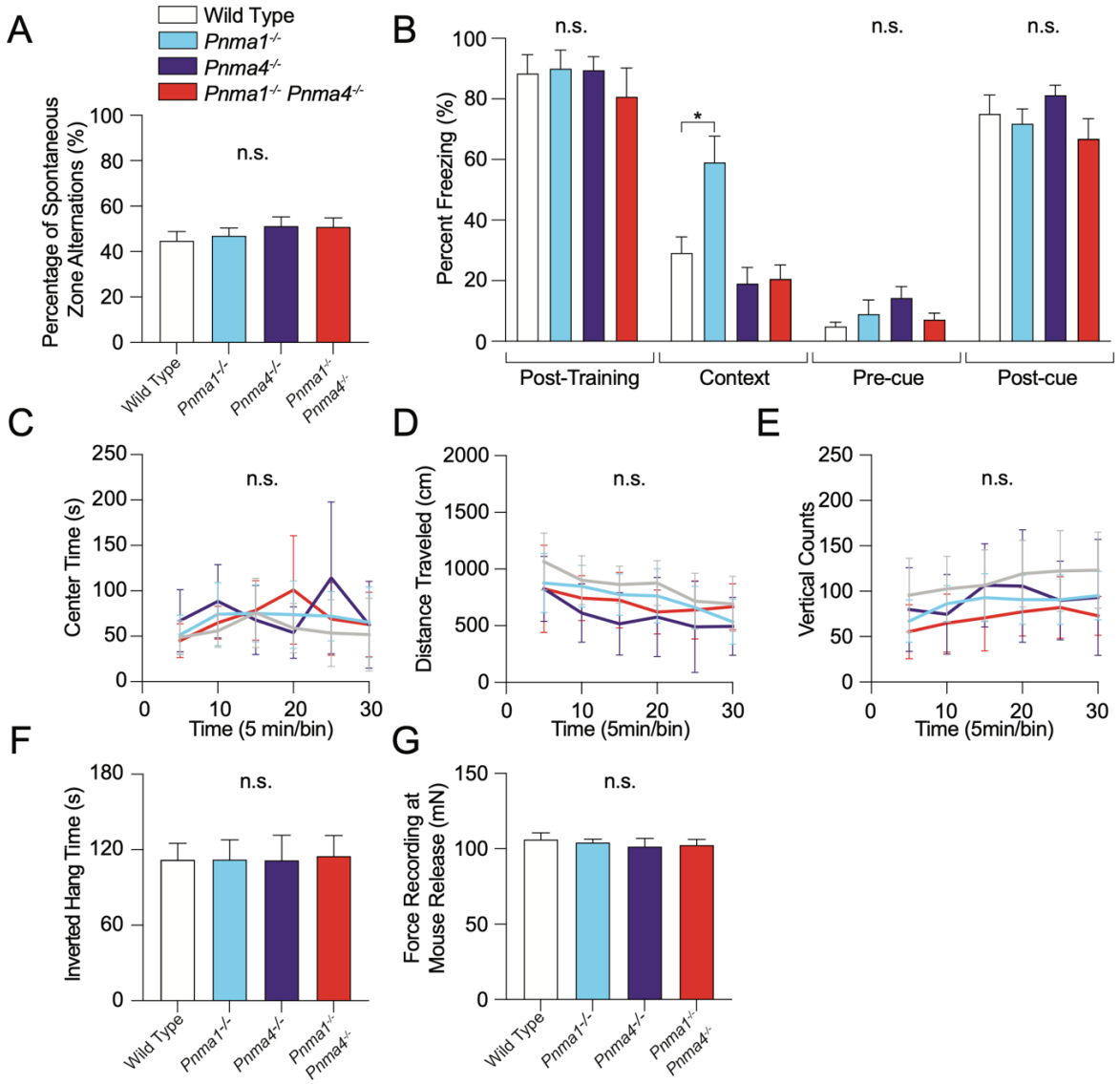

Supplementary Figure 5.7. Oocytes that enter meiosis are largely unaffected in Pnma1 and Pnma4 mutants. GV oocytes were collected from ovaries of two-month-old wild-type and Pnma1-/-; Pnma4-/- mice. We then injected mRNA encoding mClover-MAP4 (microtubule-binding protein) and H2B-mScarlet (histone) and imaged oocytes live for ~18 hours. (A) Representative images of the first meiotic division from the metaphase I (MI) plate (t = 0 min) to anaphase I (AI, t = 20 min). Misaligned chromosomes were counted at t = 0 min and lagging chromosomes at t = 12 min. (B) Time in hours (hr) between nuclear envelop breakdown to AI onset. (C) Percent of oocytes with no lagging chromosomes or more than one lagging chromosome during AI. (D) Percent of oocytes with no misaligned chromosomes or more than one misaligned chromosome in MI plate. (E) Percent of oocytes with no misaligned chromosomes or more than one misaligned chromosome at Metaphase II (MII) plate, 4 hours after meiosis I completion. (F, G) Continued data from Figure 3. As in A-E, but GV oocytes were collected at 7 months. (F) Percent of oocytes with no misaligned chromosomes or more than one misaligned chromosome in MI plate. (G) Percent of oocytes with no misaligned chromosomes or more than one misaligned chromosome at MII plate, 4 hours after meiosis I completion.



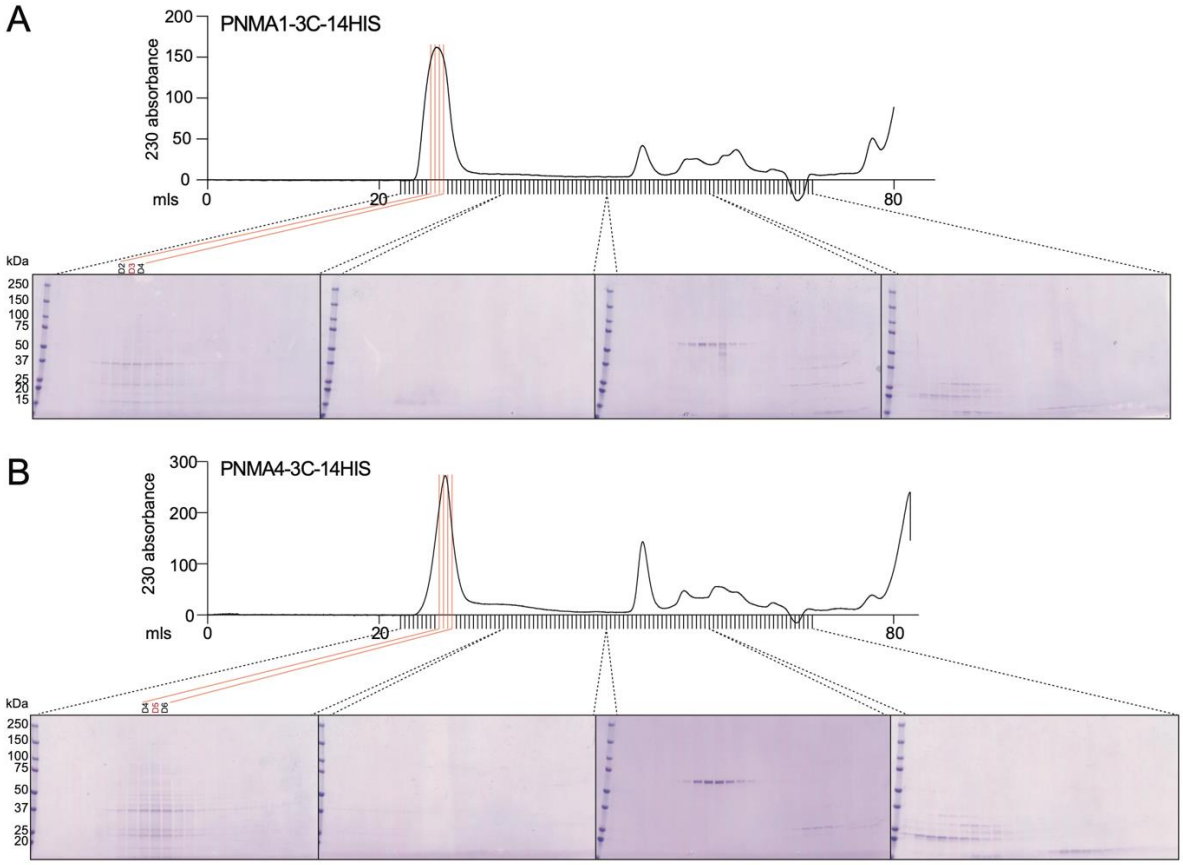

Supplementary Figure 5.8. Single-cell analysis of Pnma1 and Pnma4 expression in aging mouse ovaries. Ovaries from young (4.5 month, magenta, N = 5), peri-estropause (10.5 month, orange, N = 7), and post-estropause (15.5 month, gray, N = 5) wild type C57BL/6J mice were analyzed by single-cell RNAseq. Uniquely mapping reads for Pnma1-5 loci were assigned to ovarian tissue types based on clustering analysis. Note that the majority of values = 0 and lie underneath the x axis (error bars represent SEM). Total n of each cell type and percentage of positive cells is indicated above the plots. Statistical significance (*$p < 0.05$) was determined by Mann-Whitney test with Welch's correction. Comparisons were not run between conditions with no positive cells.



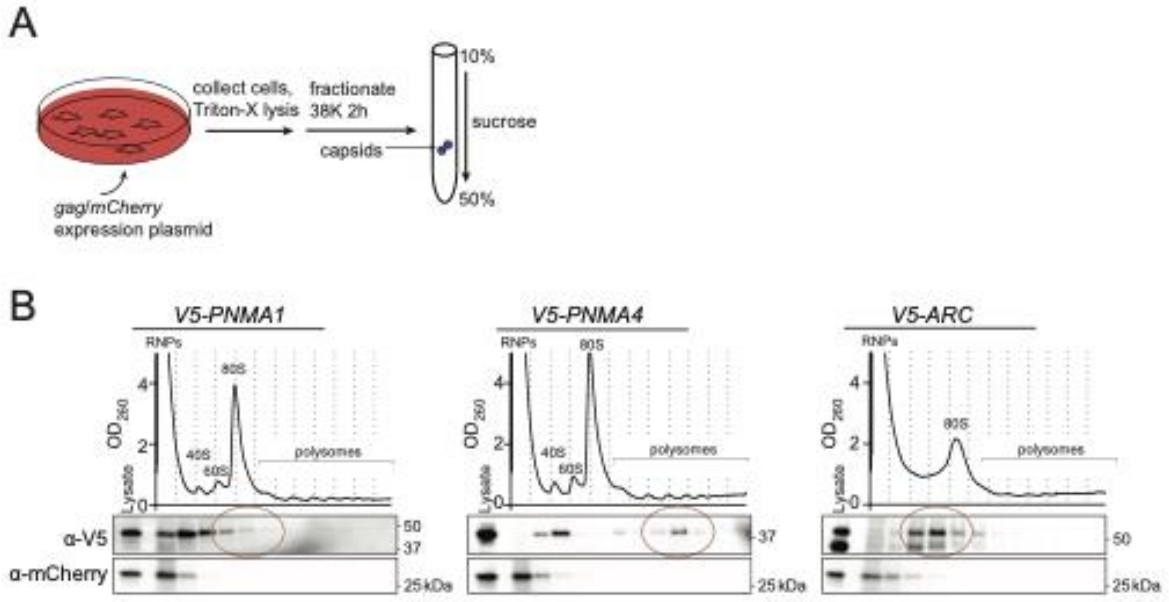

Supplementary Figure 5.9. Pnma1 and Pnma4 are expressed in subpopulations of ovarian cells. (A, B) single cell RNA-seq was performed in n = 8 ovaries from adult C57BL/6J mice. (A) UMAP plot showing distinct ovarian cell populations represented by different colors with Pnma1 and Pnma4-positive cells denoted in black. (B) Average expression of Pnma1 and Pnma4 by sample across different ovarian cell populations.




Supplementary Figure 5.10. Mice lacking Pnma1 or Pnma4 exhibit normal neurobehavioral and muscular traits. Six-month-old male wild type (N=11), Pnma1-/- (N=11), Pnma4-/- (N=8), and Pnma1-/-; Pnma4-/- double mutant mice (N = 11) were tested in a panel of behavioral assays. (A) Short-term memory was assessed using the spontaneous alternation percentage in a Y maze. (B) Long-term memory was assessed with the fear conditioning test. (C-E) The open field test was used to assess indicators of anxiety (center time) and hyperactivity (D, distance travelled and E, vertical counts). (F) We assessed muscle strength by measuring inverted hang time on a grid and (G) grip strength force. All graphs indicate mean +/- SEM. Statistical significance (*$p < 0.05$) was assessed by multiple t-test analysis with correction for multiple comparisons or one-way ANOVA with correction for multiple comparisons.




Supplementary Figure 5.11. Size exclusion chromatography of recombinant PNMA1 and PNMA4. (A, B) His-tagged PNMA1 (A) and PNMA4 (B) were expressed in E. coli and affinity purified on nickel resin. Imidazole elutions (15 ml) were concentrated and run over a Superose 6 (HiScale 80 ml) size exclusion column. Shown are UV absorbance traces of the run and Coomassie-stained SDS-PAGE of the fractions. The fractions used for TEM are noted in red.




Supplementary Figure 5.12. PNMA1 and PNMA4 expressed in human cells forms capsid-sized particles. (A, B) Experimental setup (A): *3V5*-tagged *PNMA1*, *PNMA4*, and *ARC* (control exported VLP) expression plasmids were transfected into HEK-293T cells. *mCherry* (control non-capsid) was co-expressed from the transfected plasmid. Cells were collected and lysed in Triton-X. Lysates were fractionated on 10-50% sucrose density gradients with continuous monitoring at 260 nm. PNMA1, PNMA4, ARC (anti-V5), and mCherry protein levels in each fraction were analyzed by immunoblot.

## Supplementary Tables

| Gene | Study ID | Reported Trait | Author | Date | Study N Initial | Index Variant ID | Index Variant RSID | P-Value | Beta | Beta CI Lower | Beta CI Upper | L2G Pipeline Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PNMA1 | GCST90012113 | Total testosterone levels | Ruth KS | 2/10/20 | 194453 | 14_73737983_G_C | rs72721770 | 1.00E-09 | -0.0191253 | -0.0253889 | -0.0128617 | 0.10270465 |
| PNMA1 | GCST90012108 | Sex hormone-binding globulin levels adjusted for BMI | Ruth KS | 2/10/20 | 180094 | 14_73737983_G_C | rs72721770 | 1.20E-10 | -0.0077683 | -0.0101481 | -0.0053886 | 0.07824419 |
| PNMA1 | GCST90012109 | Sex hormone-binding globulin levels | Ruth KS | 2/10/20 | 180726 | 14_73752265_T_C | rs2878733 | 6.00E-10 | -0.0084463 | -0.0111214 | -0.0057712 | 0.05756707 |
| PNMA1 | GCST90012107 | Sex hormone-binding globulin levels | Ruth KS | 2/10/20 | 189473 | 14_73783423_T_C | rs13379043 | 9.30E-11 | 0.0100214 | 0.0068 5053 | 0.0131 9227 | 0.0104 1393 |
| PNMA1 | GCST90019518 | Sex hormone-binding globulin levels | Sinnott-Armstrong N | 1/18/21 | 322484 | 14_73783107_T_C | rs11626364 | 1.00E-11 | 0.0237 | 0.0168 7535 | 0.0305 2465 | 0.0099 114 |
| PNMA1 | GCST90025958 | Sex hormone-binding globulin levels | Barton AR | 7/5/21 | 397043 | 14_73783107_T_C | rs11626364 | 2.70E-22 | 0.0255671 | 0.0200 9208 | 0.0310 4212 | 0.0090 1901 |
| PNMA1 | GCST90012106 | Sex hormone-binding globulin levels adjusted for BMI | Ruth KS | 2/10/20 | 188908 | 14_73783423_T_C | rs13379043 | 2.20E-16 | 0.0112783 | 0.0085 1833 | 0.0140 3827 | 0.0089 8683 |
| PNMA1 | GCST90012110 | Sex hormone-binding globulin levels adjusted for BMI | Ruth KS | 2/10/20 | 368929 | 14_73783423_T_C | rs13379043 | 2.60E-28 | 0.0100981 | 0.0082 2906 | 0.0119 6714 | 0.0089 8683 |
| PNMA1 | GCST90012111 | Sex hormone-binding globulin levels | Ruth KS | 2/10/20 | 370125 | 14_73783423_T_C | rs13379043 | 2.30E-21 | 0.0095 3255 | 0.0074 5146 | 0.0116 1364 | 0.0089 8683 |
| PNMA4 | GCST007078 | Menarche (age at onset) | Kichaev G | 12/27/18 | 242000 | 14_93452539_C_T | rs8007087 | 1.00E-12 | | | | 0.0122 3284 |

| Gene | Study ID | Reported Trait | Author | Date | Study N Initial | Index Variant ID | Index Variant RSID | P-Value | Beta | Beta CI Lower | Beta CI Upper | L2G Pipeline Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PNMA4 | GCST90012088 | Male puberty timing (age at voice breaking MTAG) | Hollis B | 3/24/20 | 205354 | 14_93434263_T_C | rs4900184 | 1.00E-10 | -0.021245 | -0.027684 | -0.014806 | 0.01206023 |
| PNMA4 | GCST90020232 | Body fat percentage | Martin S | 5/12/21 | 442278 | 14_93557626_G_A | rs6575340 | 9.00E-18 | 0.013 | 0.0100324 | 0.0159676 | 0.0191622 |
| PNMA4 | NEALE2_23127_raw | Trunk fat percentage | UKB Neale v2 | 8/1/18 | 354619 | 14_93557626_G_A | rs6575340 | 1.68E-10 | 0.113665 | 0.0787919 | 0.1485381 | 0.01886099 |
| PNMA4 | NEALE2_23112_raw | Leg fat mass (right) | UKB Neale v2 | 8/1/18 | 354807 | 14_93557626_G_A | rs6575340 | 3.26E-13 | 0.0281008 | 0.0205 3885 | 0.0356 6275 | 0.0184 1251 |
| PNMA4 | NEALE2_23116_raw | Leg fat mass (left) | UKB Neale v2 | 8/1/18 | 354788 | 14_93557626_G_A | rs6575340 | 1.12E-13 | 0.0279604 | 0.0205 8047 | 0.0353 4033 | 0.0182 9385 |
| PNMA4 | NEALE2_23128_raw | Trunk fat mass | UKB Neale v2 | 8/1/18 | 354597 | 14_93557626_G_A | rs6575340 | 2.25E-12 | 0.0891092 | 0.0642 2269 | 0.1139 9571 | 0.0173 1187 |
| PNMA4 | NEALE2_23100_raw | Whole body fat mass | UKB Neale v2 | 8/1/18 | 354244 | 14_93557626_G_A | rs6575340 | 2.12E-13 | 0.167209 | 0.1225 6824 | 0.2118 4976 | 0.0172 788 |
| PNMA4 | NEALE2_23099_raw | Body fat percentage | UKB Neale v2 | 8/1/18 | 354628 | 14_93557626_G_A | rs6575340 | 1.49E-11 | 0.105113 | 0.0745 8757 | 0.1356 3843 | 0.0171 3449 |
| PNMA4 | NEALE2_23119_raw | Arm fat percentage (right) | UKB Neale v2 | 8/1/18 | 354760 | 14_93557626_G_A | rs6575340 | 2.18E-12 | 0.129137 | 0.0930 9476 | 0.1651 7924 | 0.0170 2951 |
| PNMA4 | NEALE2_23123_raw | Arm fat percentage (left) | UKB Neale v2 | 8/1/18 | 354707 | 14_93557626_G_A | rs6575340 | 1.45E-12 | 0.131346 | 0.0949 8349 | 0.1677 0851 | 0.0169 737 |
| PNMA4 | NEALE2_48_raw | Waist circumference | UKB Neale v2 | 8/1/18 | 360564 | 14_93557626_G_A | rs6575340 | 3.41E-12 | 0.202993 | 0.1458 2744 | 0.2601 5856 | 0.0168 8215 |

| Gene | Study ID | Reported Trait | Author | Date | Study N Initial | Index Variant ID | Index Variant RSID | P-Value | Beta | Beta CI Lower | Beta CI Upper | L2G Pipeline Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PNMA4 | NEALE2_23115_raw | Leg fat percentage (left) | UKB Neale v2 | 8/1/18 | 354791 | 14_93557626_G_A | rs6575340 | 1.92E-11 | 0.0864554 | 0.0612 0864 | 0.1117 0216 | 0.0166 6563 |
| PNMA4 | GCST007064 | Body fat percentage | Hubel C | 12/28/18 | 155961 | 14_93654366_C_T | rs55882426 | 1.00E-08 | | | | 0.0104 3058 |
| PNMA4 | NEALE2_23124_raw | Arm fat mass (left) | UKB Neale v2 | 8/1/18 | 354673 | 14_93557626_G_A | rs6575340 | 8.44E-13 | 0.0121223 | 0.0088 0118 | 0.0154 4342 | 0.0199 7949 |

Supplementary Table 5.1. Traits associated with PNMA1 and PNMA4 variants in the human population.

# CHAPTER SIX

# A VIRALLY-ENCODED TRNA NEUTRALIZES THE PARIS ANTIVIRAL DEFENSE SYSTEM

## Contribution of Authors and Co-Authors

Manuscript in Chapter 6

Author: Nathaniel Burman

Contributions: Biochemistry and structural data in Figure 1, Figure 2, Figure 3c, manuscript review and editing

Co-Author: Svetlana Belukhina

Contributions: Plaque assays in Figure 2, biochemistry data in Figure 3a-b,e-f,h-I, Efficiency of plating assay in Figure 4a, *in vitro* translation in Figure 4d, Plaque assays and efficiency of plating assays and other data generation in Figure 5a-f, review and editing

Co-Author: Florence Depardieu

Contributions: Plaque assays in Figure 2, Figure 3d,j, Efficiency of plating assay Figure 4a, Plaque assays and efficiency of plating assays and other data generation in Figure 5a-k, review and editing

Co-Author: Royce A. Wilkinson

Contributions: Biochemistry and structural data in Figure 1, Figure 2, Figure 3c, manuscript review and editing

Co-Author: Mikhail Skutel

Contributions: Biochemistry data in Figure 3a-b,e-f,h-I, *in vitro* translation in Figure 4d, review and editing

Co-Author: Andrew Santiago-Frangos

Contributions: Biochemistry and structural data in Figure 3c, review and editing

Co-Author: Ava B. Graham

Contributions: Biochemistry and structural data in Figure 1, Figure 2, Figure 3c, review and editing

Co-Author: Alexei Livenskyi

Contributions: *In vitro* translation in Figure 4d, review and editing

Co-Author: Anna Chechenina

Contributions: Biochemistry data in Figure 3a-b,e-f,h-i, review and editing

Co-Author: Natalia Morozova

Contributions: Light microscopy in Figure 4b, review and editing

Co-Author: Trevor Zahl

Contributions: Biochemistry in Figure 3g, review and editing

Co-Author: William S. Henriques

Contributions: Biochemistry and structural data in Figure 1, Figure 2, Figure 3c, manuscript review and editing

Co-Author: Murat Buyukyoruk

Contributions: Evolutionary analysis in Figure 6, review and editing

Co-Author: Christophe Rouillon

Contributions: Plaque assays and efficiency of plating assays in Figure 5a-f, tRNA cleavage assay in Figure 5i, review and editing

Co-Author: Baptiste Saudemont

Contributions: Small RNA extractions (Figure 5i) and Sanger sequencing of the cleavage product (ED Figure 11), review and editing

Co-Author: Lena Shyrokova

Contributions: Metabolic labelling in Figure 4c, review and editing

Co-Author: Tatsuaki Kurata

Contributions: Metabolic labelling in Figure 4c, review and editing

Co-Author: Vasili Hauryliuk

Contributions: Metabolic labelling in Figure 4c, review and editing

Co-Author: Konstanin Severinov

Contributions: Review and editing

Co-Author: Justine Groseille

Contributions: Hi-C analysis in ED Figure 9, review and editing

Co-Author: Agnès Thierry

Contributions: Hi-C analysis in ED Figure 9, review and editing

Co-Author: Romain Koszul

Contributions: Hi-C analysis in ED Figure 9, review and editing

Co-Author: Florian Tesson

Contributions: Evolutionary analysis in Figure 6

Co-Author: Aude Bernheim

Contributions: Evolutionary analysis in Figure 6

Co-Author: David Bikard

Contributions: Project conceptualization, administration, original draft, review and editing

Co-Author: Blake Wiedenheft

Contributions: Project conceptualization, administration, original draft, review and editing

Co-Author: Artem Isaev

Contributions: Biochemistry data in Figure 3a-b,e-f,h-i, Plaque assays and efficiency of plating assays and other data generation in Figure 5a-f project conceptualization, administration, original draft, review and editing

Manuscript Information

Nathaniel Burman, Svetlana Belukhina, Florence Depardieu, Royce A. Wilkinson, Mikhail Skutel, Andrew Santiago-Frangos, Ava B. Graham, Alexei Livenskyi, Anna Chechenina, Natalia Morozova, Trevor Zahl, William S. Henriques, Murat Buyukyoruk, Christophe Rouillon, Baptiste Saudemont, Lena Shyrokova, Tatsuaki Kurata, Vasili Hauryliuk, Konstantin Severinov, Justine Groseille, Agnès Thierry, Romain Koszul, Florian Tesson, Aude Bernheim, David Bikard, Blake Wiedenheft, Artem Isaev

*Nature*

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

Springer Nature
Accepted 24 July 2024
Published Online 07 August 2024
Published in print 10 October 2024
Volume 634
https://doi.org/10.1038/s41586-024-07874-3

## Abstract

Viruses compete with each other for limited cellular resources, and some deliver defense mechanisms that protect the host from competing genetic parasites[1]. PARIS is a defense system, often encoded in viral genomes, that is composed of a 55 kDa ABC ATPase (AriA) and a 35 kDa TOPRIM nuclease (AriB)[2]. However, the mechanism by which AriA and AriB function in phage defense is unknown. Here we show that AriA and AriB assemble into a 425 kDa supramolecular immune complex. We use cryo-EM to determine the structure of this complex which explains how six molecules of AriA assemble into a propeller-shaped scaffold that coordinates three subunits of AriB. ATP-dependent detection of foreign proteins triggers the release of AriB, which assembles into a homodimeric nuclease that blocks infection by cleaving host $tRNA^{Lys}$. Phage T5 subverts PARIS immunity through expression of a $tRNA^{Lys}$ variant that is not cleaved by PARIS, and thereby restores viral infection. Collectively, these data explain how AriA functions as an ATP-dependent sensor that detects viral proteins and activates the AriB toxin. PARIS is one of an emerging set of immune systems that form macromolecular complexes for the recognition of foreign proteins, rather than foreign nucleic acids[3].

## Introduction

Anti-viral defense systems in bacteria and archaea are extraordinarily diverse and many are mechanistically similar to immune responses in eukaryotic cells[4]. The recent expansion of bacterial and archaeal immune systems stems from the appreciation that they tend to co-localize in the genome[5] and are often carried by mobile genetic elements, including prophages and satellite viruses[6]. The Phage Anti-Restriction Induced System (PARIS) is one of several defense

systems recently identified within a hotspot of genetic diversity carried by P4 phage satellites[2]. The PARIS system of *Escherichia coli* strain B185 protects against phage T7 infection through a mechanism triggered by sensing the T7 Ocr (Overcoming classical restriction) protein. Ocr is a DNA mimic that inhibits Type I RestrictionModification and BREX defense systems. Thus, PARIS is an 'anti-anti-restriction' system that senses a viral immune suppressor.

The PARIS defense system consists of two proteins, AriA and AriB (Fig. 1a-b). AriA contains a conserved ABC ATPase domain, while AriB is a domain of unknown function (DUF4435), which includes a TOPRIM nuclease[7,8]. AriA and AriB are usually separate proteins, though they are occasionally fused into a single polypeptide. Several defense systems share ABC ATPase and TOPRIM domains, including the OLD (Overcoming Lysogenization Defect) protein produced by the P2 prophage, which blocks infection by phage lambda[9]. OLD does not directly sense a phage protein, but rather, the inactivation of the RecBCD exonuclease[10]. Activation of OLD results in translation inhibition[9], but the underlying mechanism has not been determined.

Association between ABC ATPase and TOPRIM domains is an emerging theme in recently discovered immune systems[11]. Examples include the Gabija system[12,13], effector toxin proteins of retron immunity[14], and *bona fide* toxin-antitoxin (TA) modules[15]. This implies a broad adaptability of the conserved ABC ATPase and TOPRIM nuclease architecture to a variety of antigenic stimuli, resulting in the diversification of the immune modules likely acting through abortive mechanisms. We currently have little understanding of the molecular mechanisms underpinning how these systems sense and respond to viral infections.

Here we use cryo-EM to show that the *E. coli* B185 PARIS system components assemble into a 425 kDa propeller-shaped complex with six molecules of AriA and three subunits of AriB.

We show that AriA senses the T7 Ocr anti-restriction protein leading to the release of the AriB nuclease effector. Release of AriB from the AriA scaffold is necessary to potentiate AriB, which forms a nuclease-active dimer that degrades host lysine tRNA ($tRNA^{Lys}$), thereby blocking translation and leading to growth arrest or cell death. We find that phage T5 carries its own $tRNA^{Lys}$ isoacceptor, which rescues the phage from PARIS-mediated defense. Finally, we perform a phylogenetic analysis of PARIS systems, which provides insights into their evolutionary history and shows how they relate to other defense systems that employ ABC ATPases.

## Results

### PARIS Forms a Propeller-shaped Complex

The PARIS defense system consists of two proteins, AriA and AriB, which protect cells from phages through a previously undetermined mechanism[2]. Here, we focused on the system from the P4 satellite integrated into the *E. coli* (str. B185) genome (**Figure 6.1a-b**). Based on published experimental structures of related proteins[8,13] and AlphaFold2[16] predictions (**Figure. 6.1c, Extended Data Figure 6.1a-d**), we anticipated that AriA:AriB would assemble into a multi-subunit complex. To determine how AriA and AriB assemble, we overexpressed AriA with a C-terminally tagged AriB (AriB-Strep), pulled down on AriB, and purified the complex using size exclusion chromatography (**Figure 6.1d**). The AriA and AriB proteins co-elute in a single peak with an estimated molecular weight of 430 kDa. Next, we used cryo-EM to determine structures of the PARIS complex (**Figure 6.1e**). The structure explains how six molecules of AriA come together to form a D3 symmetric scaffold decorated by three subunits of AriB. The helical domains, which typically provide a platform for DNA binding in related ABC ATPases[17],

instead function as additional dimerization interfaces that enable a trimeric assembly of AriA homodimers (**Figure 6.1c, Extended Data Figure 6.2**).

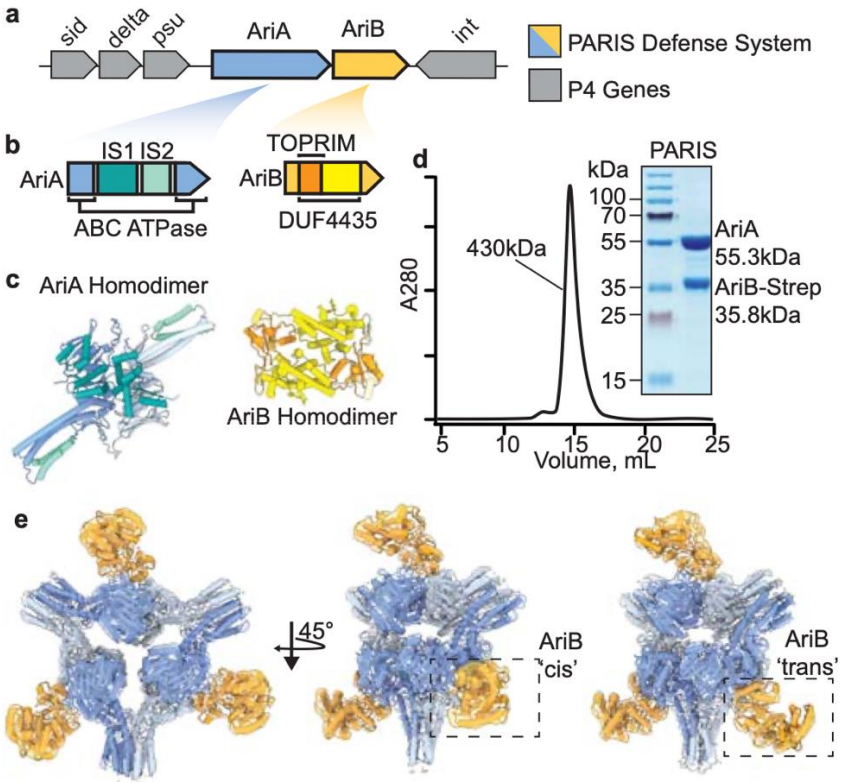



Figure 6.1. PARIS is a two-component system that assembles into a propeller-shaped supramolecular complex. a, Schematic of the integrated P4 satellite in the E. coli genome. b, PARIS genes colored according to domain organization. AriA is an ABC ATPase, while AriB a DUF4435 that includes a TOPRIM nuclease. Insertion sequence 1 (IS1) is composed of three helices flanked by the ABC ATPase domain, while Insertion sequence 2 (IS2) adds two additional helices to the coiled-coil resulting in a helical bundle. c, AlphaFold2-predicted structures of AriA and AriB as homodimers colored according to structural features shown in panel b. d, Size exclusion chromatography (SEC) reveals that AriA and AriB copurify as an ~430 kDa complex. The SEC profile is representative of three independent biological experiments. e, Experimentally determined density maps of the PARIS complex reveal a homohexameric assembly of AriA with D3 symmetry. The AriA hexamer is decorated by AriB subunits that attach to AriA homodimers in one of two possible orientations, a 'cis' arrangement where the three AriB subunits are facing the same direction, or a 'trans' arrangement where one AriB subunits is rotated by 180° relative to the other two.

Dimerization of the helical domains from adjacent AriA homodimers form three blades of the propeller (**Figure 6.1e, Extended Data Figure 6.2**). Additional density is evident between each of the AriA blades, where AriB attaches to AriA near the ATP binding sites. However, density in this region of the reconstruction reveals conformational heterogeneity, consistent with AriB attaching to either, but not both AriA molecules at the dimer interface (**Extended Data Figure 6.1e-f, Supplemental Video 1**). Using multiple rounds of focused 3-D classification, we resolved two distinct isomers of the complex (**Figure 6.1e, Extended Data Figure 6.3d**). In one reconstruction, AriB attaches to three symmetrically equivalent AriA molecules (3.8 Å-resolution, *cis* configuration), while in the other reconstruction, one of the three AriB molecules binds the opposing AriA, which flips AriB 180-degrees relative to the other two AriB subunits (4.0 Å-resolution, *trans* configuration) (**Extended Data Figure 6.3g-h, Supplementary Video 1**).

AriA Sequesters AriB

While the homohexameric assembly of AriA was apparent in the early stages of reconstruction, conformational heterogeneity limited the resolution (**Extended Data Figure 6.3, Supplementary Video 1**). To improve the resolution, we used local refinement to determine a 3.2 Å nominal resolution reconstruction of one asymmetric unit of the PARIS complex ($AriA_2$:$AriB_1$), into which we built an atomic model (PDB ID: 8UX9, **Extended Data Table 6.1**) (**Figure 6.2**). Each asymmetric unit of the PARIS complex is composed of two AriA molecules, which form a head to tail homodimer[18–20], and one AriB molecule that binds directly above the AriA dimer interface.

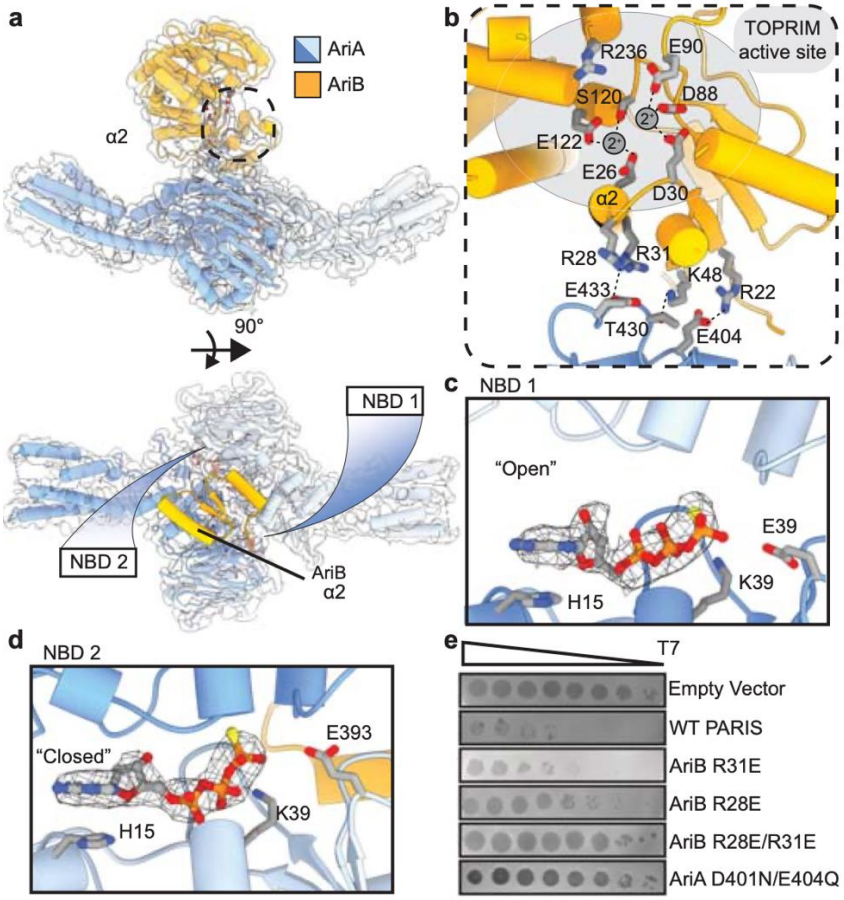



Figure 6.2. Interaction between the AriA ATPase and AriB essential for PARIS-mediated defense. a, Each asymmetric unit of the PARIS complex contains an AriA homodimer with two nucleotide binding domains (NBD). A 90-degree rotation of the asymmetric unit, showing 1-53 N-terminal residues (yellow) of AriB positioned over the ATPase active sites of AriA. b, AriB attaches to AriA through electrostatic interactions near helix α2 of AriB. TOPRIM active site residues predicted to participate in metal binding (gray spheres) and catalysis are shown as sticks. c, Close-up of NBD1 showing ATPγS bound in an "open" conformation. d, Close-up of NBD2 showing bound ligand trapped in a "closed" pre-cleavage state. e, Plaque assays were performed in duplicate with two independent clones of each of the mutant, using a serial dilution of phage T7 on E. coli MG1655 carrying the WT or mutated PARIS system.

AriB is a metal-dependent TOPRIM nuclease and the predicted effector of the PARIS defense system[2,7,8]. The structure reveals a series of spatially conserved, negatively charged residues (E26, D30, D88, E90, and E122) that are characteristic of TOPRIM domains that bind two metal ions[7]. The E26A mutation inactivates PARIS-mediated defense (**Figure 6.2b**)[2].

Glutamic acid 26 is just upstream of an N-terminal alpha helix (α2) on AriB, which includes two Arginine residues (R31 and R28) that form salt bridges with negatively charged residues on AriA that is positioned directly above the ATP binding sites (**Figure 6.2b, Extended Data Figure 6.2**, **Supplementary Video 1**). Mutations that interfere with the AriA:B interface (AriB R28E, AriB R31E, AriB R31E/R28E or AriA D401N/E404Q) limit phage protection by PARIS (**Figure 6.2e**), which suggest that the association between AriA and AriB is required to provide protection.

The AriA ABC ATPase domain resembles that of Rad50 (RMSD 1.1 Å across 86 C-alpha atom pairs), a universally conserved protein involved in dsDNA break repair[21–23]. Structural comparison of AriA to Rad50 confirms that AriA contains the sequence motifs required for ATPase activity, along with two unique insertion sequences, IS1 and IS2 (**Figure 6.1b, Extended Data Figure 6.2**). Insertion sequence 2 contains a series of aromatic residues located on one face of helix α8 (F230, F234, F238, and Y304) that coordinate interactions between AriA homodimers, and explains how AriA homodimers assemble into a trimer of dimers (**Extended Data Figure 6.2);** a rare assembly state for this family of proteins[21]. Like other ABC ATPases, the head to tail assembly of the AriA homodimer results in two symmetrically positioned nucleotide binding domains (NBD) at the dimer interface (**Figure 6.2a**). In the presence of a nonhydrolysable ATP analog (i.e., ATPγS), we observe density for the ligand in both nucleotide binding domains (NBDs), but in different coordination states (**Figure 6.2c-d**). Previous studies of ABC ATPases revealed that ATP hydrolysis occurs at the two NBDs in an alternating fashion[21,24], which is consistent with the observed “open” and “closed” conformations of NBDs (**Figure 6.2c-d, Supplementary Video 1**).

AriA Binds to Ocr and Releases AriB

The anti-restriction protein Ocr from phage T7 was previously identified as a trigger of the PARIS system, and its expression in the presence of PARIS leads to growth arrest or cell death[2]. To determine if AriA:B senses Ocr through a direct interaction, we co-expressed AriA with a non-toxic active site mutant of AriB (E26A) and Ocr-Strep. AriA, but not AriB, co-purified with Ocr-Strep and the complex migrates according to the predicted mass of an AriA hexamer (**Figure 6.3a, Extended Data Figure 6.5a-b**). This is consistent with a direct interaction between AriA and Ocr, and release of AriB. A pull-down with AriBStrep confirms that AriB dissociates from the complex in the presence of Ocr (**Figure 6.3b**).

To determine how Ocr triggers the release of AriB, we used the purified AriA:Ocr-Strep complex for structure determination. The 4.4 Å-resolution structure reveals a D3 symmetric, homohexamer of AriA with no AriB (**Figure 6.3c and Extended Data Figure 6.6**). At this resolution, the structure of AriA purified using Ocr-Strep is indistinguishable from the structure of AriA purified using AriB-Strep. We anticipated that the structure would include Ocr, since Ocr-Strep was used to pull down the AriA hexamer and the complex was stable during gel filtration. However, multiple data processing strategies failed to resolve Ocr, suggesting that Ocr has several binding sites on AriA and/or that the interaction is conformationally flexible.

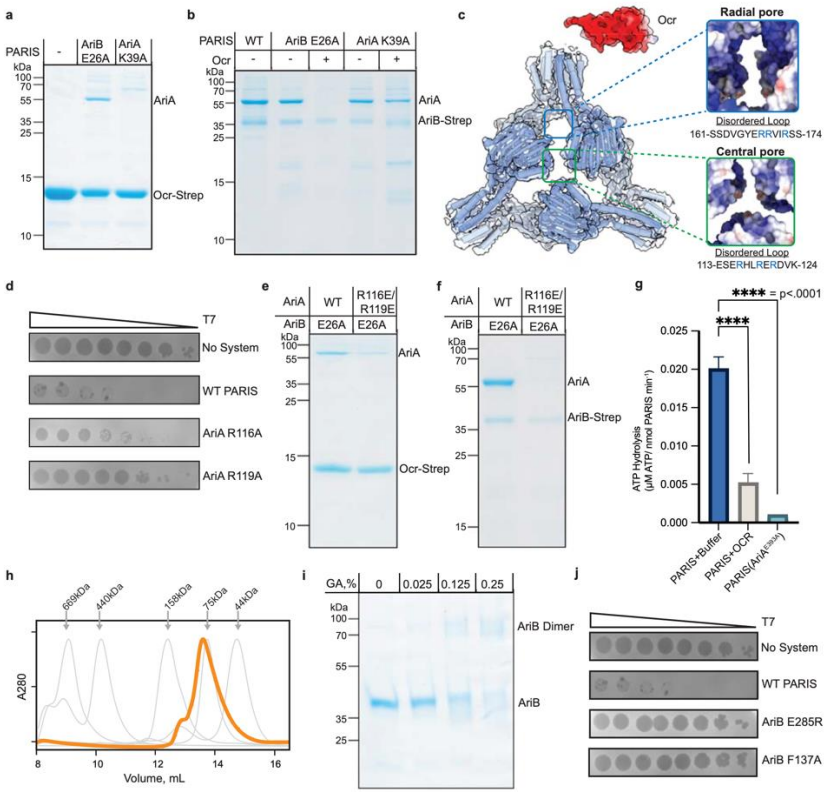



Figure 6.3. AriA interacts with Ocr in an ATPase-dependent manner, leading to the release of AriB from the PARIS complex. a, Ocr-Strep pulls down AriA, but not AriB. A mutation in the ATPase active site of AriA (K39A) blocks binding to Ocr-Strep. b, Ocr releases AriB (E26A) from wild-type (WT) AriA (lane 3), but Ocr does not trigger AriB release from an ATP binding deficient variant of AriA (K39A) (lane5). c, A 4.5 Å-resolution reconstruction of AriA that was purified using Ocr-Strep as a bait. The structure reveals three radial pores that are symmetrically positioned around a central pore. The pores contain disordered loops with several positively charged residues. Ocr (PDB: 1S7Z) is shown as an electrostatic surface. d, Mutations in the central pore (i.e., R116A or R119A) reduce the efficiency of phage protection. e, Ocr-Strep binds WT AriA and ejects AriB (E26A), while the charge swap mutations in the central pore of AriA (R116E/R119E) limit interaction with Ocr-Strep. f, AriB-Strep (E26A) pulls down WT AriA but not AriA (R116E/R119E). g, ATP and ADP were separated using thin layer chromatography (Extended Data Fig. 7) and the rates of ATP hydrolysis for PARIS, PARIS mixed with ten-fold excess Ocr, or PARIS with an ATPase mutation in AriA (E393A), were measured. The experiments were performed in triplicate and the error bars show +/- SEM. Two-sided statistical analysis performed using a post hoc Dunnett's test ****p<0.0001. h, Size exclusion chromatography of AriB-Strep (~36 kDa) after Ocr-mediated release from AriA. The column was calibrated using molecular weight standards (gray lines). AriB elutes in single peak with an estimated molecular weight of 81 kDa (Extended Data Fig. 5a). i, Glutaraldehyde (GA) cross-linking assay with activated AriB (E26A). j, Mutations predicted to block AriB dimerization (E285R and F137A) prevent PARIS-mediated defense (Extended Data Fig. 1g). Plaque assays were performed in duplicate with two independent clones of each mutant.

Since Ocr is a negatively charged DNA mimic[25,26], we hypothesized that the binding site on AriA would be positively charged. An electrostatic surface of the AriA hexamer reveals a central pore, flanked by a series of three positively charged radial pores (**Figure 6.3c**). The radial pores contain two flexible loops (one from each of the adjacent AriA homodimers) with 14 unmodelled residues (161-174, SSDVGYERRVIRSS), while the central pore contains six loops (one from each AriA) that contain 12 unmodelled residues (113-124, ESERHLRERDVK). To determine the role of these loops in phage defense, we tested the effects of alanine substitutions and charge swap mutations on positively charged residues. Arginine to alanine substitutions (i.e., R116A and R119A) in the central pore result in a partial loss of phage protection (**Figure 6.3d**). Attempts to generate other mutations in the central (R116E, R119E) or radial pores (R168E, R168A, R172E, or R172A) of AriA consistently failed when we used a vector that also included wild-type AriB. We hypothesized that these mutations in AriA mimic trigger binding, which releases AriB, and results in toxicity. To test this hypothesis, we expressed the central pore mutant (R116E/R119E), with the catalytically inactive AriB (E26A), which is non-toxic. Pull down experiments performed with Ocr-Strep (**Figure 6.3e**) or AriB-Strep (**Figure 6.3f**) demonstrate that AriA (R116E/R119E) has limited interactions with both Ocr and AriB. Collectively, these results suggest that either the central, radial, or both pores are involved in the interaction with the trigger.

Expression of AriB alone does not provide phage defense and is only mildly toxic when overexpressed by itself (**Extended Data Figure 6.4**)[2]. However, activated AriB (i.e., AriB released from PARIS) is significantly more toxic. Moreover, mutations in AriA that disrupt

interaction with AriB, limit AriB toxicity and abolish phage defense. Collectively, these data suggest that the toxicity of AriB depends on its initial association with AriA.

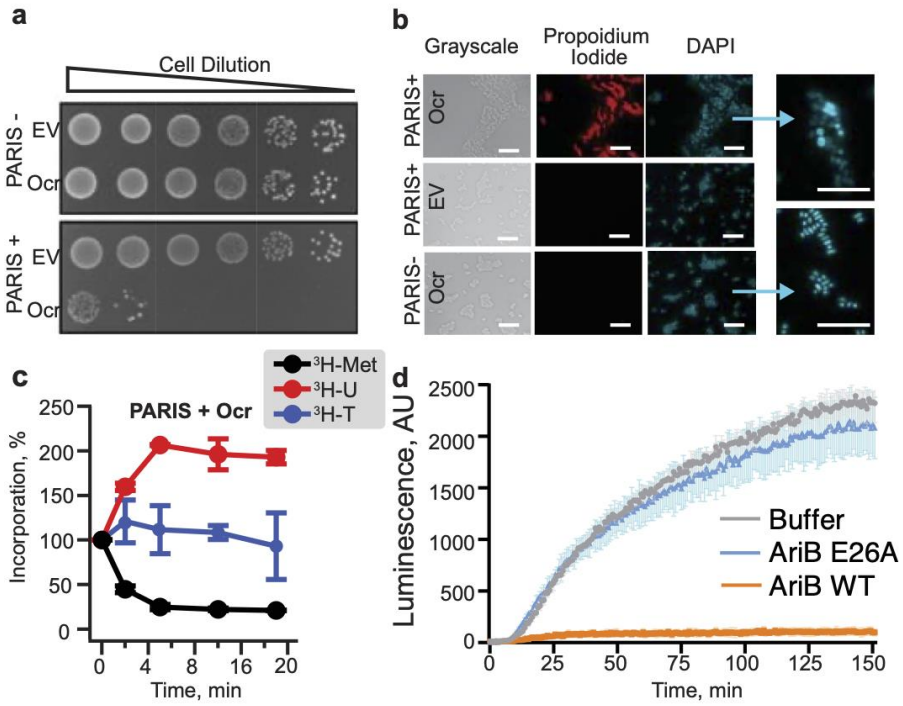



Figure 6.4. The activation of PARIS leads to cell death and inhibition of translation**.** a, Expression of Ocr from the pBAD vector in PARIS+ culture results in cell toxicity, while the empty vector (EV) has no impact. b, Ocr expression in PARIS+ cells induces cell death (propidium iodide staining, red) and DNA compactization (DAPI staining, blue). Arrows point to close-up images of the nucleoid structures from PARIS+ / Ocr (top) and PARIS- / Ocr (bottom). Cells were imaged 1 hour after induction of Ocr expression. All scale bars (white lines) represent 10 micrometers (μm). c, Metabolic labelling experiments reveal a drop in 3H methionine incorporation following induction of Ocr, indicative of translation inhibition upon PARIS activation. The mean and SD were calculated for each time point for three independent experiments. d, In vitro translation in the presence of activated AriB (WT) or active site mutant (E26A). Translation of firefly luciferase mRNA was monitored by luminescence. In the presence of activated AriB, luciferase signal is reduced dramatically, further indicating that PARIS functions through a mechanism of translational inhibition. The mean and SD were calculated for each time point for three independent experiments.

The ATPase activity of AriA is also required for phage defense[2], which suggest that the ATPase of AriA is necessary for loading AriB onto AriA, or release of activated AriB. To

differentiate between loading or release of AriB, we performed pull down assays using AriB-Strep and WT or an ATP binding defective mutant of AriA (K39A in Walker A). AriB-Strep associates with mutant and wild-type AriA with similar efficiency (**Figure 6.3b**), demonstrating that the ATPase mutant of AriA loads, but does not release AriB in the presence of Ocr. We reasoned that the lack of AriB release in the presence of AriA K39A is due to a defect in trigger recognition. To confirm the role of ATP in AriA-mediated trigger recognition, we conducted pull downs using Ocr-Strep and WT or K39A AriA. The AriA K39A mutant fails to copurify with Ocr, indicating trigger recognition is ATP-dependent (**Figure 6.3a**). To determine if or how the trigger impacts the ATPase activity of AriA, we measured the rate of PARIS-mediated ATP hydrolysis in the presence or absence of Ocr and compared these results to an ATPase defective AriA mutant (E393A in Walker B) (**Figure 6.3g**). The results demonstrated that Ocr significantly ($p=0.0001$) reduces ATP turnover by AriA (**Figure 6.3g and Extended Data Figure 6.7**).

TOPRIM-containing proteins are known to form homodimers[27,28], but a predicted homodimer of AriB generated using AlphaFold2 clashes with AriA (**Extended Data Figure 6.1f**). Thus, we hypothesized that AriB forms an active homodimer after trigger-mediated release from AriA. The toxicity of PARIS upon activation makes purification of active AriB challenging. To overcome this problem, we mixed lysates from cells expressing AriA:B-Strep with lysates from cells expressing Ocr, and then recovered activated AriB using affinity chromatography (**Extended Data Figure 6.5c**). Size exclusion chromatography indicated that activated AriB forms a homodimer (**Figure 6.3h, Extended Data Figure 6.5a**), which was further confirmed by glutaraldehyde cross-linking experiments (**Figure 6.3i**). Mutations in the AriB dimerization interface (E285R and F137A), predicted from the AlphaFold2 model (**Figure 6.1c, Extended**

**Data Figure 6.1g),** resulted in the loss of phage defense (**Figure 6.3j**), highlighting that AriB dimerization is essential for PARIS activity. Collectively, these data show that the Ocr trigger interacts with AriA in an ATP-dependent manner, leading to the release and dimerization of AriB.

Activated PARIS Inhibits Translation

We next investigated the consequences of PARIS activation by Ocr. The expression of Ocr from a plasmid in *E. coli* is sufficient to cause cell death in the presence of PARIS (**Figure 6.4a**). Monitoring of PARIS$^+$ cells under the microscope using live/dead staining with propidium iodide revealed the accumulation of cells with permeabilized membrane within 20 min of Ocr induction (**Figure 6.4b and Supplementary Video 2**). The TOPRIM domain, likely responsible for AriB toxicity, is frequently associated with DNase or RNase activities[29], but RNA or DNA preparations from AriB activated cells do not reveal evidence for indiscriminate nuclease activity (**Extended Data Figure 6.8a-c**). However, DAPI staining revealed that PARIS activation results in nucleoid compaction (**Figure 6.4b**), reminiscent of the translation inhibition phenotype[30]. To understand the nature of this compaction we performed chromosome conformation capture experiments (Hi-C). The results revealed a loss of chromosomal structures similar to that observed upon treatment with chloramphenicol, a known inhibitor of translation (**Extended Data Figure 6.9).**

To better understand the nature of PARIS-induced translational inhibition, we conducted metabolite labelling experiments and *in vitro* translation assays. Production of Ocr in PARIS$^+$ cells resulted in rapid inhibition of $^3$H methionine incorporation, consistent with translation inhibition (**Figure 6.4c**), and in a moderate increase of $^3$H thymidine uptake. This increase is a known consequence of translation inhibition that can be explained by an enhanced transcription

of rRNA genes[15]. The importance of activated AriB in translational arrest was further corroborated using a luciferase reporter in an *in vitro* translation assay. The presence of activated AriB in the reaction mixture interferes with the production of luciferase, while the active site mutant (i.e., AriB E26A) has no impact (**Figure 6.4d**). Collectively, the results demonstrate that PARIS activation results in translational inhibition.

Triggers and a Suppressor of PARIS

Comparative analysis of related phages has provided considerable insight into defense and antidefense strategies[31]. While testing the activity of PARIS against phages of the T5 family, we noticed distinct plaquing phenotypes among related phages (**Figure 6.5a and Extended Data Figure 6.10**). PARIS does not protect cells from $T5_{WT}$ phage infection, but efficiently blocks infection by a variant of T5 from the Moscow phage collection ($T5_{Mos}$). Consistent with translation arrest upon PARIS activation, infection of $PARIS^+$ cells with $T5_{Mos}$ at high MOI resulted in a rapid growth inhibition, which occurred faster than the time it takes to lyse $PARIS^-$ cells (**Extended Data Figure 6.10a**). Aside from a large deletion (*Δ30661-38625*), the $T5_{Mos}$ genome is nearly identical to the T5 archetype (**Figure 6.5b**). We hypothesized that both phages contain a trigger that activates PARIS, but the T5 archetype also contains a suppressor of PARIS that has been lost in $T5_{Mos}$. To determine how $T5_{Mos}$ triggers PARIS, we isolated and sequenced $T5_{Mos}$ mutants that escape PARIS immunity. All PARIS-escape mutants carried mutations in two genes encoding highly acidic proteins of unknown function: ORF094 and ORF103 (**Figure 6.5a-d**). Expression of these proteins from a plasmid resulted in activation of PARIS toxicity (**Figure 6.5e**), confirming that T5 carries two novel PARIS triggers. Mutated variants encoded by escape mutants did not trigger PARIS. We named these proteins PARIS trigger 1 and 2 (Ptr1 - ORF094;

Ptr2 - ORF 103)[32]. These results reveal that structurally diverse, negatively charged viral proteins can activate the PARIS system. The presence of *ptr1* and *ptr2* genes in the genome of $T5_{wt,}$ which is not affected by PARIS, further indicates that this phage contains a suppressor of PARIS immunity.

Viral TRNA Resistant to AriB Cleavage

The T5 genome encodes 23 tRNAs, 16 of which have been lost in $T5_{Mos}$. Given the inhibitory effect of activated AriB on translation, we hypothesized that phage-encoded tRNAs are responsible for overcoming PARIS toxicity. To narrow the search for a PARIS suppressor, we tested a collection of T5 deletion mutants for their sensitivity to PARIS[33]. Among the variants tested, $T5_{123}$ carries the smallest deletion in the tRNA genes region (*Δ29191-32442*) and is still sensitive to PARIS (**Extended Data Figure 6.10b**). The DNA fragment missing from $T5_{123}$ was divided into three segments, which were separately cloned on an expression plasmid and introduced in $PARIS^+$ cells. A fragment encoding $tRNA^{Pro}$, $tRNA^{fMet}$, $tRNA^{Lys}$, and $tRNA^{Val}$, partially rescued the ability of phage $T5_{123}$ to infect $PARIS^+$ cells, while other fragments had no effect (**Extended Data Figure 6.10c**). Next, we cloned each of the tRNAs individually and tested their ability to rescue $T5_{123}$ infection. Over-expression of the T5 $tRNA^{Lys}$ completely restored $T5_{123}$ infectivity (**Figure 6.5f**), demonstrating that this phage-encoded tRNA neutralizes PARIS-induced toxicity, and suggests that the host $tRNA^{Lys}$ is targeted by AriB.

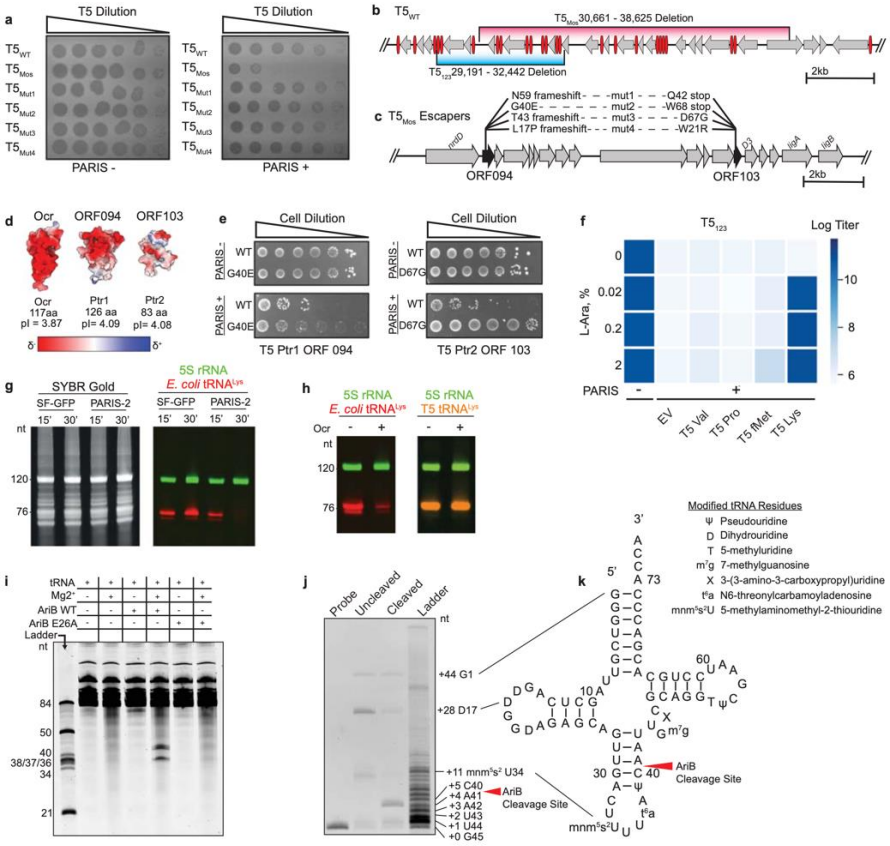


Figure 6.5. PARIS cleaves E. coli tRNALys and T5 encodes a non-cleavable tRNALys. a, Plaque assays conducted using cells with (+) or without (-) PARIS against $T5_{WT}$, $T5_{Mos}$ and $T5_{Mos}$ mutants ($T5_{Mut1-4}$). b, Graphical map of the wild-type T5 genome. ORFs are gray arrows and tRNAs are red ovals. $T5_{Mos}$ is missing an 8 kb fragment (red), while $T5_{123}$ has a smaller 3 kb deletion (blue). c, $T5_{Mos}$ phages acquire mutations in ORF094 (Ptr1) and ORF103 (Ptr2) (black arrows) to escape PARIS mediated immunity (mut1-4). d, AlphaFold2-predicted structures for Ptr1 and Ptr2 compared to T7 Ocr. Proteins surfaces colored by electrostatic potential. e, Expression of WT but not mutant triggers cause toxicity in PARIS+ cells but not PARIS- cells. f, Overexpression of T5 $tRNA^{Lys}$ in PARIS+ cells rescues infection by phage $T5_{123}$, which lacks $tRNA^{Lys}$. g, Northern Blot performed on total RNA extracted from E. coli MG1655 carrying the PARIS system or a GFP control, and Ocr under the control of a PPhlF promoter. RNA was extracted 15 or 30 minutes after induction. Identification of the 5S rRNA (green) and E. coli $tRNA^{Lys}$ (red) were identified using distinct probes. h, Northern Blot performed on total RNA extracted from E. coli carrying PARIS or GFP, Ocr, and the T5 $tRNA^{Lys}$ under the control of an araBAD promoter. RNA was extracted after 30 min of induction. The $tRNA^{Lys}$ (red) and the T5 $tRNA^{Lys}$ (yellow) probes reveal that T5 $tRNA^{Lys}$ is not degraded during PARIS-mediated defence. i, Small RNA extracted from E. coli show a cleavage product when incubated with AriB in a metal-dependent manner. j, Primer extension assay. AriB cleaves E. coli $tRNA^{Lys}$ between positions 40 and 41. k, Schematic representation of the AriB cleavage site on E. coli $tRNA^{Lys}$.

To confirm that the *E. coli* $tRNA^{Lys}$ is indeed degraded upon PARIS activation, we performed a Northern Blot with total RNA extracted from $PARIS^{+}$ cells 15 or 30 min after induction of the Ocr trigger. $tRNA^{Lys}$ was degraded, while SYBR-Gold staining did not reveal degradation of other RNAs (**Figure 6.5g**). We note that we cannot exclude that other *E. coli* tRNAs might be specifically degraded by AriB, and that $T5_{123}$ complements their loss through some of the other tRNAs that it carries. More work is necessary to fully characterize the activity of AriB.

To determine how T5 $tRNA^{Lys}$ suppresses PARIS immunity, we analyzed RNA extracted from PARIS activated cells expressing the phage tRNA from a plasmid. While the *E. coli* $tRNA^{Lys}$ was still degraded, the T5 $tRNA^{Lys}$ remained intact (**Figure 6.5h, Extended Data Figure 6.11a-b**). The phage-encoded variant of $tRNA^{Lys}$ contains mutations in the anticodon stem-loop (**Extended Data Figure 6.10d**), which could be responsible for the lack of AriB recognition. To test this hypothesis, we introduced mutations from the T5 tRNA into the *E. coli* $tRNA^{Lys}$ counterpart. The resulting chimeric tRNA rescued infection by phage $T5_{123}$ in the presence of PARIS, confirming that AriB recognizes the tRNA anticodon stem-loop (**Extended Data Figure 6.10e**). To determine the precise cleavage site of AriB, we incubated tRNAs extracted from *E. coli* with activated AriB, which resulted in cleavage products only in the presence of $Mg^{2+}$ and with the catalytically active AriB (**Figure 6.5i**). A primer extension assay performed using a probe complementary to the 3' end of the host $tRNA^{Lys}$ reveals that the AriB cleavage site lies between residues 40 and 41 of the *E. coli* $tRNA^{Lys}$ (**Figure 6.5j-k**). This cleavage site was corroborated by Sanger sequencing of the cleaved 3' fragment (**Extended Data Figure 6.11c**). Collectively, these results indicate that phage T5 encodes a non-cleavable variant

of tRNA that carries a mutation in the AriB cleavage site, compensating for the loss of the host tRNA depleted upon PARIS activation.

Evolution of the PARIS Defense System

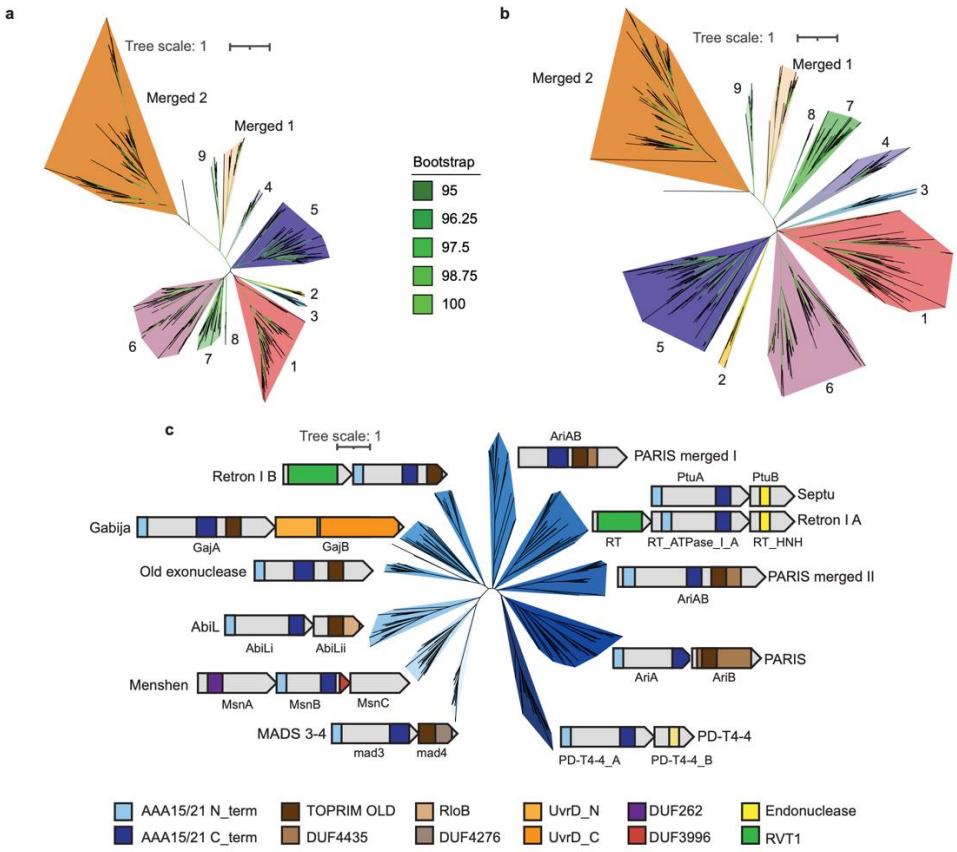



Figure 6.6. Phylogenies of PARIS and related ABC ATPase powered defense systems. a-b, Phylogenetic trees of AriA and AriB share similar branching patterns. Different colors represent 11 subclades of PARIS including two clades in which ariA and ariB merged into a single ariAB gene (Merged 1 and Merged 2). c, Phylogenetic tree of the AAA15/21 ATPase containing defense proteins. Pfam accession: AAA15 (PF13175), AAA21 (PF13304), TOPRIM-Old (PF20469), DUF4435 (PF14491), RloB (PF13707), DUF4276 (PF14103), UvrD_N (PF00580), UvrD_C (PF13361), DUF262 (PF03235), DUF3996 (PF13161), HNH endonuclease (PF14279, PF10107), RVT1 (PF00078), were used for domain annotations.

To inform our understanding of the diversity of PARIS systems and their relationship with other ABC ATPase powered defense systems, we conducted phylogenetic analyses. First,

we generated trees for AriA and AriB, revealing a diverse set of systems that can be grouped into 11 distinct clades including two in which *ariA* and *ariB* merged into a single *ariAB* gene (**Figure 6.6a-b**). *ariA* and *ariB* genes from the same system consistently fall in matching clades, showing they coevolve and are not frequently swapped between systems. A rooted version of the AriB tree shows that AriAB emerged once from clade 9 of two gene systems in a single event and was then split in two different clades (**Extended Data Figure 6.11**). Interestingly, sequence alignments of AriA reveal that residues associated with the central and radial pores of the complex are poorly conserved (**Extended Data Figure 6.11**). This suggests that the signal transduction and effector parts of PARIS immunity are shared across homologs, while the central and radial pores have evolved to recognize different viral triggers.

PARIS was recently assigned to a family of OLD-like defense systems that share an ABC ATPase and a TOPRIM nuclease[11]. This analysis identified four classes, where the single gene systems such as P2 Old are categorized as class 1, Gabija are class 2, reverse-transcriptase containing systems are class 3, and PARIS systems are class 4. We built a tree of known anti-phage defense systems that carry an ABC ATPase of the AAA15/21 family (**Figure 6.6c**). Our analysis adds AbiL and MADS3-4 systems to the list of systems sharing an ABC ATPase and TOPRIM nuclease. It also reveals how this ABC ATPase can be found in association with a variety of other effector domains, revealing the evolutionary success of this domain as the likely sensor of viral infection in diverse immune systems.

## Discussion

Our results demonstrate how AriA forms a propeller-shaped hexamer capable of binding three AriB monomers. AriA is an ATP-dependent sensor that detects viral immune suppressors

(e.g., Ocr), which trigger the release of AriB. Released AriB assembles into a homodimer that cleaves specific host tRNAs resulting in a rapid interruption of translation followed by a loss of membrane integrity. As opposed to prototypical toxin-antitoxin systems, AriB has modest toxicity when expressed alone. This suggests that AriA activates AriB through a structural modification that enables dimerization or an unidentified posttranslational modification. More work will be necessary to determine the mechanism of AriB release and activation. Interestingly, we found that the PARIS system from *E. coli* B185 is not only triggered by the phage T7 Ocr protein but also by the Ptr1 and Ptr2 proteins of phage T5. These proteins are small and negatively charged but lack obvious sequence or structural similarities. Determining how PARIS detects such diverse proteins and whether different systems have distinct trigger sensitivities presents an intriguing avenue for future research.

Our work adds to a growing list of defense systems that cause translational arrest upon infection. The PrrC and Retron I A (i.e., Ptu*AB* proteins) systems induce tRNA degradation after detecting foreign proteins, while Cas13a degrades tRNA after detection of foreign RNA[34–36]. However, translational arrest isn't limited to tRNA degradation, Mogila *et al.* recently identified a CRISPR-Cas system that cleaves mRNA associated with the ribosome, which arrests translation[37], and the Lit (Late Inhibition of T4) protease degrades the translation elongation factor EF-Tu after detection of the major capsid protein of phage T4[38].

We show that some phages circumvent PARIS-induced toxicity by expressing non-cleavable tRNA variants. Many tailed phages, such as those from *Demerecviridae* (T5-like phages), *Straboviridae* (T4-like phages) or *Ackermannviridae* families encode large arrays of tRNAs previously thought to be associated with translation optimization due to codon bias in the

host and viral genome[39]. However, estimation of the abundance of host and phage tRNAs during infection, compared to the corresponding codon frequencies in produced mRNAs does not directly support this hypothesis[39]. A recent computational analysis suggests that phages may have evolved these variants in response to bacterial tRNA-targeting toxins, as evidenced by distinct anti-codon loop structures aimed at evading host nucleases[40]. This hypothesis is further supported by a parallel work demonstrating that phage T5 $tRNA^{Tyr}$ provides protection against the PtuAB toxin of the Eco7 (Ec78) retron[36].

Interestingly, this strategy of encoding tRNA clusters is also observed in some eukaryotic viruses, like those from *Herpesviridae, Mimiviridae* and *Phycodnaviridae* families[41], possibly serving a similar function in evading host immunity. Strikingly, an anti-codon nuclease was also recently found to protect humans against Pox viruses, demonstrating that the inhibition of translation through the inactivation of specific tRNAs is a conserved anti-viral strategy used across domains of life[42].

## Methods

### Bacterial Strains, Phages, and Plasmids

Phages, bacterial strains, and plasmids used in the study are listed in Supplementary Table 6.1. Primers are listed in **Supplementary Table 6.2**. Infection with T5 phage was carried in LB supplemented with 1mM $CaCl_2$. Unless otherwise indicated, 0.2% L-arabinose, 0.1mM IPTG, and 0.2 µg/ml aTc were used for the induction.

For structural studies, the coding sequences for AriA and AriB from *E. coli* B185[2] were cloned under control of a T7 promoter into a pRSF-Duet vector with a C-terminal strep-tag II on AriB. For coexpression of PARIS with T7 Ocr trigger, the TOPRIM nuclease of AriB was

inactivated (E26A) and the strep-tag II was removed by site-directed mutagenesis. The coding sequence for T7 Ocr protein was cloned into a pET-Duet vector with a C-terminal strep-tag II. For *in vitro* binding assays, the coding sequence for T7 Ocr protein was cloned into a pET-Duet vector with a N-terminal $His_6$TwinStrep-SUMO tag. All *in vivo* experiments were performed with pFR85, encoding PARIS from *E. coli* B185 with its native promoter and an inducible $P_{tet}$ promoter upstream. For *in vivo* protein pulldowns, C-terminal AriB strep-tag II was introduced to the pFR85. Mutants of PARIS were constructed from pFR85 using the Gibson method[43] or Q5 site-directed mutagenesis kit (NEB). The PARIS triggers Ptr1 (T5 ORF094) and Ptr2 (T5 ORF103) from T5 phage as well as T5 genome fragments, T5 $tRNA^{Lys}$ and *E. coli* $tRNA^{Lys}$ were cloned on the pBAD18 vector under control of *araBAD* promoter using the Gibson method[43]. 20 bp upstream and downstream of tRNA genes were preserved during cloning. T5 nucleotide positions are provided according to the genome assembly AY543070.1. All constructions were verified using Sanger sequencing.

Expression and Purification of the PARIS Complex

AriA and AriB-strep were co-expressed from a pRSF-Duet plasmid in BL21-AI cells (Invitrogen). Cultures grown at 37°C to 0.4-0.5 $OD_{600}$ were induced with 0.1 mM IPTG and 0.1% L-arabinose. Following induction, the cultures were incubated overnight at 16°C. Cells were pelleted (3,000 × g, 10 min, at 4°C), resuspended in lysis buffer (25 mM Tris, pH=8.5, 150 mM NaCl, 1 mM EDTA) and lysed by sonication. Cellular debris was removed by centrifugation (10,000 × g, 25 min, at 4°C) and AriAB complex was purified by affinity chromatography on 5 mL streptactin resin (IBA) and eluted with 2.5 mM desthiobiotin (IBA) in lysis buffer. The eluted protein was concentrated (100k MWCO PES Spin-X UF concentrator, Corning) and

further purified by size exclusion chromatography (Sup200 column, Cytiva) in lysis buffer with 2% glycerol. Fractions of interest were combined and concentrated to 5μM (100k MWCO PES concentrator, Pierce) and used immediately for vitrification on cryo-EM grids.

*In Vivo* Protein Pull-down

To determine protein-of-interest interacting partners *in vivo*, AriB-Strep expressed from pFR85 or OcrStrep expressed from pBAD were used as bait, and pull-downs were performed in *E. coli* BW25113. 3 L of cell culture was grown in LB media at 37°C in Thomson flasks with aerated lids (1.5 L per flask). Expression of proteins was induced with 0.2 μg/ml aTc or 0.2% L-arabinose at $OD_{600}$ ~ 0.1 and cells were harvested by centrifugation after reaching $OD_{600}$ ~ 0.8-1. Cells were processed as described above, and Strep-tagged proteins were purified on two stacked 5 mL StrepTrap HP (Cytiva) columns. Identity of protein bands after SDS-PAGE was determined by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. ~1/3 of the Coomassie-stained band was cleaved from the gel and samples were prepared with Trypsin Gold (Promega) in accordance with manufacturer's instructions. Mass spectra were obtained using the rapifleX system (Bruker).

For production of activated AriB, AriB-Strep from pFR85 and non-tagged Ocr from pBAD were expressed in separate cultures and cells were grown as described. Cell cultures were mixed, and cell walls were disrupted by sonication on ice (60% power, 10'' pulse, 20'' pause, 30-60 minutes) on Qsonica sonicator with 6 mm sonotrode, followed by StrepTrap HP purification of activated AriB-Strep. The molecular weight of protein complexes was determined after size exclusion chromatography performed on Superdex 200 Increase 10/300 column (GE Healthcare) calibrated with High Molecular Weight calibration kit (GE Healthcare).

Glutaraldehyde Cross-linking Assay

To determine the oligomeric state of activated AriB in solution, we performed glutaraldehyde crosslinking assay. Concentrated AriB aliquots were mixed with glutaraldehyde (253857.1611; PanReac) to a final concentration of 0.025%, 0.125%, or 0.25% vol/vol and incubated for 30 min at room temperature. The reaction was quenched by the addition of one sample volume of 1 M Tris-HCl at pH 8.0 and a half sample volume of 5x concentrated SDS sample buffer. Samples were heated for 5 min at 60°C and separated by SDS-PAGE.

*In Vitro* Protein Pull-down

The T7 Ocr protein with an N-terminal His6-TwinStrep-SUMO was expressed from a pET-Duet plasmid in BL21(DE3) cells. Cultures grown at 37°C to 0.4-0.5 $OD_{600}$ were induced with 0.5 mM IPTG. Following induction, the cultures were incubated overnight at 16°C. Cells were pelleted (3,000 × g, 10 min, at 4°C), resuspended in lysis buffer (25 mM Tris, pH=8.5, 150 mM NaCl, 1mM EDTA) and lysed by sonication. Cellular debris was removed by centrifugation (10,000 × g, 25 min, at 4°C) and HSS-Ocr was purified by affinity chromatography on 5 mL streptactin resin (IBA) and eluted with 2.5 mM desthiobiotin (IBA) in lysis buffer. The eluted protein was incubated with SUMO protease at 4°C overnight before washing over 5 mL Ni-NTA resin (25 mM Tris pH=7.5, 300 mM NaCl, 1 mM TCEP, 20 mM Imidazole) to remove SUMO-cleaved tag. Flow through protein was concentrated (10k MWCO PES Spin-X UF concentrator, Corning) and further purified by size exclusion chromatography (Sup6 10 300 column, Cytiva) in lysis buffer with 2% glycerol. Fractions of interest were combined and concentrated (10k MWCO PES Spin-X UF concentrator, Corning). Cleaved Ocr (3.8 uM) was incubated with WT PARIS (11.3 μM) at 37°C for 15 minutes before analysis via size exclusion chromatography

(Sup6 10 300 column, Cytiva) in lysis buffer with 2% glycerol. Fractions of interest were combined, concentrated (10k MWCO PES Spin-X UF concentrator, Corning), and analyzed via SDS-PAGE (15% acrylamide, 150V for ~1 hour).

Cryo-EM Sample Preparation and Data Collection

All samples were subjected to vitrification on 1.2/1.3 Carbon Quantifoil$^{TM}$ grids which were glow discharged using an easiGlow$^{TM}$ (Pelco) glow discharge station (glow 45s hold 15s). After applying 3 μL of 5μM PARIS + 1mM ATPγS, grids were subjected to double-sided blotting using a Vitrobot Mk IV with a force of 5 for 4 seconds at 4°C and 100% relative humidity. After plunge-freezing in liquid ethane, grids were clipped and stored in liquid nitrogen. For the PARIS obtained from AriB pulldowns, cryo-EM grids were screened, and all data were collected using a Talos Arctica microscope (ThermoFisher) operating at 200kV equipped with a K3 direct electron detector (Gatan). SerialEM[44] software version 3.8.6 was used to automate data collection. For final data collection, a total of 7,340 movies were collected in super resolution mode with a pixel size of 0.552Å/pixel and a total dose of 56.42 e/Å$^2$ distributed over 50 subframes. For the structure of AriA, samples were prepared for cryo-EM analysis as described above. After screening grids, 10,340 movies were collected using a Talos Arctica microscope (Thermofisher) operating at 200kV equipped with a K3 direct electron detector (Gatan). SerialEM software version 3.8.6 was used to automate data collection. Movies were recorded in super resolution mode with a pixel size of 0.552Å/pixel and a total dose of 56.31 electrons/Å$^2$ distributed over 50 subframes.

Cryo-EM Data Processing of Inactive PARIS

All data were processed in cryoSPARC v4.41[45]. After patch motion correction and CTF estimation, micrographs with CTF fits worse that 8 Å were discarded, yielding 5,998 micrographs for particle picking. After processing a subset of 200 micrographs, the $AriA_6$:$AriB_3$ assembly of the PARIS complex was apparent. Using this low-resolution initial volume, AlphaFold2 predicted structures of AriA and AriB were fit into the map and used to generate a 20 Å lowpass filtered map using the Molmap command in ChimeraX v1.7[46]. This volume was imported to cryoSPARC and used to generate templates for particle picking. A total of 4,078,384 particles were extracted and after initial 2D classification, 1,643,515 particles were kept for further processing. These particles were sorted using a three-class *ab initio* reconstruction which revealed two junk classes and a volume corresponding to the assembled PARIS complex that contained 940,782 particles. After a second round of 2-D classification, 934,763 particles remained and were subjected to another round of *ab initio* reconstruction and heterogeneous refinement. This produced a consensus volume containing 532,010 particles that contains conformational heterogeneity for two AriB subunits, while 1 asymmetric unit of the complex is well aligned. Masked local refinement was used to determine a 3.2 Å-resolution map of one asymmetric unit of the complex. The consensus refinement with 532,010 particles was also subjected to multiple rounds of masked 3-D classification to isolate particles in the ‘cis’ and ‘trans’ arrangements. The fully assembled PARIS complex demonstrates flexibility based on 3DVA, which limited the attainable resolution of the fully assembled complex. The density map for both the local refinement and the full complex are available at EMDB-42719, 43103, and 43104 and raw micrographs are available at EMPIAR-11832.

Cryo-EM Data Processing of Ocr Pulldown

All data were processed in cryoSPARC v4.41. Of the 10,340 micrographs collected, 9,399 were selected for further processing based on the CTF fit criteria described above. Using the blob picker with a particle diameter of 200 Å, particles were extracted and subjected to 2-D classification. In the 2-D classes, there were two distinct macromolecular complexes present, one of which was interpreted to be the AriA scaffold of the PARIS complex, while the other was interpreted to correspond to RNA polymerase (RNAP).

From seven selected classes corresponding to the AriA hexamer, 667,479 particles were selected for downstream refinement. From a 2-class *ab initio* reconstruction of these particles, 392,989 were kept. Iterative sorting by multi-class *ab initio* reconstruction and heterogenous refinements yielded a volume containing 62,732 particles. AriA subunits can be fit into this density, but the poor resolution makes this structure indistinguishable to the AriA scaffold identified in the intact PARIS complex. No density is observed for Ocr, which suggests multiple binding sites may be present on the PARIS complex, or that Ocr is attached to AriA in a flexible state. Density map available at EMDB-43105 and raw micrographs are available at EMPIAR-11833.

Model Building and Refinement

To determine the structure of the asymmetric unit of the PARIS complex, density halfmaps were obtained from cryoSPARC local refinements before being provided as inputs to DeepEMhancer v0.14[47] for map sharpening. AlphaFold2 predicted structures of AriA and AriB were then docked into the map density and map to model fit was improved using ISOLDE v1.7.1[48] molecular dynamics simulation environment in ChimeraX v1.7. After initial

equilibration, the PDB files were saved and used as inputs for realspace refinement using PHENIX v1.20.1[49]. After initial refinement, problematic areas were manually inspected and modified using COOT v0.9.8.1[50] and sidechains were removed from regions with poorer than 4 Å-resolution. After addressing clashes and non-rotameric sidechain orientations, iterative refinements were carried out in Phenix until Clashscores, Ramachandran and sidechain outliers failed to improve. The final model and corresponding map have been deposited (PDB ID: 8UX9, EMDB-42719).

To generate the biological assemblies for the PARIS complex atomic model, halfmaps from C3 symmetric, and nonuniform refinements of particles corresponding to the 'cis' and 'trans' orientations were obtained from cryoSPARC and sharpened in DeepEMhancer v0.14. After fitting the three copies of the asymmetric unit into the 'cis' and 'trans' density maps using the ChimeraX v1.7 fitmap command, the find NCS tool in Phenix was then used to calculate the appropriate symmetry operators for the assembled forms of the complex. These symmetry operators were applied to the Asymmetric unit to generate biological assemblies 1 and 2 using the Apply NCS tool in Phenix. Maps for the 'cis' and 'trans' orientations of the PARIS complex are available at EMDB-43104 and EMDB43105, respectively.

ATPase Assays

ATPase activity was measured using 6 μM Ocr, 100 nM PARIS and 128 μM ATP supplemented with 10 nM [α-32 P]-ATP (PerkinElmer). The reaction buffer includes 20 mM Tris-HCl pH 8.0, 150 mM Sodium Chloride, 1 mM DTT, and 5 mM magnesium chloride. The reactions were incubated at 37ºC and 10 μL aliquots were quenched with Phenol at 1, 2, 4, 8,16, and 32 minutes. For the 'trigger' only reactions, 100 nM Ocr was mixed with 128 μM ATP

supplemented with 10 nM [α- 32 P]-ATP (PerkinElmer) in the reaction buffer. Reactions were incubated for 32 minutes at 37ºC and quenched with phenol. The ATP and ADP makers were created using T4 polynucleotide kinase (NEB). T4 PNK mixed with 128 µM ATP supplemented with 10 nM [α-32 P]-ATP, and 1 mM DNA oligo in T4 polynucleotide kinase buffer (NEB). This mixture was then incubated at 37ºC for 1 hour. All reaction products were phenol-chloroform extracted and resolved on silica TLC plates (Millipore). Reaction products were spotted 2.5 cm above the bottom of the TLC plate. The plate was placed in a TLC developing chamber filled to ~1.5 cm with developing solvent (0.2 M ammonium bicarbonate pH 9.3, 70% ethanol, and 30% water) and covered with aluminum foil for 4 hrs at room temperature. The TLC plate was exposed to a phosphor screen, imaged with Typhoon phosphor imager and quantified with the ImageQuant software package (Cytiva).

Efficiency of Plaquing Assay

The activity of PARIS mutants was measured in comparison to the wild-type system by performing efficiency of plaquing (EOP) assays with phage T7. *E. coli* K-12 MG1655 carrying each of the AriA or AriB mutants, the control plasmid pFR66 (sfGFP), or the wild-type PARIS system (pFR85) were grown overnight in LB + kanamycin 50 µg/ml. Bacterial lawns were prepared by mixing 100 µl of a stationary culture with 5 ml of LB + 0.5% agar, and the mixture was poured on Petri dish of LB + kanamycin 50 µg/ml. Ten-fold serial dilutions of high titer stock of T7 phage were spotted on each plate and incubated 5 h at 37°C. EOPs with T5 and T5-like phages were performed with *E. coli* BW25113 as described above, except that plates were incubated at 37°C overnight, and top agar was supplemented with 1 mM $CaCl_2$. Induction of

tRNA genes or phage T5 fragments was achieved by supplementing top agar with indicated amount of L-arabinose. Plaque assays were performed in at least two independent replicates.

Liquid Culture Phage Infection

To monitor dynamics of $T5_{Mos}$ phage infection of the PARIS$^+$ (pFR85) and PARIS$^-$ (pFR66) culture in liquid, we used EnSpire Multimode Plate Reader (Perkin Elmer). Overnight bacterial cultures were diluted 100-fold in LB medium with appropriate antibiotics and grown at 37°C in 10 mL of LB supplemented with 1mM $CaCl_2$. At $OD_{600}$ = 0.6, 200 µl aliquots were transferred to 96-well plates and infected at indicated Multiplicity of Infection (MOI). Optical density was monitored for 10 hours. All experiments were performed in three biological replicates.

Solid Media Toxicity Assay

PARIS toxicity in the presence of T7 Ocr or Ptr1 and Ptr2 was measured using a spot-test assay. PARIS$^+$ (pFR85) and PARIS$^-$ (pFR66) cultures carrying pBAD vector encoding indicated trigger were grown overnight in 10 ml LB at 37°C. Stationary cultures were diluted to $OD_{600}$=0.6 and plated on LB or minimal media (M9 with 5% v/v LB) agar plates supplemented with 0.2% L-arabinose by serial 10fold dilution. Control plates without induction contained 0.2% glucose to prevent leakage of *araBAD* promoter. To assess the toxicity of PARIS AriA D401N/E404Q mutant, we used either pFR85 (PARIS expression driven from native promoter, BW25113 *E. coli* strain) or pRAW464 vector (PARIS overexpression from T7 promoter, BL-21 (DE3) *E. coli* strain). The cultures were plated on M9 as described above and PARIS expression was induced with 1 mM IPTG. Ocr expression was driven from pBAD with 0.2% L-arabinose.

Liquid Culture Toxicity Assay

To assess the toxicity of AriB expression alone compared to activated PARIS, *E. coli* BL-21 AI cells were transformed with plasmids encoding PARIS (pRAW 464), AriB (pRAW 466), or PARIS (pRAW 464) and Ocr (pRAW 472). After growing overnight cultures in LB supplemented with appropriate antibiotic, cells were diluted to an $OD_{600}$ of 0.1 in LB + antibiotics before growing to an $OD_{600}$=0.5. After cells reached OD600 0.5, 200µL of culture was added to 800µL of media supplemented with antibiotics and either 0.4% glucose, 0.2% L-arabinose and 0.1 mM IPTG, or 0.5% L-arabinose and 0.1 mM IPTG. 200µL of each culture were aliquoted into a 96 well plate (Evergreen 222-8030-01F).

Growth was monitored using an Agilent BioTek Cytation5 plate reader set to 30ºC with continuous linear shaking and $OD_{600}$ readings taken every 5 minutes for 15 hours.
*E. coli* MG1655 strains carrying plasmids pFR85 ($P_{tet}$ ariAB), pFR85ΔAriA (pFR99), pFR85ΔAriB (pFR100) or pFR85-AriA(R212R*) (pFD294) in which a stop codon was introduced in AriA, truncating the protein before the domain that interacts with AriB, were diluted 1:100 from an overnight culture in LB + kanamycin 50 µg/ml and arrayed in a 96 well plate. Growth was then monitored every 10 minutes for 30 hours at $OD_{600}$ in an infinite M200 Pro plate reader (Tecan) at 37°C with shaking. Absorbance values of three biological replicates were analyzed using GraphPad Prism v10.1.2. Standard two-way ANOVA was used to calculate p-values for growth rates relative to wild-type PARIS.

Fluorescence Microscopy

$PARIS^+$ (pFR85) and $PARIS^-$ (pFR66) carrying pBAD Ocr were diluted 1:100 and cultivated in LB supplemented with appropriate antibiotics at 37°C. When $OD_{600}$ reached 0.4, an

aliquot was mixed with DAPI (Invitrogen) at 1 mg/mL final concentration and propidium iodide (PI, Invitrogen) at 1 μg/mL final concentration and incubated at room temperature in the dark for 5 minutes. LB + 1.5% agarose slabs (∼0.2 mm thick) were prepared on a 75 x 25 mm microscopy slide (Fisher Scientific). Agarose slabs contained staining dyes at the same concentration and optionally were supplemented with 0.2% L-arabinose, to induce Ocr expression. ~1 μl of stained cells was placed on agarose slab and after 1min drying the slab was covered by a small 22 × 22 mm coverslip (Fisher Scientific). Imaging was performed on a Nikon Eclipse Ti-E inverted epifluorescence microscope. Each field of view was imaged in transmitted light channel (200 ms exposure) and in DAPI (filterset Semrock DAPI-50LP-A, 200 ms exposure) and in PI (filterset TxRed-4040C, 200 ms exposure) channels.

Extraction of Total DNA

Total DNA was extracted at different time points (15, 30, 60 and 120 min) with or without DAPG to induce the expression of Ocr in the presence of PARIS. DNA was extracted using the Wizard® genomic DNA purification kit (Promega) according to the manufacturer's instructions and treated with RNase at 0.1 mg/ml final. Each sample of total DNA was loaded on 1% agarose 1X TAE gel.

TUNEL Assay

TUNEL assays was performed to estimate *in vivo* accumulation of dsDNA breaks, according to methods described previously[51]. $PARIS^+$ (pFR85) culture carrying pBAD Ocr vector was grown to $OD_{600}$=0.3, followed by Ocr induction with 0.2% L-arabinose for 2 hours. As a positive control, cells were treated with 0.1% $H_2O_2$ to induce accumulation of DNA breaks. 2 ml of cells were harvested by centrifugation and treated according to the standard protocol (TUNEL

Assay Kit – FITC, ab66108, Abcam). Measurement of FITC fluorescence was performed with CytoFLEX cytometer (Beckman Coulter) in the FITC channel. 100,000 events were collected for each sample in three biological replicates. Data was analyzed and visualized in FlowJo™ v10[52].

Hi-C Procedure and Sequencing

*E. coli* MG1655 carrying plasmids pFR85 ($P_{tet}$ ariAB) and pFD250 ($P_{PhlF}$ ocr), or the pFR66 (sfGFP) + pFD245 (PhlF sfGFP) control plasmids were diluted 1:100 from an overnight culture in LB + kanamycin 50 µg/ml + chloramphenicol 20 µg/ml and grown until $OD_{600}$ = 0.3. DAPG 50 µM was then added, followed by incubation for 15 and 30min. To compare the effect of PARIS activation with that of treatment with chloramphenicol, *E. coli* MG1655 harboring pFR66 were diluted 1:100 from an overnight culture in LB + kanamycin 50 µg/ml until $OD_{600}$ = 0.3, followed by addition of chloramphenicol 20 µg/ml and incubation for 15 and 30min. Cell fixation was performed with 4% formaldehyde (Sigma-Aldrich, Cat. F8775) as described in Cockram et al. (2021)[53]. Quenching of formaldehyde with 300 mM glycine was performed at 4°C for 20 min. Hi-C experiments were performed with the Arima kit. Samples were sonicated using Covaris (DNA 300bp).

Preparation of the samples for paired-end sequencing was performed using Invitrogen TM Colibri TM PS DNA Library Prep Kit for Illumina according to manufacturer instructions. The detailed protocol is available in Cockram et al. (2021)[53]. Reads were aligned with bowtie2 v2.4.4 and Hi-C contact maps were generated using hicstuff v3.0.3 (https://github.com/koszullab/hicstuff) with default parameters and using HpaII enzyme to digest. Contacts were filtered as described in[54], and PCR duplicates (defined as paired reads mapping at exactly the same position) were discarded. Matrices were binned 4kb. Balanced

normalizations were performed using ICE algorithm[55]. For all comparative analyses, matrices were downsampled to the same number of contacts. The Hi-C signal was computed as the contacts between adjacent 5kb bins as described in Lioy et al. (2018)[56]. In order to compare this signal with other genomics tracks, we binned it at the desired resolution.

Metabolic Labeling

Metabolic Labeling with 3H-methionine, 3H-uridine and 3H-thymidine Metabolic labelling experiments were performed as described previously[57], with minor modifications. *E. coli* BW25113 PARIS$^{+}$ (pFR85) strain was transformed with pBAD plasmid carrying Ocr trigger (for L-arabinose-inducible expression). Transformed cells were initially plated on LB plates supplemented with 100 μg/mL ampicillin, 25 μg/mL kanamycin, and 0.2% glucose. Using individual *E. coli* colonies for inoculation, 2 mL liquid cultures were prepared in defined Neidhardt MOPS minimal media (supplemented with 100 μg/mL ampicillin, 25 μg/mL kanamycin, 0.1% casamino acids, and 1% glucose, and grown overnight at 37°C with shaking. Subsequently, experimental 20 mL cultures were prepared in 125 mL conical flasks in MOPS medium, supplemented with 0.5% glycerol, 100 μg/mL ampicillin, 25 μg/mL kanamycin, as well as a set of 19 amino acids (lacking methionine), each at a final concentration of 25 μg/mL. These cultures were inoculated at $OD_{600}$ of 0.05 and grown at 37°C with shaking until the $OD_{600}$ reached 0.3. At this point, 1 mL aliquots (designated as the pre-induction zero time-point) were transferred to 1.5 mL Eppendorf tubes containing 10 μL of pre-incubated at 37°C respective radioisotope (3H methionine [0.77 μCi, Perkin Elmer], 3H uridine [0.1 μCi, Perkin Elmer] or 3H thymidine [0.32 μCi, Perkin Elmer]). Concurrently, Ocr expression in the remaining 19 mL culture was induced by adding L-arabinose to a final concentration of 0.2%. Throughout the Ocr

induction time course, 1 mL aliquots were taken from the 20 mL culture and transferred to 1.5 mL Eppendorf tubes containing 10 µL of the appropriate radioisotope. Radioisotope incorporation was halted after 8 minutes of incubation at 37°C by adding 200 µL of ice-cold 50% trichloroacetic acid (TCA) to the 1 mL cultures. Additionally, 1 mL aliquots were sampled for $OD_{600}$ measurements during induction time course. The resultant 1.2 mL TCA-stopped culture samples were loaded onto GF/C filters (Whatman) prewashed with 5% TCA and unincorporated label was removed by washing the filter twice with 5 mL of ice-cold TCA followed by a 5 mL wash with 95% EtOH (twice). The filters were placed in scintillation vials, dried for two hours at room temperature, followed by the addition of 5 mL EcoLite™scintillation cocktail (MP Biomedicals). After shaking for 15 minutes radioactivity was quantified using automatic TDCR Liquid Scintillation Counter (HIDEX). Isotope incorporation was quantified by normalizing radioactivity counts (CPM) to $OD_{600}$, with the pre-induction zero time-point serving as the reference (set to 100%). All experiments were performed as biological triplicates using three independent liquid starter cultures inoculated with different colonies.

*In Vitro* Translation

*In vitro* translation was monitored by the production of luciferase signal in a PURExpress *in vitro* protein synthesis kit (NEB), using firefly luciferase mRNA as an input. pT7-Fluc, encoding firefly luciferase (*fluc*) gene under control of T7 promoter, was used as a template for T7 *in vitro* transcription with MEGAscript kit (Thermo Scientific). mRNA was treated with DNase I (Thermo Scientific) and purified with Monarch PCR & DNA cleanup kit (NEB). PURExpress *in vitro* translation reaction was assembled in RNase-free tubes according to manufacturer's instructions, with minor modifications. Reaction mixtures were adjusted to a total

volume of 5 µl (2 µl solution A; 1.5 µl solution B; 0.3 µl RiboLock (40 U/µl, Thermo Scientific); 0.2 µl D-Luciferin (10 mM, Sigma)), supplemented with 0.5 µl of activated AriB or AriB (E26A) (100 µg/ml final concentration) or the same volume of buffer A (50 mM NaCl; 1mM EDTA; 100 mM Tris-HCl; 5mM β-ME; pH 8) as a control.

Reaction was started by the addition of 0.5 µl of purified *fluc* mRNA (500 ng), immediately placed in 384-well white plate and covered with optically clear film. Accumulation of luminescent signal was measured by EnSpire Multimode Plate Reader (PerkinElmer) for 3h at 37°C. AriB concentration was determined with Qubit Protein Broad Range Assay Kit (Invitrogen).

Isolation and Sequencing of T5 Escaper Mutants

To select PARIS escape mutants, phage $T5_{Mos}$ was continuously incubated with $PARIS^+$ (pFR85) culture for 3 days. In short, after reaching an $OD_{600} = 0.6$, cells were infected with $T5_{Mos}$ at MOI~0.1, and the culture was incubated overnight at 37°C. Phage was collected the next day and 1 ml of lysate was used to initiate the next round of infection with fresh $PARIS^+$ (pFR85) culture. Escaper mutants were purified from the single plaques obtained on $PARIS^+$ (pFR85) culture and produced on $PARIS^-$ (pFR66) culture. 8 mL of high-titer lysate (~$10^{10}$ pfu/ml) was PEG-precipitated, and phage genomic DNA was purified as described previously[58]. DNA libraries were prepared according to a standard protocol and sequenced on MiniSeq platform (Illumina) with paired-end 150 cycles (75 + 75). Genome assemblies were performed with SPAdes implemented in Unicycler[59]. To identify mutations, genomes of $T5_{Mos}$ escapers were aligned with the sequenced genomes of $T5_{Mos}$ and $T5_{wt}$ from initial stocks. T5 strains reported by

Glukhov et. al.[33] were sequenced on DNBSeq-G400 platform (BGI) to validate boundaries of deletions.

Extraction of Total RNA

Total RNA was isolated from *E. coli* strain MG1655 harboring pFR85 ($P_{tet}$ ariAB) or a control plasmid pFR66 ($P_{tet}$ sfGFP). Cells also carried pFD250 ($P_{PhlF}$ ocr) with Ocr under the control of an inducible DAPG $P_{PhlF}$ promoter (**Figure 6.5g and Extended Data Figure 6.8b**), as well as plasmid pFD287 carrying the $tRNA^{Lys}$ of phage T5 under the control of *araBAD* promoter inducible by L-arabinose (**Figure 6.5h**). Cells were diluted 1:100 in LB with the appropriate antibiotics from overnight cultures and grown with agitation at 37°C. When cultures reached an $OD_{600}$ of 0.2, aTc was added to a final concentration of 0.5 µg/ml to induce the expression of the PARIS system. For the experiment shown in **Figure 6.5h**, Larabinose was added to a final concentration of 0.2% to induce the T5-$tRNA^{Lys}$ from plasmid pFD287. Growth continued until 0.4 and the Ocr trigger was induced with DAPG added to a final concentration of 50 µM, followed by incubation for an additional 15 and 30 min post induction.

Cells were centrifuged at 4,000 g for 10 min, resuspended with 0.2 ml of lysozyme buffer (20 mM Tris-HCl, pH8.0, 2 mM EDTA, 1% Triton X-100 and 20 mg/ml lysozyme) and lysed for 30 min at 37°C. One milliliter of TRIzol reagent (Zymo Research) was added to the samples and incubated for 5 min at room temperature (RT). Lysates were extracted with 0.2 ml of cold chloroform and centrifuged at 12,000 g for 15 min at 4°C. The aqueous phase of the sample was precipitated with 0.5 ml of 100% cold isopropanol, incubated 10 min at RT and centrifuged at 12,000 x g for 10 min at 4°C. RNA pellets were washed with 1 ml of 75% ethanol, dried at RT and dissolved in 50 µl of nuclease free water. Total RNA concentration was measured with a

nanodrop spectrophotometer and then the samples were resolved in a 7 M Urea 10% acrylamide gel before visualization with SYBR-Gold.

Preparation of Enriched Small RNA Fractions from the Total RNA Samples

Cells were grown in 20 ml of media under the same conditions as those used for the extraction of total RNA. When the induced cultures reached an $OD_{600}$ of 0.8, cells were centrifuged at 5000g for 25 min, washed in 5 ml of 0.9% NaCl, and centrifuged again at 5000 g for 25 min. The following is adapted from Avcilar-Kucukgoze *et al*.[60]. The pellets were suspended in 2 ml of 50 mM sodium acetate (NaOAc), 10 mM MgOAc, pH 5.0, with the addition of 1.9 ml of commercial acidic phenol, pH 4.5 (Ambion). The resulting mixture was shaken at 200 rpm for 30 min at 37°C, then centrifuged at 5000 x g for 15 min at 4°C. The upper aqueous phases were collected and subjected to precipitation of total nucleic acids in 0.1 M NaCl, supplemented with an equal volume of 2-propanol, for an overnight incubation at room temperature (RT). After centrifugation at 14500 x g for 15 min at 4°C, pellets were washed with 80% ethanol and air-dried.

Removal of rRNA from the samples was achieved through precipitation: the pellets were suspended in 0.8 ml of cold 1M NaCl, and spun at 9500 x g for 20 min at 4°C. The supernatants were collected, mixed with 1.7 ml ethanol, incubated for 30 min at -20°C, and then centrifuged at 14500 x g for 5 min at 4°C. The pellets were washed with 80% ethanol and air-dried.
Removal of DNA from the samples was achieved through precipitation: the pellets were resuspended in 0.6 ml of 0.3 M NaOAc, pH 5.0. The mixture was then heated for 5 min at 60°C with regular pipetting, followed by the addition of 0.34 ml of 2-propanol, incubation for 10 min at RT, the solution was centrifuged at 14500 x g for 5 min at RT, and collection of supernatants.

The resulting small RNA fractions were precipitated by adding 0.23 ml of ethanol to the supernatant, centrifuging at 14500g for 15 min at 4°C, washing the pellet with 80% ethanol, and air-drying for 5–10 min. Pellets were dissolved in 0.35 ml of RNase-free water.
Deacylation: the fraction was incubated in 0.1 M Tris pH 9.0 for 45 min at 37°C. Final precipitation in 0.3 M NaOAc, pH 5.0 was achieved by adding 2.7 volume of ethanol, incubation for 30 min at -80°C, centrifugation at 16000g for 25 min at 4°C. Pellets were washed with 80% ethanol, air-dried, and dissolved in 80 µl of 1 mM sodium citrate.

Cleavage Assay by AriB

50 nM of activated AriB wt or nuclease deficient AriB (mutant E26A) dimer were incubated 30 min at 37°C with 90 ng of enriched small RNAs, extracted as described above, in a 10 µl reaction mix containing 20 mM Tris pH 8.0 and 200 mM NaCl. $MgCl_2$, when added, was at 2 mM final concentration. Reactions were stopped by adding 10 µl of 100% formamide and loaded on a 15% acrylamide / 7 M Urea denaturing medium gel ran at 22W constant. SYBR Gold-stained gels were imaged using the ChemiDoc MP imaging system (BIO-RAD).

Mapping of the AriB Cleavage Site by Primer Extension

To precisely identify the AriB cleavage site, DNA probe F693 (5'-/FAM/TGGTGGTGGGTCGTGCAGGATTCGAACCTGCGACC-3', harboring a fluorescein at the 5'-end) was used in a RT reaction on small RNAs that were incubated with or without activated AriB. In a 12 µl reaction, 2 µl of 1µM probe was incubated with 750 ng RNA on a heat block at 95°C that was then switched off and left to cool down. 8 µl of RT-PCR mix (Superscript III RT, Invitrogen), was added to the reaction mix and loaded to a thermocycler with the following conditions: 5 min 25°C/60 min 55°C/15 min 70°C. After adding 20 µl of 100%

formamide, reactions were loaded on a 15% acrylamide / 7M Urea denaturing medium gel ran at 22W constant. The fluorescein signal was recorded using the ChemiDoc MP imaging system (BIO-RAD). The ladder was generated by performing an extension reaction using fluorescent probe F693 annealed to a 106 nt DNA template (5'-CGATTGAGGCCGGTAATACGACTCACTATA*GGGTCGTTAGCTCAGTTGGTAGAGCAGTT GACTT TTAATCAATTGGTCGCAGGTTCGAATCCTGCACGACCCACCA*-3' containing the 76 bases corresponding to the $tRNA^{Lys}$. Extension was performed in a Taq DNA polymerase PCR mix incubated in the following conditions: 3 min 95°C / (30 sec 95°C/30 sec 66°C/2 min 72°C) x 5 cycles / 5 min 72°C.

Mapping the AriB Cleavage Site by Specific Reverse Transcription (RT) and Sanger Sequencing

750 ng of AriB-cleaved small RNAs were incubated for 1h at 25°C in a 30 µl ligation reaction with 3 µl of 5′ Ligation Reaction Buffer (10X), 2.5 µl of 5′ Ligation Enzyme Mix, and 1 µl of 5′ SR Adaptor for Illumina (denatured) from the NEBNext® Multiplex Small RNA Library Prep Set (E7300S). A $tRNA^{Lys}$ specific RT was then performed on 15 µl of the ligation reaction by adding 1.5 µL of 10 mM dNTPs, 6 µl of 5X First-Strand Buffer, 1.5 µl of 0.1 M DTT, 1.5 µl of SuperScript™ III RT from Invitrogen's SuperScript III kit (18080-093), and 50 pmol of the RT primer (5-TGTGCTCTTCCGATCT GGTGGGTCGTGCAGGATTCGAACCTG-3) in a total volume of 30 µl and incubation at 55°C for 1h. PCR was performed using the ThermoScientific DreamTaq Green PCR Master Mix (2X) kit with the following oligos: 5-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA-3and 5-CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

T-3. A band of approximately 120 bp was gel-purified and sent for Sanger sequencing with the primer 5-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3.

Northern Blot

RNAs were run for 1 hr at 180V in a 7M Urea 10% acrylamide gel and then transferred to a nylon membrane (Invitrogen) using XCell SureLock Mini-Cell system with the XCell II Blot Module (Invitrogen). The membrane was cross-linked 5 min with UV. After prehybridization with 30 ml of solution containing 6 X SSC, 0.5% SDS and 0.05 % casein for 1h at 45°C, the membrane was hybridized with 20 µM FAM or IR dye-labeled oligonucleotide probes (Supplementary Table 3) in 10 ml of prehybridization solution at 45°C overnight. Finally, the membrane was washed twice with 2 X SSC, 0.1% SDS for 10 min at RT and twice with 0.2 X SSC, 0.1% SDS for 10 min at 60°C and images captured using a ChemiDoc MP Imaging system (BIO-RAD). The sequence of the *E. coli* $tRNA^{Lys}$ probe (B803) and T5 $tRNA^{Lys}$ probe (B806), as well as their homology to the *E. coli* and T5 $tRNA^{Lys}$ are shown in Extended Data Figure 6.11a. To confirm the specificity of the T5 $tRNA^{Lys}$ probe (B806), a Northern Blot was performed with this probe in the absence of the T5 $tRNA^{Lys}$, showing that our conclusion cannot be affected by non-specific binding of this probe to the *E. coli* $tRNA^{Lys}$ or other tRNA (**Extended Data Figure 6.11b**). We can also confirm that the B803 probe used to detect the *E. coli* $tRNA^{Lys}$ does not detect the T5 $tRNA^{Lys}$, as the signal of this probe is lost when PARIS is activated in the presence of the T5 $tRNA^{Lys}$ (**Figure 6.5h**).

Defense System Detection

Defense system detection was performed using DefenseFinder v1.1.1[61] and DefenseFinder models v1.2.3 on all RefSeq complete genomes from July 2022.

Domain Annotation

Defense system domain annotation was done using HHpred[62] webserver with the PFAM [63] database v33.1 and standard setting. For PARIS, trigger binding domain and Coiled-coil domain were annotated using the structure information and ESMfold[64]structure prediction for AriAB.

AriA and AriB Tree

PARIS systems detected by DefenseFinder v1.1.1 on RefSeq complete genomes from July 2022 were used to build phylogenetic trees. Only hits to AriA and AriB with a coverage of more than 75% were kept to avoid pseudogenes. Hits were then clustered using Mmseqs2[65] v13.45111 on AriA and AriAB with an identity threshold of 80% and coverage threshold of 90%. A single representative of each cluster was then used to build the AriA tree, and the cognate AriB was used to build the AriB tree. The AriA tree was built using the AAA15 ATPase domain only and the AriB tree using the DUF4435 domain only. Domains were detected by HMMsearch (from HMMER 3.3.2) on the *E. coli* B185 PARIS system and a first alignment using Mafft v7.505[66] (default parameters) was used to extract the corresponding sequences in other PARIS variants. The final alignment used to make the tree was done using MUSCLE v.5.1.linux64[67] using the model super5 and trim using clipkit v1.3.0cwith option seq-gap. The DUF4435 tree using M5 Ribonuclease as outgroup was done using selected representative hits from different clades and aligned using muscle v5.1.linux64 with model super5. All trees were built using Iqtree v2.0.6[68] using ModelFinder and Ultrafast Bootstrap.

Defense System ATPase AAA15/21 Tree

ATPase AAA15/21 defense system containing detection was done using hmmsearch from HMMER 3.3.2 and Pfam Hidden Markov Model 33.1 on defense systems previously detected[69]. Random selection of 20 proteins from the detection results with a coverage of more than 75% were selected in addition to experimentally validated homologs to build the tree. The ATPase alignment was done using Mafft v7.505 and the tree was built using Iqtree v2.0.6 using ModelFinder and Ultrafast Bootstrap 1000.

Statistics and Reproducibility

Pull-down experiments presented in Figures 3a, b, e, f, i were reproduced in two or three biological replicates. Experiments presented in Figures 5g, i, and j were reproduced three times, and Figure 5h two times.

## References

1. Chevallereau, A. & Westra, E. R. Bacterial immunity: Mobile genetic elements are hotspots for defence systems. *Current Biology* **32**, R923–R926 (2022).

2. Rousset, F. *et al.* Phages and their satellites encode hotspots of antiviral systems. *Cell Host & Microbe* **30**, 740-753.e5 (2022).

3. Zhang, T. *et al.* Direct activation of a bacterial innate immune system by a viral capsid protein. *Nature* **612**, 132–140 (2022).

4. Wein, T. & Sorek, R. Bacterial origins of human cell-autonomous innate immune mechanisms. *Nat Rev Immunol* **22**, 629–638 (2022).

5. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *Journal of Bacteriology* **193**, 6039–6056 (2011).

6. Rocha, E. P. C. & Bikard, D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biology* **20**, e3001514 (2022).

7. Schiltz, C. J., Lee, A., Partlow, E. A., Hosford, C. J. & Chappie, J. S. Structural characterization of Class 2 OLD family nucleases supports a two-metal catalysis mechanism for cleavage. *Nucleic Acids Research* **47**, 9448–9463 (2019).

8. Schiltz, C. J., Adams, M. C. & Chappie, J. S. The full-length structure of Thermus scotoductus OLD defines the ATP hydrolysis properties and catalytic mechanism of Class 1 OLD family nucleases. *Nucleic Acids Research* **48**, 2762–2776 (2020).

9. Haggård-Ljungquist, E., Barreiro, V., Calendar, R., Kurnit, D. M. & Hans, C. The P2 phage *old* gene: sequence, transcription and translational control. *Gene* **85**, 25–33 (1989).

10. Wilkinson, M. *et al.* Structural basis for the inhibition of RecBCD by Gam and its synergistic antibacterial effect with quinolones. *eLife* **5**, e22963 (2016).

11. Dot, E. W., Thomason, L. C. & Chappie, J. S. Everything OLD is new again: How structural, functional, and bioinformatic advances have redefined a neglected nuclease family. *Molecular Microbiology* **120**, 122–140 (2023).

12. Cheng, R. *et al.* A nucleotide-sensing endonuclease from the Gabija bacterial defense system. *Nucleic Acids Research* **49**, 5216–5229 (2021).

13. Antine, S. P. *et al.* Structural basis of Gabija anti-phage defence and viral immune evasion. *Nature* **625**, 360–365 (2024).

14. Mestre, M. R., González-Delgado, A., Gutiérrez-Rus, L. I., Martínez-Abarca, F. & Toro, N. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Research* **48**, 12632–12647 (2020).

15. Ernits, K. *et al.* The structural basis of hyperpromiscuity in a core combinatorial network of type II toxin–antitoxin and related phage defense systems. *Proceedings of the National Academy of Sciences* **120**, e2305393120 (2023).

16. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13**, 1265 (2022).

17. Hopfner, K.-P. & Tainer, J. A. Rad50/SMC proteins and ABC transporters: unifying concepts from high-resolution structures. *Current Opinion in Structural Biology* **13**, 249–255 (2003).

18. Hopfner, K.-P. *et al.* Structural Biology of Rad50 ATPase: ATP-Driven Conformational Control in DNA Double-Strand Break Repair and the ABC-ATPase Superfamily. *Cell* **101**, 789–800 (2000).

19. Lammens, A. & Hopfner, K.-P. Structural Basis for Adenylate Kinase Activity in ABC ATPases. *Journal of Molecular Biology* **401**, 265–273 (2010).

20. Xie, T., Zhang, Z., Fang, Q., Du, B. & Gong, X. Structural basis of substrate recognition and translocation by human ABCA4. *Nat Commun* **12**, 3853 (2021).

21. Krishnan, A., Burroughs, A. M., Iyer, L. M. & Aravind, L. Comprehensive classification of ABC ATPases and their functional radiation in nucleoprotein dynamics and biological conflict systems. *Nucleic Acids Research* **48**, 10045–10075 (2020).

22. Gut, F. *et al.* Structural mechanism of endonucleolytic processing of blocked DNA ends and hairpins by Mre11-Rad50. *Mol Cell* **82**, 3513-3522.e6 (2022).

23. Seifert, F. U., Lammens, K., Stoehr, G., Kessler, B. & Hopfner, K. Structural mechanism of ATP-dependent DNA binding and DNA end bridging by eukaryotic Rad50. *The EMBO Journal* **35**, 759–772 (2016).

24. Higgins, C. F. & Linton, K. J. The ATP switch model for ABC transporters. *Nat Struct Mol Biol* **11**, 918–926 (2004).

25. Walkinshaw, M. D. *et al.* Structure of Ocr from Bacteriophage T7, a Protein that Mimics B-Form DNA. *Molecular Cell* **9**, 187–194 (2002).

26. Isaev, A. *et al.* Phage T7 DNA mimic protein Ocr is a potent inhibitor of BREX defence. *Nucleic Acids Research* **48**, 5397–5406 (2020).

27. Deep, A. *et al.* The SMC-family Wadjet complex protects bacteria from plasmid transformation by recognition and cleavage of closed-circular DNA. *Molecular Cell* **82**, 4145-4159.e7 (2022).

28. Weiß, M. *et al.* The MksG nuclease is the executing part of the bacterial plasmid defense system MksBEFG. *Nucleic Acids Research* **51**, 3288–3306 (2023).

29. Aravind, L., Leipe, D. D. & Koonin, E. V. Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Research* **26**, 4205–4213 (1998).

30. Zimmerman, S. B. Toroidal nucleoids in *Escherichia coli* exposed to chloramphenicol. *Journal of Structural Biology* **138**, 199–206 (2002).

31. Yirmiya, E. *et al.* Phages overcome bacterial immunity via diverse anti-defence proteins. *Nature* **625**, 352–359 (2024).

32. Wang, J. *et al.* Complete genome sequence of bacteriophage T5. *Virology* **332**, 45–65 (2005).

33. Glukhov, A. S., Krutilina, A. I., Kaliman, A. V., Shlyapnikov, M. G. & Ksenzenko, V. N. Bacteriophage T5 Mutants Carrying Deletions in tRNA Gene Region. *Mol Biol* **52**, 1–6 (2018).

34. Jain, I. *et al.* tRNA anticodon cleavage by target-activated CRISPR-Cas13a effector. 2021.11.10.468108 Preprint at https://doi.org/10.1101/2021.11.10.468108 (2021).

35. Jiang, Y., Meidler, R., Amitsur, M. & Kaufmann, G. Specific interaction between anticodon nuclease and the tRNALys wobble base1. *Journal of Molecular Biology* **305**, 377–388 (2001).

36. Azam, A. H. *et al.* Viruses encode tRNA and anti-retron to evade bacterial immunity. 2023.03.15.532788 Preprint at https://doi.org/10.1101/2023.03.15.532788 (2023).

37. Mogila, I. *et al.* Ribosomal stalk-captured CARF-RelE ribonuclease inhibits translation following CRISPR signaling. *Science* **382**, 1036–1041 (2023).

38. Bingham, R., Ekunwe, S. I. N., Falk, S., Snyder, L. & Kleanthous, C. The Major Head Protein of Bacteriophage T4 Binds Specifically to Elongation Factor Tu*. *Journal of Biological Chemistry* **275**, 23219–23226 (2000).

39. Yang, J. Y. *et al.* Degradation of host translational machinery drives tRNA acquisition in viruses. *cels* **12**, 771-779.e5 (2021).

40. van den Berg, D. F., van der Steen, B. A., Costa, A. R. & Brouns, S. J. Phage tRNAs evade tRNA-targeting host defenses through anticodon loop mutations. *eLife* **12**, e85183 (2023).

41. Albers, S. & Czech, A. Exploiting tRNAs to Boost Virulence. *Life (Basel)* **6**, 4 (2016).

42. Zhang, F. *et al.* Human SAMD9 is a poxvirus-activatable anticodon nuclease inhibiting codon-specific protein synthesis. *Science Advances* **9**, eadh8502 (2023).

43. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345 (2009).

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

45. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290–296 (2017).

46. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).

47. Sanchez-Garcia, R. *et al.* DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun Biol* **4**, 1–8 (2021).

48. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519–530 (2018).

49. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst D* **75**, 861–877 (2019).

50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Cryst D* **66**, 486–501 (2010).

51. Rohwer, F. & Azam, F. Detection of DNA Damage in Prokaryotes by Terminal Deoxyribonucleotide Transferase-Mediated dUTP Nick End Labeling. *Applied and Environmental Microbiology* **66**, 1001–1006 (2000).

52. FlowJo Software. Becton, Dickinson and Company (2023).

53. Cockram, C., Thierry, A. & Koszul, R. Generation of gene-level resolution chromosome contact maps in bacteria and archaea. *STAR Protocols* **2**, 100512 (2021).

54. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).

55. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999–1003 (2012).

56. Lioy, V. S. *et al.* Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell* **172**, 771-783.e18 (2018).

57. Kurata, T. *et al.* A hyperpromiscuous antitoxin protein domain for the neutralization of diverse toxin domains. *Proceedings of the National Academy of Sciences* **119**, e2102212119 (2022).

58. Sambrook, J. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989).

59. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13**, e1005595 (2017).

60. Avcilar-Kucukgoze, I., Gamper, H., Hou, Y.-M. & Kashina, A. Purification and Use of tRNA for Enzymatic Post-translational Addition of Amino Acids to Proteins. *STAR Protocols* **1**, 100207 (2020).

61. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* **13**, 2561 (2022).

62. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**, W244–W248 (2005).

63. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).

64. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

65. Mirdita, M., Steinegger, M. & Söding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **35**, 2856–2858 (2019).

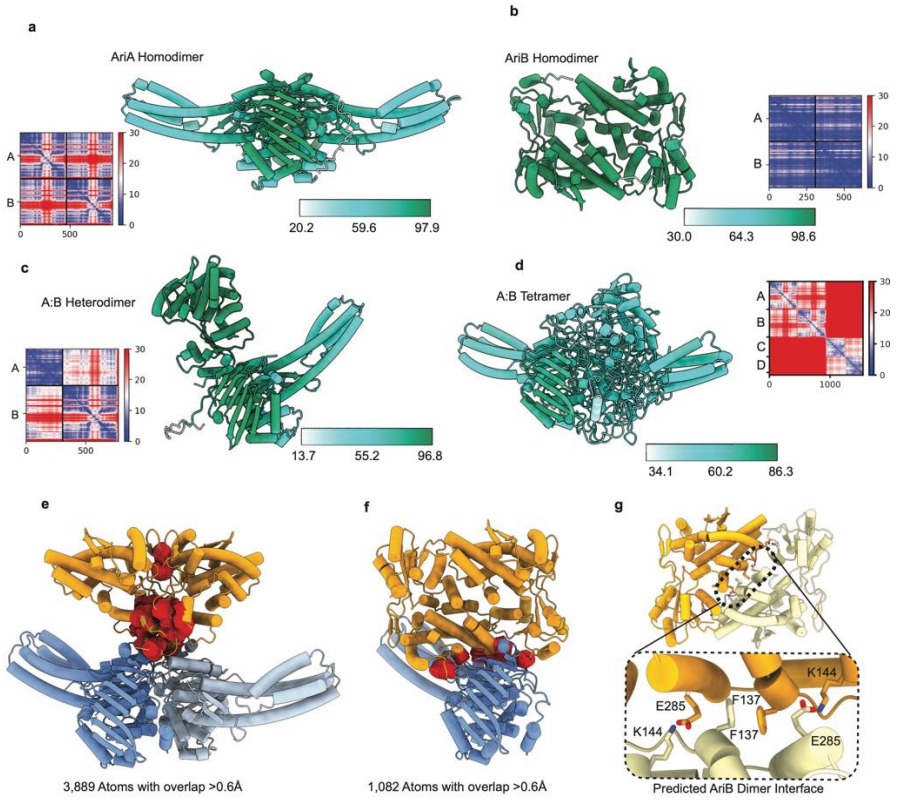
66. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).

67. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).

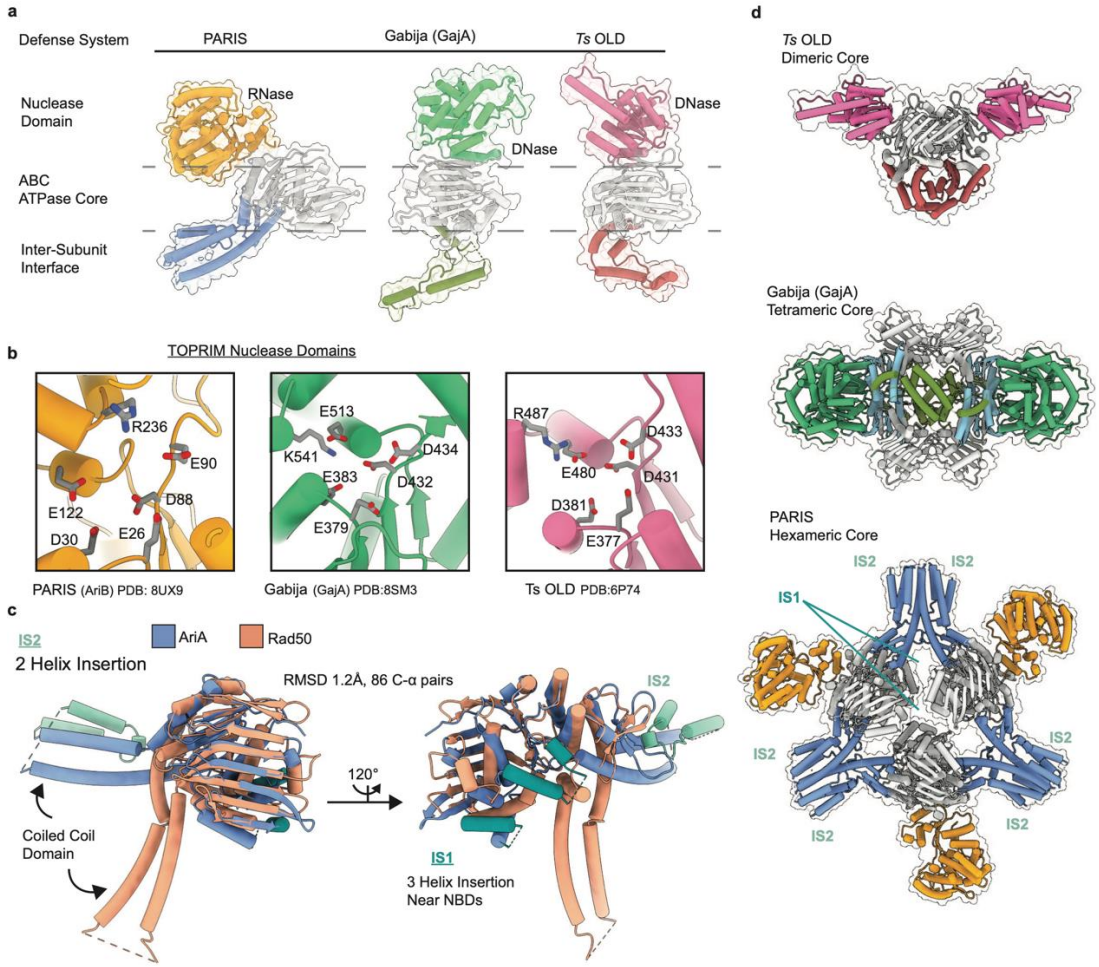
68. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* vol. 32 268–274 (2015).

69. Wiegand, T. *et al.* Functional and Phylogenetic Diversity of Cas10 Proteins. *The CRISPR Journal* **6**, 152–162 (2023).

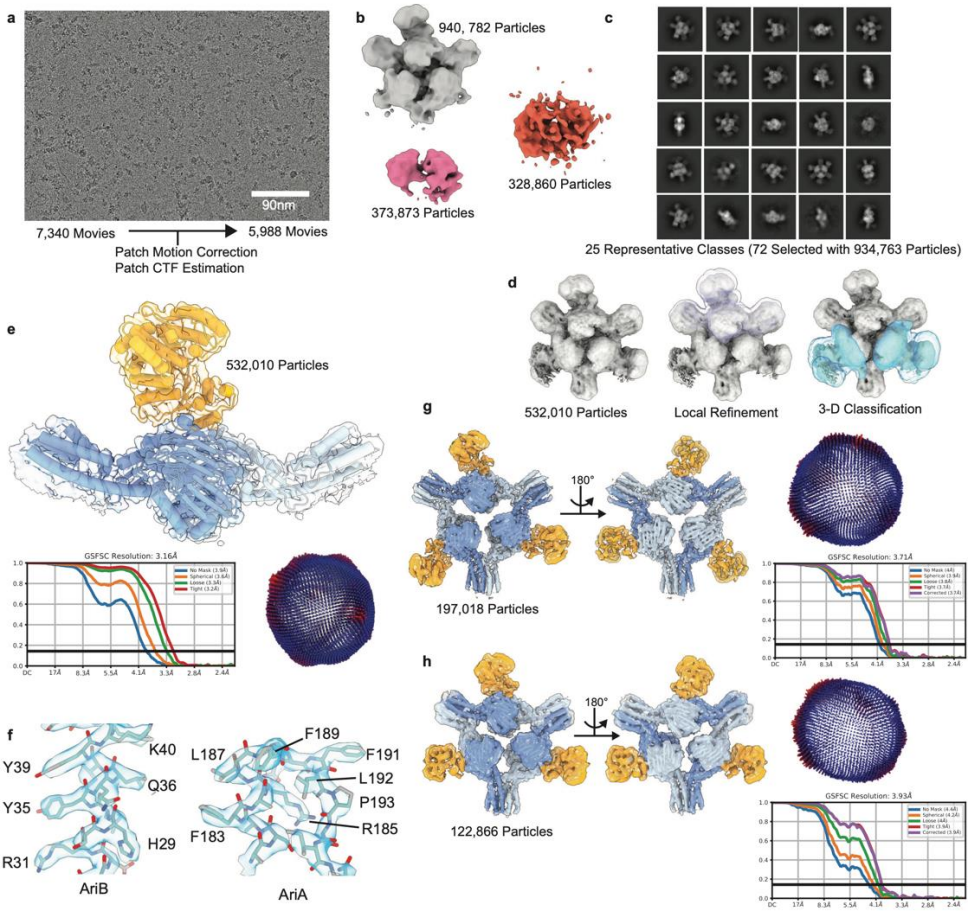
Extended Data Figures



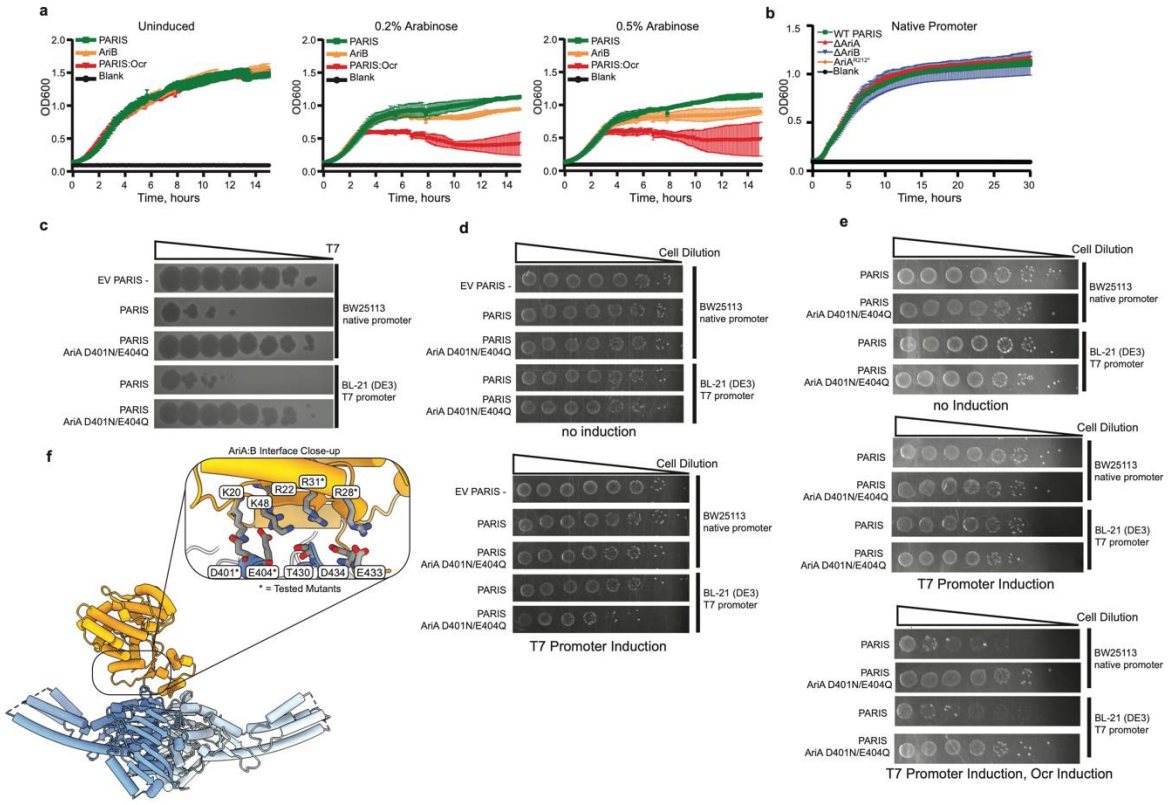

Extended Data Figure 6.1. AlphaFold2 structural predictions of AriA and AriB. AlphaFold2-predicted structures of AriA and AriB oligomers colored by the predicted local distance difference test (pLDDT) score (0-100) with associated scale bars. Values greater than 90 indicate high confidence and values below 50 indicate low confidence. Predicted Aligned Error plots (PAE) are also shown for each structural prediction. a, AlphaFold2-predicted structure of the AriA homodimer. Confidence is high (PLDDT >90, green) for the ATPase domain, and low for the insertion sequences (PLDDT<70, teal). b, AlphaFold2-predicted structure of the AriB homodimer results in a high confidence model with more >98% of the residues above a PLDDT of 90. c, AlphaFold2-predicted structure of the AriA-B heterodimer and associated PAE plot. d, AlphaFold2 returned unreliable models when attempting to predict structures for higher ordered assemblies of AriA and AriB as demonstrated by the PLDDT plot below 50 for most of the molecule. e, Clashing (red) between AriBs, prevents the assembly of two AriB subunits on a single AriA homodimer. Clashing is defined by atoms with overlap greater than 0.6 Å. f, A predicted homodimer of AriB cannot associate with a homodimer of AriA without clashing (red). Atoms with an overlap greater than 0.6 Å are highlighted with red disks signifying the clash. g, Predicted structure of the AriB dimer shown in panel b, highlighting residues involved in dimerization.



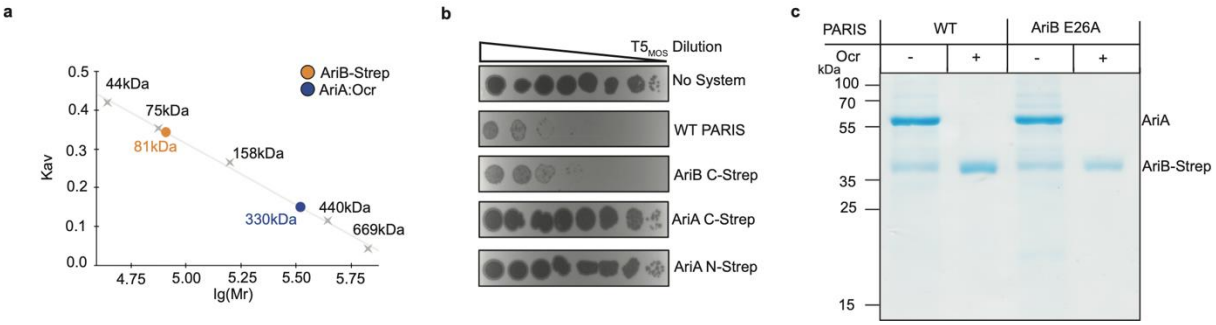

Extended Data Figure 6.2. Structural comparison of PARIS homologs. a, Domain organization of phage defense systems where an ABC ATPase domain is associated with a TOPRIM nuclease domain. While the ATPase domain is highly conserved, the relative orientation of the TOPRIM nuclease domains and dimerization domains vary between PARIS (PDB ID: 8UX9), Gabija (PDB ID: 8SM3), and the Thermus scotoductus Overcoming Lysogenization Defect (Ts OLD, PDB ID: 6P74). b, AriB is the effector of PARIS defense. The AriB TOPRIM is homologous to the TOPRIM domains of Ts OLD (RMSD 1.1 Å, 20 C-alpha pairs) and GajA (RMSD 1.02 Å, 26 C-alpha pairs). The TOPRIM domain of AriB is within the DUF4435, which plays a role in AriB dimerization. c, Structural comparison of AriA to Rad50, a universally conserved ATPase. AriA primarily differs from Rad50 at Insertion Sequence (IS) 1 (residues 122-180) and IS2 (residues 242-289). IS1 introduces three alpha helices near the nucleotide binding domains of AriA that are predicted to be involved in trigger recognition. IS2 expands the coiled coil domain with two alpha helices and enable the homohexameric assembly of AriA subunits in the PARIS complex. d, Structural comparison of PARIS and related OLD systems highlights the unique assembly state of PARIS.



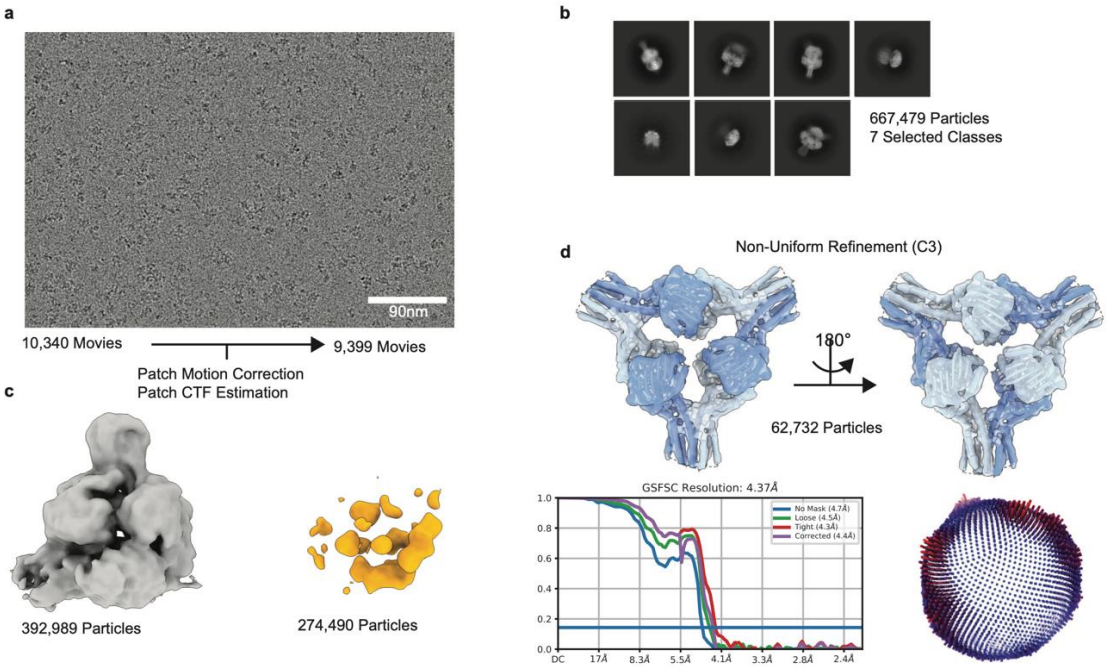

Extended Data Figure 6.3. Workflow for structural determination of the PARIS complex. Image processing was performed using cryoSPARC v4.41. a, 7,340 movies were recorded. Micrographs with CTF-fits worse than 8 Å were removed prior to downstream processing (n=5,988). Using a de novo template generated from a 200-micrograph subset of this data, 4,078,384 particles were identified and extracted. From 17 selected classes, 1,643,515 particles remained, and a 3-class ab initio reconstruction yielded volumes shown in panel b. Particles belonging to junk classes (pink and red) were discarded. c, 25 representative 2-D Classes from 940,782 particles in panel b. A total of 934,763 particles from 72 classes remained after final 2-D Classification (n=100 classes). d, After successive rounds of ab initio and heterogenous refinements, a 532,010 particle volume was obtained, corresponding to the intact PARIS complex with compositional heterogeneity at two AriB attachment sites. Masks were generated for local refinement and 3-D classification to produce the reconstructions shown in e, g, and h, respectively. e, Local Refinement of one asymmetric unit of the PARIS immune complex with FSC and 3-D orientation plot shown below. f, Close up of density map and model from E demonstrating map quality. g, C3 reconstruction of the PARIS complex at 3.71 Å-resolution with the experimental structure determined in panel e fit into the density. h, Cryo-EM reconstruction of PARIS with AriB in the ‘trans’ arrangement at approximately 3.93 Å-resolution with experimental structure determined in panel e, fit into the density.



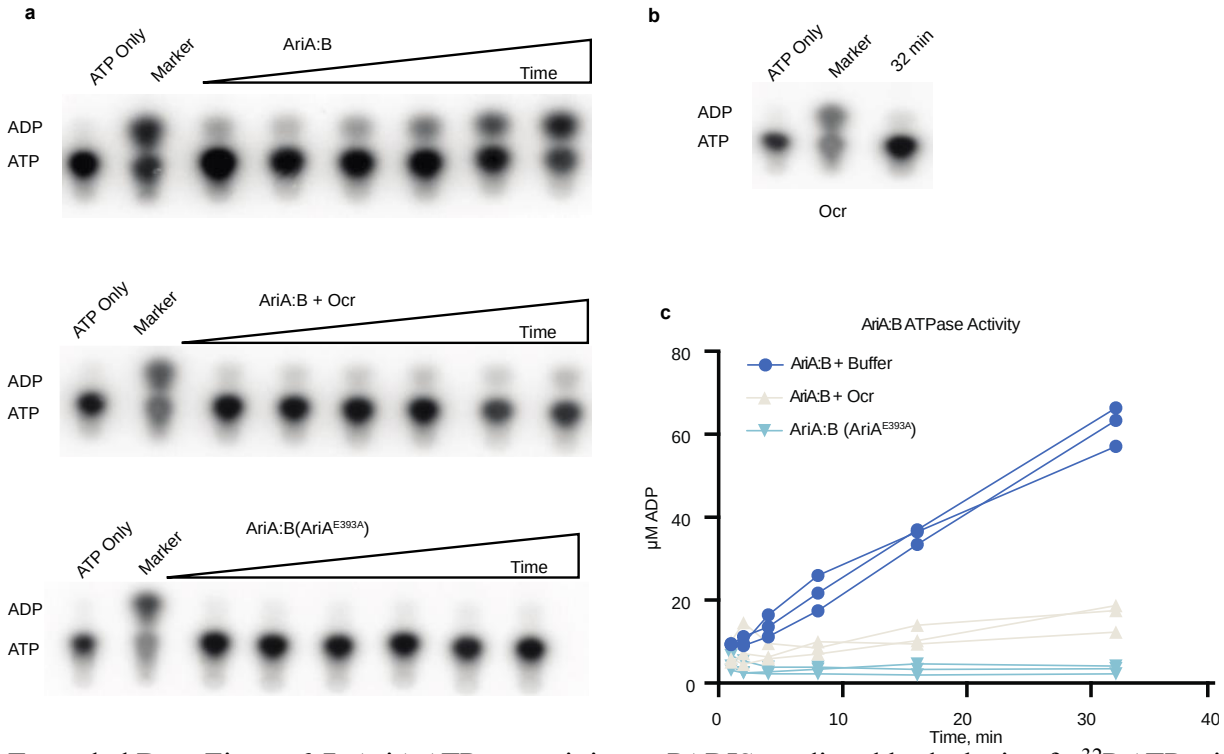

Extended Data Figure 6.4. AriB toxicity is dependent on association with AriA. a, Liquid culture toxicity assays of AriB alone or AriA:B together (PARIS). Genes are overexpressed using a T7 inducible promoter. AriB alone results in a mild growth defect, while Ocr-triggered release of AriB (PARIS Ocr) is significantly more toxic. b, Growth of *E. coli* MG1655 carrying plasmid pFR85 (pSC101 with *ariAB* under the control of the native promoter), or variants with *ariA* deleted, *ariB* deleted or a stop codon in *ariA* yielding an interrupted protein that lacks the interaction interface with AriB. c, Structure-guided mutations that disrupt the AriA:AriB interface result in a loss of defense phenotype. d, Drop-test toxicity assay with PARIS (WT) or PARIS with AriA D401N/E404Q mutation. A modest growth defect can be observed only for the PARIS AriA D401N/E404Q overexpressed from T7 promoter. e, The addition of pBAD Ocr to the PARIS AriA mutants does not increase the toxicity, indicating that AriB is not functional when AriA interactions are disrupted. f, Close-up of the AriA:AriB interaction interface. Mutated residues are labeled with asterisks.



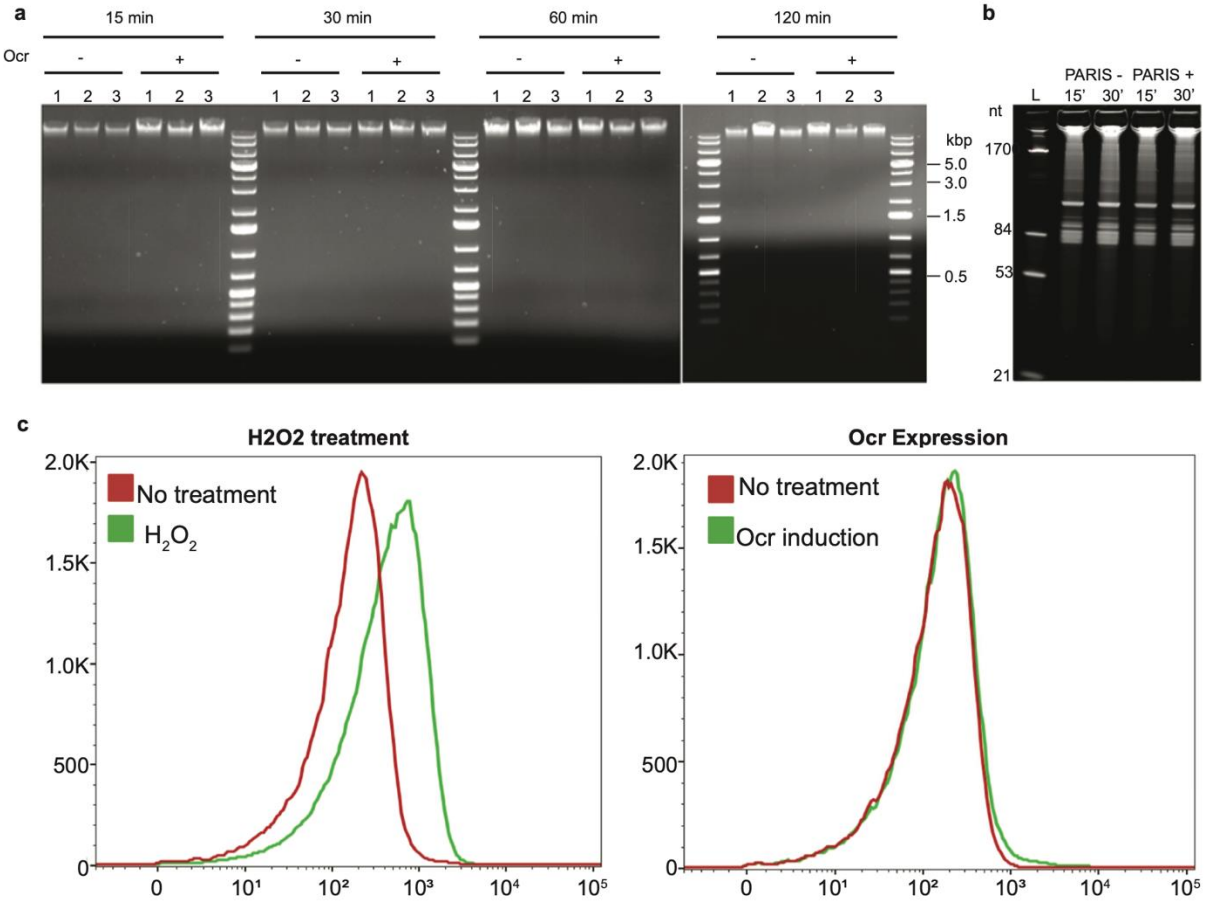

Extended Data Figure 6.5. Purification of activated AriB. a, Size exclusion chromatography calibration curve with estimated masses of AriA:Ocr and activated AriB complexes. b, Introduction of the Strep-tag at the C- or N- terminus of AriA, but not on the C-terminus of AriB, interferes with PARIS anti-phage defense. c, Purification of activated AriB after mixing cell lysates containing strep-tagged AriB and non-tagged Ocr.



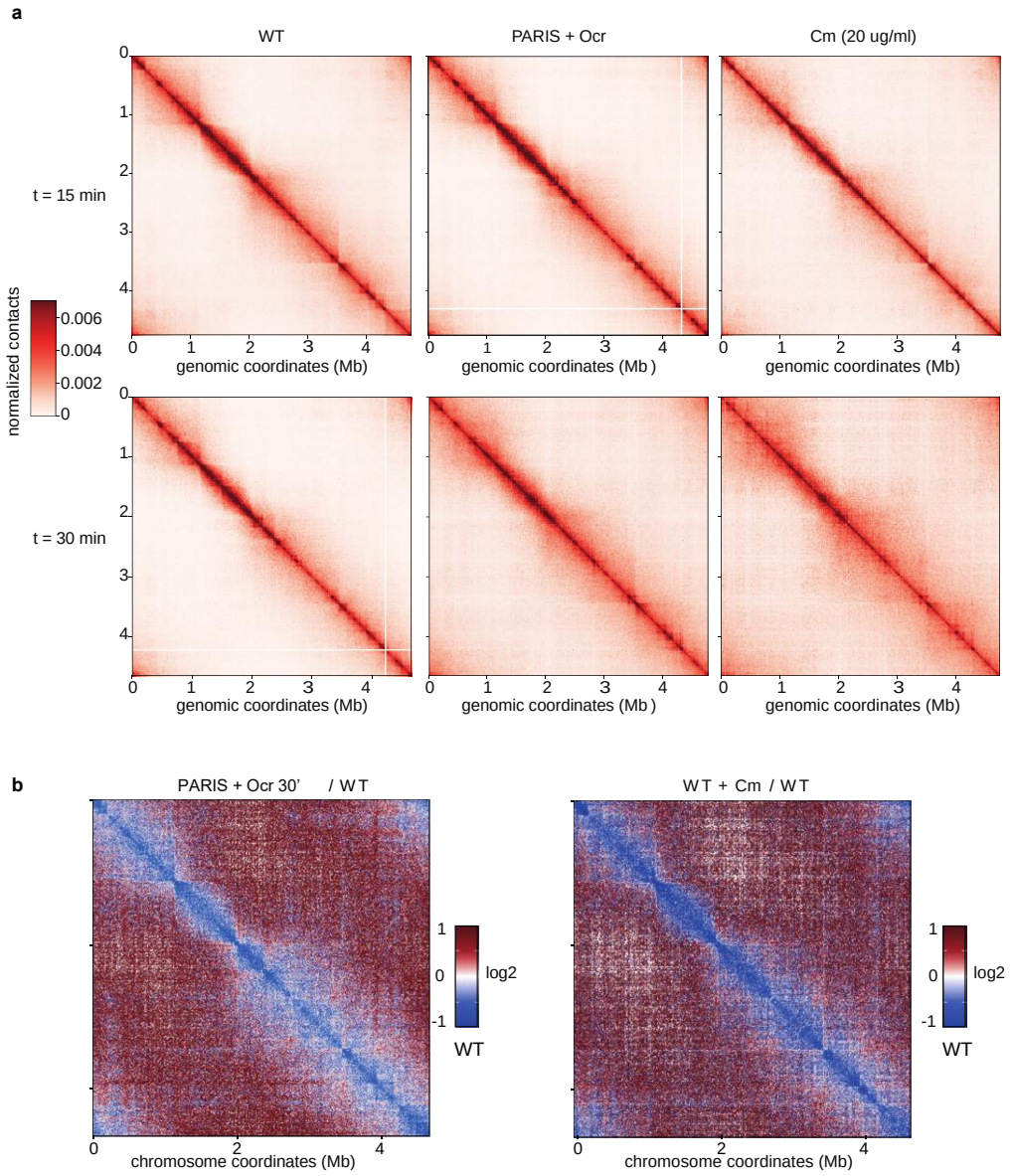

Extended Data Figure 6.6. Structure determination of AriA purified using a Ocr-Strep pulldown. Image processing and analysis was performed in cryoSPARC v4.41. a, Representative micrograph from a total of 10,340 movies. Micrographs with CTF-fits worse than 8 Å-resolution were discarded. Particle picking on 9,399 micrographs identified 4,415,782 particles which were extracted and subjected to 2-D Classification. b, Selected 2-D classes corresponding to the AriA hexamer contained 667,479 particles. c, 2-Class ab initio reconstruction of particles from panel b. The gray class corresponds to the AriA homohexamer, while particles associated with the orange class were discarded. Density for the ATPase domains is present, but one blade of the propellor is poorly resolved due to the increased flexibility of the AriA scaffold relative to the fully assembled PARIS complex. d, After iterative rounds of sorting the 392,989 particles from panel C, a stack containing 62,732 particles was isolated and C3 refinement produced a 4.4 Å-resolution reconstruction of the AriA scaffold.



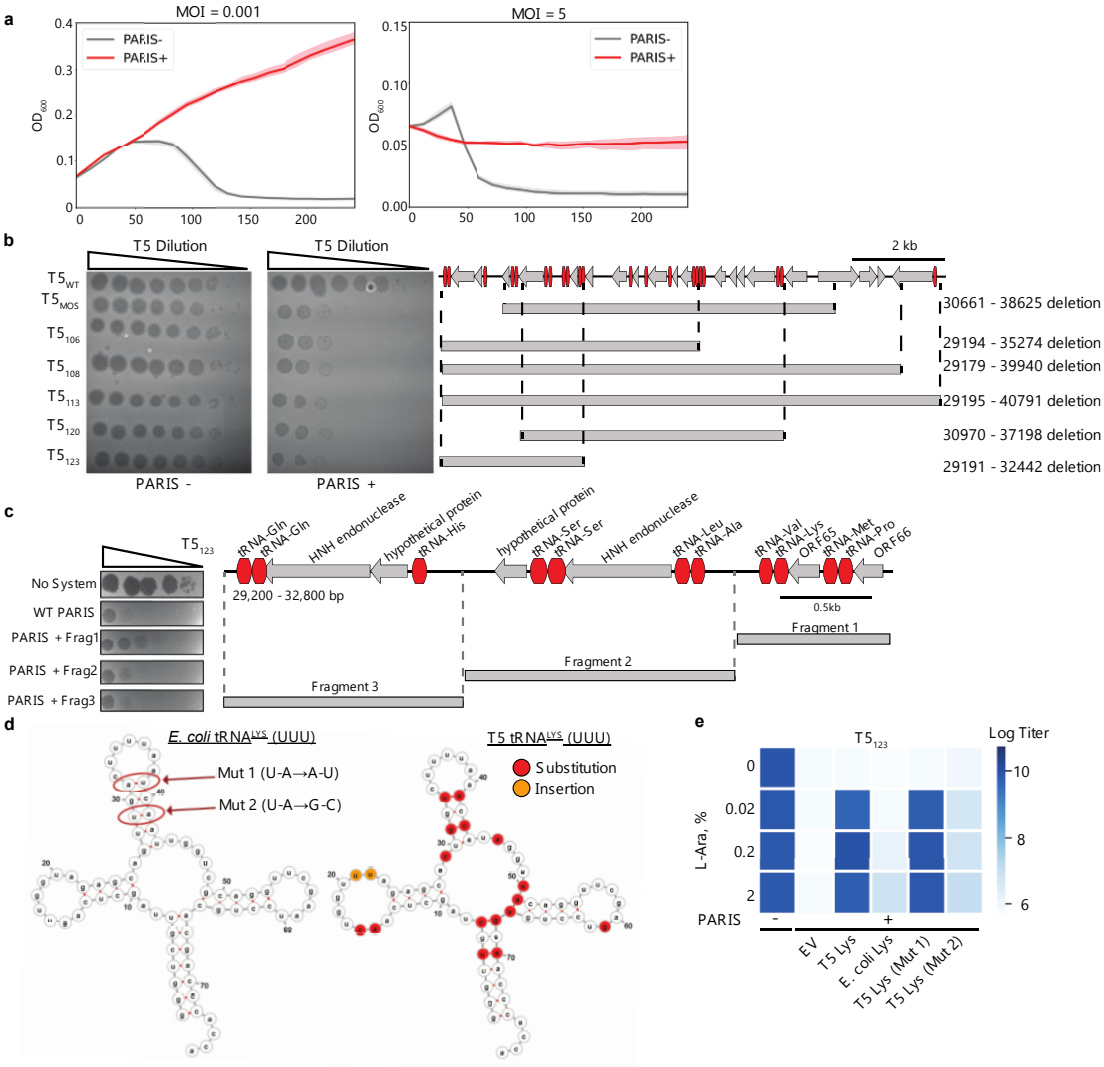

Extended Data Figure 6.7. AriA ATPase activity. a, PARIS mediated hydrolysis of $\alpha^{32}$P ATP with PARIS alone or with 10-fold excess trigger (T7 Ocr). Reactions were run from 1 to 32 minutes and products were resolved via thin layer chromatography. Reactions with $\alpha^{32}$P ATP incubated with T4 polynucleotide kinase were used as a positive control. Images are representative of reactions that were performed in triplicate. b, Ocr alone was incubated with $\alpha^{32}$P ATP for 32 minutes indicating that the trigger alone does not hydrolyze ATP. c, Concentration of ADP formed at each time point was quantified and plotted. Data points of three technical replicates are shown.



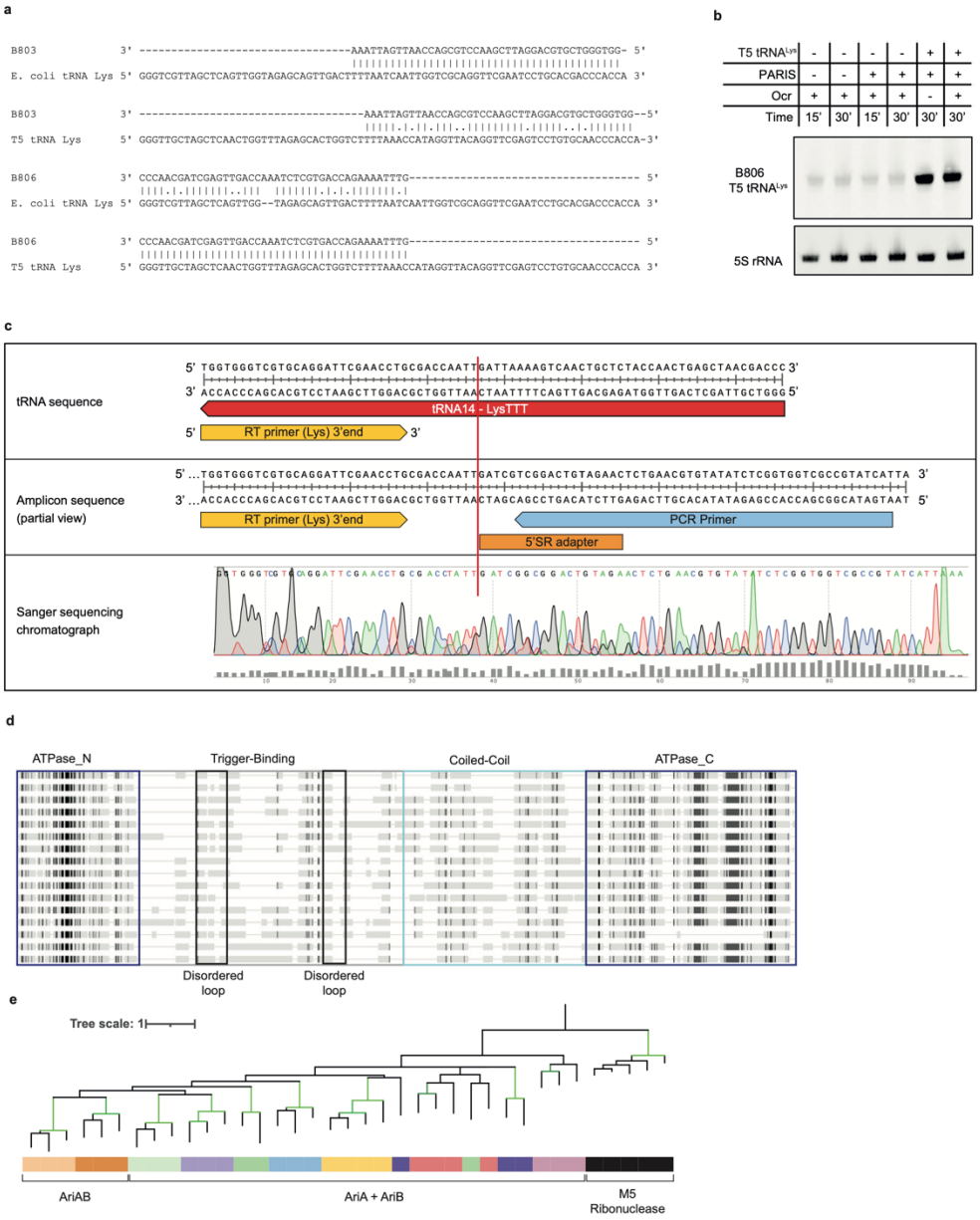

Extended Data Figure 6.8. PARIS activation does not result in total RNA or DNA degradation. a, Total DNA was extracted from E. coli MG1655 carrying plasmids pFR85 (PARIS) and pFD250 (PPhlFOcr) at different time points after Ocr induction with or without DAPG (n=3), or from a non-induced control. Samples were run on a 1% agarose TAE gel. b, Total RNA was extracted from E. coli MG1655 carrying plasmid pFR85 (PARIS) or pFR66 (sfGFP), and pFD250 (PPhlF-Ocr) at different time points after induction of Ocr expression with DAPG. Samples were run on a TBE Urea (7M) acrylamide (10%) gel and stained with SYBR-Gold. A representation of 3 replicates is shown. c, TUNEL assay, demonstrating the lack of accumulation of dsDNA breaks in PARIS+ cells 2 hours after Ocr induction, as measured through terminal deoxynucleotidyl transferase labeling of free 3'-OH groups in DNA. As a positive control of DNA damage, cells were treated with 0.1% H2O2.




Extended Data Figure 6.9. PARIS activation results in DNA compactization akin to inhibition of translation with chloramphenicol. a, Hi-C contact maps (bin: 4kb). From left to right: WT (pFR66 + pFD245 control vectors), PARIS (pFR85) + PPhlF-ocr (pFD250), and chloramphenicol treated cells. Top and bottom rows correspond to 15 and 30 minutes after induction with DAPG or treatment with chloramphenicol. b, Ratio between contact maps of WT at t = 30 min and contact maps 30 min after either PARIS activation (left) or chloramphenicol treatment (right).




Extended Data Figure 6.10. Identification of the T5 genomic region required for infection of PARIS+ cells. a, Growth of PARIS- or PARIS+ cells infected with PARIS-sensitive phage T5Mos at low or high MOI. The phage was added at t=0. b, Phage T5 variants with non-essential deletions in the region encoding tRNAs are sensitive to PARIS defense. Boundaries of the deletions are shown on the right, tRNA genes are highlighted in red. c, Overexpression of Fragment 1 (31885-32870) derived from the deletion in the phage $T5_{123}$ partially rescues $T5_{123}$ infection of the PARIS+ culture. d, Comparison of E. coli and T5 tRNALys highlights substitution mutations near the anticodon stem loop that we hypothesize prevents targeting by AriB. e, Swapping U-A for A-U (Mut1) on the E. coli tRNALys confers resistance to PARIS.




Extended Data Figure 6.11. tRNALys cleavage product and phylogenetics of PARIS. a, Alignment of the B803 and B806 Northern Blot probes to the E. coli and T5 tRNALys. b, Northern Blot with probe B806 in the presence or absence of the T5 tRNALys expressed from a pBAD. c, Mapping of the tRNALys cleavage site by reverse-transcription, adapter ligation, PCR and Sanger sequencing. The cleavage site, identified as the junction between the tRNA sequence and the ligated adapter, is marked with a vertical red line. d, Alignment of homologs of AriA representative of the different PARIS clades. Different domains of AriA are indicated. Grey scale represents % of identity between homologs. **b**, Phylogenetic tree of DUF4435 domain of AriB using M5 Ribonuclease (TOPRIM) as an outgroup.

Extended Data Tables

| | PARIS Asymmetric Unit (PDB ID: 8UX9, EMD-42719) | PARIS C3 Arrangement (EMD-43103) | PARIS 'trans' Arrangement (EMD-43104) | AriA-Ocr (EMD-43105) |
|---|---|---|---|---|
| | **Data Collection and Processing** | | | |
| Microscope | Talos Arctica | | | |
| Voltage | 200kV | | | |
| Camera | Gatan K3 | | | |
| Magnification | 36,000 X | | | |
| Pixel size (Å) | 0.552 | 0.552 | 0.552 | 0.552 |
| Total Dose ($e$/Å$^2$) | 56.42 | 56.42 | 56.42 | 56.31 |
| Exposure time (s) | 4.993 | 4.993 | 4.993 | 4.107 |
| Frames/Exposure | 50 | 50 | 50 | 50 |
| Defocus Range (μm) | -0.6, -2.5 | -0.6, -2.5 | -0.6, -2.5 | -0.6, -2.5 |
| Number of Micrographs | 7,340 | 7,340 | 7,340 | 10,034 |
| Number of Initial Particles | 4,078,384 | 4,078,384 | 4,078,384 | 667,479 |
| Number of particles for 3-D Analysis | 532,010 | 532,010 | 532,010 | 100,862 |
| Symmetry | C1 | C3 | C1 | C3 |
| Number of Final Particles | 532,010 | 197,018 | 122,866 | 62,732 |
| Resolution (0.143 FSC, Å) | 3.2 | 3.7 | 3.9 | 3.7 |
| | **Model Building and Refinement Statistics** | | | |
| Refinement Software | Phenix. real_space_refine | | | |
| Clash score | 5.75 | | | |
| MolProbity score | 1.71 | | | |
| Poor rotamers | 24 (4%) | | | |
| Favored rotamers | 529 (96%) | | | |
| Ramachandran outliers | 1 (.09%) | | | |
| Ramachandran Favored | 1019 (94%) | | | |
| Bad bonds | 2 | | | |
| Bad Angles | 4 | | | |

Extended Data Table 6.1. Cryo-EM data collection, processing, and model validation. This table lists the refinement statistics for cryo-EM structures of PARIS.

Supplementary Tables

| ***E. coli* strain** | **Comment** | | **Source** |
|---|---|---|---|
| BW25113 | *E. coli* K12 F– Δ(*araD-araB*)567 Δ*lacZ*4787(::*rrnB*-3) λ– *rph*-1 Δ(*rhaD-rhaB*)568 *hsdR*514 | | Lab stock |
| MG1655 | *E. coli* K12 F-, λ-, *ilvG*-, *rfb*-50, *rph*-1 | | Lab stock |
| XL1-Blue | *E. coli* K12 *recA*1 *endA*1 *gyrA*96 *thi*-1 *hsdR*17 *supE*44 *relA*1 *lac* [F'*proAB lacIqZ*ΔM15 Tn10], TetR | | Evrogen |
| **Phage** | **Comment** | | **Source** |
| T7 | | | Lab stock |
| T5 | | | Lab stock |
| $T5_{Mos}$ | T5 with 30661-38625 deletion | | [33] |
| $T5_{106}$ | T5 with 29194-35274 deletion | | |
| $T5_{108}$ | T5 with 29179-39940 deletion | | |
| $T5_{113}$ | T5 with 29195-40791 deletion | | |
| $T5_{120}$ | T5 with 30970-37198 and 24956-26633 deletions | | |
| $T5_{123}$ | T5 with 29191-32442 deletion | | |
| **Plasmid Name** | **Plasmid Map** | **Comment** | **Source** |
| pFR66 | https://benchling.com/s/seq-eyqx3mULGDzw04wc28W7?m=slm-KMqDsXL62sGpcvkRDd08 | sfGFP, Ptet promoter, $Kan^R$ | [2] |

| | | | |
|---|---|---|---|
| pFR85 (PARIS) | https://benchling.com/s/seq-P0S4QJUYX0D54AVNp81e?m=slm-DWv2SMSVbNQMxjbonDYa | pFR66 with PARIS from *E.coli* B185, native PARIS promoter + Ptet promoter, $Kan^R$ | [2] |
| pFR99<br>(PARIS ΔAriA) | https://benchling.com/s/seq-lq5A7pdmGcd0pMtSeUbM?m=slm-IBFWM53hTt9F2tJeiDQX | pFR85 with ΔAriA, $Kan^R$ | [2] |
| pFR100<br>(PARIS ΔAriB) | https://benchling.com/s/seq-jd3hwdpF6hXmaKBOawA6?m=slm-LHqWOEpNPycaxjqz0hX5 | pFR85 with ΔAriB, $Kan^R$ | [2] |
| pFR85 AriB C-strep | https://benchling.com/s/seq-sUB4VGJfrfQd89nt3msu?m=slm-QbAER6Gp1NJsmGBY4D3N | pFR85 with strep-tag II cloned at the C-terminus of AriB, $Kan^R$ | This work |
| pSol16<br>(pFR85 AriB E26A) | https://benchling.com/s/seq-WftnRXuL4gjKKI7r10Z7?m=slm-I4J4w4IHYg48zyPsvBdU | pFR85 with AriB E26A mutation in TOPRIM domain, $Kan^R$ | [2] |
| pFR85 AriB E26A C-strep | https://benchling.com/s/seq-OoMBvkn7Mr6FILD4XNgr?m=slm-ElvGNMgAUIA3YfwqTO9u | pFR85 AriB E26A with strep-tag II cloned at the C-terminus of AriB, $Kan^R$ | This work |
| pSol12<br>(pFR85 AriA K39A) | https://benchling.com/s/seq-BKNrelnSgCdOeAOFZlax?m=slm-1snqLus07G5RW3pKWFII | pFR85 with AriA K39A mutation in Walker A ATPase motif, $Kan^R$ | [2] |

| | | | |
|---|---|---|---|
| pFR85 AriB C-strep<br>AriA K39A | https://benchling.com/s/seq-1EuCP99koHh69XYbckwh?m=slm-6S1Ckf9EtlmjVicwcPKT | pFR85 AriB C-Strep with AriA K39A mutation in Walker A ATPase motif, $Kan^R$ | This work |
| pFR85 AriB E26A<br>AriA R116E/R119E | https://benchling.com/s/seq-o5OL7dZIrsuUnWBbaVX2?m=slm-9JcE2E6FFj28hC3OMViy | pFR85 AriB E26A<br>with AriA R116E/R119E double mutation in central pore, $Kan^R$ | This work |
| pFR85 AriB E26A<br>C-strep<br>AriA R116E/R119E | https://benchling.com/s/seq-VE3RNuSNCTiPYiJqTYNj?m=slm-TOVPtfnvGWfHGfypKY3A | pFR85 AriB E26A C-strep<br>with AriA R116E/R119E double mutation in central pore, $Kan^R$ | This work |
| pFR85 AriB C-strep<br>AriA D401N/E404Q | https://benchling.com/s/seq-wMMUEqtYd4x22oBAtkii?m=slm-zNZWu0ZQR3X0Ol69BRY9 | pFR85 AriB C-Strep with AriA D401N/E404Q mutation in AriA/B interface, $Kan^R$ | This work |
| pRAW-464 | https://benchling.com/s/seq-WJDrKW8L3zayPXdI2ODo?m=slm-mOeMAJe57kDbC2CEN9BV | pRSFDuet-1 with PARIS-2 (*ariA* & *ariB* from *E. coli* B185 under T7 promoter) with strep-tag cloned at C-terminus of AriB, $Kn^R$ | This work |
| pAG-14 | https://benchling.com/s/seq-OGitMlIvIoTlpTeVW1qh?m=slm-p7zmhMZRoQwJbLFcFsVM | pRSFDuet-1 with PARIS-2 (*ariA* & *ariB* from *E. coli* B185 under T7 promoter) AriB E26A mutation in TOPRIM domain, $Kn^R$ | This work |
| pRAW-464<br>AriA E393A | https://benchling.com/s/seq-RQYgRALvYJ5NiZ6PJMc3?m=slm-ewc53iABrXLYPzGQVz7O | pRSFDuet-1 with PARIS-2 (*ariA* & *ariB* from *E. coli* B185 under T7 promoter) AriA E393A mutation in ATPase domain, $Kn^R$ | This work |
| pRAW-464<br>AriB C-strep<br>AriA D401N/E404Q | https://benchling.com/s/seq-1if9CepVd6OtxfUFSsCd?m=slm-ho0slpR1DWyGqpmBwRTd | pRSFDuet-1 with PARIS-2 (*ariA* & *ariB* from *E. coli* B185 under T7 promoter) with strep-tag cloned at C-terminus of AriB, AriA D401N/E404Q mutation in AriA/B interface, $Kn^R$ | This work |

| | | | |
|---|---|---|---|
| pRAW-466 | https://benchling.com/s/seq-jgkJNm7zf8yn5j7uzCpZ?m=slm-ieq96cdQSEAoZ9RjnBWH | pRSFDuet-1 with PARIS-2 ΔAriA (*ariB* from *E. coli* B185 under T7 promoter) with strep-tag cloned at C-terminus of AriB, $Kn^R$ | This work |
| pRAW-472 | https://benchling.com/s/seq-mer4TEqPr6m955fB8P9R?m=slm-h24zJ4KIVSRMYl2orneK | pRSFDuet-1 with Ocr (Gene 0.3 from T7 under T7 promoter) $Amp^R$ | This work |
| pT7-Fluc | https://benchling.com/s/seq-2wtCLqR3DGsVXZWhvU4G?m=slm-WEbdjqe6kiooqPfwdM1D | Firefly luciferase gene (*fluc*) cloned under control of T7 promoter, $Amp^R$ | A gift from Dr. Sergiev |
| pBAD Ocr | https://benchling.com/s/seq-Mcb9hM3HgRdwZo0A3YP2?m=slm-2UNmUVUYQHtIqvg3LuOL | Phage T7 Ocr cloned under control of *araBAD* promoter, $Amp^R$ | [26] |
| pBAD Ocr C-Strep | https://benchling.com/s/seq-qECZj1604QHIwJGIK9XJ?m=slm-e7OSyufN44pgTS6RvpIT | pBAD Ocr with strep-tag II cloned at the C-terminus of Ocr, $Amp^R$ | [26] |
| pBAD T5-ORF094 | https://benchling.com/s/seq-luQGd3suHCjMDMMNlRz4?m=slm-SlNU3D9NRKmvQAe3dmGp | Phage T5 Ptr1 (ORF094) cloned under control of *araBAD* promoter, R Amp | This work |
| pBAD T5-ORF103 | https://benchling.com/s/seq-MCvxxnzaSNvA9vKhqUsF?m=slm-F6rI9hH0RgcwinY1Hhqs | Phage T5 Ptr2 (ORF103) cloned under control of *araBAD* promoter, $Amp^R$ | This work |

| | | | |
|---|---|---|---|
| pBAD T5_123_frag1 | https://benchling.com/s/seq-TDfBPYerlxgpUzxTgUWM?m=slm-tTXMo5qzOHRS2PyIfZGa | Fragment #1 (31885-32870) of the deletion found in $T5_{123}$ cloned under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5_123_frag2 | https://benchling.com/s/seq-0f9arhOZSmPh2UIVFkAw?m=slm-7azJAuH418UYDTXaaD46 | Fragment #2 (30456-31884) of the deletion found in $T5_{123}$ cloned under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5_123_frag3 | https://benchling.com/s/seq-zbxfbPNLcJyAiRfuKFT2?m=slm-YHWQhlE8xpywqjU5XFv7M | Fragment #3 (29203-30455) of the deletion found in $T5_{123}$ cloned under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD *E.coli* $tRNA^{Lys}$ | https://benchling.com/s/seq-PBfQldqJYlgCeL11vHOq?m=slm-GesuVCZ1c4UPdrwixsVG | *E.coli* $tRNA^{Lys}$ (UUU) cloned together with 20 bp upstream and downstream under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5 $tRNA^{Lys}$ | https://benchling.com/s/seq-vMWoZV7cTRTJmxHolzoL?m=slm-lwz3eDeiC18WokES0QGM | T5 $tRNA^{Lys}$ (UUU) cloned together with 20 bp upstream and downstream under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5 ORF65 | https://benchling.com/s/seq-kepEtQyKZHSpfPTBXnZY?m=slm-aE9t7dx2o22kvxnr0br4 | Phage T5 ORF065 cloned under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5 ORF66 | https://benchling.com/s/seq-8UTEgAmBE0Ak0Nn2SyHs?m=slm-h2ZccJmYur0Uzi51Uk9U | Phage T5 ORF066 cloned under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5 $tRNA^{fMet}$ | https://benchling.com/s/seq-LfKR2YCCL2qQiZf22DBo?m=slm-CLh9Vitt22QkOpqcIJbm | T5 $tRNA^{fMet}$ (CAU) cloned together with 20 bp upstream and downstream under control of *araBAD* promoter, $Amp^R$ | This work |

| | | | |
|---|---|---|---|
| pBAD T5 $tRNA^{Pro}$ | https://benchling.com/s/seq-Myxb3aGPLIU0aY5A6s17?m=slm-vJzCOuUPlrka2cLyxp1k | T5 $tRNA^{Pro}$ (UGG) cloned together with 20 bp upstream and downstream under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD T5 $tRNA^{Val}$ | https://benchling.com/s/seq-RtF7XESubeGYeh21XQ1i?m=slm-H1QDnkm5B94iE16Nopvn | T5 $tRNA^{Val}$ (UAC) cloned together with 20 bp upstream and downstream under control of *araBAD* promoter, $Amp^R$ | This work |
| pBAD *E.coli* $tRNA^{Lys}$ mut1 | https://benchling.com/s/seq-lYEcOBedbkp0FTsFySax?m=slm-XPkoZgQVNSNHZU93axVR | pBAD *E.coli* $tRNA^{Lys}$ with mutations in the anticodon stem-loop mimicking T5 $tRNA^{Lys}$ (31A->U, 39U->A), $Amp^R$ | This work |
| pBAD *E.coli* $tRNA^{Lys}$ mut2 | https://benchling.com/s/seq-dvmea4IbqeqorFmoBUOA?m=slm-JOeMISpTgEnySwizxegx | pBAD *E.coli* $tRNA^{Lys}$ with mutations in the anticodon stem-loop mimicking T5 $tRNA^{Lys}$ (29U->G, 41A->C), $Amp^R$ | This work |
| pAG-8 | https://benchling.com/s/seq-P3pzQbxfrdntx84QtBr8?m=slm-0ZcOzje1Rziy7fEq823w | pRSFDuet-1 with His6-TwinStrep-SUMO at the N-terminus of Ocr, T7 promoter, $Kn^R$ | This work |
| pAG-11 | https://benchling.com/s/seq-n2KnqpVRlv8x4lo4NEWk?m=slm-rNrKpvYLgs1HJlslvuIV | pET-Duet with Ocr carrying Strep-tag on C-terminus, T7 promoter, $Amp^R$ | This work |
| pRAW-401 | https://benchling.com/s/seq-CoC6yFNJ755I6Pp32O2l?m=slm-NOM1jk1TNNivclNE8nXd | Vector backbone that has His6-TwinStrep-SUMO | [69] |

| | | | |
|---|---|---|---|
| pFD200 | https://benchling.com/s/seq-rUkLXHvua674pynDfnm5?m=slm-8YQXlfp0QQ1Xu8necS3s | pFR56 with dCas9 *ccdB* cloned under the control of $P_{PhlF}$ promoter, $Cm^R$ | This work |
| pFR94 | https://benchling.com/s/seq-pitmlcFiBDtqOCLy6jUe?m=slm-ARwDRtOCZuH4gdkL4YtB | pBAD18 with *0.3* gene encoding Ocr under the control of *araBAD* promoter | [2] |
| pFD245 | https://benchling.com/s/seq-IO4LSvihp3CzlnznYfk4?m=slm-EkG77x6ylh1VTUABLQky | pFD200 with *sfGFP* cloned under the control of $P_{PhlF}$ promoter, $Cm^R$ | This work |
| pFD250 | https://benchling.com/s/seq-3Y0xK3fj0qRzXKjSqkrH?m=slm-WXcGNkiVOWzJvLAupoOb | pFD200 with *0.3* gene encoding Ocr under the control of $P_{PhlF}$ promoter, $Cm^R$ | This work |
| pFD287 | https://benchling.com/s/seq-PHnTolybLsHXKBFWv4UH?m=slm-Sny163elYSt2RsiAMsdr | T5 $tRNA^{Lys}$ cloned with 20 bp upstream and under the control of *araBAD* promoter, $Amp^R$ | This work |
| pDB288 | | pFR85 with AriA R116E mutation in central pore, $Kan^R$ | This work |
| pDB289 | | pFR85 with AriA R119E mutation in central pore, $Kan^R$ | This work |
| pDB290 | | pFR85 with AriA R116A mutation in central pore, $Kan^R$ | This work |
| pDB291 | | pFR85 with AriA R119A mutation in central pore, $Kan^R$ | This work |
| pDB292 | | pFR85 with AriA E393A mutation in ATPase domain, $Kan^R$ | This work |
| pDB293 | | pFR85 with AriB R31E mutation in AriA-interacting interface, $Kan^R$ | This work |

| | | | |
|---|---|---|---|
| pDB294 | | pFR85 with AriB E285R mutation in AriB dimerization interface, $Kan^R$ | This work |
| pDB295 | | pFR85 with AriB F137A mutation in AriB dimerization interface, $Kan^R$ | This work |
| pFD290 | | pFR85 with AriA R168A mutation in radial pore, $Kan^R$ | This work |
| pFD292 | | pFR85 with AriA R172A mutation in radial pore, $Kan^R$ | This work |
| pFD293 | | pFR85 with AriA R168E mutation in radial pore, $Kan^R$ | This work |
| pFD295 | | pFR85 with AriA R172E mutation in radial pore, $Kan^R$ | This work |
| pFD296 | | pFR85 with AriB R28E mutation in AriA-interacting interface, $Kan^R$ | This work |
| pFD297 | | pFR85 with AriB R28E R31E mutations in AriA-interacting interface, $Kan^R$ | This work |
| pFD294 | | pFR85 with a stop codon R212* in AriA truncating the protein before the domain interacting with AriB | This work |

Supplementary Table 1. Bacterial strains, phages, and plasmids used in the study

| **Name** | **Sequence (5′->3′)** | **Template** | **Comment** |
|---|---|---|---|
| pFR66_check_F | GAGCCTTTCGTTTTATTTGATG | pFR85 | Sequencing primers |
| pFR66_check_R | GAGTGATATCCGGAGGCAT | | |
| B185_AriB_Strep_F | CGGGTGGCTCCACTCAATATTTTTCAAAAGCGAG | pFR85 | Introduction of AriB C-strep-tag II to the pFR85 through KLD cloning |
| B185_AriB_Strep_R | CAGTTCGAAAAATGATAATTGGCCAGCCTAG | | |
| B185_AriA_check_R | TAGTGGGCGGCAAATTATC | pFR85 | Sequencing primer |
| AriA_ATPK39A_F | ACCTGCACCATTCTTTC | pFR85 | Introduction of mutations through KLD cloning |
| AriA_ATPK39A_R | GCGTCTACATTAATACATGTTATAGC | | |
| AriA_R116A_R119A_F | AGTGAGCTTCACTTTCTATCTCTCTC | pFR85 | |
| AriA_R116A_R119A_R | TGGCGGAACGAGATGTTAAATC | | |
| AriA_R116E_R119E_F | AGTGTTCTTCACTTTCTATCTCTCTC | pFR85 | |
| AriA_R116E_R119E_R | TGGAAGAACGAGATGTTAAATC | | |
| AriA_D401N/E404Q_F | CCAATTGATATGTAATGAAATC | pFR85 or pRAW | |
| AriA_D401N/E404Q_R | CAACAAGATCTATTGAAAAG | | |

| | | | |
|---|---|---|---|
| T5_094_F | TTTGGGCTAACAGGA GGAAGAATTCATGACTTT AAAACCGTATATAGTG | T5 phage DNA | Cloning of T5 PARIS triggers through Gibson Assembly |
| T5_094_R | CAGCCAAGCTTGCGGCCGCGAGC TCTTAGTCATTATTAAACTCCACT G | | |
| T5_103_F | TTTGGGCTAACAGGAGGAAGAAT TCGTGAGTAGACTAACTGATTTA CTTA | T5 phage DNA | |
| T5_103_R | CAGCCAAGCTTGCGGCCGCGAGC TCTTACTCCA CCAGGCCA | | |
| pBAD_For | ATGCCATAGCATTTTTATCC | pBAD18 | pBAD sequencing primers |
| pBAD_Rev | GATTTAATCTGTATCAGGCTG | | |
| pBAD256b_F | TTTTTATAACCTCCTTAGAGCTCG | pBAD18 | Amplification of pBAD backbone for Gibson Assembly |
| pBAD256b_R | GTCGACCTGCAGGCATGC | | |
| pBAD256_T5del_123 F | AGCTCTAAGGAGGTTATAAAAAC GTGGGGTTGGTAGCCTC | T5 phage DNA | Cloning of fragments found in $T5_{123}$ deletion through Gibson Assembly |
| pBAD256_T5del_123_ 1 R | TTGCATGCCTGCAGGTCGACATT CAGTTGCCGCTGAATCTAGTG | | |
| pBAD256_T5del_123 R | TTGCATGCCTGCAGGTCGACTAC TCAGCGAGGCTTGGAAGG | T5 phage DNA | |
| pBAD256_T5del_123_ 3 F | CTCTAAGGAGGTTATAAAAACCC GACATGCGTTCGGGGGTA | | |
| pBAD256_T5del_123_ 2 F | CTCTAAGGAGGTTATAAAAAAGT GTATTGGGTTTGCGAGA | T5 phage DNA | |

| | | | |
|---|---|---|---|
| pBAD256_T5del_123_2 R | TTGCATGCCTGCAGGTCGACAAAGTCTTTCCAAGCGCATC | | |
| pBAD256_T5_66_F | AGCTCTAAGGAGGTTATAAAAAGTGAAGTATAAAATTGCATC | T5 phage DNA | Cloning of ORFs found in $T5_{123}$ deletion through Gibson Assembly |
| pBAD256_T5_66_R | TTGCATGCCTGCAGGTCGACTTATTTCTTCAGCCACAAC | | |
| pBAD256_T5_65_F | AGCTCTAAGGAGGTTATAAAAAATGAAAGTATTTCGATTAG | T5 phage DNA | |
| pBAD256_T5_65_R | TTGCATGCCTGCAGGTCGACTTAGTAACCATTATATACTTC | | |
| pBAD256_T5_Pro_F | TTTTTGGGCTAGCGAATTCGAGCTCGTGGCTGAAGAAATAATTTTC | T5 phage DNA | Cloning of tRNAs found in $T5_{123}$ deletion through Gibson Assembly |
| pBAD256_T5_Pro_R | CAAGCTTGCATGCCTGCAGGTCGACAGATCTAACCCGCAATTG | | |
| pBAD256_T5_fMet_F | TTTTTGGGCTAGCGAATTCGAGCTCTTCATTGGAGACCAAAC | T5 phage DNA | |
| pBAD256_T5_fMet_R | CAAGCTTGCATGCCTGCAGGTCGACTTCGAAATCCTCTATTAAAATG | | |
| pBAD256_T5_Lys_F | TTTTTGGGCTAGCGAATTCGAGCTCAAGTATATAATGGTTACTAAGGG | T5 phage DNA | |

| | | | |
|---|---|---|---|
| pBAD256_T5_Lys_R | CAAGCTTGCATGCCTGCAGGTCG ACATACTAACCGAGCAAAAATTT G | | |
| pBAD256_T5_Val_F | TTTTTGGGCTAGCGAATTCGAGC TCGTGCAACCCACCAAATTTTTG | T5 phage DNA | |
| pBAD256_T5_Val_R | CAAGCTTGCATGCCTGCAGGTCG ACTTTTTGAAATTTCTTGGACTTG G | | |
| pBAD256_T5_LysColi _F | TTTTTGGGCTAGCGAATTCGAGC TCACCGAAAGCAACGAAAAAGT G | BW25113 genomic DNA | Cloning of *E. coli* $tRNA^{Lys}$ through Gibson Assembly |
| pBAD256_T5_LystCol i_R | CAAGCTTGCATGCCTGCAGGTCG ACACCGCTTCCACCTTAGCG | | |
| pBAD_tRNALysE.coli _mut1_F | TAAACAATTGGTCGCAGGTT | pBAD *E.coli* $tRNA^{Lys}$ | Mutagenesis of *E. coli* $tRNA^{Lys}$ through KLD reaction |
| pBAD_tRNALysE.coli _mut1_R | AAAGACAACTGCTCTACCAACT | | |
| pBAD_tRNALysE.coli _mut2_F | TAATCCATTGGTCGCAGGTT | pBAD *E.coli* $tRNA^{Lys}$ | |
| pBAD_tRNALysE.coli _mut2_R | AAAGTCCACTGCTCTACCAACTG | | |
| F507-F | CAAATAAAACGAAAGGCGTCGG GTAGCACCAGAAGTCTATAG | pFD200 | Cloning of sfGFP cassette using Gibson assembly leading to **pFD245** |
| F508-R | GCCTCCGGATATCACTTTTGCCA CCTGCACTCGTCCTTTG | | |
| F509-F | GTGCAGGTGGCAAAAGTGATATC CGGAGGCATATCAAATGAC | pFR66 | |
| F510-R | CTGGTGCTACCCGACGCCTTTCG TTTTATTTGATGCCTGG | | |

| | | | |
|---|---|---|---|
| F511-F | GTCGGGTAGCACCAGAAGTCTATAG | pFD200 | Cloning of ocr trigger using Gibson assembly leading to **pFD250** |
| F512-R | TTTTGCCACCTGCACTCGTCC | | |
| F523-F | CATATCAAAGGACGAGTGCAGGTGGCAAAAATGGCTATGTCTAACATGACTTAC | pFR94 | |
| F524-R | CATGATACGAAACGTACCGTATCGTTAAGGTCGAGATTTACTCAGGAGGTTAATTTTAATGGATATTAATACTGAAACTGAGATC | | |
| V318-F | CGGGTATGGAGAAACAGTAGAGAG | pBAD18 | Cloning of T5 $tRNA^{Lys}$ using Gibson assembly leading to **pFD287** |
| F593-R | CAAGCTTGGCTGTTTTGGC | | |
| F617-F | TCGCAACTCTCTACTGTTTCTCCATACCCGAAGTATATAATGGTTACTAAGGGTTGCTAG | T5 phage DNA | |
| F605-R | CTTCTCTCATCCGCCAAAACAGCCAAGCTTGAAATTTGGTGGGTTGCACAGGACTC | | |
| Sol35-F | ATGGTGCAGGTGCATCTACATTAATACATG | pFR85 | Mutation in AriA-K39A of pFR85 (PARIS) using Gibson assembly leading to **pSol12** |
| LC191-R | GTCTAGGGCGGCGGATTTG | | |
| LC192-F | CGCTCTCCTGAGTAGGACAAAT | pFR85 | |
| Sol36-R | TATTAATGTAGATGCACCTGCACCATTCTT | | |
| Sol41-F | CGTCTGTTAGTTGCGGGGAGGCATGACCGT | pFR85 | Mutation in *ariB*-E26A of pFR85 (PARIS) using Gibson assembly leading to **pSol16** |
| LC191-R | GTCTAGGGCGGCGGATTTG | | |

| | | | |
|---|---|---|---|
| LC192-F | CGCTCTCCTGAGTAGGACAAAT | pFR85 | |
| Sol42-R | GTCATGCCTCCCCGCAACTAACA GACGCTT | | |
| B765-F | GAGATAGAAAGTGAAGAACACT TGAGGGAACGAGATGTTA | pFR85 | Mutation in *ariA*-R116E of pFR85 (PARIS) using Gibson assembly leading to **pDB288** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B766-R | CTCGTTCCCTCAAGTGTTCTTCAC TTTCTATCTCTCTCAC | | |
| B767-F | GAACGTCACTTGGAAGAACGAG ATGTTAAATCCATGCTTG | pFR85 | Mutation in *ariA*-R119E of pFR85 (PARIS) using Gibson assembly leading to **pDB289** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B768-R | ATGGATTTAACATCTCGTTCTTCC AAGTGACGTTCACTTTC | | |
| B769-F | GAGATAGAAAGTGAAGCGCACT TGAGGGAACGAGATGTTA | pFR85 | Mutation in *ariA*-R116A of pFR85 (PARIS) using Gibson assembly leading to **pDB290** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |

| | | | |
|---|---|---|---|
| B770-R | CTCGTTCCCTCAAGTGCGCTTCACTTTCTATCTCTCTCAC | | |
| B771-F | GAACGTCACTTGGCGGAACGAGATGTTAAATCCATGCTTG | pFR85 | Mutation in *ariA*-R119A of pFR85 (PARIS) using Gibson assembly leading to **pDB291** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B772-R | ATGGATTTAACATCTCGTTCCGCCAAGTGACGTTCACTTTC | | |
| B773-F | TGCTAATTGATGCGCCAGAGATTTCATTACATATCGATTG | pFR85 | Mutation in *ariA*-E393A of pFR85 (PARIS) using Gibson assembly leading to **pDB292** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B774-R | GATATGTAATGAAATCTCTGGCGCATCAATTAGCACAATTG | | |
| B777-F | TAACAACCGCTTTAGCTCGCAGTTGGGTTAGGAAAATTGG | pFR85 | Mutation in *ariB*-E285R of pFR85 (PARIS) using Gibson assembly leading to **pDB294** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B778-R | TTTCCTAACCCAACTGCGAGCTAAAGCGGTTGTTAATGAT | | |

| | | | |
|---|---|---|---|
| B779-F | AATAGATGCATTTCAGGCGCTCT CACCTTCGGAATATAAATATAAG | pFR85 | Mutation in *ariB*-F137A of pFR85 (PARIS) using Gibson assembly leading to **pDB295** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B780-R | TATTCCGAAGGTGAGAGCGCCTG AAATGCATCTATTATTATAG | | |
| F633-F | GATGTAGGCTATGAGGCGAGAGT TATTCGCTCATCATTTTATAACC | pFR85 | Mutation in *ariA*-R168A of pFR85 (PARIS) using Gibson assembly leading to **pFD290** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| F634-R | GAGCGAATAACTCTCGCCTCATA GCCTACATCAGAAGATGAAC | | |
| F637-F | GAGCGAAGAGTTATTGCGTCATC ATTTTATAACCGTAAAGCCAGTG | pFR85 | Mutation in *ariA*-R172A of pFR85 (PARIS) using Gibson assembly leading to **pFD292** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| F638-R | CTTTACGGTTATAAAATGATGAC GCAATAACTCTTCGCTCATAGCC TACATCAG | | |
| F639-F | GATGTAGGCTATGAGGAAAGAG TTATTCGCTCATCATTTTATAAC | pFR85 | Mutation in *ariA*-R168E of pFR85 (PARIS) using |

| | | | |
|---|---|---|---|
| LC327-R | CGCCTTCTTGACGAGTTCTT | | Gibson assembly leading to **pFD293** |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| F640-R | GAGCGAATAACTCTTTCCTCATA GCCTACATCAGAAGATGAAC | | |
| F643-F | GAGCGAAGAGTTATTGAATCATC ATTTTATAACCGTAAAGCCAGTG | pFR85 | Mutation in *ariA*-R172E of pFR85 (PARIS) using Gibson assembly leading to **pFD295** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| F644-R | CTTTACGGTTATAAAATGATGAT TCAATAACTCTTCGCTCATAGCC TACATC | | |
| F645-F | CTGTTAGTTGAAGGGGAACATGA CCGTTCGCATCTGTATC | pFR85 | Mutation in *ariB*-R28E of pFR85 (PARIS) using Gibson assembly leading to **pFD296** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| F646-R | GATGCGAACGGTCATGTTCCCCT TCAACTAACAGACGCTT | | |
| B775-F | TTGAAGGGAGGCATGACGAATC GCATCTGTATCAATTAATTTA | pFR85 | Mutation in *ariB*-R31E of pFR85 (PARIS) using Gibson assembly leading to **pDB293** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |

| | | | |
|---|---|---|---|
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| B776-R | ATTGATACAGATGCGATTCGTCATGCCTCCCTTCAACTAAC | | |
| F647-F | CTGTTAGTTGAAGGGGAACATGACGAATCGCATCTGTATCAATTAATTTA | pFR85 | Mutation in *ariB*-R28E/R31E of pFR85 (PARIS) using Gibson assembly leading to **pFD297** |
| LC327-R | CGCCTTCTTGACGAGTTCTT | | |
| TG99-F | GGATCTATCAACAGGAGTC | pFR85 | |
| F648-R | ATTGATACAGATGCGATTCGTCATGTTCCCCTTCAACTAACAGACGCTT | | |
| HiFi PARIS-2 AriB-strep-F | GTATAAGAAGGAGATATACATATGAGTTCATGTGCTTAC | pFR85 | Cloning AriB with C-strep-tag leading to **pRAW-464** |
| HiFi PARIS-2 AriB-strep-R | CAGCGGTTTCTTTACCAGACTCGAGTCATTTTTCGAACTGCGG | | |
| HiFi PARIS-2 AriA-F | ACTTTAATAAGGAGATATACCATGGCGATTAGAACAATAAG | pFR85 | Cloning AriA to **pRAW-464** |
| HiFi PARIS-2 AriA-R | TTAAGCATTATGCGGCCGCAAGCTTTCATAAAGATTCATCCTCTTC | | |
| Tagless AriB-F | TGACTCGAGTCTGGTAAAG | pRAW-464 | Deletion of C-term strep tag from AriB in pRAW-464, for co-expression experiments with strep-tagged trigger |
| Tagless AriB-R | CTCAATATTTTTCAAAAGCGAG | | |

| | | | |
|---|---|---|---|
| AriB nuclease KO -F | TCTGTTAGTTGCGGGGAGGCATG | pRAW-464 | Introduction of mutation in AriB (E26A) |
| AriB nuclease KO -R | CGCTTTTTAGAGCTCATAG | | |
| HiFi pAG8 Backbone-F | TAATGCTTAAGTCGAACAGAAAG | pRAW-401 | Amplification of vector backbone that has His6-TwinStrep-SUMO leading to **pAG-8** |
| HiFi-pAG8 Backbone-R | TCCACCAATCTGTTCTCTG | | |
| HiFi-pAG-8 Insert-F | ACAGAGAACAGATTGGTGGAAT GGCTATGTCTAACATGAC | pBAD-L18-Ocr-C-Strep | Cloning of T7 Ocr leading to **pAG-8** |
| HiFi-pAG-8 Insert-R | TCTGTTCGACTTAAGCATTACTCT TCATCCTCCTCGTAC | | |
| HiFi-pAG11 Backbone-F | TTAACCTAGGCTGCTGCC | pET-Duet1 | Amplification of vector backbone of pET-Duet1 leading to **pAG-11** |
| HiFi-pAG11 Backbone-R | GGTATATCTCCTTCTTAAAGTTA AACAAAATTATTTC | | |
| HiFi-pAG11 Insert-F | CTTTAAGAAGGAGATATACCATG GCTATGTCTAACATG | pBAD Ocr C-Strep | Cloning of T7 Ocr gene with strep-tag on C-terminus leading to **pAG-11** |
| HiFi-pAG11 Insert-R | GTGGCAGCAGCCTAGGTTAATTA TTTTTCGAACTGCGG | | |
| AriB-only-F | GTATAAGAAGGAGATATACATAT GAGTTCATGTGCTTAC | pRAW-464 | Mutagenesis leading to **pRAW 466** |
| AriB-only-R | CAGCGGTTTCTTTACCAGACTCG AGTCATTTTTCGAACTGCGG | | |

| | | | |
|---|---|---|---|
| Ocr-F | TAATTAACCTAGGCTGCTG | T7 | Leading to **pRAW 472** |
| Ocr-R | CTCTTCATCCTCCTCG | | |

Supplementary Table 2. Primers used for cloning

| **Name** | **Probe** | **Dye labeled** | **Sequence 5'-> 3'** |
|---|---|---|---|
| **B803** | **tRNA$^{Lys}$** | IRD700 | /5IRD700/GGTGGGTCGTGCAGGATTCGAACCTGCGACCAATTGATTAAA |
| **B806** | **T5 tRNA$^{Lys}$** | IRD800 | /5IRD800/GTTTAAAAGACCAGTGCTCTAAACCAGTTGAGCTAGCAACCC |
| **B811** | **5S rRNA** | FAM | /56-FAM/TCGGCATGGGGTCAGGTGGGACCACCGCGCTACTGCCGCCAGGCA |
| **F693** | **tRNA$^{Lys}$RT** | FAM | /56-FAM/TGGTGGTGGGTCGTGCAGGATTCGAACCTGCGACC |

Supplementary Table 3. Nucleotide sequence of the probes used for Northern blot and in primer extension

# CHAPTER SEVEN

# STRUCTURE REVEALS WHY GENOME FOLDING IS NECESSARY FOR SITE-SPECIFIC INTEGRATION OF FOREIGN DNA INTO CRISPR ARRAYS

## Contribution of Authors and Co-Authors

Manuscript in Chapter 7

Author: Andrew Santiago-Frangos

Contributions: Conceptualization, Data Curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft.

Co-Author: William S. Henriques

Contributions: Data curation, Investigation, Methodology, Visualization, Writing – review & editing

Co-Author: Tanner Wiegand

Contributions: Data curation, Investigation, Methodology, Visualization, Writing – review & editing

Co-Author: Colin C. Gauvin

Contributions: Software, Resources, Writing – review & editing

Co-Author: Murat Buyukyoruk

Contributions: Data curation, Investigation, Methodology, Visualization, Writing – review & editing

Co-Author: Ava B. Graham

Contributions: Investigation, Methodology

Co-Author: Royce A. Wilkinson

Contributions: Investigation, Methodology

Co-Author: Lenny Triem

Contributions: Visualization

Co-Author: Kasahun Neselu

Contributions: Investigation, Resources

Co-Author: Edward T. Eng

Contributions: Investigation, Resources

Co-Author: Gabriel C. Lander

Contributions: Methodology, Supervision, Visualization, Writing – review & editing

Co-Author: Blake Wiedenheft

Contributions: Funding acquisition, Project administration, Resources, Supervision, Visualization, Writing – review & editing.

Manuscript Information

Andrew Santiago-Frangos, William S. Henriques, Tanner Wiegand, Colin C. Gauvin, Murat Buyukyoruk, Ava B. Graham, Royce A. Wilkinson, Lenny Triem, Kasahun Neselu, Edward T. Eng, Gabriel C. Lander, Blake Wiedenheft

*Nature Structural and Molecular Biology*

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

Nature Springer
Submitted: 22 March, 2023
Publisehd: 14 September, 2023
https://doi.org/10.1038/s41594-023-01097-2

## Abstract

Bacteria and archaea acquire resistance to viruses and plasmids by integrating fragments of foreign DNA into the first repeat of a CRISPR array. However, the mechanism of site-specific integration remains poorly understood. Here, we determine a 560 kDa integration complex structure that explains how *Pseudomonas aeruginosa* Cas (Cas1-2/3) and non-Cas proteins (IHF) fold 150 base-pairs of host DNA into a U-shaped bend and a loop that protrude from Cas1-2/3 at right angles. The U-shaped bend traps foreign DNA on one face of the Cas1-2/3 integrase, while the loop places the first CRISPR repeat in the Cas1 active site. Both Cas3s rotate 100-degrees to expose DNA binding sites on either side of the Cas2 homodimer, that each bind an inverted repeat motif in the leader. Leader sequence motifs direct Cas1-2/3-mediated integration to diverse repeat sequences that have a 5'-GT. Collectively, this work reveals new DNA binding surfaces on Cas2 that are critical for DNA folding, and site-specific delivery of foreign DNA.


## Introduction

Vertebrates, bacteria, and archaea have domesticated transposases (e.g., RAG1 and Cas1) for adaptive immunity[1,2]. Integrases, transposases and recombinases often co-opt additional DNA-bending proteins (e.g., IHF, HU, H-NS, or HMGB1) that facilitate DNA integration and excision[3–7]. However, the structural role of DNA folding during this mobilization of DNA remains largely enigmatic.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are essential components of an adaptive immune system that stores DNA-based molecular memories of past infections[8]. CRISPR-associated proteins, Cas1 and Cas2, integrate fragments of foreign DNA

("spacers") into CRISPRs. Integration duplicates a repeat sequence, which thereby maintains the characteristic repeat-spacer-repeat architecture (**Figure 7.1a**). Cas1 and Cas2 form a heterohexameric complex that consists of two Cas1 homodimers (Cas1a-a* and Cas1b-b*) flanking a Cas2 homodimer (**Figure 7.1a-b**) [9–11]. Foreign DNA fragments bind across one face of the Cas2 homodimer, which positions the 3'-ends into Cas1 active sites on either end of the complex (i.e., Cas1a* and Cas1b*) [8–10,12]. The CRISPR repeat sequence wraps around the opposing face of Cas2, sandwiching the Cas2 homodimer between the foreign and repeat DNA duplexes. Opposing Cas1 subunits (Cas1a* and Cas1b*) catalyze two successive strand-transfer reactions, linking the 3'-ends of the foreign DNA to opposite ends of the repeat [5,11]. CRISPR integration complexes sense a 2-5 bp 3' overhang called a protospacer-adjacent motif (PAM) in the foreign DNA to determine the integration orientation[9]. Correct spacer orientation is necessary to produce a functional CRISPR RNA that guides the CRISPR interference machinery (i.e., Cascade) to complementary targets[8,13]. Integration occurs in a stepwise manner. First, the non-PAM end of the foreign DNA is integrated at the leader-side of the repeat[14]. Second, the PAM is cleaved by Cas or non-Cas nucleases and the trimmed 3' end is integrated at the spacer-side of the repeat[14–16]. These integration events tie a non-covalent knot around the Cas2 homodimer (foreign DNA on one side and repeat DNA on the other) that is held together by complementary base-pairing in the foreign DNA.

New foreign DNA is preferentially integrated at the first repeat in a CRISPR locus, ensuring efficient transcription and processing of CRISPR RNAs that target the most recently encountered genetic parasites (**Figure 7.1a**)[17,18]. Cas1-2 is thought to recognize a palindromic sequence within the CRISPR repeat, similar to target site recognition by many DNA transposases

[5,11,19–21]. However, Cas1-2 recognition of the palidromic repeat does not explain how the first repeat is differentiated from downstream repeat sequences in a CRISPR. Thus, polarized integration often relies on additional proteins and DNA sequence motifs upstream of the CRISPR (i.e., leader) [3–5,11,18,22–27]. Integration Host Factor (IHF) facilitates polarized integration in the type I-E CRISPR system from *Escherichia coli*[3]. A structure of the I-E integration complex revealed that IHF bends the leader DNA to bring an upstream sequence motif into contact with Cas1, and IHF further stabilizes the Cas1-2 integrase at the first repeat through direct Cas1-IHF interactions[3,5].

Cas1 and Cas2 are conserved components of CRISPR-mediated immune systems. However, the type I-F CRISPR system has a unique fusion of the Cas2 subunit to the Cas3-nuclease/helicase found in many type I systems (**Figure 7.1a-b**) [28,29]. Cas3 degrades Cascade-bound DNA into fragments with PAM-containing termini, that are captured by Cas1-2 and integrated into the CRISPR locus, in a process called "primed acquisition" [4,30–37]. However, the structural mechanism for primed acquisition is unclear. Additionally, in contrast to the *E. coli* I-E system, which relies on one IHF to bend DNA and recruit a DNA motif found ~50 bp upstream of the CRISPR repeat, most I-F, I-C, and some I-E CRISPR leaders, contain multiple IHF binding sites and multiple subtype-specific DNA motifs found up to 200 bp upstream of the CRISPR repeat (**Figure 7.1a and Extended Data Figure 7.1a**) [22].

To determine how DNA sequence motifs in the CRISPR leader regulate the Cas1-2/3 integrase, we determined the structure of a ~560 kDa CRISPR integration complex from *Pseudomonas aeruginosa*. The structure reveals that Cas1-2/3 cannot interact with the leader without first undergoing a large conformational change that may be induced by foreign DNA

binding. Further, the structure explains how the I-F leader and IHF proteins guide Cas1-2/3 to deliver and integrate foreign DNA at the first repeat in the CRISPR array (**Figure 7.1**). Cas1-2/3 and IHF interact with all five DNA sequence motifs (i.e., two Inverted Repeats, two IHF binding sites, and the CRISPR repeat) primarily through a shape-based readout [38,39]. The shape of the folded I-F CRISPR leader is similar to that of the λ-phage excision complex, suggesting that DNA is often used as a flexible scaffold to regulate DNA mobilization [3–7,40–53]. The structure suggests that site-specific integration relies on protein-induced folding of the upstream DNA rather than sequence-specific recognition of the repeat. To test this idea, we perform a series of integration reactions demonstrating that efficient integration relies on conserved sequences in the leader and a 5' GT dinucleotide in the repeat. We show that 5' GT dinucleotides are broadly conserved in repeats derived from different CRISPR types, suggesting that they play a conserved role in integration across diverse CRISPR systems. In addition, the I-F CRISPR integration complex suggests a structural mechanism for interactions of the Cas1-2/3 integrase with the Cascade surveillance complex, that may be neccesary for rapid adaptation to phage escape mutants[4,30–37].

## Results

### Cryo-EM Structure of Type I-F CRISPR Integration Complex

To understand how the Cas1-2/3 integrase cooperates with IHF and CRISPR leader motifs to integrate foreign DNA at the first CRISPR repeat, we purified the heterohexameric Cas1-2/3 integrase and the IHFα-β heterodimer, incubated these proteins with a DNA substrate representing a half-site integration intermediate, and isolated the assembled complex using Size

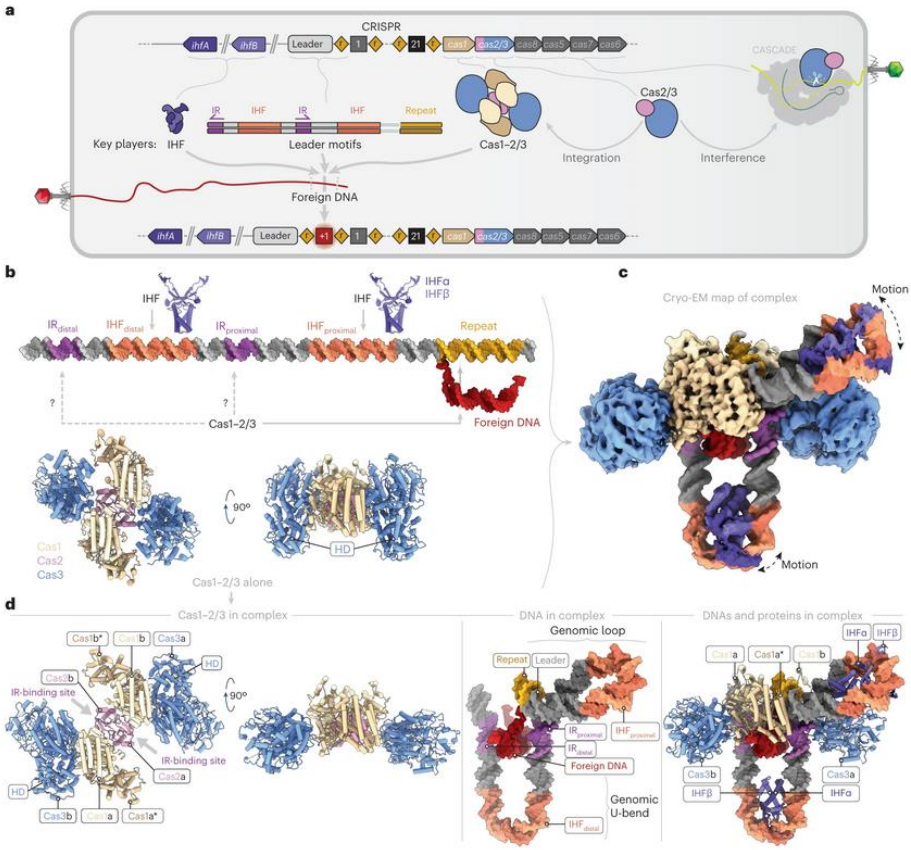



Figure 7.1. Cryo-EM structure of the type I-F CRISPR integration complex. a, Scheme of the I-F CRISPR system of Pseudomonas aeruginosa PA14. A CRISPR is composed of repeated DNA sequences (diamonds) interspersed with unique spacer sequences (black squares). The CRISPR is adjacent to six cas genes (arrows). Four Cas1 and two Cas2/3 proteins assemble into a heterohexamer in which the Cas3 and Cas1 subunits surround the central Cas2 homodimer like petals of a closed flower (Cas14-Cas2/32). Cas1-2/3 and integration host factor (IHF) proteins cooperate with DNA upstream of the CRISPR (leader) to integrate foreign DNA at the first repeat. The leader sequence contains two IHF binding sites and two Inverted Repeats (IRs) that are necessary for integration of foreign DNA at the leader-repeat junction. In addition to playing a central role in integration, the Cas2/3 fusion is recruited to DNA-bound Cascade CRISPR surveillance complex to degrade foreign genetic parasites. b, The Cas1-2/3 heterohexamer and IHF proteins were mixed with a half-site DNA integration intermediate consisting of a foreign DNA linked to one strand of the CRISPR DNA at the leader-repeat junction. c, Cryo-EM density map of the type I-F CRISPR integration complex at ~3.5 Å resolution (Extended Data Figure 7.1g-k, Table 7.1). d, Atomic model of the type I-F CRISPR integration complex. Cas1-2/3 proteins alone (left) are shown in cartoon representation. Cas3 domains rotate by 100° simulating the motion of a bloomed flower and exposing DNA binding sites on Cas2 that interact with each of the IRs (Extended Data Figure 7.2a and Supplementary Video 7.1). DNA alone is shown in the middle (surface representation). Proteins (cartoon representations) and DNAs of the integration complex are shown on the right.

Exclusion Chromatography (SEC) (**Extended Data Fig 7.1a-f**). The purified I-F integration complex was applied to cryo-EM grids and vitrified. We recorded 10,740 movies and picked 366,794 particles to determine a ~3.48 Å-resolution structure of the integration complex. The reconstructed density was of sufficient to model 90.7% of the 10 polypeptides and 88.4% of the 396 nucleotides of DNA (**Figure 7.1b-d, Extended Data Figure 7.1g-k, Table 7.1**). The model explains how the Cas1-2/3 subunits cooperate with two IHF heterodimers to kink and twist ~150 base-pairs of host DNA into a structure that precisely positions foreign DNA for integration at the first repeat of the CRISPR (**Figure 7.1b-d**).

The Cas1 and Cas2 subunits adopt a familiar quaternary arrangement that binds a foreign DNA on one face of the Cas2 homodimer and CRISPR repeat DNA on the other face (**Figures 7.1d, 7.2a, 7.3a**)[8]. A previously determined structure of Cas1-2/3 alone revealed that the Cas3 and Cas1 domains surround the central Cas2 homodimer like petals of a closed flower ($Cas1_4$:$Cas2/3_2$) (**Figure 7.1b**)[54]. While this structure explained how Cas1 regulates the Cas3 nuclease, the role of Cas3 during integration remained unclear[54]. Here, we show that the addition of DNA drives a series of conformational changes in both the DNA and proteins. The Cas3 domains rotate ~100° to align in a planar configuration with Cas2, simulating the motion of a bloomed flower, and exposing equivalent surfaces on opposite sides of the Cas2 homodimer that recognize an inverted repeat (IR) that is conserved in I-F leaders (**Figure 7.1d, Extended Data Figure 7.2a and Supplementary Video 1**) [54]. Thus, the new planar conformation of Cas1-2/3 enables the simultaneous coordination of four DNA helices ($IR_{distal}$, $IR_{proximal}$, foreign DNA, and CRISPR repeat) around the central Cas2 homodimer (**Figure 7.1d and 7.3**). Further, this Cas3 rotation flips the nuclease domain from an interaction with Cas1 that suppresses the Cas3

nuclease activity, to the opposite side of the complex, where the back of the Cas3 nuclease domain docks onto a groove created at the Cas1-Cas1 interaface (**Figure 7.1b,d and Extended Data Figure 7.2a,b**).

| | I-F Integration Complex (EMDB-**29280**) (PDB **8FLJ**) |
|---|---|
| **Data collection and processing** | |
| Magnification | 46,860x |
| Voltage (kV) | 300 |
| Electron exposure ($e^-/Å^2$) | 69.09 |
| Defocus range (μm) | 0.7 – 2.1 |
| Pixel size (Å) | 1.067 |
| Symmetry imposed | *C*1 |
| Initial particle images (no.) | 5,846,923 |
| Final particle images (no.) | 366,794 |
| Map resolution (Å) | 3.47 |
| FSC threshold | 0.143 |
| Map resolution range (Å) | 2.3 – 9.9 |
| | |
| **Refinement** | |
| Initial model used (PDB code) | 1IHF; 3GOD |
| Model resolution (Å) | 2.50; 2.17 |
| FSC threshold | - |
| Model resolution range (Å) | - |
| Map sharpening *B* factor ($Å^2$) | -55 |
| Model composition | |
| Non-hydrogen atoms | 31,794 |
| Protein residues | 3,549 |
| DNA nucleotides | 350 |
| Ligands | - |
| *B* factor ($Å^2$) | 51.4 |
| R.m.s deviations | |
| Bond lengths (Å) | 0.005 |
| Bond angles (°) | 0.835 |
| Validation | |
| MolProbity score | 1.54 |
| Clashscore | 4.67 |
| Poor rotamers (%) | 0.26 |
| Ramachandran plot | |
| Favored (%) | 95.72 |
| Allowed (%) | 4.28 |
| Disallowed (%) | 0 |

Table 7.1. Cryo-EM data collection, refinement, and validation statistics

The structure reveals two prominent DNA bends that protrude at right angles from Cas1-2/3 (**Figure 7.1c,d**). An IHF heterodimer is wedged at the apex of each DNA bend, consistent with IHF's well-defined role in DNA bending[38]. These two DNA protrussions extend ~75 Å from the Cas1-2 core. Flexablity of these DNA extentions limits the resolution of the regions to 4-8 Å (**Figure 7.1c,d and Supplementary Video 2**). IHF-mediated bending of the $IHF_{distal}$ site

positions the flanking IR sequences as symmetrical DNA pillars, which are recognized by equivalent surfaces on opposite sides of the Cas2 homodimer (**Figure 7.3 and Extended Data Figure 7.3b and 7.4c**)[22]. Cas2 binding to these DNA pillars traps foreign DNA on one face of the Cas1-2/3 integrase. Further, Cas2 bends the IRs and steers downstream DNA away from Cas1-2/3, which would project the downstream CRISPR repeat away from the Cas1-2/3 integrase (**Fig. 1d**). However, Cas1-2/3 and IHF cooperate to constrict the DNA around the $IHF_{proximal}$ site, forming a loop that places the CRISPR repeat into the Cas1a* active site (**Figures 7.1d and 7.3**).

Foreign DNA Constrains Cas2/3 Linker against Cas1

The type I-F Cas2 and Cas3 subunits are connected by a 20 amino acid disordered linker (residues 90-110) [4,28,29,55]. The structure explains how foreign DNA constrains the Cas2/3 linker against conserved surfaces of Cas1, which suggests that foreign DNA-binding either initiates, or stabilizes the Cas3 rotation (**Figure 7.2a**) [54,55]. The constrained Cas2/3 linker positions the HD nuclease domain of Cas3 (residues 111-374) against the Cas1-Cas1 interface, and facilitates Cas3 interactions with the IRs (**Figure 7.2a and Extended Data Figure 7.2 and 7.4c**). The foreign DNA and amino acids in the Cas2/3 linker contact conserved residues in type I-F Cas1 proteins (**Figure 7.2d,e and Extended Data Figure 7.2c**). Polar residues in the Cas2/3 linker may assist the binding or splaying of the foreign DNA duplex at the conserved histidine wedge (H25) in Cas1 (**Figure 7.2b,c and Extended Data Figure 7.3**). Mutation of the histidine wedge ($Cas1^{H25A}$) decreases Cas1-2/3 integration activity of foreign DNA that has either fully complementary or splayed DNA ends (**Extended Data Figure 7.5 and 7.6a,b**). The integration defect on substrates with splayed ends suggest that H25 is more than a simple wedge that pries

apart the ends for foreign DNA [10]. The histidine steers the 3'-ends down a positively charged channel that positions each 3'-hydroxyl into Cas1 active sites on opposite ends of the complex

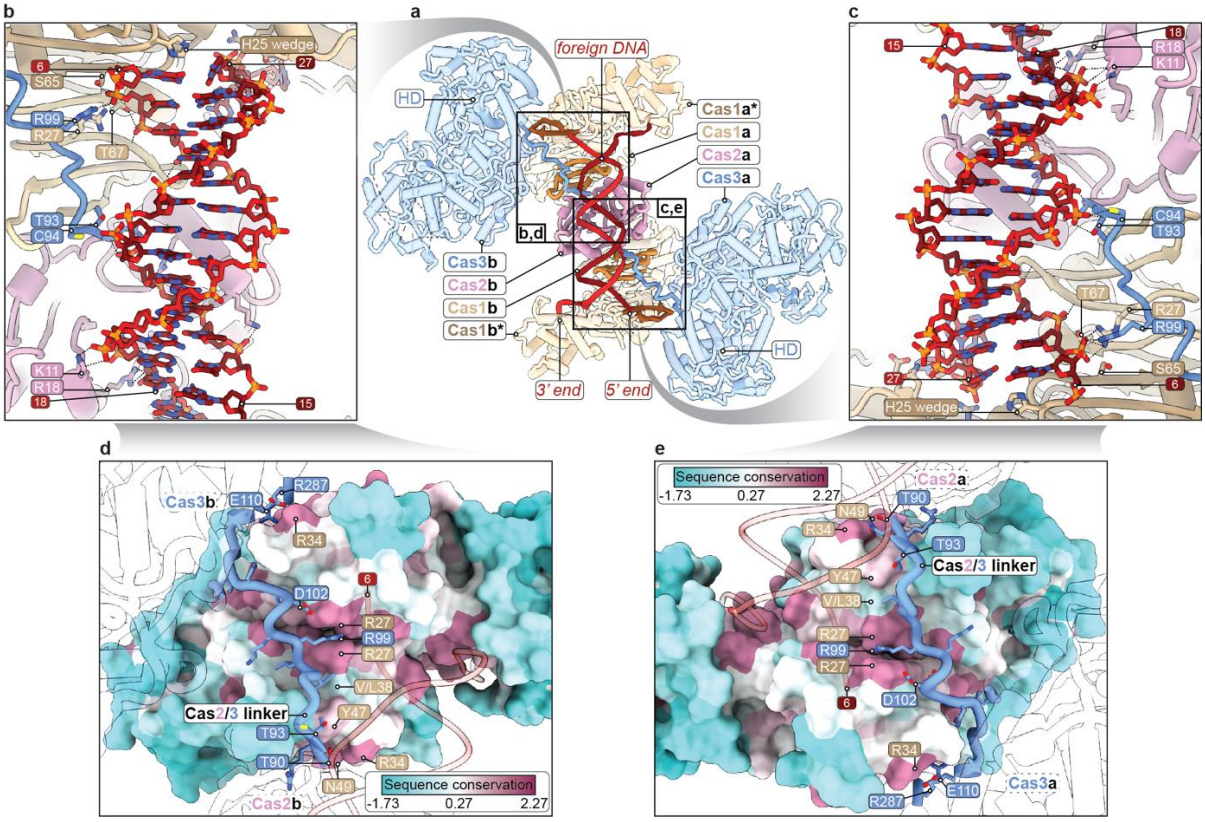



Figure 7.2. Foreign DNA constrains the Cas2/3 linker against conserved Cas1 residues. **a**, View of the foreign DNA-bound face of Cas1-2/3. The foreign DNA, Cas2 subunits, Cas2/3 linker and Cas1 beta hairpins that contact the start and end of the Cas2/3 linker are shown in solid while other parts of the complex are shown at 40% transparency for clarity. Insets outline locations of close-up views shown in panels b-e. b,c, The foreign DNA constrains the Cas2/3 linker against each Cas1 subunit (Extended Data Figure 7.2b). Cas2, the Cas2/3 linker and, Cas1 cooperate to bind the foreign DNA body and to splay the ends of the foreign DNA. Histidine wedges in Cas1 measure out a central foreign DNA duplex of 22 base-pairs. Most DNA-binding residues are conserved or undergo conservative mutations (Extended Data Figure 7.3). d,e, Conserved Cas2/3 linker residues (blue, sticks) contact residues conserved in Cas1 proteins from type I-F CRISPR systems (mauve, surface) (Extended Data Figure 2b,c). Cas2 and Cas3 domains are shown at 90% transparency for clarity. Inset shows the Cas1 sequence conservation color key.

(**Figure 7.2 and Figure 7.4a,d**), whereas the 5'-ends of the protospacer DNA are directed towards the back face of the Cas3 HD domain (**Figure 7.2**) [9,10,56].

## Cas2 Homodimers Recognize and Bend Inverted Repeat Sequences

The structure of the type I-F CRISPR integration complex reveals that Cas2 is the homodimer that binds the IRs (**Figure 7.3a**). Mutations that scramble the order of nucleotides in either the $IR_{distal}$ or $IR_{proximal}$ motifs limit Cas1-2/3-mediated integration[22]. While Cas2 doesn't make extensive sequence-specific contacts with nucleobases of the IR, a single residue ($Cas2^{R55}$) intercalates in the minor groove, and may participate in recognizing two conserved bases in the 10 bp-long motif (**Figure 7.3a and Extended Data Figure 7.4c,e**). However, there is insufficient density for the R55 sidechain to confidently assign contacts. Other conserved Cas2 residues (i.e., K11, R12, and N56) form additional hydrogen bonds with the phosphate backbone of one DNA strand in each IR (**Figure 7.3a and Extended Data Figure 7.4c**). Mutation of these Cas2 residues ($Cas2^{K11D,R12E}$, $Cas2^{R55E,N56D}$, $Cas2^{K11D,R12E,R55E,N56D}$) prevents Cas1-2/3-mediated DNA integration (**Extended Data Figures 7.5 and 7.6c,d**). Cas2 acts as a wedge that induces a 25-35° bend in the DNA upstream of $IR_{distal}$ and downstream of $IR_{proximal}$ (**Figure 7.3a**). These flared IRs lean against basic residues (K381, R393, K397) on the back surface of Cas3 (**Figure 7.3 and Extended Data Figures 7.3 and 7.4c**). In sum, these observations reveal that the IR DNA sequences are primarily recognized by Cas1-2/3 through shape readout rather than base readout[39].

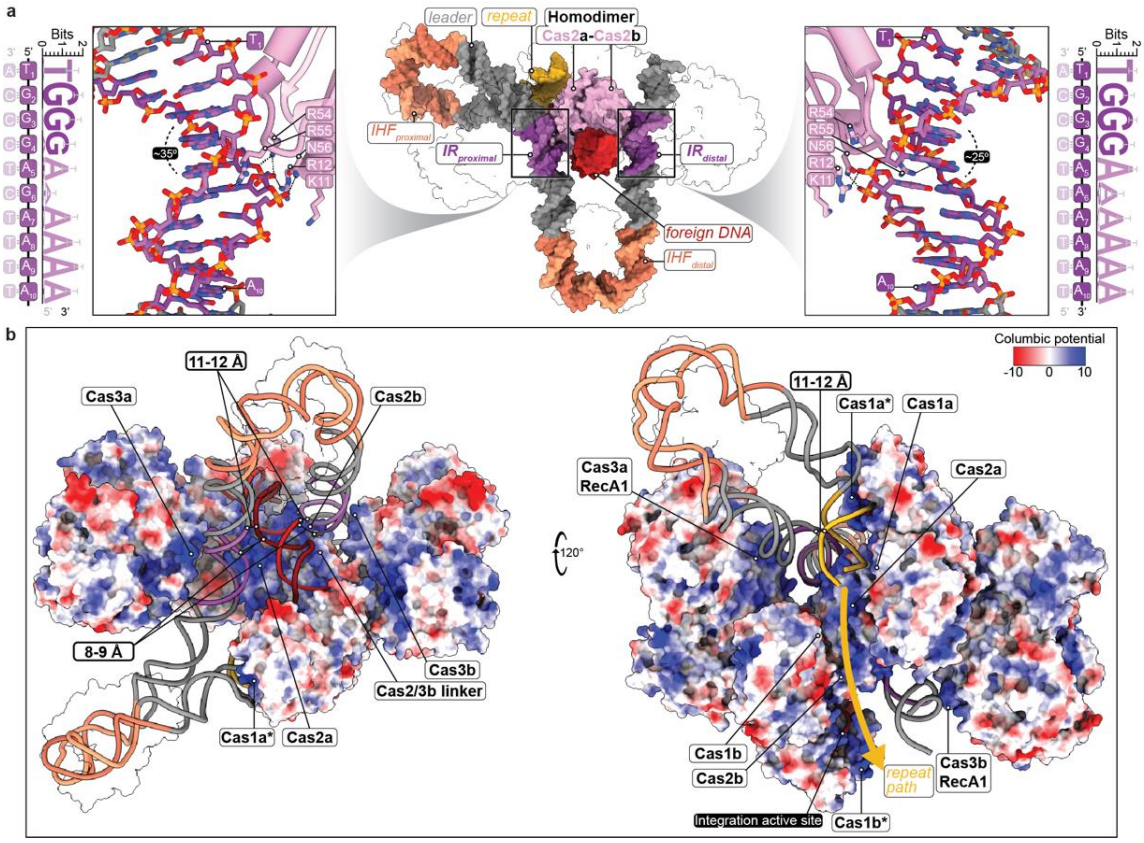



Figure 7.3. The Cas2 homodimer simultaneously coordinates four dsDNA helices critical to CRISPR integration. a, The Cas2 homodimer (pink surface) is flanked by DNA on four sides. Prior structures have shown that the CRISPR repeat (yellow) and foreign DNA (red) are bound to opposite faces of Cas2. Here, we show that symmetrical surfaces on Cas2 also bind inverted repeat (IR, left and right) motifs in the leader. Surface representations of the IHF heterodimers, Cas1 homodimers, Cas2/3 linkers, Cas3 domains, and the 3' overhang of the foreign DNA are shown in 100% transparency for clarity. Each Cas2 inserts an arginine (R55) into the center of the IRs, which stack between deoxyribose sugars, and additional polar residues (R54, N56, R12, and K11) contact the DNA backbone. Cas2 induces 25-35° bends in the DNA (Extended Data Figure 7.4c,e). The sequence logos of the type I-F IRproximal (left) and IRdistal motifs (right), and the IR sequences present in the P. aeruginosa PA14 CRISPR leader are shown. b, Views of the foreign DNA- (left) and repeat-bound (right) faces of Cas1-2/3 are shown in surface representation and colored by columbic potential. For clarity the highly electronegative DNA is shown in cartoon representation. Labels highlight highly basic and conserved surfaces of each Cas1-2/3 subunit that accommodate the packing of four dsDNA helices in proximity around the Cas2 homodimer (Extended Data Figure 7.3). IHF heterodimers are shown in 100% transparency for clarity. The phosphate-to-phosphate distances of DNA helices packed around Cas2 are noted.

Cas2 Homodimer is Surrounded by Four DNA Helices

Cas2 is a cube-shaped homodimer at the center of the Cas-integrase. The Cas2 cube is flanked by Cas1 homodimers to form an elongated DNA binding platform that interacts with the CRISPR repeat on one face and the foreign DNA on the other (**Figure 7.3**). Unique to the type I-F Cas1-2/3 integration complex, the IRs occupy the last two accessible surfaces of the Cas2 cube (**Figure 7.3a**). Positively charged surfaces on Cas1-2/3 bind and shield negatively charged DNA, which enables the packing of four DNA helices around the small Cas2 homodimer (**Figure 7.3b and Extended Data Figure 7.3**). The foreign DNA-binding face of Cas2 has two electronegative pillars of leader DNA that straddle the foreign DNA, such that major grooves of the leader DNA pillars are clamped against major grooves of the foreign DNA. The two DNA pillars continue past Cas2 to flank the Cas1 active sites (**Figure 7.3b**). At the $IHF_{proximal}$ loop, Cas3 packs the leader against the Cas1-bound repeat, decreasing the phosphate-to-phosphate distances between these helices to ~11-12 Å. Although the latter two-thirds of the CRISPR repeat could not be resolved, the repeat's trajectory suggests it will follow a path that threads between the distal leader DNA duplex and the 3'-hydroxyl of the foreign DNA that rests in the Cas1b* active site (**Figure 7.3b**). Collectively, Cas1, the Cas2/3 linker, and Cas3 accommodate four DNA helices ($IR_{distal}$, $IR_{proximal}$, foreign DNA, and repeat) around the central Cas2 homodimer to facilitate site-specific integration.

IHF and the Leader Direct Integration into Diverse Sequences

The structure suggests that Cas1-2/3 is guided to the first repeat of the CRISPR by IHF-mediated folding of the I-F leader, rather than direct recognition of the repeat sequence (**Figure 7.1d and Extended Data Figure 7.4b,d**). To determine if or how the repeat sequence impacts

integration, we measured the efficiency of Cas1-2/3-catalyzed integration into DNAs containing either a I-F, I-E, I-C or II-A repeat downstream of a I-F leader (**Figure 7.4a-c**). The type I-F leader supports Cas1-2/3-catalyzed leader-side integration at repeats derived from I-E, I-C, and II-A CRISPR loci (**Figure 7.4a-c and Extended Data Figures 7.7 and 7.8**). Integration efficiency at non-native repeats is not correlated with sequence similarity to the I-F repeat or with GC-content (**Extended Data Figure 7.8b**). Instead, integration efficiency is correlated to the length of the repeat. I-F and I-E repeats are similar in length (28 and 29 bps, respectively), whereas the I-C and II-A repeats are 0.5 to 1 full DNA turns longer than the I-F repeat (33 and 37 bp, respectively) (**Figure 7.4b and Extended Data Figure 7.9a,b**). While leader-side integration is robust with different repeats, spacer-side integration is ~3.5-fold slower for the I-E repeat, and undetectable for the longer repeats (**Extended Data Figure 7.7 and 7.9a,b)**.

As expected, Cas1-2/3 does not catalyze integration at a I-F repeat downstream of a scrambled I-F leader, nor does Cas1-2/3 catalyze integration at I-E, I-C or II-A repeats downstream of their respective leaders[22] (**Extended Data Figure 7.7**). Cas1-2 sequences are diverse, such that leader-interacting residues are only conserved in a subset of proteins within a given CRISPR subtype[22]. For example, the I-F Cas1 protein lacks residues required to interact with the I-E leader[22]. Further, the I-E, I-C and I-F leaders have distinct nucleotide spacings between the leader motifs and the repeat. These nucleotide spacings impart a unique shape to the DNA, which is critical for integration.

These experiments were performed with foreign DNA substrates either with or without a PAM (**Extended Data Figure 7.7**). Cas1-2/3 integration of PAM-containing DNA is more specific, but the conclusions are otherwise consistent between the two substrates. The PAM must

be trimmed by an ancillary nuclease before Cas1-2/3 can catalyze spacer-side integration, therefore we focused our discussion on results from the trimmed foreign DNA to compare differences in spacer-side integration (**Extended Data Fig. 7a-d**) [14–16]. Collectively, these

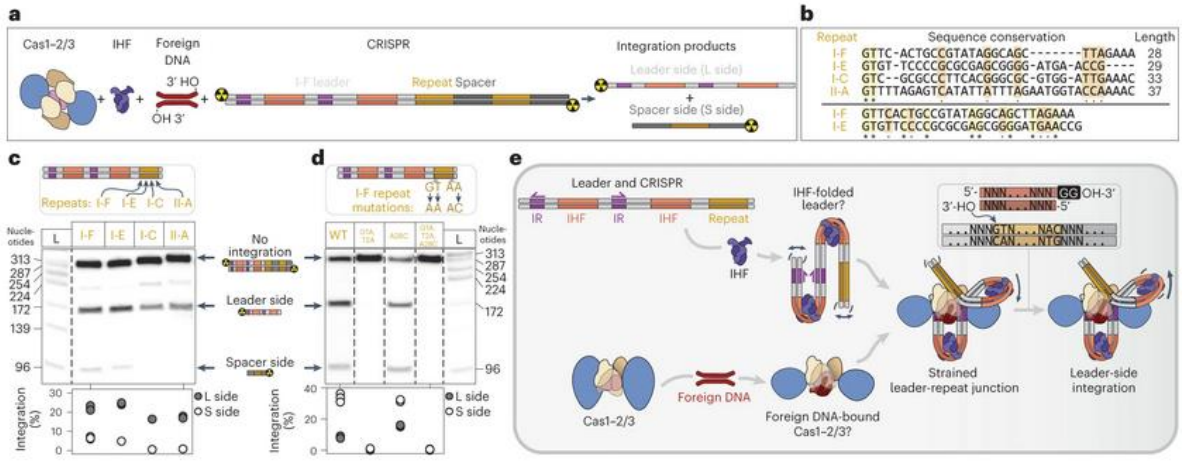



Figure 7.4. Sequence motifs in the leader and IHF proteins facilitate Cas1-2/3-based integration into diverse repeat sequences. a, Scheme of reactants and products of *in vitro* CRISPR integration assays (Extended Data Fig. 7.6,7.7 and 7.9). b, Four CRISPR repeats used in the integration assays. A gapped sequence alignment highlights two identical (asterisks) and six similar (dots) positions. An un-gapped sequence alignment reveals nine identical nucleotide positions between the I-F and I-E repeats. All four repeats have different internal palindromes and GC-content (Extended Data Figure 7.8b). c, Endpoint integration reactions with CRISPR repeat-swapped mutants, resolved on denaturing polyacrylamide gels. One of three representative gel images is shown (Extended Data Figure 7.7). Quantification of leader- (grey circles) or spacer-side (white circles) integration events from all three replicate gels (Extended Data Figure 7.7). The reactions were performed in triplicate, each dot represents one reaction, and some dots overlap. d, 4-minute time point of time-course integration reactions with I-F repeat mutants, resolved on denaturing polyacrylamide gels. One of three representative images is shown (Extended Data Figure 7.9). Quantification of leader- (grey circles) or spacer-side (white circles) integration events from all three replicate gels (Extended Data Figure 7.9). e, CRISPR integration model. IHF-mediated folding of the genome presents IRs as symmetric DNA pillars that recruit foreign DNA-bound Cas1-2/3. Cas3 domains of Cas1-2/3 must rotate away from Cas2 to expose IR binding sites on Cas2. Cas1-2/3 and IHF cooperate to fold DNA into a loop, docking the leader-repeat junction at the Cas1 active site. Foreign DNA integration at the leader-repeat junction nicks the DNA duplex, releasing tension in the DNA duplex and inhibiting the reverse disintegration reaction (Extended Data Figure 7.4)[61,62]. 5' GT dinucleotides are required for efficient leader- and spacer-side integration, but no strict sequence requirements are necessary in the rest of the repeat.

integration experiments indicate that leader sequences and host factors dictate site-specific integation of foreign DNA at diverse DNA target sites.

5’ GT is Critical for Cas1-mediated Integration

Repeat sequences are strongly conserved within CRISPR subtypes, but vary in sequence and length between subtypes[57,58]. However, in the small subset of repeats tested above, we noticed that the 5’ GT is conserved. To determine if conservation of the 5’ GT is a coincidence or a more widely conserved feature of repeats, we performed a bioinformatic analysis consisting of 24,940 CRISPRs. This bioinformatic analysis reveals that a 5’ GT dinucleotide is broadly conserved at the leader-side of the repeat, and conserved in some CRISPR systems at the spacer-side of the repeat (**Extended Data Figure 7.4g**). Therefore we hypothesized that the 5’ GT dinucleotide is a base-specific determinant for leader-side integration. To test this hypothesis, we mutated the 5’ GT (G1A, T2A) and repeated the integration assays. The 5’ GT to AA mutation ablates both leader- and spacer-side integration, indicating that the 5’ GT is essential and that leader-side integration is a prerequisite for spacer-side integration (**Figure 7.4d**). Since Cas1 requires a 5’ GT at the leader-side of the repeat, we hypothesized that introducing a 5’ GT at the spacer-side of the repeat would increase spacer-side integration efficiency. To test this hypothesis, we replaced adenosine 28 of the I-F repeat with cytosine (A28C) and repeated the integration assays. The A28C mutation increases the rate and amount of spacer-side integration by ~2-fold, relative to the WT I-F repeat (**Figure 7.4d**). We do not detect integration into a I-F repeat that lacks a 5’ GT at the leader-side even if the repeat contains a 5’ GT at the spacer-side. This result further supports our conclusion that leader-side integration is a prerequisite for spacer-side integration (**Figure 7.4d**). We examined the structure of the Cas1 active site to

determine whether the 5' G is directly recognized by protein contacts. The Cas1 residue E184 is within 4 Å of the 5' G (**Extended Data Figure 7.4b,d**). However, a Cas1$^{E184A}$ mutation destabilizes the complex, decreasing the amount of Cas1 subunits per Cas1-2/3 complex (**Extended Data Figure 7.5**). Therefore, the decrease in integration activity of the Cas1 $^{E184A}$-2/3 complex cannot be solely attributed to a decrease in recognition of the repeat (**Extended Data Figure 7.9d,f**). Collectively, these data reveal that the 5' GT is a conserved feature necessary for integration in most CRISPR systems, though no available structure provides a mechanism for direct recognition of the 5' GT of the repeat [5,11,15,59].

## Discussion

Here we demonstrate that Cas1-2/3 and IHF fold DNA into a structure that is necessary for site-specific integration of foreign DNA into CRISPRs. IHF proteins are highly expressed and most IHF binding sites are thought to be occupied *in vivo*[60]. Therefore, IHF may pre-fold the CRISPR leader into a "landing pad" that recruits foreign DNA-bound Cas1-2/3 (**Figure 7.4e**) [22,38]. The Cas3 and Cas1 domains of the Cas1-2/3 complex are arranged like petals of a closed flower around the central Cas2 homodimer, such that the Cas3 domains occlude two of the four DNA binding surfaces on Cas2, which precludes interactions with the leader (**Figure 7.1b,d**). Foreign DNA binding to Cas1-2/3 physically constrains the Cas2/3 linker against the Cas1 homodimer, pulling the Cas3 HD domain against the Cas1-Cas1 interface (**Figure 7.2, Extended Data Figure 7.2 and Supplementary Video 1**). The 100-degree rotation of each Cas3 simulates the motion of a bloomed flower and exposes DNA binding sites on Cas2 that interact with each of the inverted repeat (IR) motifs in the leader (**Figure 7.1, 7.3, Extended Data Figure 7.2, 7.3**). Cas1-2/3 and IHF proteins fold DNA around the leader-repeat junction into a 260° loop that

docks the first CRISPR repeat into the Cas1 active site under tension (**Figure 7.1, 7.4, Extended Data Figure 7.3, 7.4**). The structure suggests that Cas1-mediated strand-transfer releases tension in this DNA loop, which may prevent disintegration of an otherwise isoenergetic strand-transfer reaction, and thereby favour complete integration (**Extended Data Figure 7.4f**). A similar

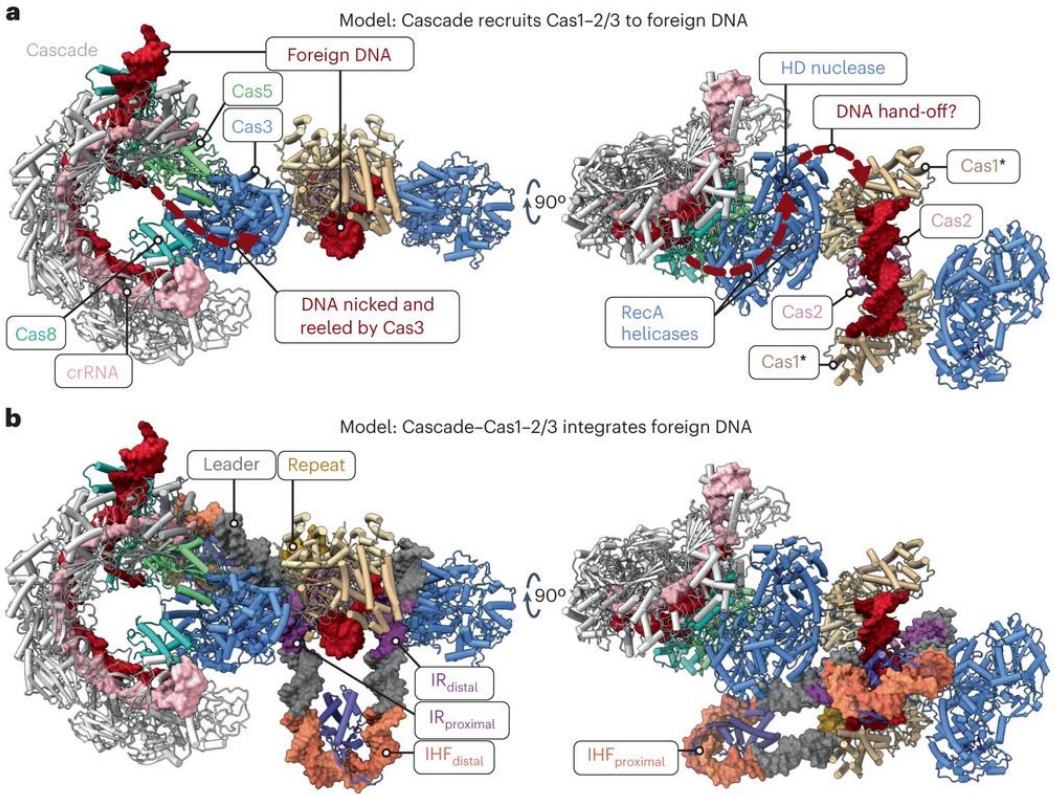



Figure 7.5. I-F CRISPR integration complex suggests a mechanism for primed acquisition and Cascade impact on integration. a, Cascade (grey) bound to foreign DNA (red) displays a Cas8 helical bundle (turquoise) that recruits Cas2/3. Cascade-recruited Cas3 degrades foreign DNA into fragments that are captured by the Cas1-2 integrase for subsequent integration [14,30,31,54,55,67,76–78]. The Cas1-2/3 integration complex can be docked onto dsDNA-bound Cascade with minimal clashing and suggests a model for the formation of a primed acquisition complex that facilitates rapid adaptation to genetic parasite variants[34,68,76]. b, Recruitment of the Cascade-Cas1-2/3 complex to IHF-folded CRISPR leader DNA is not predicted to form any new clashes, suggesting a mechanism for the role of Cascade in facilitating integration in vivo [69,70]. A total of two Cascade complexes can be docked onto the Cas1-2/3 integration complex (one per each Cas3) without introducing additional clashing, a single Cascade is shown above for clarity.

mechanism has been proposed to favour complete integration in other systems, where both strand transfer events occur simultaneously [61,62].

We show that Cas1-2/3, IHF and the I-F leader facilitate leader-side integration at four different repeat sequences (**Figure 7.4b-c**). These repeats are diverse in sequence identity, length, palindrome and GC-content, but they share a 5' GT (**Figure 7.4b and Extended Data Figure 7.8b**). To determine whether the 5' GT is a universal feature of CRISPR repeats we analyzed 24,940 CRISPRs. This analysis reveals that CRISPR repeats contain a strongly conserved 5' GT at the leader-end, and that a 5' GT is also conserved at the spacer-end of repeats from several CRISPR systems (**Extended Data Figure 7.4g**). We demonstrate the 5' GT is critical for leader-side integration and that introducing a 5' GT at the spacer-end of the repeat increases spacer-side integration. The broad conservation of 5' GT is consistent with previous reports that type I-A, II-A, and I-E systems require a 5' G for integration[11,63]. Similarly, the putative evolutionary ancestor to Cas1 enzymes, casposase, requires a conserved 5' dinucleotide in target site DNA for integration[64]. Collectively, these data suggest that Cas1 proteins retain a shared sequence preference for a 5' G or a 5' GT, and lack strict sequence requirements for the central body of the repeat (**Figure 7.4b,c**) [65]. The lack of strict sequence requirements may be advantageous because the CRISPR repeat is at the nexus of foreign DNA integration, processing of the transcribed CRISPR, and loading the mature crRNA into the surveillance complexes (e.g., Cascade, Cas9).

Genetic parasites commonly escape CRISPR-based immunity through point mutations [66]. To counter escape mutants, many CRISPR-Cas systems use existing spacer sequences to enhance the acquisition of new spacers from the same foreign genetic element via "primed" acquisition[4,30–37]. In some examples of primed acquisition, the Cas3 nuclease/helicase degrades

CRISPR-targeted DNAs into ssDNA fragments enriched in PAM-containing termini[67]. Cas1-2 has been proposed to anneal complementary ssDNA fragments and integrate these into CRISPRs[14,30,31,67]. Single-molecule co-localization and bulk immunoprecipitation suggest that the type I-E Cas1-2 integrase is recruited to a Cas3-Cascade-target DNA complex to facilitate primed acquisition [34,68]. The structure of the type I-F integration complex reveals conformational changes in Cas3 that may enable interactions with DNA-bound Cascade (**Figure 7.5a**)[36,69]. Cascade improves integration efficiency and fidelity *in vivo* [69,70], and the structure suggests a model for the formation of a primed acquisition complex (Cas1-2/3-Cascade-target DNA) that transfers new foreign DNA fragments to the integrase. Additional structures will be necessary to clarify the mechanism(s) of primed adaptation.

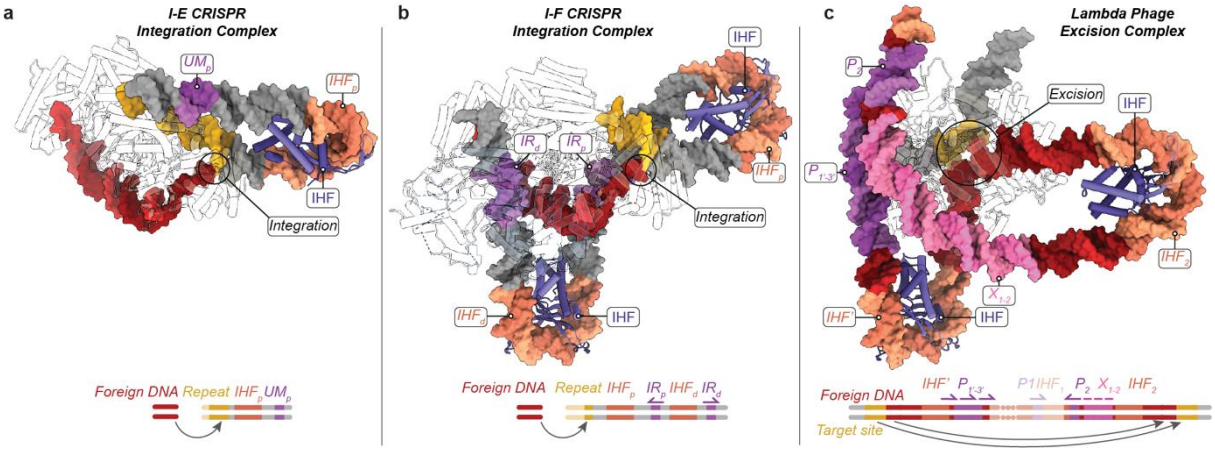



Figure 7.6. DNA is a flexible scaffold that controls DNA mobilization. Structures for the I-E CRISPR integration complex (a), I-F CRISPR integration complex (b), and Lambda phage excision complex (c). DNA shown as a surface, IHF (purple) and all other proteins shown as transparent cartoons. Integration and excision sites, along with DNA motifs that regulate DNA mobilization, are labeled and colored according to the schematic (bottom).

A comparison of the I-E and I-F integration complexes with a structure of the λ-phage excision complex, reveals structural similarities and differences. In I-E systems, the IHF protein folds the leader DNA to present an upstream motif to a lobe of Cas1. In contrast, the I-F structure

highlights extensive cooperation between IHF and Cas1-2/3 in bending the leader into an energetically strained conformation that may increase the specificity of CRISPR recognition (**Figure 7.6a,b**). This cooperation includes Cas1-2/3 kinking the IR DNA motifs presented as parrallel DNA pillars by IHF, and Cas1-2/3 constricting the $IHF_{proximal}$ 180° bend ito a 260° (**Figure 7.6b**). Further, the sequestration of the IR-binding surface of Cas2 suggests a unique structural mechanism that prevents Cas1-2/3 interactions with the leader until foreign DNA binding induces a rotation of Cas3 (**Figure 7.4e, Supplementary Video 1 and 3**) [54].

Many CRISPR leaders contain multiple IHF binding sites and subtype-specific motifs that are reminiscent of motifs found at the ends of transposons, phages, and plasmids, which facilitate integration, excision, or recombination complex assembly (**Extended Data Figure 7.8a**)[3–7,22,40–53]. For example, the left and right ends of the λ-phage genome contain three IHF binding sites, five copies of a second DNA motif (P1-2,1'-3'), and three copies of third DNA motif (X1,1.5,2), that recruit λ-phage (Int, Xis) and non-λ-phage (IHF, Fis) proteins (**Figure 7.6c**)[7]. These proteins use DNA as a flexible scaffold that organizes enzyme active sites and DNA substrates to facilitate either the integration, or excision, of λ-phage DNA from the bacterial genome (**Figure 7.6c**). Similarly, the I-F CRISPR leader motifs recruit Cas (Cas1-2/3) and non-Cas (IHF) proteins, that bend DNA into a flexible scaffold to organize enzyme active sites and DNA substrates in space, facilitating the integration of foreign DNA into the bacterial genome (**Figure 7.6b**). Diverse systems thus use DNA as a flexible scaffold to regulate the isoenergetic mobilization of DNA (**Figure 7.6a-c**).

Diverse DNA mobilizing enzymes across the tree of life co-opt DNA folding to regulate DNA mobilization [3–7]. In sum, these data provide a mechanistic understanding for the role of

DNA as a flexible scaffold that controls DNA mobilization. These insights are critical to developing applications of DNA-mobilizing enzymes in gene therapy, genetic engineering, and chronological DNA recordings [6,53,71–75].

## Methods

### Nucleic Acid Preparation

Four single-stranded DNAs (**Supplementary Table 7.1**) were synthesized (IDT) and resuspended in 1x TE buffer (10 mM Tris-HCl pH 8, 1 mM EDTA) before being used to assemble the structure of the type I-F integration complex, the assembly is detailed in a below section. The splayed foreign DNAs used in integration assays (**Supplementary Table 7.1**) were synthesized and resuspended in 1x TE buffer before use. To make $^{32}$P-labeled CRISPR integration substrates, the sequences consisting of the leader and CRISPR arrays were first synthesized and cloned into pUC57 (Genscript). These plasmids have been made available on Addgene (**Supplementary Table 7.1**). These plasmids were transformed into chemically competent *E. coli* DH5α cells and the transformed cells were plated onto LB agar plates containing 100 μg/mL Ampicillin. These cells were cultured in LB media and plasmids were purified using ZymoPURE II Plasmid Midiprep kit (Zymo Research). Each plasmid was then digested with EcoRI-HF and BamHI-HF (NEB) restriction enzymes, and the 294-383 bp inserts of interest were separated from the vector backbone by agarose gel electrophoresis. The gel segments containing the DNA inserts of interest were excised and DNA was purified using a Zymoclean Gel DNA Recovery kit (Zymo Research) (**Supplementary Table 7.1**). The 5' ends of the CRISPR leader and array fragments were dephosphorylated using Quick Calf Intestinal alkaline Phosphatase (NEB), and the DNAs were purified away from protein using a DNA Clean

& Concentrator kit (Zymo Research). Both 5' ends of 1 pmole of the CRISPR leader and array fragments were then labelled with $^{32}P$, by incubation with 4 pmoles of [$\gamma$-$^{32}P$]ATP (PerkinElmer) by polynucleotide kinase (NEB) in 1x PNK buffer at 37°C for 45 minutes. PNK was heat-denatured by incubation at 65°C for 20 minutes. Spin column purification (G-25, GE Healthcare) was used to remove unincorporated radioactive nucleotides and to buffer exchange DNAs into 1x TE buffer.

Cas1 and Cas2/3 Mutagenesis

The plasmid used to express Cas1 and Cas2/3 (Addgene #89240) was PCR amplified with mutagenic primer pairs using Q5 polymerase (NEB) (**Supplementary Table 7.1**). The parental template plasmid was digested with DpnI (NEB) and the PCR amplified products were purified using the DNA Clean and Concentrate kit (Zymo). The purified DNA was 5' phosphorylated with T4 Polynucleotide Kinase (NEB) and ligated with homemade T4 DNA ligase. The ligation reactions were transformed into DH5α cells. The Cas2 K11D, R12E, R55E, N56D mutant was prepared using the Cas2 K11D, R12E mutant as the template with the PCR primers for Cas2 R55E, N56D in the mutagenic PCR reaction. Plasmids for expressing the mutant Cas1-2/3 complexes, Cas1H25A-2/3, Cas1E184A-2/3, Cas1-2K11D,R12E/3, Cas1-2R55E,N56D/3, and Cas1-2K11D,R12E,R55E,N56D/3 have been deposited at Addgene (#200213, #200214, #200216, #200217, #200218).

Protein Purification

*P. aeruginosa* IHF heterodimer was purified as previously described[22]. Briefly, 6xHis-tagged IHFα and StrepII-tagged IHFβ were co-expressed in *E. coli* BL21(DE3). Cell pellets were lysed by sonication in IHF lysis buffer (25 mM HEPES-NaOH pH 7.5, 500 mM NaCl, 10 mM

Imidazole, 1 mM TCEP, 5% Glycerol), supplemented with 0.3x Halt Protease Inhibitor Cocktail (ThermoFisher), at 4°C. Lysate was clarified by two rounds of centrifugation at 10,000 x g for 15 minutes, at 4°C. His-tagged IHF was captured on HisTrap HP resin (Cytiva), and eluted with 500 mM Imidazole. Affinity tags were cleaved using PreScision protease, and the PreScision protease and remaining 6x-His-IHFα were removed by affinity chromatography using HisTrap HP resin (Cytiva). Untagged IHF heterodimer was then further purified on Heparin Sepharose (Cytiva) and eluted with a linear gradient to a buffer containing 2 M NaCl. Fractions containing IHF heterodimer were concentrated and further purified by SEC (Size Exclusion Chromatography) on a Superdex 75 column (Cytiva) equilibrated in IHF buffer (25 mM HEPES-NaOH pH 7.5, 200 mM NaCl, 5 % Glycerol) (**Extended Data Figure 7.1a**). IHF overexpression plasmids have been previously described and deposited at Addgene (#149384, #149385) [22].

*P. aeruginosa* Cas1-2/3 heterohexamer complexes (wildtype and mutants) were purified as previously described [22]. Briefly, StrepII-tagged Cas1-Cas2/3 was overexpressed in *E. coli* BL21(DE3). Cell pellets were lysed via sonication in Cas1-2/3 lysis buffer (50 mM HEPES pH 7.5, 500 mM KCl, 10% Glycerol, 1 mM DTT), supplemented with 0.3x Halt Protease Inhibitor Cocktail (ThermoFisher), at 4°C. Lysate was clarifed as above. StrepII-tagged Cas1-Cas2/3 complexes were affinity purified on StrepTrap HP resin (GE Healthcare) and eluted with Cas1-2/3 Lysis Buffer containing 3 mM desthiobiotin (Sigma-Aldrich). Eluate was concentrated at 4°C (Corning Spin-X concentrators), before purification over a Superdex 200 size-exclusion column (Cytiva) equilibrated in 10 mM HEPES pH 7.5, 500 mM Potassium Glutamate, and 10% Glycerol (**Extended Data Figure 7.1c and 7.5e**).

*In Vitro* Integration Assays

End-point integration reactions were performed in triplicate using 300 nM of splayed foreign DNA fragments, containing or lacking a PAM (IDT), 200 nM of Cas1-2/3, 300 nM of IHF heterodimer and roughly 1 nM of a given $^{32}$P-labelled CRISPR variant fragment in Integration Buffer (20 mM HEPES pH 7.5, 150 mM potassium glutamate, 5 mM $MnCl_2$, 1 mM TCEP, 1% Glycerol) (**Supplementary Table 7.1)**. Reactions were assembled on ice and then incubated at 35°C for 20 minutes.Timecourse integration reactions were performed in triplicate using 300 nM of splayed, or fully complementary, foreign DNA fragments lacking a PAM (IDT), 200 nM of Cas1-2/3, 300 nM of IHF heterodimer and roughly 1 nM of a given $^{32}$P-labelled CRISPR variant fragment in modified Integration Buffer (20 mM HEPES pH 7.5, 150 mM potassium glutamate, 5 mM $MnCl_2$, 10 mM TCEP, 1% Glycerol) (**Supplementary Table 7.1)**. We noticed that IHF and Cas1-2/3 protein stocks exhibited a propensity to precipitate when diluted into pre-chilled buffer to form working dilution stocks. Therefore all working dilutions of IHF and Cas1-2/3 prepared for timecourse assays were made by first diluting the proteins into room-temperature buffer, mixed, and then chilled on ice. Reactions were assembled on ice and then incubated at 20 minutes. Timepoints were taken at 0, 1, 2, 4 and 8 minutes. Reactions were stopped by the addition of phenol. The aqueous (nucleic acid-containing) layer was mixed 1:1 with 2x formamide loading buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol) and then denatured at 95°C for 5 minutes, before resolving the $^{32}$P-labeled CRISPR substrates and integration products on a 7% (w/v) (29:1 mono:bis) polyacrylamide Urea gel in 1x TBE (100 mM Tris-Borate pH 8.3, 2 mM EDTA). Gels were dried and quantified using a Typhoon phosphorimager (GE Healthcare). The intensity of full-length CRISPR variant, leader-side integration fragments and spacer-side integration fragments were quantified with

Multi Gauge v3 (Fujifilm). These readings were then used to calculate leader- and spacer-side integration events as percentages of all events. Images of all gels that resolved integration reactions are shown **Extended Data Figure 7.6,7.7 and 7.9**), and additional control gels show that Cas1-2/3 is required for integration, and show how the custom $^{32}$P-labelled ladder was generated by restriction enzyme digestion of $^{32}$P-labelled CRISPR variant DNAs (**Extended Data Figure 7.8**). Timecourse integration data was fit to a plateau followed by one phase association (GraphPad Prism v10).

Assembly and Purification of I-F Integration Complex

A total of four single-stranded DNAs (ssDNAs) synthesized to mimic a half-site integration intermediate were annealed in a step-wise manner. 2 nanomoles of ssDNAs mostly corresponding to the sense and anti-sense strands of the CRISPR leader ("strand_1" and "strand_2" were denatured at 100°C and then slow annealed using a PCR program that cooled the samples to 25°C in 6°C over an hour, in 100 µL of hybridization buffer (20 mM Tris-HCl pH 7.5, 100 mM monopotassium glutamate, 5 mM EDTA, 1 mM TCEP). 2 nanomoles of ssDNAs mostly corresponding to the corresponding to the sense and anti-sense of the strands of the foreign DNA ("strand_3" and "strand_4") were slow annealed using the same protocol (**Extended Data Figure 7.1a and Supplementary Table 7.1**). The two sets of annealed DNAs (tube 1: "Strand_1" and "Strand_2"; tube 2: "Strand_3" and "Strand_4") were mixed together, heated to 80°C and then slow annealed using a PCR program that cooled the samples to 25°C in 6°C over an hour, to anneal the complementary sense and anti-sense regions of the CRISPR repeat included in Strand_2 and Strand_3 together. Next, 6 nanomoles of IHF heterodimer in 50 µL of hybridization buffer was warmed to 25°C and then mixed and incubated with the annealed

DNAs at 25°C for 10 minutes. Next, 3 nanomoles of Cas1-2/3 in 250 µL of hybridization buffer was warmed to 25°C and mixed with the prepared DNA and IHF mixture, and incubated at 25°C for 10 minutes. The total concentration of monopotassium glutamate in the mixture at this stage was ~200 mM, due to carryover from the stored protein stocks. This sample was centrifuged at 22,000 g, 4°C for 20 minutes to remove precipitates. The type I-F CRISPR integration complex was then purified on a Superdex 200 10/300 column (Cytiva) equilibrated in SEC buffer (20 mM Tris-HCl pH 7.5, 200 mM monopotassium glutamate, 5 mM EDTA, 1 mM TCEP, 2 % Glycerol). 0.5 ml fractions were individually concentrated and stored. The sixth SEC fraction contained all DNAs and proteins of interest and was further analyzed by cryo-EM (**Extended Data Figure 7.1d-f**).

Cryo-EM Sample Preparation and Data Acquisition

Purified integration complex was diluted to a concentration of 1 µM in SEC buffer lacking glycerol (20 mM Tris-HCl pH 7.5, 200 mM monopotassium glutamate, 5 mM EDTA, 1 mM TCEP), such that the final glycerol concentration was 0.2% within 1 hour of freezing. Sample was applied to Quantifoil R2/2 Cu 200 mesh grids that were glow discharged using 15 mA for 15 seconds with a 10 second hold (easiGlow, Pelco). 4 µl of diluted integration complex was applied to the grids, and then the grids were blotted for 5-6 seconds using Vitrobot™ Filter paper (Electron Microscopy Sciences) with a blot force of 6, at 100% humidity, 8°C, followed by plunge freezing into liquid ethane using a Vitrobot (Mk. IV, ThermoFisher Scientific). A preliminary dataset of 230 movies was collected on Montana State University's Talos Arctica transmission electron microscope (ThermoFisher Scientific), with a field emission gun operating at an accelerating voltage of 200 kV using parallel illumination conditions [76]. Movies were

acquired using a Gatan K3 direct electron detector, operated in electron counting mode applying a total electron exposure of 50 e-/Å$^2$ over 50 frames (3.995 s exposure, 0.08 s frame time). The SerialEM data collection software was used to collect micrographs at 36,000-fold nominal magnification (1.152 Å/pixel at the specimen level) with a nominal defocus set to 0.5 µm - 2.0 µm [77]. Stage movement was used to target the center of four 2.0 µm holes for focusing, and image shift was used to acquire high magnification images in the center of each of the holes. An preliminary reconstruction was determined from a curated set of 160 images that had CTF fits less than 9 Å and a full-frame motion less than 40 pixels. Briefly, a round of blob picking (150-270 Å) followed by 2D classification was used to identify 2D classes used as templates for template picking in cryoSPARC [78]. Template picking identified 33,403 initial particles from the above 160 images. 2D classification of these 33,403 particles into 50 classes was used to identify 5 classes with strong structural features containing 4,002 particles. Non-uniform refinement of these 4,002 particles resulted in a ~14.7Å reconstruction that appeared to contain a complete integration complex, therefore new grids were prepared as described above and shipped to National Center for CryoEM Access and Training (NCCAT) and the Simons Electron Microscopy Center located at the New York Structural Biology Center (NYSBC) for additional data collection (**Table 7.1**). At NCCAT, grids were imaged using a 300 kV Titan Krios G3i (Thermo Fisher Scientific) equipped with a GIF BioQuantum and K3 camera (Gatan). 10,740 images were recorded with Leginon v3.5 [79] (Suloway et al., 2005) with a calibrated pixel size of 0.5335 Å/px (micrograph dimension of 11520 x 8184 px) over a nominal defocus range of −0.7 µm to −2.1 µm and 20 eV slit. Movies were recorded in “super-resolution mode” (native K3

camera binning 1) with subframes of 50 ms over a 2.5 s exposure (50 frames) to give a total exposure of ~69 e-/Å$^2$ (**Table 7.1**).

Cryo-EM Image Processing

Patch motion correction and patch CTF correction were performed in cryoSPARC [78]. 3,792 of 10,740 total images (*CTF < 8Å, Full-frame motion < 30Å*) were processed first to build an initial template. Blob picking was used to pick particles with diameters ranging from 120-280 Å. These ~1.8 million particles were extracted, Fourier-binned 2x2, and then subjected to 2D classification (*Custom parameters: Initial classification uncertainty factor = 3; Number of online-EM iterations = 30; Batchsize per class = 200*) (**Extended Data Figure 7.1g**). Particles from 82 of the 200 2D classes were selected for an initial round of *ab initio* reconstruction and heterogeneous refinement. Particles from 1 of 5 of these classes were selected for a second round of *ab initio* reconstruction and heterogeneous refinement. Particles from 1 of 3 of these classes (174,000 particles) were selected for Non-uniform refinement (*Custom parameters: Optimize per-particle defocus = true; Optimize per-group CTF params = true*) to create an initial reconstruction with a resolution of ~3.7Å, that was used to calculate templates [80]. These templates were used to choose particles from 9,858 of 10,740 total images (CTF < 8Å cutoff). These ~5.85 million particles were classified into a total of 6 classes by Heterogeneous refinement, that were seeded with 1 good volumes and 5 junk volumes taken from the above Heterogeneous refinement analyses. The ~1.31 million particles in the single selected class, were passed through a round of 2D classification (*Custom parameters: Batchsize per class = 200*) (**Extended Data Figure 7.1h**). ~1.29 million particles from 49 of the 50 2D classes were selected for a round of *ab initio* reconstruction followed by heterogeneous refinement into two

classes. ~1.1 million particles from one of these two classes were subjected to 3D classification into 4 classes (*Custom parameters: Batchsize per class = 20,000; Initialization mode = PCA; Target resolution = 2Å; Particles per reconstruction = 500; Class similarity = 0.3*) , followed by separate Non-uniform refinements of particles from each of these four classes (*Custom parameters: Optimize per-particle defocus = true; Optimize per-group CTF params = true*)[80]. The final set of 366,794 particles were re-extracted and re-centered, Fourier-binned 2x2, and subjected to Non-uniform refinement to generate a final reconstruction refined to a global resolution of 3.48 Å based on the 0.143 FSC cutoff (**Extended Data Figure 7.1h-k and Table 7.1**)[81]. The 3D FSC was calculated using webserver (3dfsc.salk.edu)[82].

Model Building and Validation

The map was sharpened from two half-maps using the local anisotropic sharpening job in Phenix[83]. The published structure of the *P. aeruginosa* Cas1 homodimer was used as starting models[56], because Colabfold consistently failed to predict the alternative fold that one Cas1 subunit adopts to form the assymetric homodimer interface[84], even when provided template structures. Whereas the Colabfold-predicted models for the *P. aeruginosa* IHF heterodimer, and the *P. aeruginosa* Cas2/3 subunit were used as starting models. The conformation of the DNA sequences within the *E. coli* IHF-DNA co-crystal structure (PDB: 1IHF) [38] was used as starting models for DNA segments within the $IHF_{distal}$ and $IHF_{proximal}$ DNA bends. For all other double stranded DNA segments, B-form DNA was used as a starting model. Single-stranded DNA segments were built in *de novo*. Protein and DNA segments were individually rigid-body fitted into the EM density map. The relative orientation of the Cas2, Cas2/3 linker and Cas3 domains were corrected by real-space refinement into the EM density map in WinCoot [85]. The ReadySet

job in Phenix was used to generate hydrogens on all proteins and nucleic acids and prepare the model for further refinement. Then protein and DNA segments were real-space refined in WinCoot [85], restrained to ideal geometry, secondary structure and German McClure distance restraints generated in ProSMART from the input models[86]. The models were iteratively real-space refined in WinCoot and in Phenix using Ramachandran and secondary structure restraints [83,85]. The starting model was used as a reference model, and harmonic restraints on the starting coordinates were enabled. MolProbity[87] and the PDB validation service server (https://validate-rcsb-1.wwpdb.org/) were used to identify problem regions subsequently corrected in WinCoot [85]. For regions of the reconstruction where side chains are not visible (resolution >4.0Å) the atomic model was truncated to the peptide backbone. For regions of the reconstruction where the backbone was ambigous the sections of the peptide or DNA model were removed. Contacts and hydrogen bonds between residues were identified by ChimeraX v1.4 using the "contacts" and "hbonds" commands respectively, with default parameters [88,89]. The DNAproDB webserver (https://dnaprodb.usc.edu/) was further used to analyze DNA-protein contacts (**Extended Data Figure 7.4d,e**) [90]. Structure-guided mutagenesis was used to further validate key Cas1-2/3-DNA contacts in the above biochemical assays.

Cas1, Cas2/3, and Repeat Conservation Analysis

To build a list of type I-F Cas1 sequences, CRISPRDetect v2.4 with default parameters was used to identify CRISPR arrays within a total of 18,225 bacterial and 376 archaeal complete genomes accessed from the NCBI Assembly database on June 10$^{th}$ of 2019 as previously described [22,91]. 15,274 high-confidence CRISPR arrays were classified with a CRISPR subtype by CRISPRDetect v2.4 (by matching to a list of repeats with known subtype annotations), and by

genetic proximity to subtype-specific *cas* genes (within 20,000 bp). To identify *cas* genes, the 20,000 bp flanking the CRISPR were submitted to PRODIGAL v2.6.3 (default parameters) to predict all potential open reading frames (ORFs) [92]. This ORF database was then used as input to search for *cas* gene clusters with MacsyFinder v1.0.5 [93]. The following parameters were used: "*macsyfinder --sequence-db <peptide_database> --db-type gembase -d <CRISPR_subtype_definitions> -p <HMM_profiles> -w 50 -vv all*". HMM profiles and classification definitions used in MacsyFinder were acquired from the local version of CRISPRCasFinder v4.2.20 [94]. Next, the first repeat and 200 nucleotides upstream of CRISPR arrays (leader) which were classified as Type I-F (1,683 arrays) were collected. A non-redundant list of I-F CRISPR leaders (536 leaders) was generated using CD-HIT v4.8.1 with a 95% identity cutoff [95]. A local copy of FIMO was used to find significant matches to the position weight matrix representing I-F IHF binging site as previously described [22,96]. I-F CRISPR arrays that possess more than one IHF site ($IHF_{proximal}$ and/or $IHF_{distal}$) in the leader sequences were extracted for downstream analyses. Cas1 homologs were identified within the 20,000 base-pair flanking regions of extracted 444 I-F CRISPR arrays by using PRODIGAL and MacsyFinder with the same parameters described above. 371 Cas1 homologs associated with Type I-F CRISPRs and possessing at least one IHF site in the leader sequences were identified. A non-redundant list of Cas1 sequences was generated CD-HIT v4.8.1 with a 95% identity cutoff, resulting in 222 sequences [95]. Sequences smaller than 200 residues and larger than 500 residues were removed, and the remaining 205 sequences were further curated with MaxAlign, which selected a list of 144 unique type I-F Cas1 sequences [97]. The *P. aeruginosa* PA14 Cas1 sequence was then added to a final list of 145 type I-F Cas1 sequences. To build a list of type I-F Cas2/3

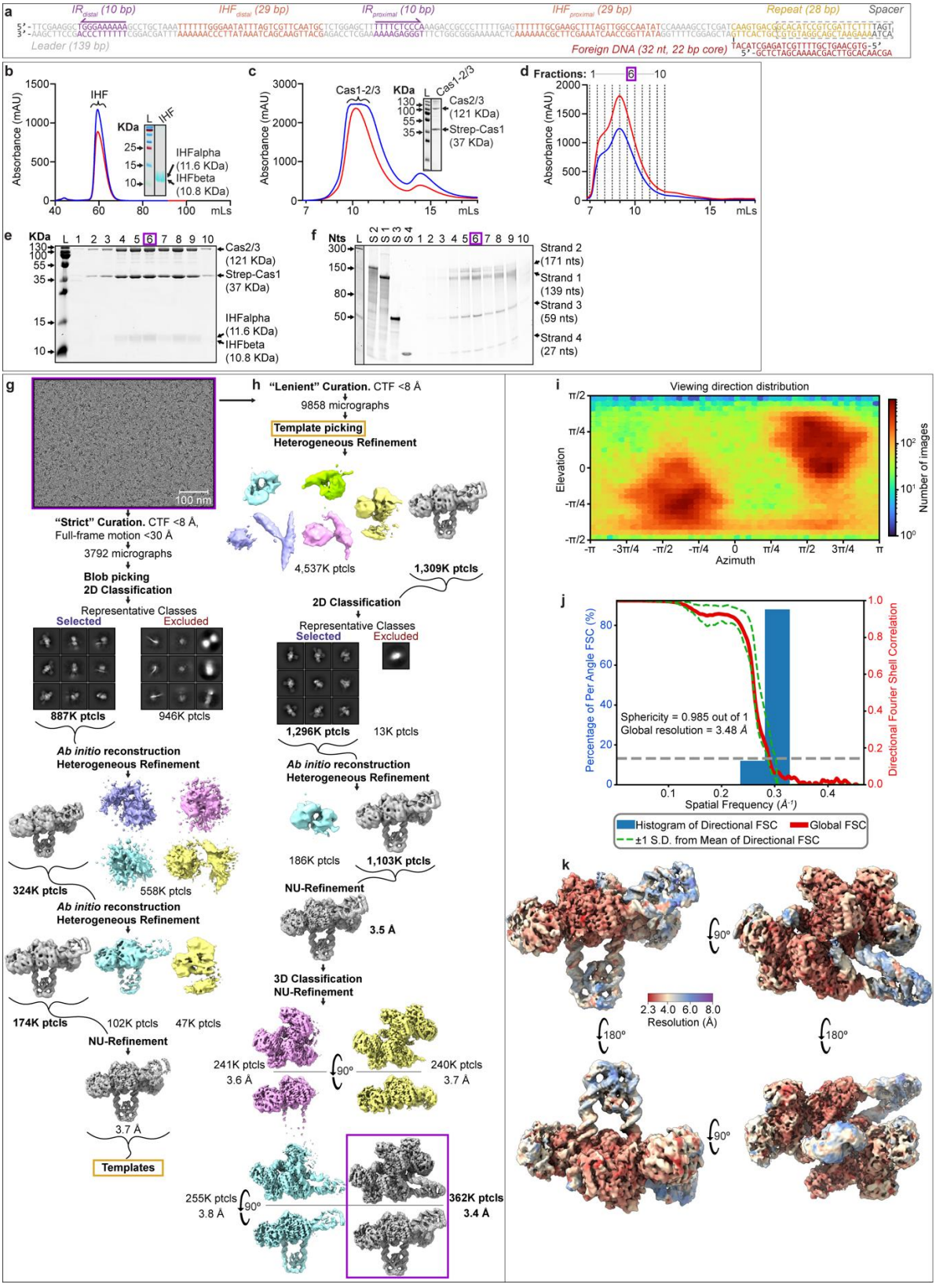
sequences, the *P. aeruginosa* PA14 Cas2/3 sequence was used as an input for HHMER for a search for homologs using 3 iterations, an E-value cutoff of 0.0001, against the UNIREF-90 database [98,99]. A list of 500 representative sequences was further curated with MaxAlign, to generate a final list of 458 unique Cas2/3 sequences. Type I-F Cas1 and Cas2/3 sequences were aligned using the MAFFT webserver with the E-INS-I iterative refinment methods to result in alignments with the highest number of gap-free sites [100].

To build an updated list of CRISPR repeat sequences, CRISPRDetect v3.0 with default parameters was used to identify CRISPR arrays within a total of 25,502 bacterial and 398 archaeal complete genomes and chromosomes accessed from the NCBI RefSeq Assembly database (accessed on June 10$^{th}$, 2021)[91]. This search identified CRISPR loci within 58,864 genomic and plasmid sequences, resulting in 24,940 high-confidence CRISPR loci predictions (array quality score >3). Similar to above, CRISPRDetect annotated the subtype of 14,446 of these CRISPR loci, based on the sequence similarity of the repeats in these loci to known CRISPR repeats. The subtypes of the remaining 10,494 CRISPR loci were determined by their proximity to subtype-specific *cas* genes as described above. 5,321 of the 10,494 unclassified CRISPR loci were assigned a subtype using this protocol, such that 5,173 CRISPR loci remained unclassified. The consensus repeat for each of the 24,940 CRISPR loci, as reported by CRISPRDetect, were used for downstream analyses. To ensure the repeats were arranged in the correct orientation, the 24,940 repeats were grouped by subtype, and each group was individually aligned by MAFFT using the "*--adjustdirection*" parameter. Sequence logos of the first and last three bps of CRISPR repeats were made using Weblogo v3.7.1 for CRISPR subtypes and across all subtypes[101,102] (**Extended Data Figure 7.4g**).

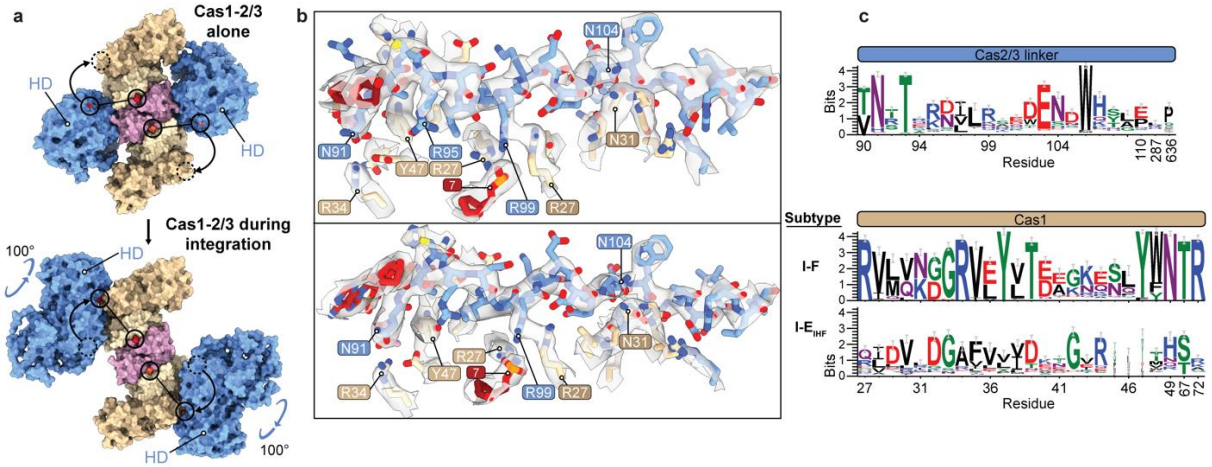
## Data Availability

The data that support the findings of this study are available from the corresponding author Blake Wiedenheft upon request. Cryo-EM maps were deposited in the Electron Microscopy Data Bank under accession number EMD-29280. The atomic model of the type I-F integration complex was deposited in the PDB under accession number 8FLJ. Plasmids generated in this study are available from Addgene. Source data are provided with this paper.

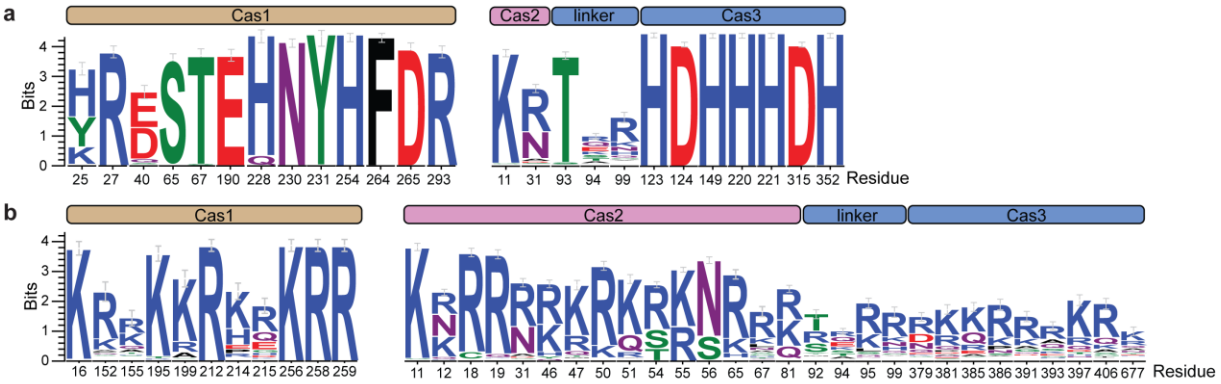
## Extended Data Figures


a
IR distal (10 bp)
IHF distal (29 bp)
IR proximal (10 bp)
IHF proximal (29 bp)
Repeat (28 bp)
Spacer
Leader (139 bp)
Foreign DNA (32 nt, 22 bp core)
b
IHF
KDa
IHFalpha (11.6 KDa)
IHFbeta (10.8 KDa)
Absorbance (mAU)
mLs
c
Cas1-2/3
Cas2/3 (121 KDa)
Strep-Cas1 (37 KDa)
d
Fractions:
e
Cas2/3 (121 KDa)
Strep-Cas1 (37 KDa)
IHFalpha (11.6 KDa)
IHFbeta (10.8 KDa)
f
Nts
Strand 2 (171 nts)
Strand 1 (139 nts)
Strand 3 (59 nts)
Strand 4 (27 nts)
g
100 nm
"Strict" Curation. CTF <8 Å, Full-frame motion <30 Å
3792 micrographs
Blob picking
2D Classification
Representative Classes
Selected
Excluded
887K ptcls
946K ptcls
Ab initio reconstruction
Heterogeneous Refinement
324K ptcls
558K ptcls
174K ptcls
102K ptcls
47K ptcls
NU-Refinement
3.7 Å
Templates
h
"Lenient" Curation. CTF <8 Å
9858 micrographs
Template picking
Heterogeneous Refinement
4,537K ptcls
1,309K ptcls
2D Classification
1,296K ptcls
13K ptcls
186K ptcls
1,103K ptcls
3.5 Å
3D Classification
241K ptcls
3.6 Å
240K ptcls
3.7 Å
255K ptcls
3.8 Å
362K ptcls
3.4 Å
i
Viewing direction distribution
Elevation
Azimuth
Number of images
j
Percentage of Per Angle FSC (%)
Directional Fourier Shell Correlation
Sphericity = 0.985 out of 1
Global resolution = 3.48 Å
Spatial Frequency (Å⁻¹)
Histogram of Directional FSC
Global FSC
±1 S.D. from Mean of Directional FSC
k
Resolution (Å)

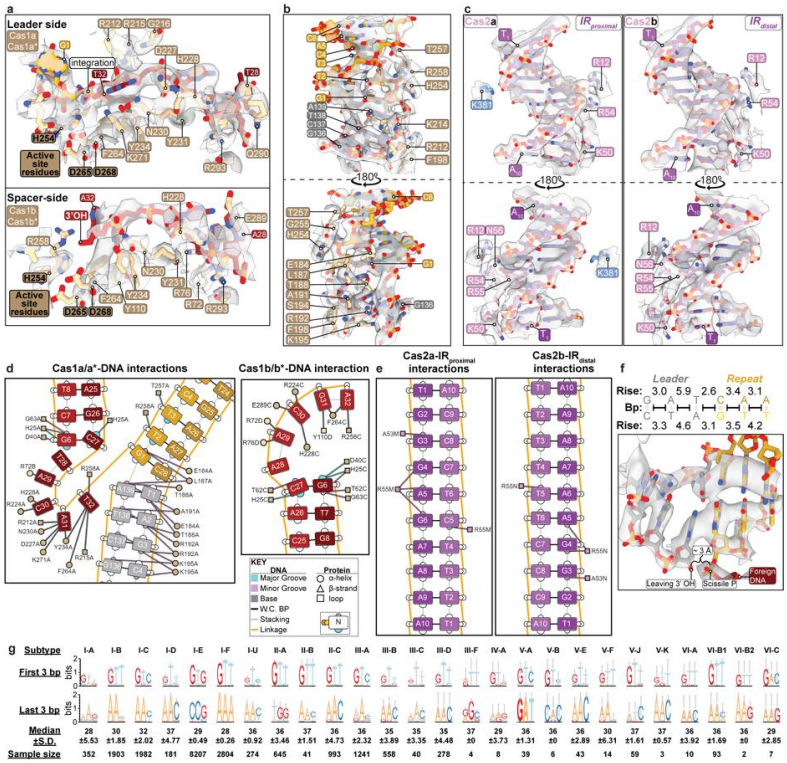
Extended Data Figure 7.1. Cryo-EM sample preparation, imaging and processing for type I-F integration complex. a, Sequence-level schematic of DNA used to assemble the integration complex. The length of each motif is listed. The latter two thirds of the CRISPR repeat and second spacer (grey dashed box) could not be resolved in the cryo-EM reconstruction. See also Supplementary Table 1. b, Size-exclusion chromatography (SEC) profile (Superdex 75 16/600, Cytiva) of IHF heterodimer purified as described in methods section, and SDS-PAGE gel (inset). c, SEC profile (Superdex 200 10/300, Cytiva) of Cas1–2/3 heterohexamer purified as described in methods section, and SDS-PAGE gel (inset). d, I-F integration complex was assembled from purified DNAs, IHF and Cas1–2/3 as described in the methods section, and the assembled complex was further purified by size-exclusion chromatography (SEC) (Superdex 200 10/300, Cytiva). Individual fractions were collected along the elution profile, and were concentrated and stored separately for further analysis and imaging. e, Individual SEC fractions were analyzed by SDS-PAGE to determine which fractions contained all the proteins necessary for a complete complex. f, Individual SEC fractions were phenol-chloroform extracted, and the aqueous layer was analyzed by Urea-PAGE to determine which fractions contained all four DNA strands necessary for a complete complex. The fraction chosen for cryo-EM analysis is indicated with a dotted purple box. g, Image processing pipeline for a small subset of 10,740 total micrographs for the type I-F integration complex, to generate an initial model for template picking. Scale bar represents 100 nm. h, Final image processing pipeline for the type I-F integration complex. i, Viewing direction distribution plot depicting particle orientations present in final reconstruction. More populated views are shown in red, and less populated views are shown in blue. j, 3D Fourier Shell Correlation (3DFSC) of the final I-F integration complex reconstruction. The global resolution at 0.143 is indicated by a dashed line, 3.48 Å. k, Local resolution estimation of the cryo-EM reconstruction calculated by cryoSPARC81. The purification of proteins, assembly of the integration complex, and analysis of these samples by SDS-PAGE or Urea page was performed once. Micrographs were collected on two separate occasions with similar results.



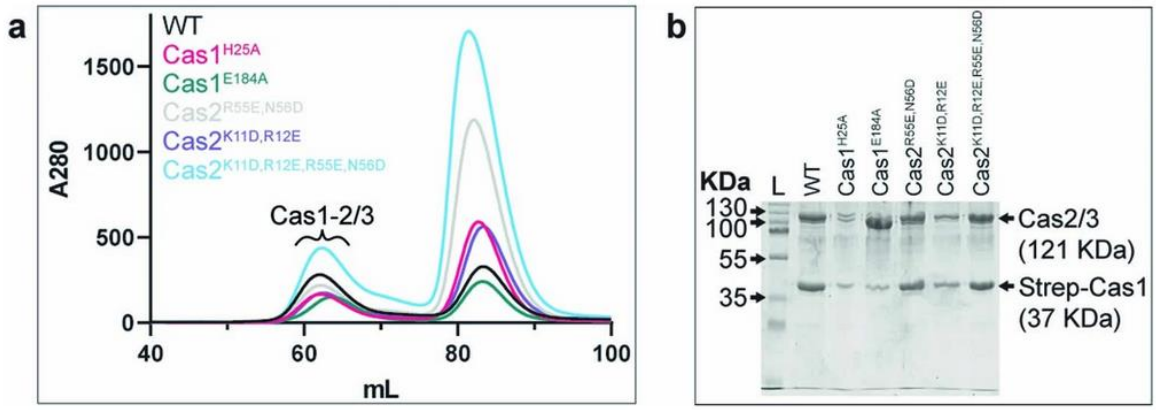

Extended Data Figure 7.2. Cas1-2/3 undergoes a large structural rearrangement during integration. a, The Cas3 domains of the Cas1-2/3 complex have undergone a ~ 100° rotation in the structure of the integration complex as compared to Cas1-2/3 alone54. The positions of the first (90) and last residues (110) of the Cas2/3 linker are shown (Cas2/3a, red; Cas2/3b, tomato). The linker residues were not resolved for the previously determined pseudo-atomic model of Cas1-2/3 alone, and so are not shown for either complex for clarity[54]. The rotation of the Cas3 domains outwards unveils new DNA binding sites on two opposing faces of the Cas2 dimer. b, Zoom-in on the atomic fit of the Cas2/3 linker to the cryo-EM map. The Cas2/3 linker is disordered in the absence of Cas1 (PDB:5B7I)[55], but has become ordered in the I-F integration complex structure due to packing by the foreign DNA against the Cas1 beta hairpins. c, A sequence logo depicting the conservation of Cas2/3 linker residues (top). The Cas1 residues that contact the Cas2/3 linker are conserved in Cas1 proteins associated with type I-F CRISPR loci (middle), but are not conserved in the closely related Cas1 proteins associated with type I-E CRISPR loci that have similar IHF motif-containing leaders[22]. Residues are numbered according to P. aeruginosa PA14 Cas1 and Cas2/3 proteins.



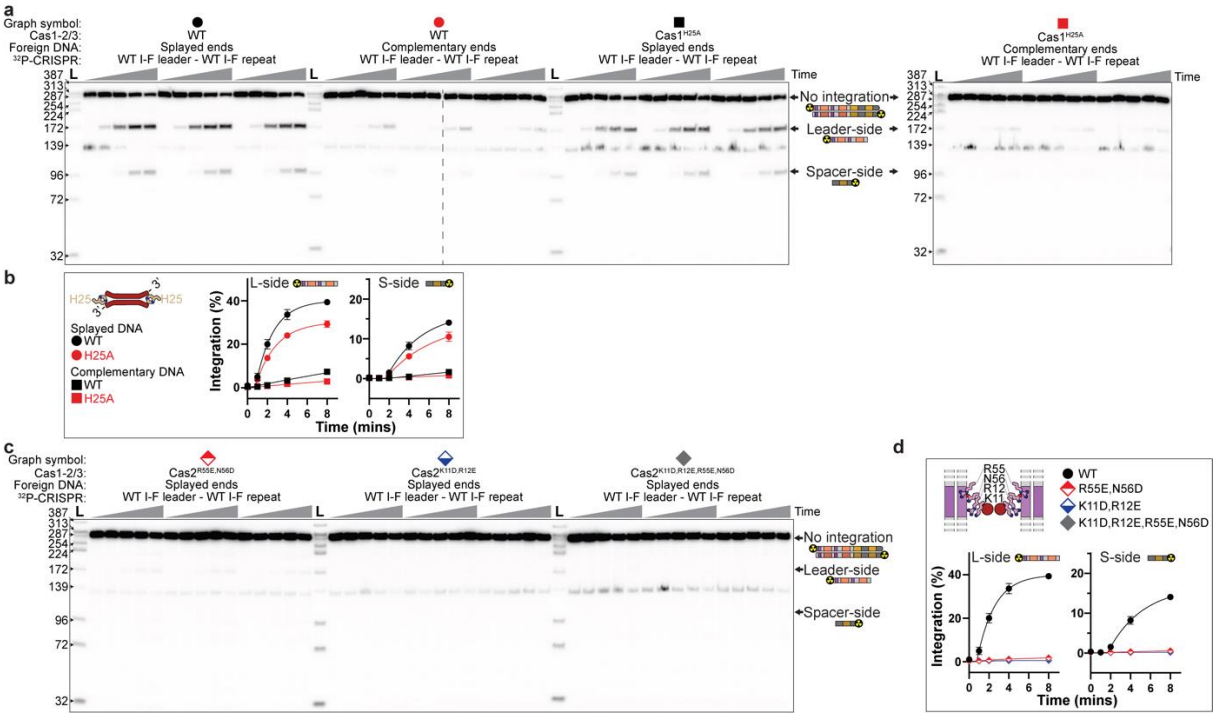

Extended Data Figure 7.3. Conservation analysis of Cas1-2/3 residues involved in DNA binding and integration. a, Conservation of Cas1 and Cas2/3 residues involved in binding the foreign DNA, or catalyzing the strand transfer reaction, or catalyzing the degradation of nucleic acids. See Figure 7.2. b, Conservation of basic and polar Cas1 and Cas2/3 residues involved in accommodating the DNA duplexes bound by the Cas1-2/3 complex during integration. See Figure 7.3. Residues are numbered according to P. aeruginosa Cas1 and Cas2/3 proteins.



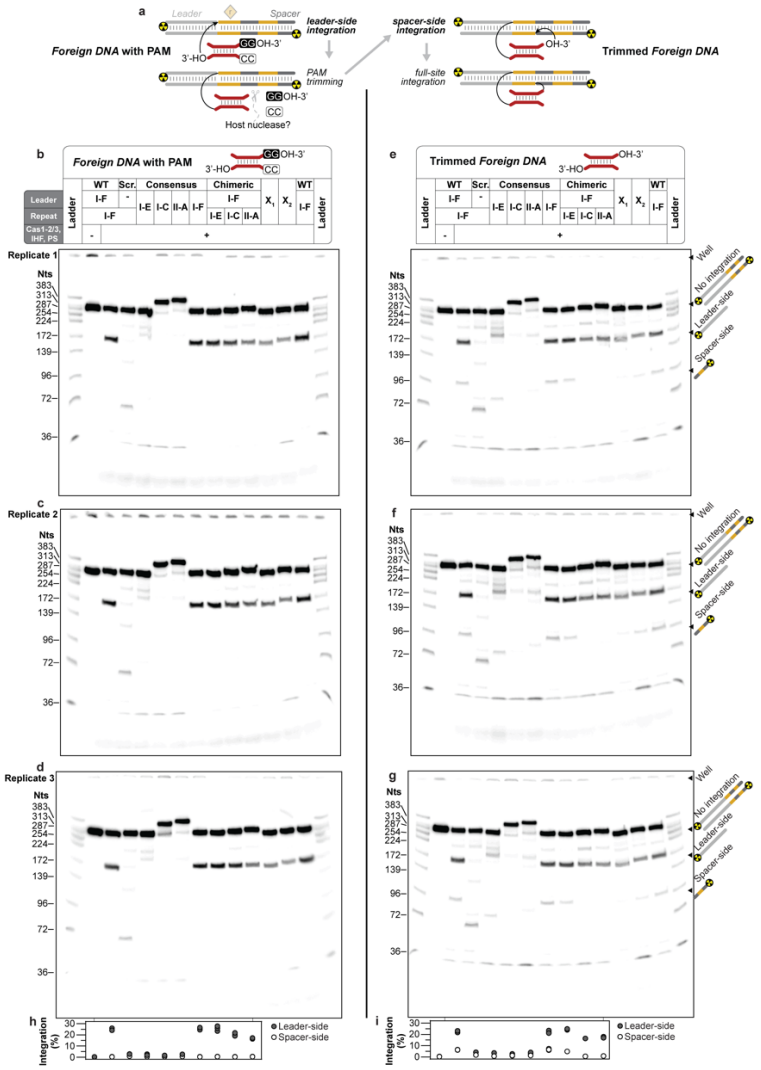
Extended Data Figure 7.4. Cas1-2/3 predominantly recognizes IR motifs, CRISPR repeat and foreign DNA through non-sequence specific interactions. a, Splayed 3′ ends of the foreign DNA are directed into the Cas1 transesterification active site. The product of the first strand-transfer reaction is shown in the Cas1a* active site (top), and the 3′ OH of the other end of the foreign DNA is positioned in the Cas1b* active site (bottom). The cryo-EM map is shown in transparent grey. b, Zoom-in on the Cas1-2/3 contacts to the CRISPR repeat (ChimeraX contacts command with default parameters). Most protein contacts occur to the DNA backbone and minor groove. Cas1 residue E184 appears to probe nucleotide G1 of the repeat. c, Zoom-in on the Cas1-2/3 contacts to the IR leader motifs. Most protein contacts occur to the DNA backbone and minor groove. d, DNAproDB analysis of Cas1-2/3 interactions with the 3′ ends of foreign DNA and the CRISPR leader-repeat junction. For clarity, only protein interactions to the nucleobases are shown[93]. e, DNAproDB analysis of Cas1-2/3 interactions with the IR leader motifs[93]. For clarity, only protein interactions to the nucleobases are shown. f, Zoom-in on the atomic fit of the base-pairs around the leader-repeat junction (-3 and +3 bps, coordinated by Cas1a*), to the cryo-EM map. Tension in the DNA loop at the leader-repeat junction has been released in the post-integration structure by a physical separation of base-pairs, as measured by an increase in base step rise. This tension may further pull the leaving 3′OH out of the Cas1 transesterification active site, to inhibit disintegration of the foreign DNA from the repeat. The approximate local base step rise was calculated using the http://web.x3dna.org/ webserver. g, A bioinformatic analysis of the first repeat from 24,940 CRISPR loci reveals that a 5′ GT dinucleotide is strongly conserved across most CRISPR subtypes. Similarly, a 5′ GT is present at the spacer-end of the repeat (seen as AC-3′ on the sense strand) within certain CRISPR subtypes (I-D, II-C, III-C, III-D, V-B, V-E, V-K, VI-A), but it is not broadly conserved.



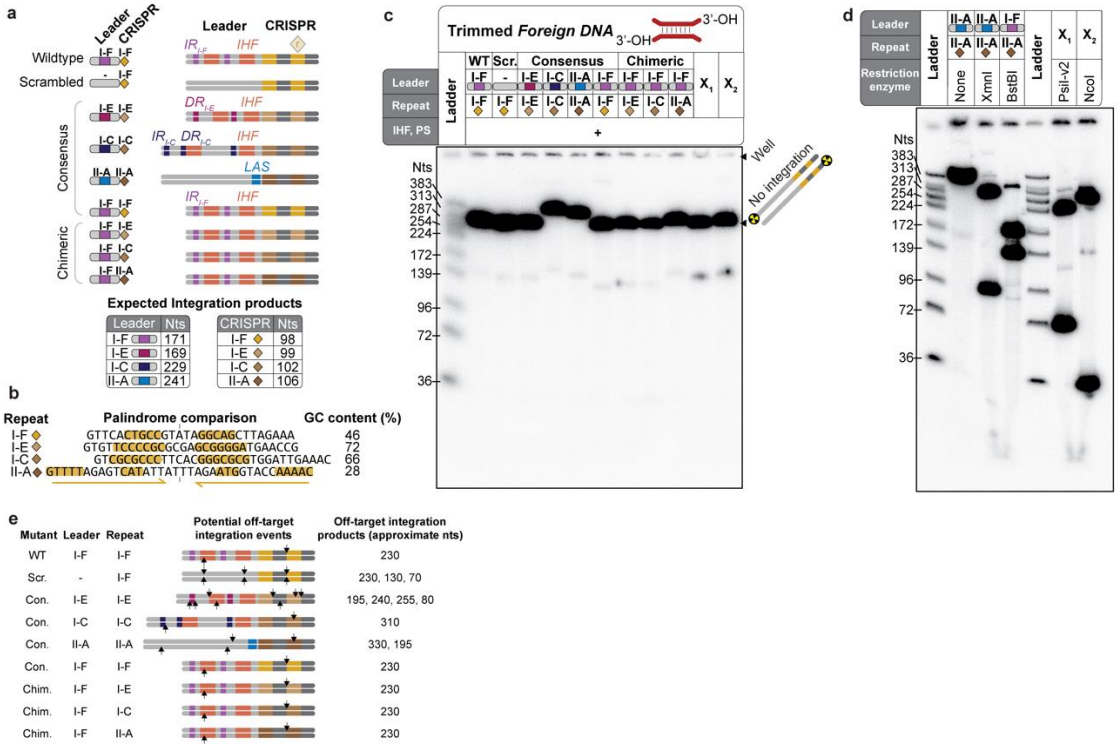

Extended Data Figure 7.5. Purification of structure-guided mutants of Cas1-2/3. a, SEC profile of a new preparation of wildtype Cas1-2/3 and all variants purified in the same manner on a Superdex 200 16/600 (Cytiva). An excess of free Strep-tagged Cas1 elutes at approximately 82 mL. b, SDS-PAGE gel of the Cas1-2/3 hetero-hexamer peak for all purified Cas1-2/3 variants. The SDS-PAGE gel of all Cas1-2/3 samples was run twice with similar results.



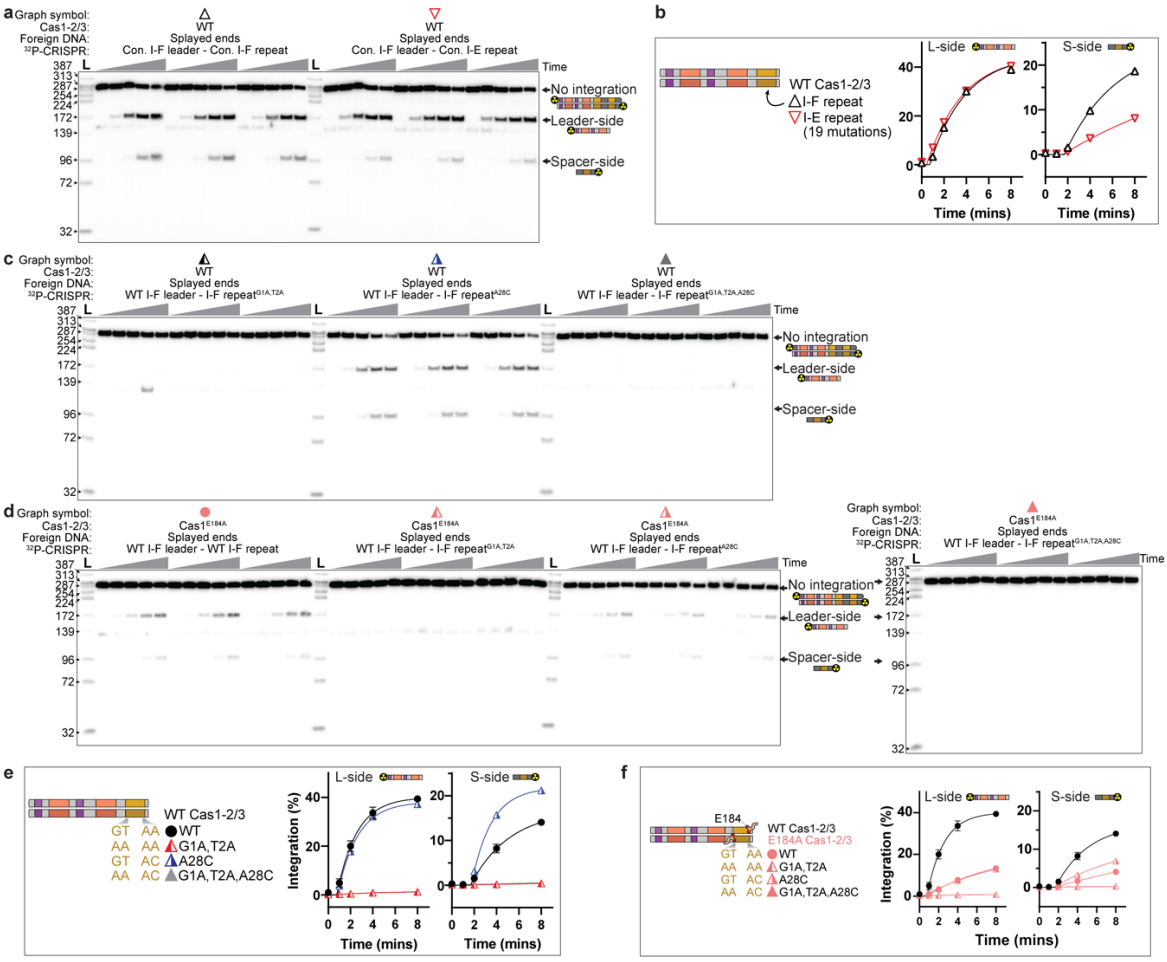

Extended Data Figure 7.6. Validation of Cas1-2/3 interactions with the foreign DNA and IR motifs. a, Time-course integration reactions to test the role of Cas1H25 in splaying the foreign DNA ends. Integration reactions were performed with trimmed foreign DNA (lacking a PAM) in triplicate, resolved on denaturing polyacrylamide gels. Timepoints were taken at 0, 1, 2, 4 and 8 minutes. Reactions were stopped by the addition of phenol. A 32P-labelled DNA that is shorter (140-160 bp) than the full length CRISPR is present in some DNA preparations (also see Extended Data Fig. 6c). Full-length CRISPR DNA, leader- and spacer-side integration products, do not overlap with this band. Further, Cas1-2/3, foreign DNA and IHF are in excess over the 32P-labelled DNA. The 140-160 bp band does not interfere with the quantification or generation of integration products. b, Quantification of time-course experiments to determine the role of Cas1 residue H25 in integration. The mean and standard deviation of three replicate experiments are shown. The Cas1H25A mutant integrates splayed and fully complementary foreign DNA fragments less efficiently that WT Cas1-2/3, suggesting that H25 steers the non-nucleophilic DNA strand away from the Cas1 active site. These results mirror the previously published effect of type I-E Cas1-2 tyrosine wedge mutation10. c, Time-course integration reactions to test the role of Cas2 residues in recognition of the IR motifs in the leader, performed as in panel a. d, Quantification of time-course experiments to determine the role of Cas2 residues K11, R12, R55 and N56 in integration. The mean and standard deviation of three replicate experiments are shown. The Cas2R55E,N56D/3 mutant retains a small amount of integration activity. The Cas2K11D,R12E/3 and Cas2K11D,R12E,R55E,N56D/3 mutants do not integrate DNA into the I-F CRISPR. Quantification of leader- (grey circles) or spacer-side (white circles) integration events from all three replicate gels. Individual dots for each triplicate reaction are shown, and some dots overlap.




Extended Data Figure 7.7. PAM blocks Cas-mediated integration of foreign DNA into CRISPR repeat. a, Schematic summarizing the step-wise strand-transfer reactions catalyzed by a Cas integrase. In the absence of putative host nucleases that trim PAMs, the foreign DNA fragments that contain a PAM stall the reaction at leader-side integration. However, the foreign DNA fragments that have been trimmed (no PAM) proceed through leader- and spacer-side integration. b-d, Endpoint integration reactions performed with a PAM-containing foreign DNA in triplicate, resolved on denaturing polyacrylamide gels. The X1 and X2 lanes signify lanes that were not further analyzed for this manuscript. e-g, Endpoint integration reactions performed with a trimmed foreign DNA in triplicate, resolved on denaturing polyacrylamide gels. The X1 and X2 lanes signify integration substrates that were not analyzed for this manuscript. h, i, Quantification of leader- (grey circles) or spacer-side (white circles) integration events from all three replicate gels. Individual dots for each triplicate reaction are shown, and some dots overlap. Three independent gels were run for PAM-containing or trimmed Foreign DNA integration reactions with similar results.




Extended Data Figure 7.8. Control reactions for integration assay and generation of 32P-labelled ladder. a, Scheme of nine CRISPR fragments used for in vitro integration assays. Each CRISPR locus contains two repeats and two spacers. Leader motifs are color-coded and annotated (IR, inverted repeat; DR, direct repeat; IHF, IHF binding site; LAS, Leader anchoring site). To simplify the pictograms shown here and in subsequent panels, a single-colored rectangle was used to represent a given collection of leader motifs, and a single diamond was used to represent a CRISPR locus composed of two repeats and two spacers. b, The four CRISPR repeats tested in the integration assays have diverse palindromes (yellow), and a wide range of GC-content. c, Control reactions in which all components necessary for integration, except Cas1-2/3, were incubated. The overexposed gels show that the majority of the 32P signal for a given integration substrate DNA corresponds to the full-length strands. d, A custom 32P-labelled DNA ladder was made by mixing the degradation products generated by individual restriction enzyme digests of different 32P-labelled integration substrate DNAs. X1 and X2 signify integration substrates that were not analyzed for this manuscript. e, Schematics of the nine CRISPRs tested in Extended Data Figure 7.5. The arrows identify locations of off-target integration reactions. Most off-target integration reactions occur by spurious integration at DNA motifs (for example, second CRISPR repeat, IHFdistal site, or upstream motifs) found near the ends of the CRISPR DNA target. Previous deep-sequencing of similar integration reactions has shown that the second repeat is a common off-target integration site, and that IHF blocks integration at the IHF binding site[22]. Urea-PAGE gels were run once.




Extended Data Figure 7.9. Validation of Cas1-2/3 interactions with the repeat. a, Time-course integration reactions to compare rate of integration into I-E and I-F repeats downstream of a I-F leader. b, Quantification of time-course experiments to determine the impact of 19 mutations associated with swapping the I-F repeat for the I-E repeat, on integration. Leader-side integration is indistinguishable. But spacer-side integration is slower into the I-E repeat. The mean and standard deviation of three replicate experiments are shown. c, Time-course integration reactions to measure the impact of I-F repeat mutations on integration rate. d, Time-course integration reactions to measure the impact of I-F repeat mutations on integration rate, in the context of a Cas1E184A mutation. The Cas1E184A mutation is expected to disrupt 5′ G recognition, but also impacts stability of the Cas1-2/3 complex (Extended Data Fig. 5). e, Quantification of time-course experiments to determine the impact of I-F repeat mutations on integration rate, in the context of WT Cas1-2/3. The mean and standard deviation of three replicate experiments are shown. f, Quantification of time-course experiments to determine the impact of I-F repeat mutations on integration rate, in the context of Cas1E184A-2/3. In panels e and f, no integration occurs into either the ‘G1A,T2A,A28C’ or ‘G1A,T2A’ repeat mutations, and the datapoints for these plots overlaps at roughly Y = 0 over the time course. The mean and standard deviation of three replicate experiments are shown. Each Urea-PAGE gel was run once.

References

1. Koonin, E. V. & Krupovic, M. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat. Rev. Genet. 16, 184–192 (2015).

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

3. Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell 62, 824–833 (2016).

4. Fagerlund, R. D. et al. Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proc. Natl. Acad. Sci. 114, 201618421 (2017).

5. Wright, A. V. et al. Structures of the CRISPR genome integration complex. Science 357, 1113–1118 (2017).

6. Hickman, A. B. & Dyda, F. Mechanisms of DNA transposition. Mob. DNA III 529–553 (2015) doi:10.1128/9781555819217.ch25.

7. Laxmikanthan, G. et al. Structure of a holliday junction complex reveals mechanisms governing a highly regulated DNA transaction. eLife 5, 1–23 (2016).

8. Lee, H. & Sashital, D. G. Creating memories: molecular mechanisms of CRISPR adaptation. Trends Biochem. Sci. 1–13 (2022) doi:10.1016/j.tibs.2022.02.004.

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

10. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature 527, 535–538 (2015).

11. Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature 550, 137–141 (2017).

12. Jackson, S. A. et al. CRISPR-Cas: Adapting to change. Science 356, eaal5056 (2017).

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

14. Kim, S. et al. Selective loading and processing of prespacers for precise CRISPR adaptation. Nature 579, 141–145 (2020).

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

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

17. Liao, C. et al. Spacer prioritization in CRISPR–Cas9 immunity is enabled by the leader RNA. Nat. Microbiol. (2022) doi:10.1038/s41564-022-01074-3.

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

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

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

21. Linheiro, R. S. & Bergman, C. M. Testing the palindromic target site model for DNA transposon insertion using the Drosophila melanogaster P-element. Nucleic Acids Res. 36, 6199–6208 (2008).

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

23. Kieper, S. N., Almendros, C. & Brouns, S. J. J. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. FEMS Microbiol. Lett. 366, 2016–2020 (2019).

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

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

26. Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Res. 43, 1749–1758 (2015).

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

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

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

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

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

32. Xiao, Y. et al. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. 0839, 1–12 (2018).

33. Nicholson, T. J. et al. Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol. 16, 566–576 (2019).

34. Brown, M. W. et al. Assembly and translocation of a CRISPR-Cas primed acquisition complex. bioRxiv 41, 1–11 (2017).

35. Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. 42, 2483–2492 (2014).

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

37. Fineran, P. C. et al. Degenerate target sites mediate rapid primed CRISPR adaptation. Proc. Natl. Acad. Sci. U. S. A. 111, (2014).

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

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

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

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

42. Nash, H. A. & Robertson, C. A. Purification and properties of the Escherichia coli protein factor required for lambda integrative recombination. J. Biol. Chem. 256, 9246–9253 (1981).

43. Lavoie, B. D. & Chaconas, G. Site-specific HU binding in the Mu transpososome: conversion of a sequence-independent DNA-binding protein into a chemical nuclease. Genes Dev. 7, 2510–2519 (1993).

44. Chalmers, R., Guhathakurta, A., Benjamin, H. & Kleckner, N. IHF Modulation of Tn10 Transposition: Sensory Transduction of Supercoiling Status via a Proposed Protein/DNA Molecular Spring. Cell 93, 897–908 (1998).

45. Haniford, D. B. Transpososome Dynamics and Regulation in Tn10 Transposition. Crit. Rev. Biochem. Mol. Biol. 41, 407–424 (2006).

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

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

48. van Gent, D. C., Hiom, K., Paull, T. T. & Gellert, M. Stimulation of V(D)J cleavage by high mobility group proteins. EMBO J. 16, 2665–2670 (1997).

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

50. Alonso, J. C., Weise, F. & Rojo, F. The Bacillus subtilis Histone-like Protein Hbsu Is Required for DNA Resolution and DNA Inversion Mediated by the β Recombinase of Plasmid pSM19035. J. Biol. Chem. 270, 2938–2945 (1995).

51. Petit, M.-A., Ehrlich, D. & Jannière, L. pAMβ1 resolvase has an atypical recombination site and requires a histone-like protein HU. Mol. Microbiol. 18, 271–282 (1995).

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

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

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

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

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

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

58. Nethery, M. A. et al. CRISPRclassify: Repeat-Based Classification of CRISPR Loci. CRISPR J. 4, 558–574 (2021).

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

60. Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth Phase-Dependent Variation in Protein Composition of the Escherichia coli Nucleoid. J. Bacteriol. 181, 6361–6370 (1999).

61. Montaño, S. P., Pigli, Y. Z. & Rice, P. A. The Mu transpososome structure sheds light on DDE recombinase evolution. Nature 491, 413–417 (2012).

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

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

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

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

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

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

68. Musharova, O. et al. Prespacers formed during primed adaptation associate with the Cas1–Cas2 adaptation complex and the Cas3 interference nuclease–helicase. Proc. Natl. Acad. Sci. 118, e2021291118 (2021).

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

70. Vorontsova, D. et al. Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. Nucleic Acids Res. 43, 10848–10860 (2015).

71. Cavazzana-Calvo, M. et al. Gene Therapy of Human Severe Combined Immunodeficiency (SCID)-X1 Disease. Science 288, 669–672 (2000).

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

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

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

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

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

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

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

79. Herzik, M. A., Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. Nat. Commun. 10, 1–9 (2019).

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

81. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods 14, 290–296 (2017).

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

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

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

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

86. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. Sect. Struct. Biol. 75, 861–877 (2019).

87. Mirdita, M. et al. ColabFold: making protein folding accessible to all. Nat. Methods 19, 679–682 (2022).

88. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. Sect. D 66, 486–501 (2010).

89. Nicholls, R. A. Conformation-independent comparison of protein structures. (2011).

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

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

92. Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 30, 70–82 (2021).

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

94. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics 17, 356 (2016).

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

96. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS ONE (2014) doi:10.1371/journal.pone.0110726.

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

98. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659 (2006).

99. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011).

100. Gouveia-Oliveira, R., Sackett, P. W. & Pedersen, A. G. MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics 8, 312 (2007).

101. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37 (2011).

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

103. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780 (2013).

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

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

Supplementary Tables

| Name | Description | Sequence (5' to 3') |
|---|---|---|
| Strand_1 | I-F leader sense strand, for I-F complex assembly - ssDNA | AAGCTTCCGACCCTTTTTTCGGACGATTTAAAAAA CCCTTATAAATCAGCAAGTTACGAGACCTCGAAAA AAGAGGGTTTCTGGCGGGAAAAACTCAAAAAACG CTTCGAAATCAACCGGTTATAGGTTTTCGGAGCTA |
| Strand_2 | I-F leader and repeat anti-sense strand, for I-F complex assembly - ssDNA | TGATTTTCTTAGCTGCCTACACGGCAGTGAACTAG CTCCGAAAACCTATAACCGGTTGATTTCGAAGCGT TTTTTGAGTTTTTCCCGCCAGAAACCCTCTTTTTTC GAGGTCTCGTAACTTGCTGATTTATAAGGGTTTTTT AAATCGTCCGAAAAAAGGGTCGGAAGCTT |
| Strand_3 | I-F repeat and prespacer sense strand, for I-F complex assembly - ssDNA | GTGCAAGTCGTTTTGCTAGAGCTACATGTTCACTG CCGTGTAGGCAGCTAAGAAAATCA |
| Strand_4 | Foreign DNA anti-sense strand, for I-F complex assembly - ssDNA | GCTCTAGCAAAACGACTTGCACAACGA |
| Splayed, Trimmed foreign DNA | DNA with 22 bp core and 5 nt overhangs on all 5' and 3' ends -dsDNA, for integration assays | TACATGCTCTAGCAAAACGACTTGCACAACGA *and* ATTAAGTGCAAGTCGTTTTGCTAGAGCTACAT |
| Fully complementary, Trimmed foreign DNA | 32 bp DNA, for integration assays | TTAAT GCTCTAGCAAAACGACTTGCAC TTAAT *and* ATTAA GTGCAAGTCGTTTTGCTAGAGC ATTAA |

| | | |
|---|---|---|
| Splayed, PAM-containing foreign DNA | DNA with 22 bp core, 5 nt overhangs on one set of ends, and 7 nts overhangs on the other end (which contains the PAM) – dsDNA, for integration assays | TACATGCTCTAGCAAAACGACTTGCACAACGAGG<br>*and*<br>CCATTAAGTGCAAGTCGTTTTGCTAGAGCTACAT |
| Wildtype I-F Leader and wildtype I-F Locus DNA | EcoRI and BamHI digest of Addgene plasmid 192207, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT<br>CCATGGACCCTTTTTTCGGACGATTTCTTACGCCCT<br>TATAAATCAGCAAGTTACGAGACCTCGAAAAAAG<br>AGGGTTTCTGGCGGGAAAAACTCGGTATTTCTTTT<br>TCCTTCAAATGGTTATAGGTTTTCGGAGCTAGTTCA<br>CTGCCGTGTAGGCAGCTAAGAAAATCAGCCGGAC<br>GTTGTAGTAGTCGAGCGCGGTGTTCACTGCCGTGT<br>AGGCAGCTAAGAAAGCCGGGTGGTGACTTCTGGA<br>CGGCGTTCGTCAAGGATC |
| Scrambled I-F Leader and wildtype I-F Locus DNA | EcoRI and BamHI digest of Addgene plasmid 192208, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT<br>CCATGGTCTTATCTCTCGGACGATTTTATTCTTGTA<br>CCACGTCACACGGTCAAAAAGACCTCGAAGAGAG<br>TGAAATTCTGGCGGGAAAAACTCTGTATTAACTCC<br>GTTATGTATTTTTGACTGGTTTTCGGAGCTAGTTCA<br>CTGCCGTGTAGGCAGCTAAGAAAATCAGCCGGAC<br>GTTGTAGTAGTCGAGCGCGGTGTTCACTGCCGTGT<br>AGGCAGCTAAGAAAGCCGGGTGGTGACTTCTGGA<br>CGGCGTTCGTCAAGGATC |

| | | |
|---|---|---|
| Consensus I-F Leader and consensus I-F Locus DNA | EcoRI and BamHI digest of Addgene plasmid 192209, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT CCATGGACCCTTTTTTCGGACGATTTAAAAAACCC TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA AAGGGTTTCTGGCGGGAAAAACTCAAAAAACGCT TCGAAATCAACCGGTTATAGGTTTTCGGAGCTAGT TCACTGCCGTATAGGCAGCTTAGAAAATCAGCCGG ACGTTGTAGTAGTCGAGCGCGGTGTTCACTGCCGT ATAGGCAGCTTAGAAAGCCGGGTGGTGACTTCTGG ACGGCGTTCGTCAAGGATC |
| Consensus I-F leader and consensus I-E locus DNA | EcoRI and BamHI digest of Addgene plasmid 192210, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT CCATGGACCCTTTTTTCGGACGATTTAAAAAACCC TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA AAGGGTTTCTGGCGGGAAAAACTCAAAAAACGCT TCGAAATCAACCGGTTATAGGTTTTCGGAGCTAGT GTTCCCCGCGCGAGCGGGGATGAACCGATCAGCC GGACGTTGTAGTAGTCGAGCGCGGTGTGTTCCCCG CGCGAGCGGGGATGAACCGGCCGGGTGGTGACTT CTGGACGGCGTTCGTCAAGGATC |
| Consensus I-F leader and consensus I-C locus DNA | EcoRI and BamHI digest of Addgene plasmid 192211, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT CCATGGACCCTTTTTTCGGACGATTTAAAAAACCC TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA AAGGGTTTCTGGCGGGAAAAACTCAAAAAACGCT TCGAAATCAACCGGTTATAGGTTTTCGGAGCTAGT CGCGCCCTTCACGGGCGCGTGGATTGAAACATCAG CCGGACGTTGTAGTAGTCGAGCGCGGTGTCGCGCC |

| | | |
|---|---|---|
| | | CTTCACGGGCGCGTGGATTGAAACGCCGGGTGGTG<br>ACTTCTGGACGGCGTTCGTCAAGGATC |
| Consensus I-F leader and consensus II-A locus DNA | EcoRI and BamHI digest of Addgene plasmid 192212, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT<br>CCATGGACCCTTTTTTCGGACGATTTAAAAAACCC<br>TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA<br>AAGGGTTTCTGGCGGGAAAAACTCAAAAAACGCT<br>TCGAAATCAACCGGTTATAGGTTTTCGGAGCTAGT<br>TTTAGAGTCATATTATTTAGAATGGTACCAAAACA<br>TCAGCCGGACGTTGTAGTAGTCGAGCGCGGTGTTT<br>TAGAGTCATATTATTTAGAATGGTACCAAAACGCC<br>GGGTGGTGACTTCTGGACGGCGTTCGTCAAGGATC |
| Consensus I-E leader and consensus I-E locus DNA | EcoRI and BamHI digest of Addgene plasmid 192213, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT<br>CCATGGCGGTGGACTTTGATGGCTTGAATCTGGTC<br>GCCTTTCAGCCCTTGGAAAACCCTGATCTTTAACA<br>AGGAAAAGTTGGTAGATTTTTAGCGGCTAATTTTC<br>CCCTTTGGGGACAATTGGTTACGCTAAGAGTGTTC<br>CCCGCGCGAGCGGGGATGAACCGATCAGCCGGAC<br>GTTGTAGTAGTCGAGCGCGGTGTGTTCCCCGCGCG<br>AGCGGGGATGAACCGGCCGGGTGGTGACTTCTGG<br>ACGGCGTTCGTCAAGGATC |
| Consensus I-C leader and consensus I-C locus DNA | EcoRI and BamHI digest of Addgene plasmid 192214, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT<br>CCATGGGAGTGCGAACGTGTAGCGACCGGGATTTT<br>ACGGGGAGGTTCGCGGATTTGCACGGCGTTGAATT<br>TGTTAGGGATTTTTTTAAGCACGGTGGGAATTTGA<br>GGACGGAGGCGTGTTTGGGCCACAGAAGAGCAGA<br>GGTTCGCGGAATTGGGCGGTTTTTGTCTGGCAGTA |

| | | |
|---|---|---|
| | | CAGTGGGTTAGGTTAGGTGGGTCGCGCCCTTCACG GGCGCGTGGATTGAAACATCAGCCGGACGTTGTAG TAGTCGAGCGCGGTGTCGCGCCCTTCACGGGCGCG TGGATTGAAACGCCGGGTGGTGACTTCTGGACGGC GTTCGTCAAGGATC |
| Consensus II-A leader and consensus II-A locus DNA | EcoRI and BamHI digest of Addgene plasmid 192215, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT CCATGGATCTCTTTAAATAATGTGGATGTCCTGTTT TTAGAACAAAGGGTGGTCCAGAACAGATTCCAAT ATATTTTGGACGAAAACTTTTATTTGAGTTATGAA AAAGCTTAAATTGTTACTGATTAGTGGTTCATTCTA AACTGAAATCTAGCTATGGATAAGTGATGCGAGTA CGGAACTTTGGAAAAAAATAATTCTCCGAGGTTTT AGAGTCATATTATTTAGAATGGTACCAAAACATCA GCCGGACGTTGTAGTAGTCGAGCGCGGTGTTTTAG AGTCATATTATTTAGAATGGTACCAAAACGCCGGG TGGTGACTTCTGGACGGCGTTCGTCAAGGATC |
| I-F leader and I-F repeat construct not studied further in this paper (X1) | EcoRI and BamHI digest of Addgene plasmid 192216, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT CCATGGACCCTTTTTTCGGACTTAAAAAACCCTTAT AAATCAGCAAGTTACGAGACCTCGAAAAAAAAGG GTTTCTGGCGGGAAAAACTCAAAAAACGCTTCGAA ATCAACCGGTTATAGGTTTTCGGAGCTAGTTCACT GCCGTGTAGGCAGCTAAGAAAATCAGCCGGACGT TGTAGTAGTCGAGCGCGGTGTTCACTGCCGTGTAG GCAGCTAAGAAAGCCGGGTGGTGACTTCTGGACG GCGTTCGTCAAGGATC |

| | | |
|---|---|---|
| I-F leader and I-F repeat construct not studied further in this paper (X2) | EcoRI and BamHI digest of Addgene plasmid 192217, $^{32}$P-labelled for integration assays | AATTCGAGCTCGGTACCTCGCGAATGCATCTAGAT CCATGGACCCTTTTTTCGGACGATTTAAAAAACCC TTATAAATCAGCAAGTTACGAGACCTGATCGAAAA AAAAGGGTTTCTGGCGGGAAAAACTCAAAAAACG CTTCGAAATCAACCGGTTATAGGTTTTCGGAGCTA GTTCACTGCCGTGTAGGCAGCTAAGAAAATCAGCC GGACGTTGTAGTAGTCGAGCGCGGTGTTCACTGCC GTGTAGGCAGCTAAGAAAGCCGGGTGGTGACTTCT GGACGGCGTTCGTCAAGGATC |
| I-F leader and I-F repeat construct with G1A,T2A mutation in repeat | EcoRI and BamHI digest of Addgene plasmid 200183, $^{32}$P-labelled for integration assays | GAATTCGAGCTCGGTACCTCGCGAATGCATCTAGA TCCATGGACCCTTTTTTCGGACGATTTCTTACGCCC TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA GAGGGTTTCTGGCGGGAAAAACTCGGTATTTCTTT TTCCTTCAAATGGTTATAGGTTTTCGGAGCTAAATC ACTGCCGTGTAGGCAGCTAAGAAAATCAGCCGGA CGTTGTAGTAGTCGAGCGCGGTAATCACTGCCGTG TAGGCAGCTAAGAAAGCCGGGTGGTGACTTCTGG ACGGCGTTCGTCAAGGATCC |
| I-F leader and I-F repeat construct with A28C mutation in the repeat | EcoRI and BamHI digest of Addgene plasmid 200184, $^{32}$P-labelled for integration assays | GAATTCGAGCTCGGTACCTCGCGAATGCATCTAGA TCCATGGACCCTTTTTTCGGACGATTTCTTACGCCC TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA GAGGGTTTCTGGCGGGAAAAACTCGGTATTTCTTT TTCCTTCAAATGGTTATAGGTTTTCGGAGCTAGTTC ACTGCCGTGTAGGCAGCTAAGAACATCAGCCGGA CGTTGTAGTAGTCGAGCGCGGTGTTCACTGCCGTG |

| | | |
|---|---|---|
| | | TAGGCAGCTAAGAACGCCGGGTGGTGACTTCTGGA CGGCGTTCGTCAAGGATCC |
| I-F leader and I-F repeat construct with G1A,T2A, A28C mutation in repeat | EcoRI and BamHI digest of Addgene plasmid 200185, $^{32}$P-labelled for integration assays | GAATTCGAGCTCGGTACCTCGCGAATGCATCTAGA TCCATGGACCCTTTTTTCGGACGATTTCTTACGCCC TTATAAATCAGCAAGTTACGAGACCTCGAAAAAA GAGGGTTTCTGGCGGGAAAAACTCGGTATTTCTTT TTCCTTCAAATGGTTATAGGTTTTCGGAGCTAAATC ACTGCCGTGTAGGCAGCTAAGAACATCAGCCGGA CGTTGTAGTAGTCGAGCGCGGTAATCACTGCCGTG TAGGCAGCTAAGAACGCCGGGTGGTGACTTCTGGA CGGCGTTCGTCAAGGATCC |
| Cas1_H25A_mut_F | One of two primers pairs used to generate the Cas1$^{H25A}$ mutant | CTACCTGCAAgccTGCCGGGTACTG |
| Cas1_H25A_mut_R | One of two primers pairs used to generate the Cas1$^{H25A}$ mutant | TACAGGTTGGCACGCTTG |
| Cas1_E184A_mut_F | One of two primers pairs used to generate the Cas1$^{E184A}$ mutant | CCCAATCATGcGCATCTACTGACCG |
| Cas1_E184A_mut_R | One of two primers pairs used to generate the Cas1$^{E184A}$ mutant | AGCCTGCTCCAACGCTCG |
| Cas2/3_K11D,R12E_mut_F | One of two primers pairs used to generate the Cas2/3$^{K11D,R12E}$ mutant | GCAATGCGAAgatgaaGCCCTGAGCGAAAC |

| | | |
|---|---|---|
| Cas2/3_K11D,R12E _mut_R | One of two primers pairs used to generate the Cas2/3$^{K11D,R12E}$ mutant | GACACCAGCAGGATGTTC |
| Cas2/3_R55E,N56D _mut_F | One of two primers pairs used to generate the Cas2/3$^{R55E,N56D}$ mutant | AAGCGCACGGgaggacACCGCCGTAG |
| Cas2/3_R55E,N56D _mut_R | One of two primers pairs used to generate the Cas2/3$^{R55E,N56D}$ mutant | TTCTTCAGCAGGCGTCGC |

Supplementary Table 1. Oligonucleotides used in this study.

# CHAPTER EIGHT

# STRUCTURES REVEAL HOW A CRISPR INTEGRASE CAPTURES, DELIVERS, AND INTEGRATES FOREIGN DNA INTO THE CRISPR ARRAY

## Contribution of Authors and Co-Authors

Manuscript in Chapter 8

Author: William S. Henriques

Contributions: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, editing and review

Co-Author: Jarrett Bowman

Contributions: Formal Analysis, Investigation, Visualization, Writing – editing and review

Co-Author: Laina Hall

Contributions: Formal Analysis, Investigation, Visualization, Writing – editing and review

Co-Author: Colin C. Gauvin

Contributions: Software, Data acquisition, Writing – editing and review

Co-Author: Andrew Santiago-Frangos

Contributions: Conceptualization, Data Curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – editing and review

Co-Author: Blake Wiedenheft

Contributions: Conceptualization, Data Curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, editing and review

Manuscript Information

William S. Henriques, Jarrett Bowman, Laina Hall, Colin C. Gauvin, Andrew Santiago-Frangos, Blake Wiedenheft

Status of Manuscript:

☑ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

## Abstract

Cas1 and Cas2 are the hallmark proteins of prokaryotic adaptive immunity. Cas1 and Cas2 integrate foreign DNA sequences into the CRISPR array and are occasionally fused to other proteins critical for CRISPR adaptation. In type I CRISPR systems, Cas3 unwinds and degrades foreign DNA during interference, but the mechanistic link between interference and adaptation remains poorly understood. To explore this functional connection, we purify the type I-F Cas1-2/3 fusion from *Pseudomonas aeruginosa* and determine four cryogenic electron microscopy structures at distinct stages of CRISPR adaptation. Collectively, these structures reveal how the interference protein Cas3 guides the integration of foreign DNA. A prominent, conserved channel of the Cas1-2/3 complex captures foreign DNA. Foreign DNA binding triggers conformational changes that reposition the Cas3 domain to expose new conserved DNA binding surfaces that home the DNA-bound integrase to the CRISPR locus. Upon docking to the CRISPR, the integrase distorts the CRISPR repeat through a conserved channel to catalyzes two sequential transesterification reactions that covalently link a discrete fragment of foreign dsDNA to the first repeat of the CRISPR locus. Taken together, these structures explain how Cas3 rearrangement coordinate foreign DNA capture, site specific delivery, and integration of new DNA into the bacterial genome.


## Introduction

Adaptive immunity requires genomic recombination to generate molecular memories of pathogens. In both V(D)J recombination in eukaryotes and CRISPR-Cas adaptive immunity in prokaryotes, viral-derived transposases have been independently domesticated to fight foreign

invaders[1–6]. However, the mechanism of adaptive immunity differs significantly. Eukaryotic V(D)J recombination splices together genomic fragments, but these adaptive rearrangements are not heritable[1,7–9]. In contrast, CRISPR integrases capture and insert short fragments of foreign DNA into CRISPR loci, creating an updated and heritable record of circulating viral predators[10–13].

Cas1 is a striking example of a selfish transposase gene repurposed for host defense[3,3,14]. CRISPR-Cas systems, nearly all of which contain Cas1, are found in ~90% of archaea and 40% of bacteria[3,14–16]. Cas1 no longer mobilizes a selfish genetic element, but protects the cellular host by providing the CRISPR interference machinery with updated guide sequences derived from selfish genetic elements[17–21]. During the adaptation stage of CRISPR-Cas adaptive immunity, Cas1 assembles into a heterohexameric complex with Cas2 to capture and integrate fragments of foreign DNA generated by host helicases and nucleases into the CRISPR array via sequential transesterification[6,10,17,18,21–26]. These fragments often contain a 2-5 bp protospacer adjacent motif (PAM) that functions to distinguish self from non-self DNA and to ensure directional integration into the CRISPR array[27]. In some CRISPR systems, dedicated Cas proteins recognize and trim the PAM, while in other systems this function is performed by host proteins such as DnaQ exonucleases involved in DNA replications or repair[28–33]. However, the mechanism of PAM trimming remains unknown in many CRISPR systems that lack Cas4[22,32,33].

Cas3 is a superfamily 2 (SF2) helicase/nuclease found in all type I CRISPR systems that generates foreign DNA fragments by processively degrading foreign DNA targeted by the Cascade complex (i.e. Csy complex)[23,34–37] Cas3 functionally couples the interference and adaptation stages of CRISPR immunity by generating fragments enriched in PAMs[22,38–40].

Underscoring this functional link is the fusion of Cas2 to Cas3 in type I-F CRISPR systems[20,38,40]. Cas1 and the Cas2/3 fusion assemble into a symmetric propeller-shaped heterohexamer where Cas1 represses Cas2/3-driven DNA degradation[20,38]. However, the Cas1-2/3 complex exists in at least two distinct conformations: a Cas3-repressed state, and a foreign DNA integration state[20,38,41]. Due to a lack of high-resolution structures of Cas1-2/3, the structural basis for Cas3 repression and the role of Cas3 in PAM trimming and integration remain poorly understood.

Here we determine cryoEM structures of the type I-F Cas1-2/3 integration complex from *Pseudomonas aeruginosa* in multiple states of foreign DNA integration into the CRISPR array. The structures explain how Cas1-2/3 recruits short fragments of foreign DNA to a positively charged channel on one face of the complex. Foreign DNA binding to this face functions as an allosteric switch that triggers conformational rearrangements, exposing three DNA binding sites previously occluded by Cas1 and Cas3. The rearranged complex is asymmetrically stabilized. The presence of a PAM blocks extension of the CRISPR repeat through the repeat channel. In the absence of a PAM, the repeat is distorted as it extends across the repeat channel from the first transesterification site to the second. Taken together, these cryoEM structures illustrate how Cas1-2/3 functions to both repress a helicase-nuclease and orchestrate foreign DNA capture, site specific delivery, and integration of new DNA into the bacterial genome.

## Results

### Cas1-2/3 Forms a DNA Capture Complex

To determine how the Cas1-2/3 integrase captures fragments of foreign DNA, we set out to determine the first high-resolution structure of a Cas1-2/3 complex in the absence of nucleic

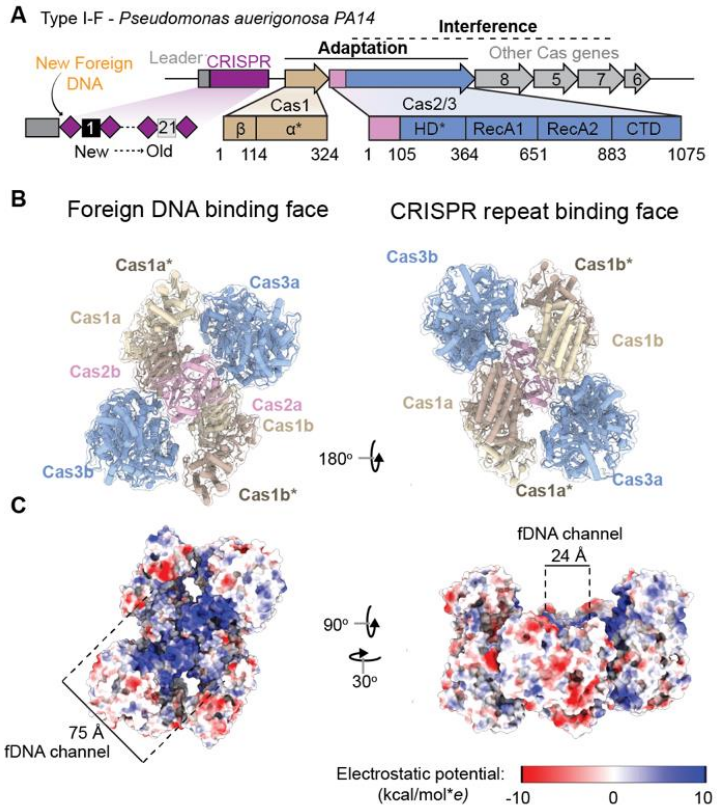



Figure 8.1. One side of the Cas1-2/3 integrase has a positively charged foreign DNA-binding channel. **(**A) Schematic of the CRISPR locus (purple) and *cas* genes in the type I-F from *Pseudomonas aeuroginosa.* The asterisks (*) indicates nuclease domains (B) A 3.3 Å resolution cryoEM structure of the Cas1-2/3 heterohexamer reveals a propeller shaped complex with C2 symmetry, and a repeat face of Cas2 face occluded by two Cas1 subunits. *indicates the Cas1 subunits where transesterification will occur. Density map displayed at 50% transparency, with the atomic model docked inside and colored by subunit. (C) A large basic patch of residues is present on one face of the complex. This patch sits in a basin created by the HD domains of Cas3 and forms a channel 24 Å wide and 75 Å long. Atomic model is colored by electrostatic potential (kcal/mol*$e$ at 298 K).

acid. Cas1 and Cas2/3 proteins from the *PA14* strain of *Pseudomonas aeruginosa* were co-expressed in *Escherichia coli* BL21 cells. An amino-terminal Strep tag on Cas1 pulls down untagged Cas2/3 in a stable complex that elutes from a size exclusion column as single peak with an estimated molecular weight of ~380 kDa (**Figure 8.1A, Supplementary Figure 8.1**)[20,38]. Purified Cas1-2/3 was concentrated to ~12 uM, spotted on cryoEM grids, blotted, and plunge

frozen in liquid ethane. Frozen grids were imaged using a 300 kEV cryo-electron microscope equipped with an energy filter. Grid squares with thin ice and a dense particle distribution were targeted to collect 3944 movies. After preliminary motion correction, CTF-correction, blob picking, and 2D classification, *ab inito* reconstruction and non-uniform refinement were used to reconstruct an initial volume that was used as a template for particle picking. 303,250 template-picked particles were extracted from 3293 cryoEM movies to determine a 3.3 Å resolution structure of the complex (**Figure 8.1B-C, Supplementary Figure 8.1)**. To build an atomic model of the complex, an Alphafold3 predicted structure of the complex was fit into the electron density map in ChimeraX, followed by iterative refinement in Isolde, Phenix, and Coot[42–44]. This approach resulted in an atomic model across 96% of Cas1-2/3 residues with 3138 unambiguously positioned sidechains (**Supplementary Table 8.1**).

The Cas1-2/3 structure reveals an assembly shaped like a four-bladed propeller. The Cas1 dimers and Cas3 subunits project away from a Cas2 homodimer at the center (**Figure 8.1B**). Cas2 is connected to Cas3 by a linker, unresolved for ten residues due to flexibility. Cas1 forms an asymmetric homodimer, highly similar to previously determined crystal structures (**Supplementary Figure 8.2**)[45,47]. To understand if the heterohexameric architecture is conserved across type I-F systems, the amino acid sequence of Cas2/3 was used to query the non-redundant database of protein sequences through iterative PSI-BLAST. The top 1000 hits were aligned using MAFFT, Cas2 and Cas3-only sequences were manually removed, and an unrooted phylogenetic tree of the remaining Cas2/3 fusions was reconstructed using Fasttree (see Materials & Methods, **Supplementary Figure 8.3**). Four representative sequences from across the tree were selected for AlaphaFold predictions. Each of these predicted structures aligned to

>1000 Cas2/3 residues (RMSDs < 10 Å) and >300 Cas1 residues (RMSDs < 12 Å) in the experimentally determined structure (**Supplementary Figure 8.3**). All predicted structures of the Cas1-2/3 complex preserved the four-bladed propeller architecture observed experimentally (**Supplementary Figure 8.3)**. These comparisons reveal broad conservation of the global Cas1-2/3 architecture across type I-F CRISPR systems.

To understand if the fusion of Cas3 to the N-terminus of Cas2 changes the assembly of the Cas1-2 core, we compared the Cas1-2/3 structure to previously determined Cas1-2 structures in the absence of DNA[26,46,48]. In all structures of Cas1-2, the anti-parallel beta sheet of the N-terminal ferredoxin fold of Cas2 forms the conserved dimer interface (**Supplementary Figure 8.4**, **Supplementary Table 8.3**). However, the variable C-terminus of Cas2 determines the extent of buried surface area at both the Cas2 homodimer interface and the Cas1-Cas2 interface in experimentally determined apo structures[18,26,48–53], presumably determining how foreign DNA binding influences conformational rearrangements. In the Cas1-2/3 structure, the buried surface area at the Cas2 dimer interface is reduced, while the surface area at the Cas1-Cas2 interfaces are increased relative to previously determined structures (**Supplementary Figure 8.4**, **Supplementary Table 8.3)**. DNA binding residues in both the invariant ferredoxin fold (K11/R18) and the variable Cas2 C-terminus (T93) form two parallel edges of a 25 Å wide positively charged channel (**Figure 8.1C, Supplementary Figure 8.4**). Ten amino acids linking Cas2 to Cas3 are unresolved from positions 95 to 105 in the structure. This positively charged channel, slightly wider than B-form DNA, is the only face of the Cas2 homodimer accessible to foreign DNA fragments.

To understand if this channel is a conserved feature of type I-F systems, we used CasFinder to identify unique occurrences of type I-F CRISPR systems in NCBI's database of complete bacterial and archaeal genomes. We aligned Cas1 (n=679) and Cas23 (n=745) sequences and used these alignments to calculate conservation scores, which were mapped onto the atomic model of the Cas1-2/3 complex (**Supplementary Figure 8.5)**. This analysis revealed that the DNA binding residues K11 and T93 are well conserved. Additional positively charged residues form an electrostatic funnel for ssDNA to thread from the foreign DNA binding face to the Cas1 active site. This analysis reveals that the foreign DNA binding channel is a conserved feature of the type I-F Cas1-2/3 complex.

### Structural Basis for Cas3 HD Nuclease and RecA Helicase Inhibition

In the Cas1-2/3 complex, all four domains of Cas3 are well resolved and provide explanations for previously described biochemical activities[38]. Cas3 is oriented perpendicular to the foreign DNA-binding surface of the complex. The HD domain stacks on top of the RecA1 and RecA2 domains, and the CTD wraps around the RecA1 and RecA2 domains like a belt (**Figure 8.2**). The CTD appears to be unique to Cas3 – queries with both Dali and Foldseek revealed no significant structural homologues outside of Cas3 (data not shown).

Previously published biochemical data describes inhibition of Cas3 HD nuclease activity in the presence of Cas1[38]. However, the Cas1-2/3 structure reveals that the HD domain is not in contact with Cas1, and the active site residues coordinate two metal cations (**Figure 8.2**)[34,54], so neither position nor absence of catalytic metals can explain this suppressed activity. A small domain consisting of two helices connected by a flexible loop in the RecA1 domain (residues 483-511) are positioned over the upper half of the ssDNA channel that runs from the dsDNA

splitting wedge to the HD active site, with the short loop inserted into the HD active site. This loop begins with the SF2 helicase DNA translocation motif 1b (**Supplementary Figure 8.6**). The conserved motif (GSES, residues 498-501) may inhibit the HD nuclease by physically obstructing access to the HD active site (**Supplementary Figure 8.6**). Alignment of Cas2/3 with structures of Cas3 in complex with ssDNA from both type I-E and I-F systems reveals that the equivalent residues are displaced by ssDNA (**Supplementary Figure 8.6**). Thus, in the DNA capture complex, the HD active site of Cas3 is physically obstructed by a loop from the RecA1 domain. This inhibition is relieved when the nuclease is recruited to the target-bound Cascade complex for targeted destruction of foreign DNA (**Supplementary Figure 8.6)**. In type I-C and I-A CRISPR systems, recruitment of Cascade and full R-loop formation displaces a similarly placed loop, suggesting that an inhibitory loop gating access to the HD active site is a conserved mechanism of Cas3 nuclease regulation across type I CRISPR systems[36,55,56].

ATP hydrolysis and DNase activity are coupled in Cas3[58]. The 3' ssDNase activity of the Cas3-HD domain stimulates ATP binding and hydrolysis at the interface between the RecA1 and RecA2 domains, which powers the ratcheting mechanism of the SF2 helicase domains along ssDNA[34,57,58]. To understand the conformational state of the helicase domains in the Cas1-2/3 complex, ATP binding sites across previously determined type I-E and type I-F structures were aligned and compared. The ATP binding site of AcrIF3-inhibited type I-F Cas3 (ADP bound, closed) and the ATP bound structures from the type I-E Cas3 of *T. fusca* capture distinct conformations of the ATP binding site. In the Cas1-2/3 complex, Cas3 is in the closed, ADP bound conformation (RMSD of 0.497 Å across 496 trimmed C*a* of the aligned RecA1 and RecA2 domains). No density is observed that corresponds with an ATP molecule in the ATP

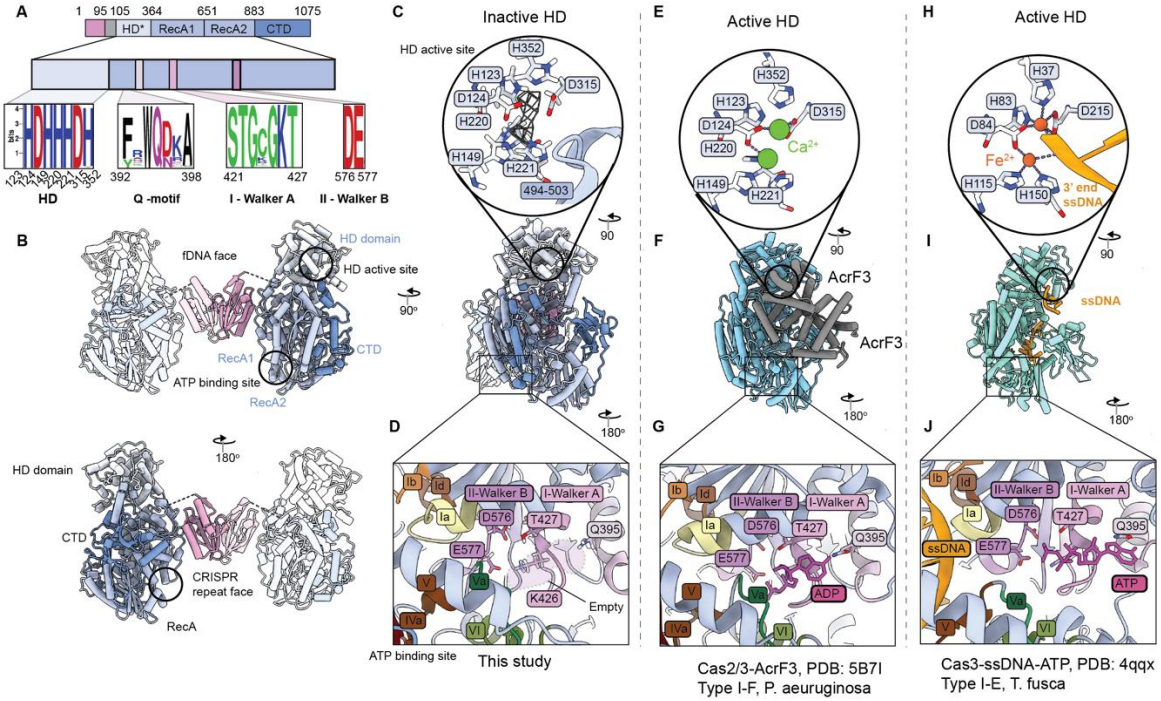



Figure 8.2. The Cas3 active sites are blocked in the DNA capture complex. A. Schematic of Cas2/3 with nuclease (HD) and ATP interaction motifs (Q, I, II) highlighted. WebLogos are built from the 685-sequence alignment from Figure S2 at each motif, and coordinates represent the Cas2/3 sequence modelled in this study. B. Atomic model of the Cas2/3 dimer with Cas1 dimers hidden for clarity and one subunit shown at 50% transparency. The Cas3 domains are colored as in panel A, with the HD active site, ATP binding site, and complex faces indicated C. The active site residues of the Cas3 HD domain coordinate unmodelled density. A loop (494-503) is inserted into the Cas3 active site at the top of two helices is inserted into the HD domain active site. D. The ATP binding site is empty. ATP binding motifs colored in various shades of pink. DNA interacting motifs (Ia, Ib, Id, V) are colored in warm colors (from light yellow to dark brown). Motifs in the RecA2 domain at the base of the ATP binding site (V, VI) are colored in shades of green E-G. Previously determined crystal structure of Cas2/3 from PA14 (PDB: 5B7I). AcrF3-bound Cas2/3 is capable of processive nucleic acid degradation (as demonstrated in Rollins et al. 2017) but not interaction with Csy. The structure reveals coordination of two Ca2+ ions in the HD active site and ADP bound in the ATP binding site. No loop obstructs the catalytic site, and residues 494-503 are unresolved in the crystal structure. H-J Previously determined crystal structure of type I-E Cas3 from T. fusca bound to ssDNA and ATP (PDB: 4QQX).

active site. These structure data provide a mechanistic explanation for previous observations that Cas1 inhibits the helicase activity of Cas3 (**Figure 8.2**)[38,54]. In the absence of DNA, Cas1 traps Cas3 in a closed conformation, preventing ATP turnover.

Two interactions on opposing sides of the RecA and CTD domains trap Cas3 in this inactive conformation (**Supplementary Figure 8.7**). The first interface is between two poorly resolved regions in the CTD of Cas3 and a flexible loop that caps the Cas1* active site. The CTD sits on top of the RecA2 domain, and is positioned to block movement of the RecA2 domain required for helicase ratcheting, thus locking the HD-blocking inhibitory domain in place. The second interface is between the RecA1 domain and the inactive Cas1 on the opposing side of Cas3. We speculate that this locks the RecA1 domain in place, providing a second block on the ratcheting mechanism of the helicase (**Figure S7**). Notably, the face of Cas3 that interacts with the Cascade complex is solvent exposed, providing a mechanism by which the I-F Cascade complex may recruit Cas3 and trigger the conformational changes required to activate the helicase and nuclease functions of the protein (**Supplementary Figure 8.6**)[34,35,54].

Cas3 Blocks Foreign DNA Integration

In this inactivated conformation, residues of Cas2 (R55 and K50) that bind to the upstream leader motifs, are sequestered by the RecA1 domain of Cas3 (**Supplementary Figure 8.7**). Residues R55 and K50 of Cas2 form a network of hydrogen bonds with residues in the RecA1 domain of Cas3: N373, D517, and D520 (R55) and D472 and D473 (K50). The surface of Cas2 involved in binding to the CRISPR repeat is obscured by two salt bridges. In addition, the RecA2 and CTD of Cas3 further block access to the repeat channel. While Cas2 has four possible faces that can bind DNA during integration, three of the four faces are blocked. These steric blocks ensure Cas1-2/3 functions as a DNA capture complex poised to bind fragments of foreign DNA prior to functioning as an integrase (**Supplementary Figure 8.7)**.

Foreign DNA Capture Exposes Genome Binding Sites

We hypothesized that capture of foreign DNA fragments would initiate conformational changes that license foreign DNA integration. To test this hypothesis, purified Cas1-2/3 complex was incubated with a 32 bp fragment of foreign DNA with 5 bp splayed ends, and a GG PAM sequence on one end (**Figure 8.3A**). The DNA fragment bound Cas1-2/3 co-elutes from a size exclusion column as a single peak with an estimated molecular weight of 415 kDa, corresponding to a 1:1 stoichiometry of Cas1-2/3 and fDNA (**Supplementary Figure 8.8)**. An aliquot of the peak fraction was spotted on cryoEM grids, blotted, and plunge frozen in liquid ethane. Grids were imaged using a 300 keV cryo-electron microscope equipped with an energy filter. Using multi-class ab initio reconstructions and heterogeneous refinements, we identified five distinct volumes, but subsequent processing only led to a high-resolution structure of one volume due to substantial conformational heterogeneity in the remaining four volumes. The final 3.31 Å resolution reconstruction includes 28,665 particles extracted from 5,403 movies (**Figure 8.3A, Supplementary Figure 8.8)**. Initially, an Alphafold3 prediction of the complex was docked into the density. While the Cas1-2 and dsDNA components of the prediction fit well into the density Cas3 in the Alphafold prediction was clearly in the wrong conformation. To build an atomic model of the complex, the Cas3 domain and the Cas1-2 hexamer bound to dsDNA were independently rigid-body fit into the density in ChimeraX followed by jiggle fitting in Coot. This chimeric model was used as a starting point for iterative refinement in Isolde, Phenix, and

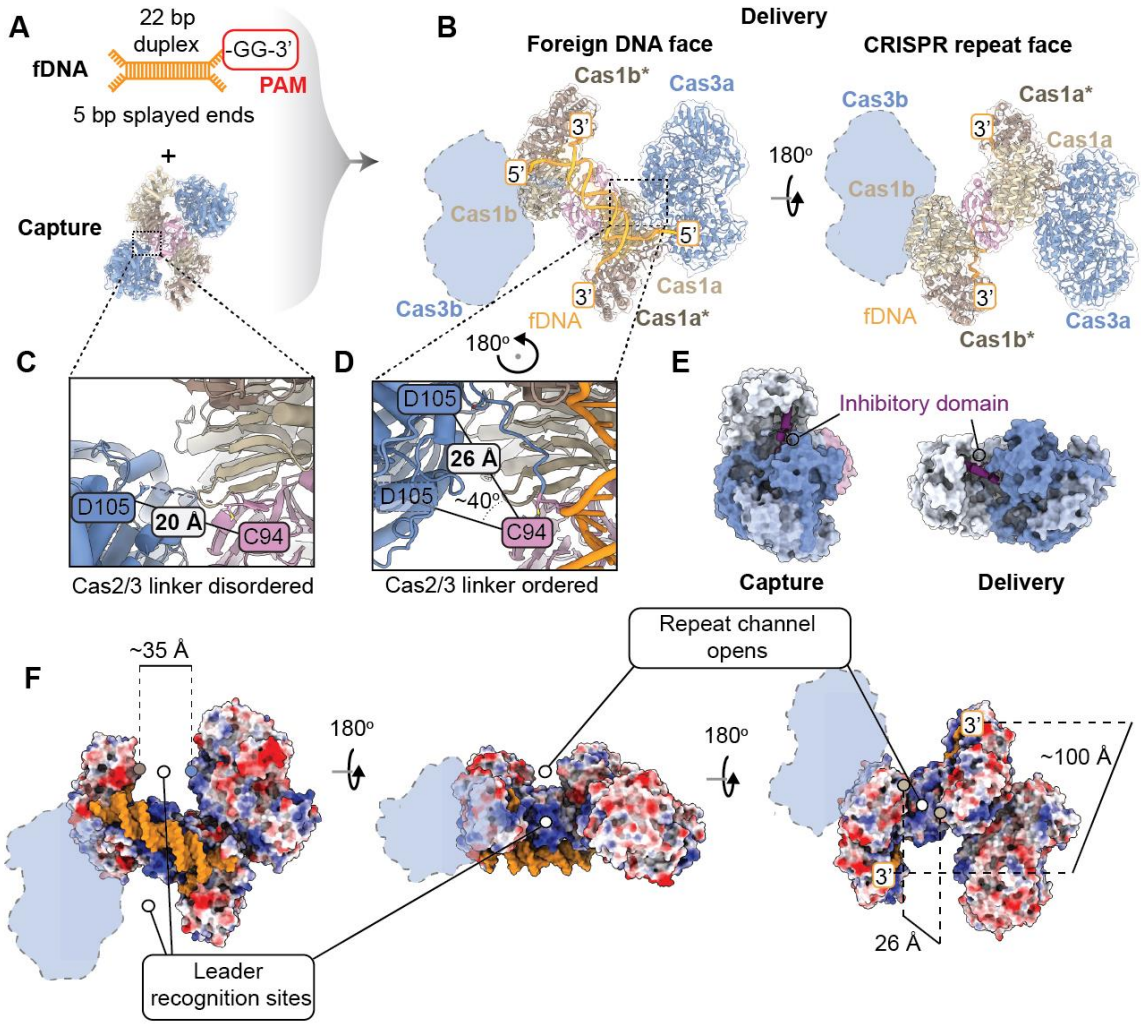



Figure 8.3. Foreign DNA triggers a conformational rearrangement that exposes leader recognition sites and preserves Cas3 inhibition. A. Purified Cas1-2/3 was incubated with a 22 bp duplex flanked on both sides by 5 nts of ssDNA substrate with a GG dinucleotide PAM on one 3' end. B. Atomic model colored by subunit docked into ~3.5 Å density map from a local refinement shown at 80% transparency. Cas3b, averaged out by conformational heterogeneity, outlined. C. Foreign DNA binding lengthens and orders the 10 amino acid linker between the Cas2 (last residue: C94) and Cas3 (first residue: D105), resulting in a 40º shift in orientation between C94 and D105. E. Despite the global repositioning of Cas3 relative to Cas1-2, the inhibitory domain remains in the HD active site in both capture and delivery conformations. F. Foreign DNA binding exposes two basic folded genome recognition and a basic channel for the CRISPR repeat between the 3' ends of the fDNA. Atomic model is colored by electrostatic potential (kcal/mol*e at 298 K). Distances measured in ChimeraX between Cα of H271 in Cas1a/b* (length of the repeat channel), R169 in Cas1a/b (distance across repeat channel), and D388 on Cas3a and G236 on Cas1b* (width of leader recognition site).

Coot[42–44]. This approach resulted in an atomic model with in which 69% of main chain residues could be confidently modeled. This atomic model captures both Cas1 dimers, the Cas2 dimer, foreign DNA bound to one face of the complex, and one of the two Cas3 lobes. There is no density for the second Cas3 lobe, presumably due to conformational heterogeneity. Twenty-two base pairs of dsDNA fit between the histidine wedge (Cas1 H25) that flank either end of the foreign DNA-binding channel. dsDNA in the channel is positioned by the DNA binding residues of the foreign DNA face. K11 in the ferredoxin fold of Cas2 contacts the phosphate backbone between nucleotide 16 and nucleotide 17, precisely bisecting the 32 bp foreign DNA fragment. T93 and C94 in the variable C-terminal domain of Cas2 coordinate the phosphate groups on either side of nucleotide 20. This series of interactions positions the minor groove against Cas2. One full helical turn in either direction positions the minor groove at the histidine wedge (H25), which splits the DNA duplex and diverts the 5 bp of the 3' end down the conserved electrostatic tunnel to the Cas1 active sites (**Supplementary Figure 8.5**).

Foreign DNA binding triggers a large rotation of Cas3, hinging at the linker between the Cas2 and Cas3 domain (residues 95-104). The individual Cas2 and Cas3 subunits hardly change in conformation (RMSD < 1 Å across 94 unpruned pairs for Cas2 domain, 972 unpruned pairs for Cas3 domain), but the global conformation changes dramatically. Interactions with fDNA and the N-terminus of Cas1 extend the linker by 6 Å and rotate the linker ~40º (**Figure 8.3**). In the DNA capture complex, Cas3 is oriented perpendicular to the fDNA binding face (**Figure 8.1-8.2 Supplementary Figure 8.4**). The extension, rotation, and rigidification of the linker breaks all of the Cas3 contacts from the capture complex, and repositions the Cas3 domain horizontally, flipped down to be in plane with the fDNA binding face. This rearrangement moves the HD

active site by ~67 Å and the RecA ATP-binding site by ~30 Å (**Figure 8.3A**, **Supplementary Figure 8.8**). This Cas3 rotation coincides with a rotation that breaks the salt bridge between Cas1 dimers and opens a channel ~26 Å wide on the CRISPR repeat binding face of Cas2. Notably, while we observed substantial conformational heterogeneity in the sample, we did not observe any classes corresponding to the capture complex, indicating that in the presence of foreign DNA fragments, the DNA-bound conformation is more likely than the unbound complex.

The observed conformational rearrangements expose three additional basic DNA binding faces on Cas2 (**Figure 8.3B,F, Supplementary Figure 8.8-8.9**). Two of these faces are diametrically opposed, and Cas3 rotation exposes R55 and K50 on $\alpha$2 of Cas2, both of which intercalate into the minor groove of the CRISPR leader[41]. Based on how these sites interact with the conserved DNA sequence motifs in the leader, we designated these two faces "leader binding sites".

Superposition of a single Cas2 subunit from the capture complex and the delivery complex reveals that foreign DNA binding reveals the overall conformation of the Cas2 subunit hardly changes upon foreign DNA binding (RMSD 0.730 A across 93 Ca). DNA binding also exposes a second positively charged channel on the back side of the Cas2 dimer. The Cas2 subunits form the floor this channel, while the sides are formed by the two Cas1 subunits that are not involved in transesterification. The enzymatically active Cas1 subunits are positioned at either end of this channel, while the inactive Cas1 subunits form the walls of the channel. Based on the position of CRISPR repeat in previously determined structures, we designated this channel the “repeat-binding channel” (**Figure 8.3F**).

Notably, the fragment of DNA only binds to the foreign DNA face of the complex with a 1:1 stoichiometry, despite exposing three additional DNA binding faces. Additionally, no density

is observed for nucleic acid in either the HD domain or helicase domain of Cas3. This result is consistent with more selective DNA-binding at the surfaces necessary for engaging with the repeat and the leader.

Structure Precludes a Role for Cas3 in PAM Trimming

The PAM is critical for directional insertion of foreign DNA into the CRISPR array. But to insert foreign DNA into the CRISPR array, the protospacer-adjacent motif must first be trimmed off[29,30,33,41]. However, the mechanism of PAM trimming is unknown in the type I-F CRISPR systems. Though Cas3 contains a nucleolytic HD domain and undergoes substantial rearrangement relative to Cas2 and Cas1 upon foreign DNA binding, the conformational changes within Cas2 and Cas3 are minimal. After foreign DNA binding, inhibitory loop (GSES, residues 498-501) is still positioned in the HD active site (**Figure 8.3E**). The 5' end of the foreign DNA interacts with the Cas1a, then becomes too flexible to resolve. While the back side of the HD domain is in contact with Cas1a, the active site is ~45 Å away from the 5' tail of the foreign DNA and faces away from this 5' tail. The 3' end of the same strand is over 100 Å away from the HD domain. Though Cas3 is a 3' exonuclease and endonuclease capable of nicking ssDNA, it's positioning in the Cas1-2/3 complex precludes a role for it in PAM trimming without additional conformational rearrangements. The structural rearrangements do expose the faces of Cas1 and Cas2 that have been shown to interact with both Cas4 and other PAM-trimming exonucleases, suggesting that an undetermined host nuclease is recruited to the DNA-delivery complex to trim the PAM prior to integration in the type I-F system[31,33], though the current study cannot rule out that the observed conformations are due to the split ends substrate used.

Foreign DNA Binding Exacerbates Asymmetric Instability

In the presence of foreign DNA, 90% of particles lack density for one of the two Cas3 lobes, one of the Cas1 homodimers, or both, predominantly on one side of the complex. This contrasts with Cas1-2/3 in the absence of DNA, in which ~30% of Cas1-2/3 particles were missing density for one Cas3 domain (**Supplementary Figure 8.1**). Masked 3D classification and refinement of the regions missing Cas3 and Cas1 on the unstable side of the complex identified many different low resolution conformational states for these missing subunits. Notably, the foreign DNA substrate contained asymmetry: one end contained an additional GG-3' dinucleotide, to simulate a PAM (**Figure 8.3A**). We hypothesized that the PAM contributes to this asymmetry. However, density for the PAM could not be identified in a global refinement of the complex fDNA-bound Cas1-2/3 to 3.3 Å, presumably due to local motion of the PAM nucleotides. To further investigate whether the PAM was responsible for asymmetrically stabilizing the complex, each Cas1 dimer was refined separately using two separate local masks, one for each Cas1 dimer. Surprisingly, the two Cas1 dimer volumes did not differ significantly. This result suggests that though the PAM may affect asymmetric complex stability, the substrate provided did not lock the PAM in a specific orientation for trimming.

The PAM Changes the Stability of the Integration Complex

Previous biochemical evidence suggests that the presence of a PAM prevents full integration of the foreign DNA by blocking the second transesterification reaction[41]. To determine the structural basis for this inhibition, the structures of Cas1-2/3 integrating fDNA sequences with and without a PAM were determined. Purified Cas1-2/3 complex was incubated with a 34 bp DNA fragment containing 22 bp of dsDNA and 5 bp of splayed ends on either end.

To simulate a PAM, a 5' 'CC' PAM and the corresponding 3' 'GG' dinucleotide on one splayed end (PAM) was added to one end of the fragment. fDNA-bound Cas1-2/3 was subsequently added to an IHF-bound DNA fragment containing the type I-F CRISPR leader and the first

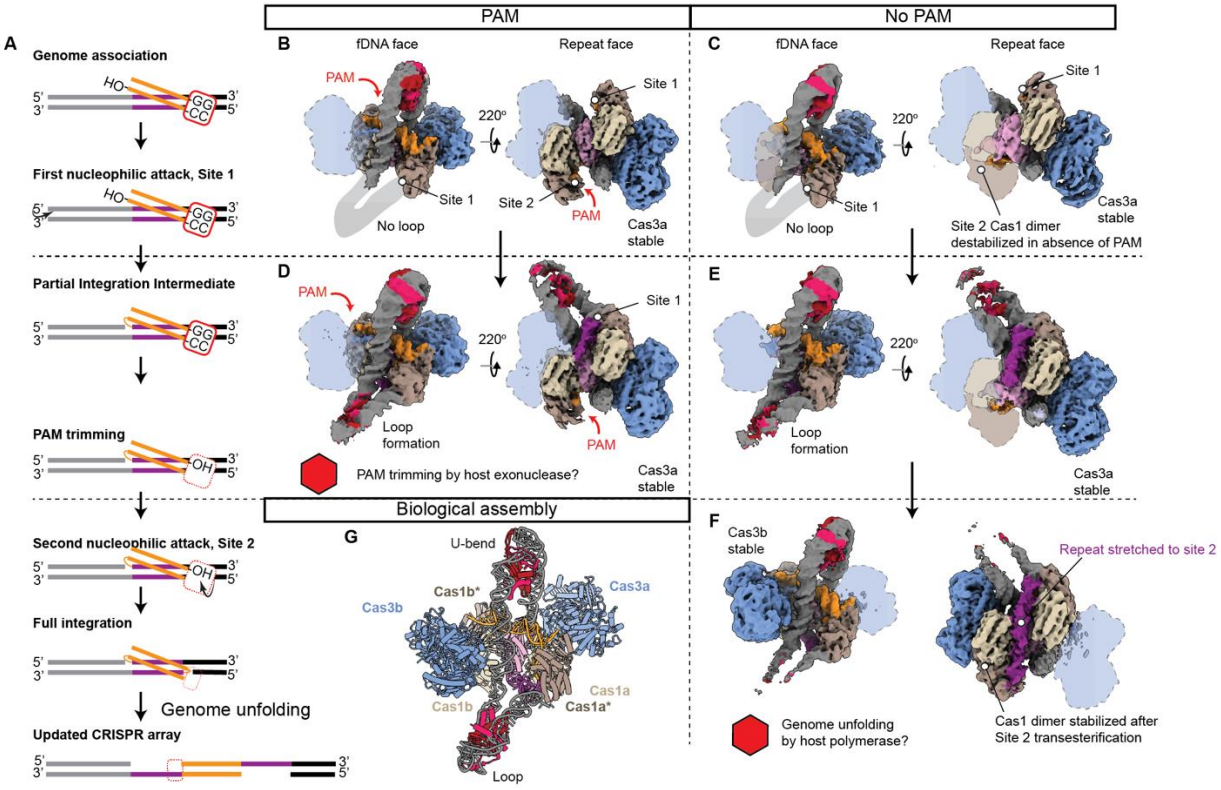



Figure 8.4. The PAM blocks the second transesterification by altering complex stability. A. Schematic overview of integration by two-step transesterification. B-F. cryoEM density maps of Cas1-2/3 mixed with either 34 bp fDNA with a GG dinucleotide PAM or 32 bp of fDNA without a PAM, a 200 bp fragment containing the CRISPR leader and locus, and IHF, and purified by size exclusion chromatography. B-C. Density maps from non-uniform refinement of 82,751 particles (PAM) and 121,485 particles (trimmed) captured in a genome-associated, pre-transesterification state in the presence/absence of a PAM. The loop has not formed and the repeat channel is empty. D-E Non-Uniform refinement of 94,2045 particles (PAM) and 97,437 particles (trimmed) captured during the first transesterification. The loop has formed to position the leader-repeat junction in Site 1 and the repeat channel is partially occupied. F Non-Uniform refinement of 93,957 particles captured during the second transesterification. From the trimmed dataset, not observed in the PAM dataset. The repeat stretches across the channel, and the Site 2 Cas1 dimer and Cas3b are stabilized. G. Biological assembly built from combining maps from E and F.

repeat and spacer of the CRISPR array (**Figure 8.4A**).

Previous structures have taken advantage of half-site intermediates, in which one strand of foreign DNA is covalently linked to the CRISPR repeat prior to incubation with Cas1-2[17,31,33,41]. To capture a snapshot of the structural intermediates formed during integration, covalent intermediates were not used. Rather, the integration reaction was purified by size exclusion chromatography 25 minutes after mixing fDNA-bound Cas1-2/3 and IHF-bound genomic fragment. The purified complex was immediately spotted on grids and vitrified. Grids were visualized in a 200 kEV Talos Arctica. Grid squares with thin ice and dense particle distribution were used to collect two datasets of 6,677 movies (PAM-containing) and 5,462 movies (PAM-trimmed). After preliminary motion correction and CTF-correction, PDB 8flj was used as a template for particle picking without low-pass filtering[41]. This template picking strategy identified 791,845 unique particles in the PAM-containing dataset and 719,626 unique particles in the trimmed dataset. From these particles sets, mid-resolution electron density maps were refined for two integration intermediates in the PAM-containing dataset, and three integration intermediates in the trimmed dataset (**Figure 8.4B-F, Supplementary Figure 8.9-8.10**).

These five volumes provide illuminating snapshots of CRISPR adaptation. First, short fragments of DNA bind to the capture complex, driving conformational changes that reveal DNA binding sites necessary for DNA delivery. The CRISPR leader does not displace these fragments from the foreign DNA binding face, as no classes corresponding to the long fragment of folded or unfolded leader interacting with the Cas1-2/3 complex at the foreign DNA face are observed, despite a 25-minute co-incubation. In all volumes, short, linear fragments of DNA bind to the

fDNA face (**Figure 8.3B, Figure 8.4B-F),** providing structural confirmation that Cas1-2/3 has a length preference[20]. Second, the delivery complex associates with specific motifs in the leader and the CRISPR. Leader motifs are presented to the Cas2 binding sites by an IHF-induced U-bend that traps foreigm DNA bound on one face of the complex. The U-bend can form without formation of the second IHF-induced loop, but the loop never forms without the U-bend (**Figure 8.4B-E**). No classes are observed in which the U-bend or loop form without a short linear fragment of foreign DNA first bound to the foreign DNA face of the complex, revealing that the conformational changes triggered by fDNA binding are pre-requisite for genome association **(Figure 8.4C-D**). The asymmetry observed in the delivery complex is preserved during integration, with one Cas3 averaged out due to conformational flexibility in all volumes, and one Cas1 dimer missing in the absence of a PAM. Third, the IHF-induced loop contorts the leader-repeat junction into Site 1, positioning the first few nucleotides of the repeat in the repeat channel. However, the presence of a PAM prevents the repeat from completely occupying the repeat channel. Only in the absence of a PAM does the repeat stretch across the channel to Site 2, destabilizing the Site 1 side of the complex and stabilizing the Site 2 side of the complex. Fourth, the repeat is always occupied directionally from Site 1 to Site 2. No fragments of DNA are observed in the CRISPR repeat channel that do not protrude from the Cas1 active site (**Figure 3E**). These structural observations reveal that foreign DNA binding, genome association, and PAM trimming each function as structural checkpoints during foreign DNA integration into the genome.

## The CRISPR Repeat Channel Stretches and Distorts The Repeat

The full repeat is only resolved across the length of the repeat channel in the PAM-trimmed integration dataset, corroborating previous biochemical evidence demonstrating that in

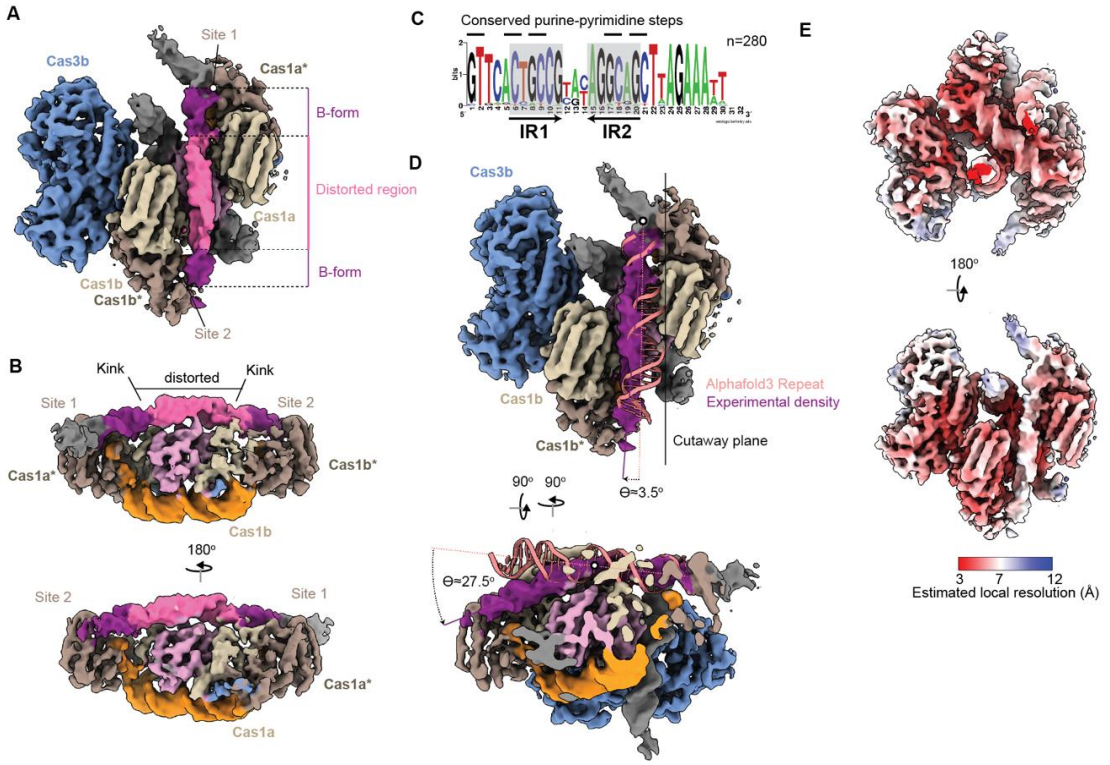



Figure 8.5. The repeat channel distorts the CRISPR repeat. A. The repeat is distorted in the middle of the repeat channel. The distorted region is displayed in pink, while the B-form region of the repeat is displayed in purple. B. Side view of the distorted repeat (top). Two kinks in the density flank the distorted region. C. Logo plot showing the conservation of the type I-F CRISPR repeat across 280(There are five di-nucleotide purine-pyrimidine steps in the CRISPR repeat, four of which are found in the inverted-repeat region. D. An Alphafold3 prediction fit into the density map starting at site 1 reveals that the repeat is twisted 3.5º off-center from the predicted conformation and bent ~27.5º down into the second transesterification site (Site 2). E. Local resolution estimation of the repeat-containing volume (n = 27,573 particles)

the presence of a PAM, the first but not the second transesterification occurs[41]. After association with the folded genome to form the U-bend, the second IHF-induced loop forms proximal to the leader-repeat-junction. This loop positions the leader repeat junction into Site 1 at the conserved

GT dinucleotide sequence[41]. The strand transfer breaks the leader-repeat junction, leaving the 3' strand kinked out away from the helix, and forming a covalent linkage between the 3' end of the foreign DNA and the 5' end of the repeat (**Figure 8.4, Figure 8.5**). Leaving the Cas1 active site, the repeat adopts an approximately B-form helical conformation for approximately ~25 Å, or 6-7 bp. A kink in the density marks the beginning of a distorted region in which the helical pitch is extended and the DNA is bent ~27.5º over the top of the Cas2 dimer. A second kink marks a transition back to approximately B-form as the repeat approaches Site 2. This kinking corresponds with the narrowest part of the channel formed by the inactive Cas1 subunits of each Cas1 dimer. This kinking also corresponds to conserved purine-pyrimidine steps in the type I-F repeat, at bases 6 and 7 and 20 and 21 in the repeat. The repeat distortion also introduces an ~3º shift from the trajectory of a linear piece of DNA of the same sequence as predicted with Alphafold (**Figure 8.5D**). This shift positions the repeat close to the Site 2 active site. We find no evidence that the Cas2 dimer at the core of the complex disassembles after the second transesterification, as evidenced by the lack of Cas1-2/3 volumes corresponding to the propeller-shaped fDNA capture complex. Rather, the cryoEM structures reveal the complex is trapped by two covalent linkages that creates a genomic knot at the site of integration.

## Discussion

In this cryoEM study of CRISPR adaptation, we show that Cas1-2/3 assemble into a complex that creates a positively charged foreign DNA binding channel. In this DNA capture conformation, Cas3 is inhibited by clamps on the RecA domains of Cas3 and an inhibitory loop inserted into the Cas3 HD domain. Foreign DNA binding induces conformational changes in the complex, exposing genome binding sites on two faces of the central Cas2 dimer, opening the

repeat channel on the opposing face of the complex, and exposing the Cas1 active sites. In the presence of foreign DNA, one side of the complex is stabilized, while the opposing side of the complex is poorly resolved, presumably due to conformational flexibility. On the stabilized side of the complex, an inhibitory loop (residues 483-511) remains in the HD active site, presumably suppressing HD nuclease activity. Formation of the proximal loop stabilizes the complex on the side of the first transesterification reaction. In the absence of a PAM, the repeat is distorted by Cas1 and Cas2 as it stretches across the repeat channel to the second Cas1 active site. Successful integration traps Cas1-2/3 between two strands of dsDNA (the repeat and the foreign DNA). The mechanism by which this knot is resolved remains unknown, but the DNA distortion in the repeat channel suggests that DNA distortion is a first step in the process.

The domestication of Cas1 from casposons for adaptive immunity highlights how the catalytic core of the transposase is preserved to generate target site duplications around a payload, but this functionally is regulated by diverse DNA binding and DNA bending proteins[6,17,41,59,60]. Intriguingly, the host DNA bending protein integration host factor (IHF) is co-opted by the lamba phage integrase to integrate the lambda phage genome into the host chromosome[63]. The co-option by lambda phage and the use of IHF for CRISPR adaptation highlight how genomic architecture is critical in modulating the conflict between parasite nad host.

Previous studies of Cas1 and Cas2 fusions reveal how Cas1 has fused with reverse transcriptases to acquire RNA sequences or Cas2 has been fused to an exonuclease to trim the PAM[60]. Cas4 and DEDDh family DNA-Q-like exonucleases trim the PAM in a sequence-specific manner[30,33]. In each case, the fusion highlights how Cas1 or Cas2 functionally coordinates with

the ancillary protein in adaptation. The Cas2/3 fusion combines a key component of the type I interference response with the adaptation machinery[16]. Previous studies demonstrate that Cas1 regulates the interference activity of Cas3, and functions in interference in the ICP1 phage type I-F system[22,23,38,52]I h. Though these findings highlight the role of the core adaptation machinery in interference, the question of how the interference machinery regulates adaptation remains open[21].

This structural study of integration in the type I-F CRISPR system adds an additional dimension to the connection between adaptation and interference by revealing that Cas3 plays a critical role in gating access to the genome binding sites that recognize the leader. Intriguingly, Alphafold3 structural predictions of Cas3 and the Cas1-2 heterohexamer from *E. coli* reveal consistent predictions of Cas3 associating with Cas1 in a conformation that would block the unique fold required in type I-E IHF-mediated integration, even in the absence of fusion to Cas2 (**Figure SX**). These predictions suggest that Cas3 may play a similar role in regulating the accessibility of folded genome interaction sites (**need supplementary figure here**), suggesting that Cas3 and Cas1 function as opposing regulators of xxxx.

Targeted delivery of foreign nucleic acids is an important step for developing gene editing and molecular recording tools that precisely paste foreign sequences at new locations in the genome[61,62,64,65]. This study advances our understanding of how Cas3 gates access to the CRISPR integrase to create structural checkpoints in CRISPR adaptation, and provides an explanation for how access to putative transposase insertion sites is regulated by a nuclease/helicase. This provides an exciting avenue for future studies to explore the regulation of precise genomic integration in parasitic transposase and for the development of gene therapies and genome recording tools.

## Materials & Methods

### Conservation Analysis of Cas2/3 and Cas1

The Cas2/3 sequence from *Pseudomonas aeruginosa* strain UCBPP-PA14 (100% identical to multi-species accession WP_003139227.1) was used to query the nr_70_Mar12 database using the ProtBLAST/PSI-BLAST tool on the MPI Bioinformatics Toolkit server (Scoring Matrix: BLOSUM62; E-value cutoff for reporting: 1e-3; E-Value cutoff for inclusion: 1e-6; Max target hits: 1000).[66] The resulting alignment of 1000 sequences was downloaded and 170 sequences missing Cas2 or Cas3 domains were manually removed. MaxAlign was used to remove 145 additional poorly aligned sequences (flags: (-v=1, -a).[67] Gappy columns were removed using trimAL (flags: -gt 0.7) to generate an alignment of Cas2/3 across 1071 positions.[68] This alignment was used to build a phylogenetic tree using Fasttree (flags: -pseudo, -wag, -gamma), which was visualized in R using ggtree (**Figure S1**)[69,70].

Complete genomes from NCBI were downloaded on July 11, 2023 and open-reading frames annotated using Prodigal.[71] These open reading frames were then queried for *cas* genes using MacsyFinder v1.0.5[72] with the following parameters: *"macsyfinder --sequence-db <open_reading_frames> --db-type gembase -d <CRISPR_subtype_definitions> -p <HMM_profiles> -w 50 -vv all"*. HMM profiles and classification definitions (*CRISPR_subtype_definitions*) used in MacsyFinder are acquired from the edited version of CasFinder v3.1.0[73] to include definitions for the CAST systems (available at https://github.com/macsy-models/CasFinder). Type I-F Cas2/3 sequences (any sequence annotated with "cas3f","cas3","cas3HD" and the I-F, n=1922) and Cas1 sequences (n=1838) were extracted from the database of Prodigal-annotated open-reading frames and identical

sequences removed with CD-hit (-c 1). The remaining 739 (Cas1) or 970 (Cas2/3) sequences were aligned in MAFFT (mafft --genafpair --maxiterate 1000), and poorly aligned sequences were removed using MaxAlign (flags: -v=1, -a). The resulting alignments of 679 (Cas1) and 745 (Cas2/3) sequences were submitted to Consurf with the atomic model to calculate conservation scores and map conservation onto the structure.

Nucleic Acid Preparation

Foreign DNA fragments (32-34 bp) were either ordered as dsDNA fragments or ordered as 100 nmole ssDNA oligonucleotides and resuspended to a working concentration of 100 uM in 1x TE buffer. To generate dsDNA foreign DNA fragments, 2 nmole of ssDNA fragments were incubated in 0.5X hybridization buffer (20 mM Tris-HCl pH 7.5, 220 mM K-Glutamate, 5 mM EDTA, 1 mM TCEP, 2 %glycerol) and were incubated at 95 C for 5 minutes, followed by 25 C for 10 minutes. Samples were then denatured at 100 C for 5 minutes, and slow cooled back to 25 C at a rate of 6 degrees/5 minutes for an hour, to a final temperature of 25 C.

To assemble a 200 bp fragment of the CRISPR leader and first repeat and spacer, 4 nmol of 200 bp ssDNA fragments were ordered and resuspended in a modified hybridization buffer (20 mM HEPES pH 7.5, 200 mM K-glutamate, 2% glycerol, 1 mM TCEP) to 200 uM. ssDNA fragments were combined in a 1:1 ratio to generate 100 uM dsDNA, heated to 100 C for 5 minutes, and then slow-cooled to 25 C over the course of an hour. Annealed oligos were stored at -80 C.

Protein Purification

*P. aeruginosa* IHF heterodimer was purified as previously described[19]. Briefly, 6xHis-tagged IHFα and StrepII-tagged IHFβ were co-expressed in *E. coli* BL21(DE3). Cell pellets were

lysed by sonication in IHF lysis buffer (25 mM HEPES-NaOH pH 7.5, 500 mM NaCl, 10 mM Imidazole, 1 mM TCEP, 5% Glycerol), supplemented with 0.3x Halt Protease Inhibitor Cocktail (ThermoFisher), at 4°C. Lysate was clarified by two rounds of centrifugation at 10,000 x g for 15 minutes, at 4°C. His-tagged IHF was captured on HisTrap HP resin (Cytiva), and eluted with 500 mM Imidazole. Affinity tags were cleaved using PreScision protease, and the PreScision protease and remaining 6x-His-IHFα were removed by affinity chromatography using HisTrap HP resin (Cytiva). Untagged IHF heterodimer was then further purified on Heparin Sepharose (Cytiva) and eluted with a linear gradient to a buffer containing 2 M NaCl. Fractions containing IHF heterodimer were concentrated and further purified by SEC (Size Exclusion Chromatography) on a Superdex 75 column (Cytiva) equilibrated in IHF buffer (25 mM HEPES-NaOH pH 7.5, 200 mM NaCl, 5 % Glycerol). IHF overexpression plasmids have been deposited at Addgene (#149384, #149385).

*P. aeruginosa* Cas1-2/3 heterohexamer complexes (wildtype and mutants) were purified as previously described[19]. Briefly, StrepII-tagged Cas1-Cas2/3 was overexpressed in *E. coli* BL21(DE3). Cell pellets were lysed via sonication in Cas1-2/3 lysis buffer (50 mM HEPES pH 7.5, 500 mM KCl, 10% Glycerol, 1 mM DTT), supplemented with 0.3x Halt Protease Inhibitor Cocktail (ThermoFisher), at 4°C. Lysate was clarifed as above. StrepII-tagged Cas1-Cas2/3 complexes were affinity purified on StrepTrap HP resin (GE Healthcare) and eluted with Cas1-2/3 Lysis Buffer containing 3 mM desthiobiotin (Sigma-Aldrich). Eluate was concentrated at 4°C (Corning Spin-X concentrators), before purification over a Superdex 200 size-exclusion column (Cytiva) equilibrated in 10 mM HEPES pH 7.5, 500 mM Potassium Glutamate, and 10% Glycerol.

Assembly and Purification of the Delivery Complex

A 1:1 mixture of purified Cas1-2/3 and dsDNA was created by mixing 2 nmoles of purified Cas1-2/3 thawed on ice then warmed to 25 C with 2 nmoles of annealed dsDNA containing a single 3' GG extension. This mixture was incubated at 25 C for 20 minutes, then centrifuged at 4 C, 22000 g for 20 minutes to pellet any aggregate. The supernatant was collected in a 1 mL syringe and injected onto a Superdex 200 10/300 column (Cytiva) equilibrated with 20 mM Tris-HCl pH 7.5, 200 mM monopotassium glutamate, 5 mM EDTA, 1 mM TCEP, 2 % Glycerol. 0.25 ml fractions were individually concentrated using 100 kDA MWCO 0.5 mL concentrator units at 4 C, 15000 g, to ~20 uL. All fractions were analyzed on SDS-PAGE gels (15% resolving) to determine protein composition and 8% Urea PAGE gels to determine nucleic acid composition The fifth SEC fraction contained all DNAs and proteins of interest and was further analyzed by cryo-EM.

Assembly and Purification of the Integration Complex

3.75 nmoles of purified IHF (at 1234 uM) was diluted to a final volume of 125 uL (30 uM) in SEC buffer (20 mM HEPS pH 7.5, 200 mM K-Glutamate, 5 mM $MnCl_2$ 7.5 mM spermidine, 2% glycerol, 1 mM TCEP) at 25 C to avoid precipitation of IHF. 1.5 nmoles of annealed CRISPR fragment was diluted to a final volume of 95 uL (15 uM) in the same SEC buffer. Both the pre-diluted IHF solution and pre-diluted CRISPR solution were briefly warmed to 25 C prior to mixing (220 uL final volume) and incubation for 25 C for 10 minutes.

1.5 nmoles of purified Cas1-2/3 was mixed with 1.875 nmoles of foreign DNA fragment and 1.4 uL of 1 M $MnCl_2$ and 2.1 uL of 1 M spermidine in pH neutralized buffer and incubated for 10 minutes at 25 C.

The CRISPR-IHF reaction was mixed with the Cas1-2/3-fDNA reaction and incubated for 20 minutes at 35 C. Precipitates were removed by centrifugation at 4 C, 20,000xg, prior to size exclusion. A Superose 6 10/300 column was equilibrated to SEC buffer (20 mM HEPS pH 7.5, 200 mM K-Glutamate, 5 mM $MnCl_2$ 7.5 mM spermidine, 2% glycerol, 1 mM TCEP) at 4 C. 0.5 mL fractions were collected, analyzed on SDS-PAGE (15%) and UREA-PAGE (8%) gels and concentrated to 10 uL (~4 uM for PAM, ~3 uM for trimmed).

CryoEM Sample Preparation and Data Acquisition – Cas1-2/3 Capture Complex

Frozen aliquots of purified Cas1-2/3 complex (12.7921 uM) were shipped on dry ice to the National Center for CryoEM Access and Training (NCCAT) and the Simons Electron Microscopy Center located at the New York Structural Biology Center (NYSBC). Sample was spotted on Quantifoil R0.6/1 Cu 300 glow discharged mesh grids (Quantifoil) and blotted for 2 s at 10 C and 100% humidity using a Leica GP2 (Leica Microsystems) prior to plunge freezing in liquid ethane. The dataset was collected on a Titan Krios at NYSBC (Thermofisher Scientific) operating at 300 keV with a Gatan K3 direct electron detector and energy filter at a nominal magnification of 81000. 40 frame movies were collected at a 50 ms frame rate and a dose rate of 25.26 $e^-/Å^2/s$ for a total dose of 52.51 $e^-/Å^2$ with a pixel size of 1.0691 Å/px over a defocus range of 0.8-2.5 um. Holes were identified and targeted using the Leginon software, and preliminary dataprocessing was done on the fly in cryoSPARC Live.

CryoEM Sample Preparation and Data Acquisition – Cas1-2/3 Delivery Complex

Frozen aliquots of purified Cas1-2/3 delivery complex bound to fDNA (45.4 uM) were shipped on dry ice to the National Center for CryoEM Access and Training (NCCAT) and the

Simons Electron Microscopy Center located at the New York Structural Biology Center (NYSBC). Sample was thawed and diluted 1:10 in SEC buffer to a concentration of 4.54 uM. 3 uL of diluted sample was spotted on glow discharged Quantifoil R1.2/1.3 Cu 300 mesh grids (Quantifoil), blotted for 4.5 s at force 0 and 4 C, 100% humidity, and plunge frozen in liquid ethane using a Vitrobot (Mk. IV, ThermoFisher Scientific). The dataset was collected on a Titan Krios at NYSBC (Thermofisher Scientific) operating at 300 keV with a Falcon 4EC direct electron detector at a nominal magnification of 96000. 40 frame movies were collected at a 150 ms frame rate and a dose rate of 8.63 $e^-/Å^2/s$ for a total dose of 51.78 $e^-/Å^2$ with a pixel size of 0.833 Å/px over a defocus range of 0.9-2.8 um. Holes were identified and targeted using the Leginon software, and preliminary dataprocessing was done on the fly in cryoSPARC Live.

CryoEM Sample Preparation and Data Acquisition – Cas1-2/3 Integration Complex

Purified integration complex with either 34 bp of dsDNA that included a ‘GG’ PAM or 32 bp dsDNA fragment was diluted to a concentration of ~1 µM in SEC buffer lacking glycerol (20 mM HEPS pH 7.5, 200 mM K-Glutamate, 5 mM $MnCl_2$ 7.5 mM spermidine, 2% glycerol, 1 mM TCEP). Sample was spotted on Quantifoil Au 300 R2/1 + 2 nm C grids that were glow discharged for 45s seconds with a 10 second hold (easiGlow, Pelco). 3 µl of diluted integration complex was spotted on the grids, and then the grids were blotted for 5 seconds, with a blot force of "5", at 100% humidity and 4°C followed by plunge freezing using a Vitrobot (Mk. IV, ThermoFisher Scientific). The datasets for both the trimmed and PAM-containing integration complexes were collected Montana State University's Talos Arctica transmission electron microscope (ThermoFisher Scientific), with a field emission gun operating at an accelerating voltage of 200 keV using parallel illumination conditions. Movies were acquired using a Gatan

K3 direct electron detector, operated in electron counting mode targeting a total electron exposure of 65.67 e-/Å$^2$ over 50 frames (5.992 s exposure, 0.12 s frame time). The SerialEM data collection software was used to collect micrographs at 36,000-fold nominal magnification (1.152 Å/pixel at the specimen level) with a nominal defocus set to cycle from 0.5 µm - 2.0 µm. Stage movement was used to target the center of four 2.0 µm holes for focusing, and image shift was used to acquire high magnification images in the center of each of the holes. Preliminary dataprocessing was done on the fly in cryoSPARC Live.

Cryo-EM Image Processing

Movies were motion and CTF-corrected in cryoSPARC using patch motion and patch CTF correction. For details, see Supplementary Figure 1 (Capture complex), Supplementary Figure X (delivery complex), Supplementary Figure X (integration complex with PAM fDNA), and Supplementary Figure X (integration complex with trimmed DNA) . In brief, initial rounds of blob picking (capture/delivery) or template (integration complexes) were used to generate 3D volumes from the dataset for a second round of template picking. Template-picked particles were subject to multiple rounds of multi-class Ab initio reconstructions and heterogenous refinements to identify the major 3D classes of the complex. Particle stacks corresponding to individual complex were further refined through iterative multi-class Ab Initio reconstructions and heterogeneous refinements, 3D classification, and 3D variability analysis in cryoSPARC.

Model Building and Validation

Alphafold3 was used to generate a prelimnary model of the capture complex. However, Alphafold3's prediction does not match the structures of the remaining complexes, so in these instances, 8FLJ was used as a preliminary model. Protein and DNA segments were individually

rigid-body fit into the EM density map using the “fit <model> inMap <map>” command in UCSF ChimeraX. Isolde was intitially used to morph the rigid body-fit initial model into the density map, followed by Real Space Refinement with morphing (no secondary structure restraints, ignore symmetry conflicts) in Phenix. The resulting map and model were re-opened in Isolde/ChimeraX and any gross conformational issues were corrected with the Isolde simulation running. After a second round of Real Space refinement (no morphing, no secondary structure restraints), problem areas were inspected in Coot and restrained to ideal geometry, using secondary structure and German McClure distance restraints generated in ProSMART. MolProbity and the PDB validation service server (https://validate-rcsb-1.wwpdb.org/) were used to identify problem regions subsequently corrected in Coot. Contacts and hydrogen bonds between residues were identified by ChimeraX v1.4 using the “contacts” and “hbonds” commands respectively, with default parameters.

## References


1. Kapitonov, V.V., and Koonin, E.V. (2015). Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biology Direct *10*. https://doi.org/10.1186/s13062-015-0055-8.

2. Martin, E.C., Le Targa, L., Tsakou-Ngouafo, L., Fan, T.-P., Lin, C.-Y., Xiao, J., Huang, Z., Yuan, S., Xu, A., Su, Y.-H., et al. (2023). Insights into RAG Evolution from the Identification of “Missing Link” Family A RAGL Transposons. Molecular Biology and Evolution *40*, msad232. https://doi.org/10.1093/molbev/msad232.

3. Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biology *12*, 36. https://doi.org/10.1186/1741-7007-12-36.

4. Fugmann, S.D., Messier, C., Novack, L.A., Cameron, R.A., and Rast, J.P. (2006). An ancient evolutionary origin of the Rag1/2 gene locus. Proceedings of the National Academy of Sciences *103*, 3728–3733. https://doi.org/10.1073/pnas.0509720103.

5. Carmona, L.M., and Schatz, D.G. (2017). New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. FEBS J *284*, 1590–1605. https://doi.org/10.1111/febs.13990.

6. Béguin, P., Charpin, N., Koonin, E.V., Forterre, P., and Krupovic, M. (2016). Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Research *44*, 10367–10376. https://doi.org/10.1093/nar/gkw821.

7. Little, A.J., Corbett, E., Ortega, F., and Schatz, D.G. (2013). Cooperative recruitment of HMGB1 during V(D)J recombination through interactions with RAG1 and DNA. Nucleic Acids Research *41*, 3289–3301. https://doi.org/10.1093/nar/gks1461.

8. Liu, C., Zhang, Y., Liu, C.C., and Schatz, D.G. (2022). Structural insights into the evolution of the RAG recombinase. Nat Rev Immunol *22*, 353–370. https://doi.org/10.1038/s41577-021-00628-6.

9. Chi, H., Pepper, M., and Thomas, P.G. (2024). Principles and therapeutic applications of adaptive immunity. Cell *187*, 2052–2078. https://doi.org/10.1016/j.cell.2024.03.037.

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

11. Lee, H., and Sashital, D.G. (2022). Creating memories: molecular mechanisms of CRISPR adaptation. Trends in Biochemical Sciences *47*, 464–476. https://doi.org/10.1016/j.tibs.2022.02.004.

12. Siezen, R.J., Starrenburg, M.J.C., Boekhorst, J., Renckens, B., Molenaar, D., and Vlieg, J.E.T. van H. (2008). Genome-Scale Genotype-Phenotype Matching of Two Lactococcus lactis Isolates from Plants Identifies Mechanisms of Adaptation to the Plant Niche. Appl. Environ. Microbiol. *74*, 424–436. https://doi.org/10.1128/AEM.01850-07.

13. Sternberg, S.H., Richter, H., Charpentier, E., and Qimron, U. (2016). Adaptation in CRISPR-Cas Systems. Molecular Cell *61*, 797–808. https://doi.org/10.1016/j.molcel.2016.01.030.

14. Krupovic, M., Béguin, P., and Koonin, E.V. (2017). Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Current Opinion in Microbiology *38*, 36–43. https://doi.org/10.1016/j.mib.2017.04.004.

15. 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. Nature Reviews Genetics *21*, 119–131. https://doi.org/10.1038/s41576-019-0172-9.

16. 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. (2020). Evolutionary

classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol *18*, 67–83. https://doi.org/10.1038/s41579-019-0299-x.

17. 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. https://doi.org/10.1126/science.aao0679.

18. 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. https://doi.org/10.1038/nature14237.

19. Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P., and Wiedenheft, B. (2021). Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. Current Biology *31*, 3515-3524.e6. https://doi.org/10.1016/j.cub.2021.05.068.

20. Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L., et al. (2017). Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences *114*, E5122–E5128. https://doi.org/10.1073/pnas.1618421114.

21. Wiegand, T., Semenova, E., Shiriaeva, A., Fedorov, I., Datsenko, K., Severinov, K., and Wiedenheft, B. (2020). Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. The CRISPR Journal *3*, 378–387. https://doi.org/10.1089/crispr.2020.0069.

22. Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., and Joo, C. (2020). Selective loading and processing of prespacers for precise CRISPR adaptation. Nature *579*, 141–145. https://doi.org/10.1038/s41586-020-2018-1.

23. Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Molecular Cell *63*, 852–864. https://doi.org/10.1016/j.molcel.2016.07.011.

24. Bernheim, A., and Sorek, R. (2020). The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol *18*, 113–119. https://doi.org/10.1038/s41579-019-0278-2.

25. 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. https://doi.org/10.1038/nature15760.

26. 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. https://doi.org/10.1038/nsmb.2820.

27. 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. https://doi.org/10.1099/mic.0.023960-0.

28. 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. https://doi.org/10.1016/j.celrep.2018.02.103.

29. 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. https://doi.org/10.1016/j.molcel.2018.03.003.

30. Lee, H., Dhingra, Y., and Sashital, D.G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. eLife *8*, e44248. https://doi.org/10.7554/eLife.44248.

31. Dhingra, Y., Suresh, S.K., Juneja, P., and Sashital, D.G. (2022). PAM binding ensures orientational integration during Cas4-Cas1-Cas2-mediated CRISPR adaptation. Molecular Cell *82*, 4353-4367.e6. https://doi.org/10.1016/j.molcel.2022.09.030.

32. 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*, e45543. https://doi.org/10.15252/embr.201745543.

33. Wang, J.Y., Tuck, O.T., Skopintsev, P., Soczek, K.M., Li, G., Al-Shayeb, B., Zhou, J., and Doudna, J.A. (2023). Genome expansion by a CRISPR trimmer-integrase. Nature *618*, 855–861. https://doi.org/10.1038/s41586-023-06178-2.

34. Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, F.D., Zhou, S., Rajashankar, R., Kurinov, I., et al. (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol *21*, 771–777. https://doi.org/10.1038/nsmb.2875.

35. Xiao, Y., Luo, M., Dolan, A.E., Liao, M., and Ke, A. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science *361*, eaat0839. https://doi.org/10.1126/science.aat0839.

36. Kim, D.Y., Lee, S.Y., Ha, H.J., and Park, H.H. (2024). Structural basis of Cas3 activation in type I-C CRISPR-Cas system. Nucleic Acids Research *52*, 10563–10574. https://doi.org/10.1093/nar/gkae723.

37. Redding, S., Sternberg, S.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C. (2015). Surveillance and Processing of Foreign DNA by the *Escherichia coli* CRISPR-Cas System. Cell *163*, 854–865. https://doi.org/10.1016/j.cell.2015.10.003.

38. Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C., and Wiedenheft, B. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc Natl Acad Sci U S A *114*, E5113–E5121. https://doi.org/10.1073/pnas.1616395114.

39. Jackson, R.N., Lavin, M., Carter, J., and Wiedenheft, B. (2014). Fitting CRISPR-associated Cas3 into the Helicase Family Tree. Current Opinion in Structural Biology *24*, 106–114. https://doi.org/10.1016/j.sbi.2014.01.001.

40. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. Nat Rev Microbiol *9*, 467–477. https://doi.org/10.1038/nrmicro2577.

41. Santiago-Frangos, A., Henriques, W.S., Wiegand, T., Gauvin, C.C., Buyukyoruk, M., Graham, A.B., Wilkinson, R.A., Triem, L., Neselu, K., Eng, E.T., et al. (2023). Structure reveals why genome folding is necessary for site-specific integration of foreign DNA into CRISPR arrays. Nat Struct Mol Biol, 1–11. https://doi.org/10.1038/s41594-023-01097-2.

42. 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. https://doi.org/10.1107/S2059798318002425.

43. Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallographica Section D *66*, 486–501. https://doi.org/10.1107/S0907444910007493.

44. 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 Cryst D *75*, 861–877. https://doi.org/10.1107/S2059798319011471.

45. Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. (2009). Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. Structure *17*, 904–912. https://doi.org/10.1016/j.str.2009.03.019.

46. Tang, D., Li, H., Wu, C., Jia, T., He, H., Yao, S., Yu, Y., and Chen, Q. (2021). A distinct structure of Cas1–Cas2 complex provides insights into the mechanism for the longer spacer acquisition in *Pyrococcus furiosus*. International Journal of Biological Macromolecules *183*, 379–386. https://doi.org/10.1016/j.ijbiomac.2021.04.074.

47. Wilkinson, M.E., Nakatani, Y., Staals, R.H.J., Kieper, S.N., Opel-Reading, H.K., McKenzie, R.E., Fineran, P.C., and Krause, K.L. (2016). Structural plasticity and in vivo activity of Cas1 from the type I-F CRISPR–Cas system. Biochemical Journal *473*, 1063–1072. https://doi.org/10.1042/BCJ20160078.

48. Wang, J.Y., Hoel, C.M., Al-Shayeb, B., Banfield, J.F., Brohawn, S.G., and Doudna, J.A. (2021). Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex. Nat Commun *12*, 2571. https://doi.org/10.1038/s41467-021-22900-y.

49. Ka, D., Hong, S., Jeong, U., Jeong, M., Suh, N., Suh, J.-Y., and Bae, E. (2017). Structural and dynamic insights into the role of conformational switching in the nuclease activity of the Xanthomonas albilineans Cas2 in CRISPR-mediated adaptive immunity. Structural Dynamics *4*, 054701. https://doi.org/10.1063/1.4984052.

50. Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., et al. (2008). A Novel Family of Sequence-specific Endoribonucleases Associated with the Clustered Regularly Interspaced Short Palindromic Repeats *. Journal of Biological Chemistry *283*, 20361–20371. https://doi.org/10.1074/jbc.M803225200.

51. Wu, C., Tang, D., Cheng, J., Hu, D., Yang, Z., Ma, X., He, H., Yao, S., Fu, T.-M., Yu, Y., et al. (2021). Mechanisms of spacer acquisition by sequential assembly of the adaptation module in Synechocystis. Nucleic Acids Research *49*, 2973–2984. https://doi.org/10.1093/nar/gkab105.

52. Zhang, L., Wang, H., Zeng, J., Cao, X., Gao, Z., Liu, Z., Li, F., Wang, J., Zhang, Y., Yang, M., et al. (2024). Cas1 mediates the interference stage in a phage-encoded CRISPR–Cas system. Nat Chem Biol, 1–11. https://doi.org/10.1038/s41589-024-01659-5.

53. Ka, D., Jang, D.M., Han, B.W., and Bae, E. (2018). Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. Nucleic Acids Research *46*, 9805–9815. https://doi.org/10.1093/nar/gky702.

54. Wang, X., Yao, D., Xu, J.-G., Li, A.-R., Xu, J., Fu, P., Zhou, Y., and Zhu, Y. (2016). Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. Nat Struct Mol Biol *23*, 868–870. https://doi.org/10.1038/nsmb.3269.

55. Hu, C., Myers, M.T., Zhou, X., Hou, Z., Lozen, M.L., Nam, K.H., Zhang, Y., and Ke, A. (2024). Exploiting activation and inactivation mechanisms in type I-C CRISPR-Cas3 for genome-editing applications. Molecular Cell *84*, 463-475.e5. https://doi.org/10.1016/j.molcel.2023.12.034.

56. Hu, C., Ni, D., Nam, K.H., Majumdar, S., McLean, J., Stahlberg, H., Terns, M.P., and Ke, A. (2022). Allosteric control of type I-A CRISPR-Cas3 complexes and establishment as effective nucleic acid detection and human genome editing tools. Molecular Cell *82*, 2754-2768.e5. https://doi.org/10.1016/j.molcel.2022.06.007.

57. 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. https://doi.org/10.1038/nsmb1246.

58. Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO Journal *30*, 1335–1342. https://doi.org/10.1038/emboj.2011.41.

59. Hickman, A.B., Kailasan, S., Genzor, P., Haase, A.D., and Dyda, F. (2020). Casposase structure and the mechanistic link between DNA transposition and spacer acquisition by CRISPR-Cas. eLife *9*, e50004. https://doi.org/10.7554/eLife.50004.

60. Watts, E.A., Garrett, S.C., Catchpole, R.J., Clark, L.M., Sanders, T.J., Marshall, C.J., Wenck, B.R., Vickerman, R.L., Santangelo, T.J., Fuchs, R., et al. (2023). Histones direct site-specific CRISPR spacer acquisition in model archaeon. Nat Microbiol *8*, 1682–1694. https://doi.org/10.1038/s41564-023-01446-3.

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

62. Lampe, G.D., King, R.T., Halpin-Healy, T.S., Klompe, S.E., Hogan, M.I., Vo, P.L.H., Tang, S., Chavez, A., and Sternberg, S.H. (2023). Targeted DNA integration in human cells without double-strand breaks using CRISPR-associated transposases. Nat Biotechnol, 1–12. https://doi.org/10.1038/s41587-023-01748-1.

63. Craig, N.L., and Nash, H.A. (1984). E. coli integration host factor binds to specific sites in DNA. Cell *39*, 707–716. https://doi.org/10.1016/0092-8674(84)90478-1.

64. Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2016). Molecular recordings by directed CRISPR spacer acquisition. Science *353*, aaf1175. https://doi.org/10.1126/science.aaf1175.

65. 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. https://doi.org/10.1038/s41586-022-04994-6.

66. Alva, V., Nam, S.-Z., Söding, J., and Lupas, A.N. (2016). The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res *44*, W410–W415. https://doi.org/10.1093/nar/gkw348.

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

68. 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. https://doi.org/10.1093/bioinformatics/btp348.

69. Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution *8*, 28–36. https://doi.org/10.1111/2041-210X.12628.

70. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

71. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

72. Abby, S.S., Néron, B., Ménager, H., Touchon, M., and Rocha, E.P.C. (2014). MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLOS ONE *9*, e110726. https://doi.org/10.1371/journal.pone.0110726.

73. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research *46*, W246–W251. https://doi.org/10.1093/nar/gky425.

74. Xiao, Y., Luo, M., Hayes, R.P., Kim, J., Ng, S., Ding, F., Liao, M., and Ke, A. (2017). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell *170*, 48-60.e11. https://doi.org/10.1016/j.cell.2017.06.012.

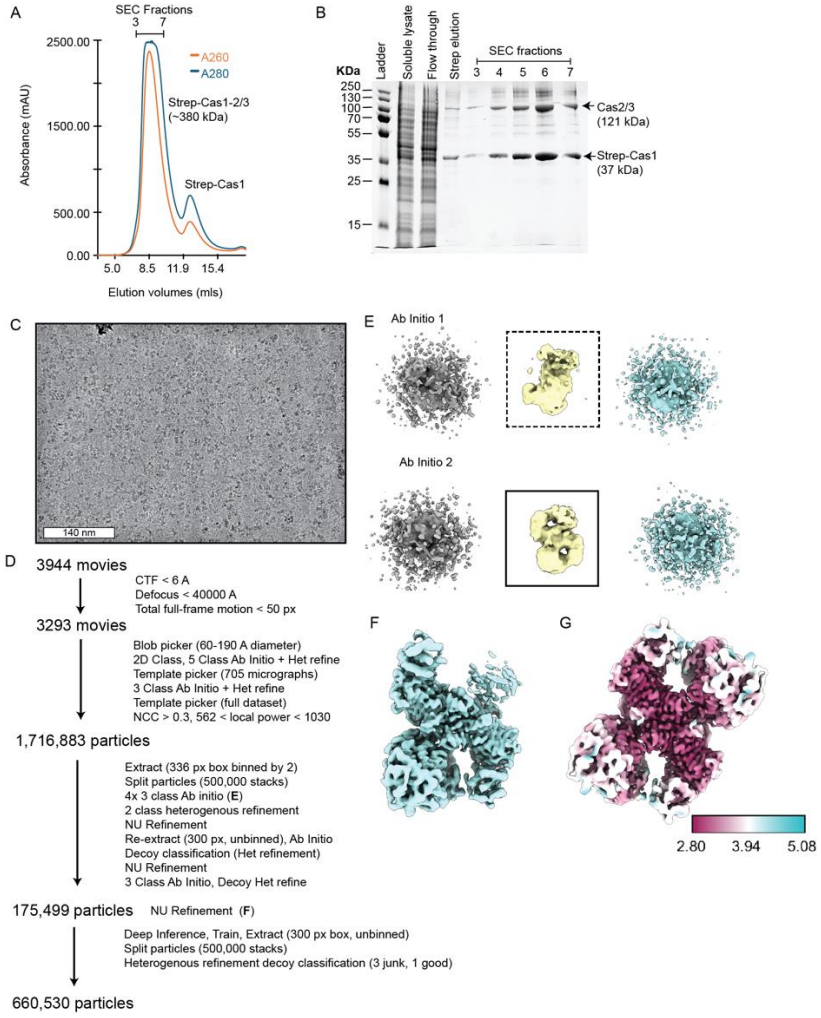
75. Mulepati, S., and Bailey, S. (2013). *In Vitro* Reconstitution of an *Escherichia coli* RNA-guided Immune System Reveals Unidirectional, ATP-dependent Degradation of DNA Target*. Journal of Biological Chemistry *288*, 22184–22192. https://doi.org/10.1074/jbc.M113.472233.

76. 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. https://doi.org/10.1126/science.1159689.

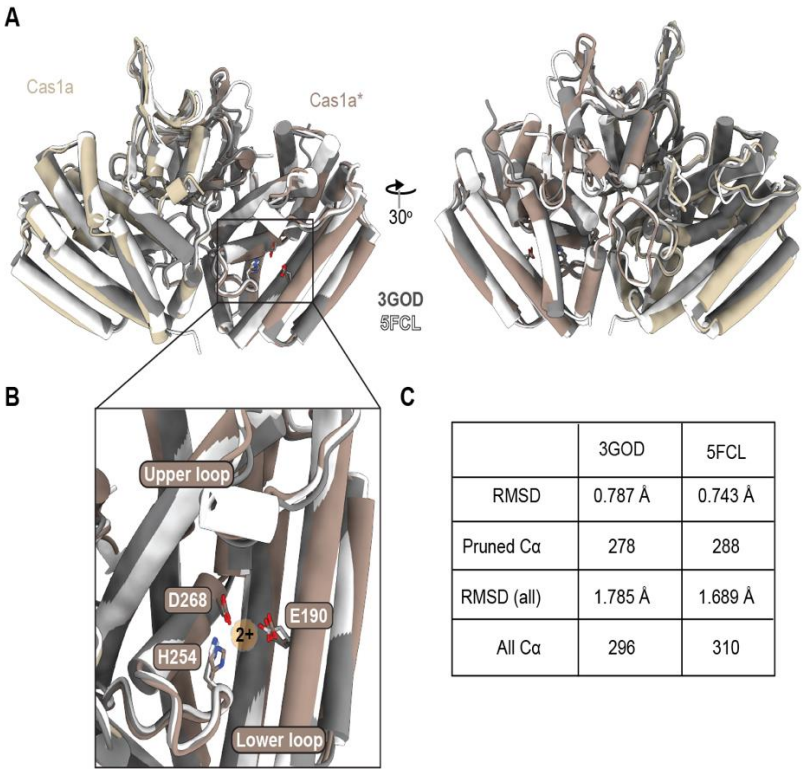
77. Béguin, P., Chekli, Y., Sezonov, G., Forterre, P., and Krupovic, M. (2019). Sequence motifs recognized by the casposon integrase of Aciduliprofundum boonei. Nucleic Acids Research *47*, 6386–6395. https://doi.org/10.1093/nar/gkz447.

78. Dickerson, R.E. (1998). DNA bending: The prevalence of kinkiness and the virtues of normality. Nucleic Acids Research *26*, 1906–1926. https://doi.org/10.1093/nar/26.8.1906.

Supplementary Figures

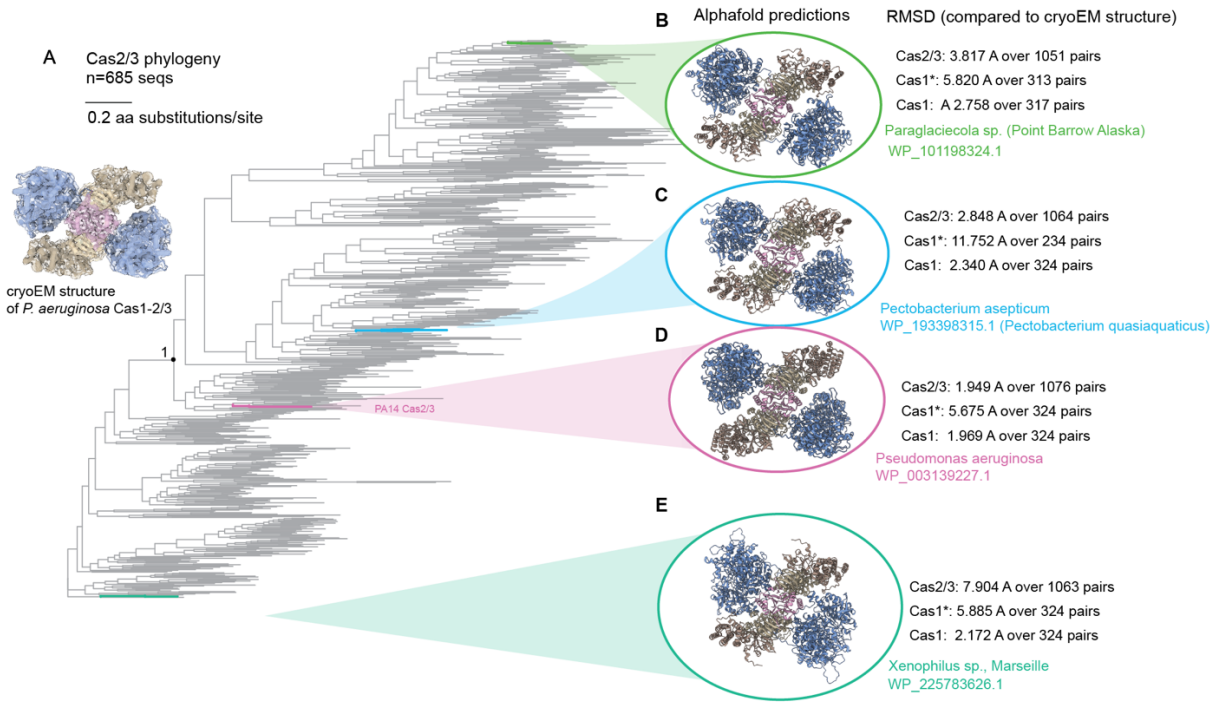



Supplementary Figure 8.1. Purification and data processing of the Cas1-2/3 capture complex. A. Size exclusion chromatography (SEC) of Cas1-2/3 results in a monodispersed complex with an estimated molecular weight of 380 kDa. B. SDS-PAGE of the fractions following affinity purification, and size exclusion. C. Sample micrograph at a nominal magnification of 81,000. D. Data processing summary E. Results from two of four total, 3-class Ab initio reconstructions and heterogenous refinements provided two volumes (boxed) for baited heterogeneous refinement (dashed box indicates volume with severe anisotropy, solid box indicates a volume with well-distributed views). F. Non-uniform refinement of 175,499 particles used to train cryoSPARC's Deep Inference job. G. Local resolution at estimation (FSC 0.5) reveals most of the complex is resolved below 3 Å

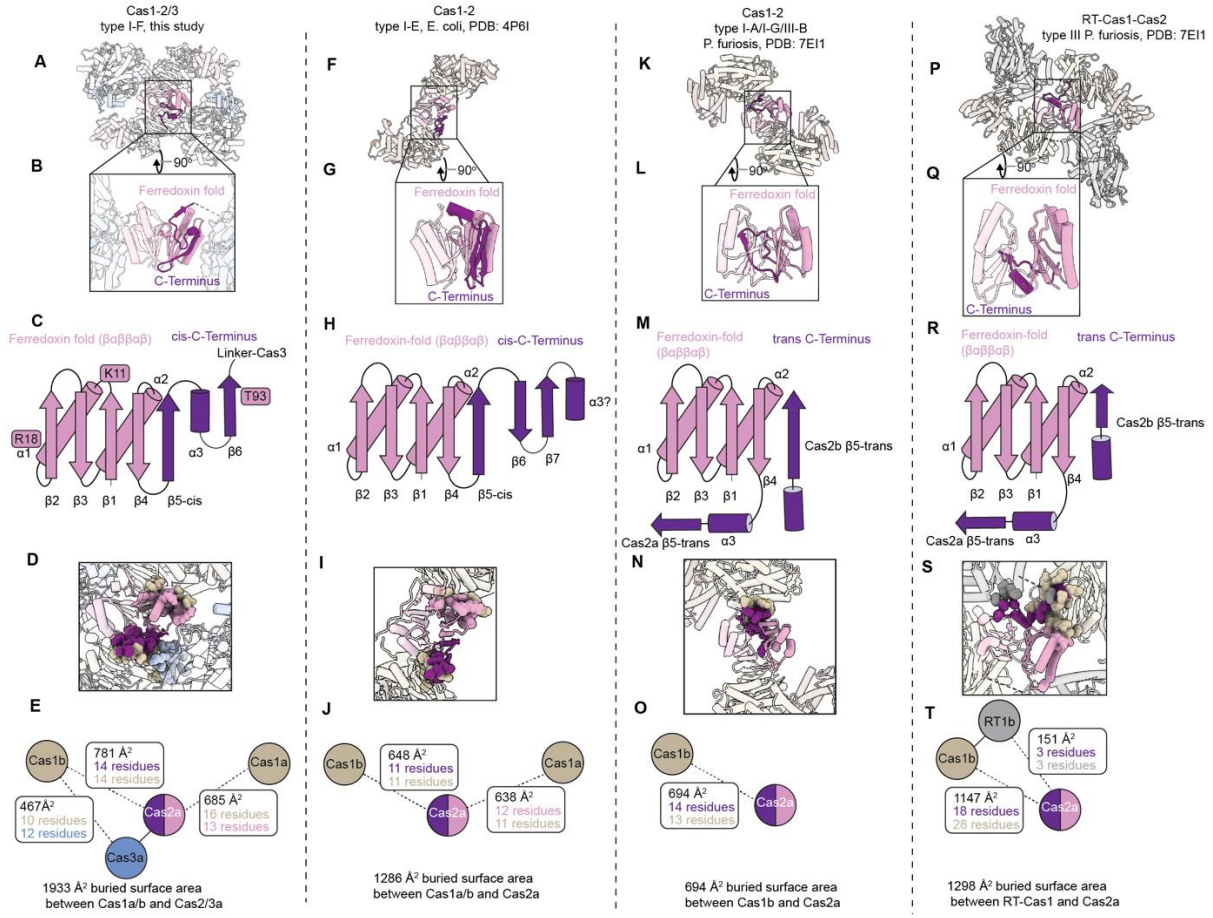



| | 3GOD | 5FCL |
|---|---|---|
| RMSD | 0.787 Å | 0.743 Å |
| Pruned Cα | 278 | 288 |
| RMSD (all) | 1.785 Å | 1.689 Å |
| All Cα | 296 | 310 |

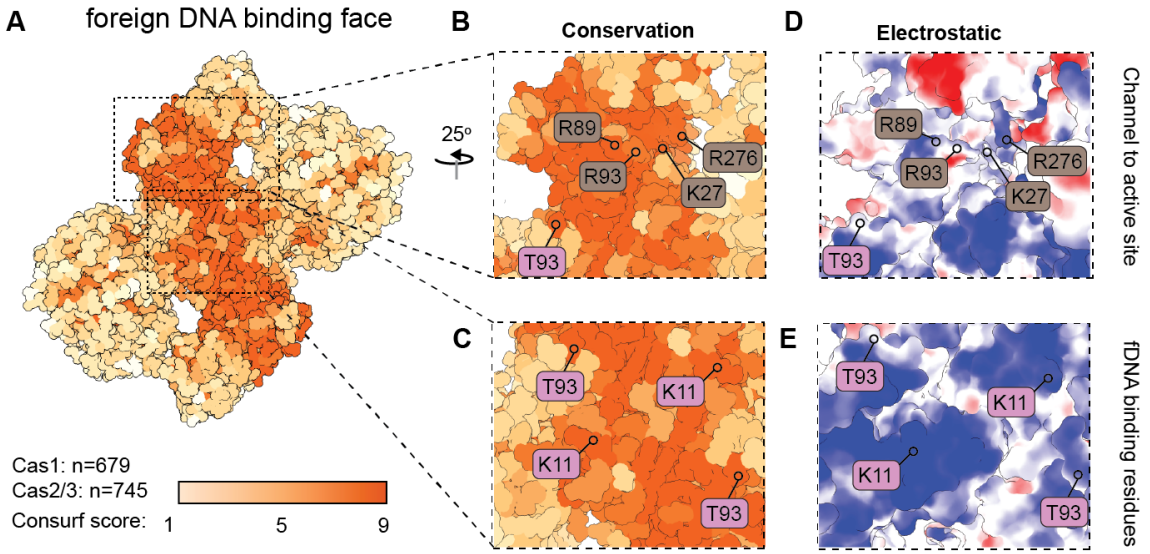
Supplementary Figure 8.2 Cas1 in complex with Cas2/3 preserves dimeric architecture. A. Structural overlay of previously determined Cas1 dimer crystal structures (PDB: 3GOD, P. aeruginosa in grey, PDB: 5FCL, P. atrosepticum in white) onto one Cas1 dimer from the Cas1-2/3 complex (colored as in Fig 1). B. Close-up of the Cas1* active site shows nearly perfect overlap of active site residues involved in coordinating a divalent metal ion (not observed in the cryoEM structure due to EDTA in the purification buffer) C. Table of root mean square deviations across pruned and unpruned C $\alpha$ atoms, from the matchmaker (mmaker) command in UCSF ChimeraX.



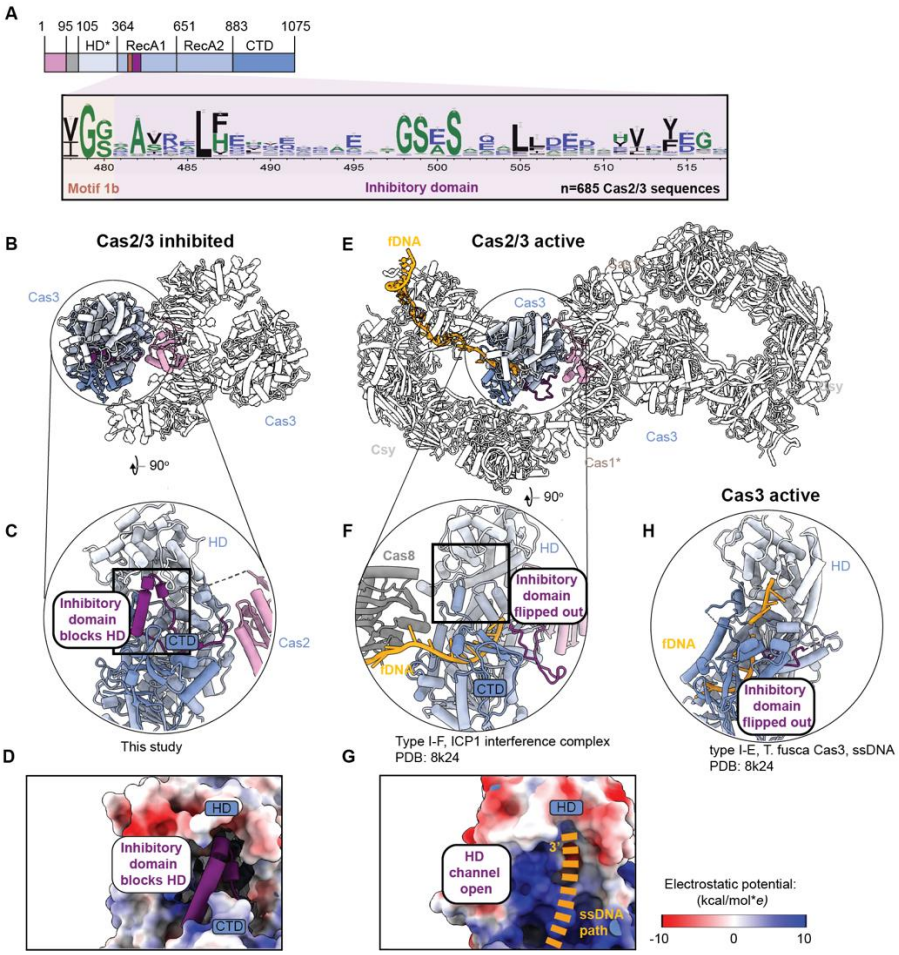

Supplementary Figure 8.3. Conservation of the four-bladed propeller Cas1-2/3 architecture across type I-F systems A. Phylogenetic tree of 685 Cas2/3 sequences. Sequences were identified through 3 iterations of PSI-BLAST, aligned in MAFFT, and the phylogenetic tree built in FastTree (see Methods). B-E. Alphafold predictions of Cas1-Cas2/3 heterohexamers reveal atomic-level structural conservation, as evidenced by overall complex architecture and the root-mean square deviation of subunits across diverse type I-F systems. Alphafold prediction and alignment statistics for each Alphafold 3 predicted structure when superimposed on the cryoEM-determined structure using mmaker in ChimeraX.



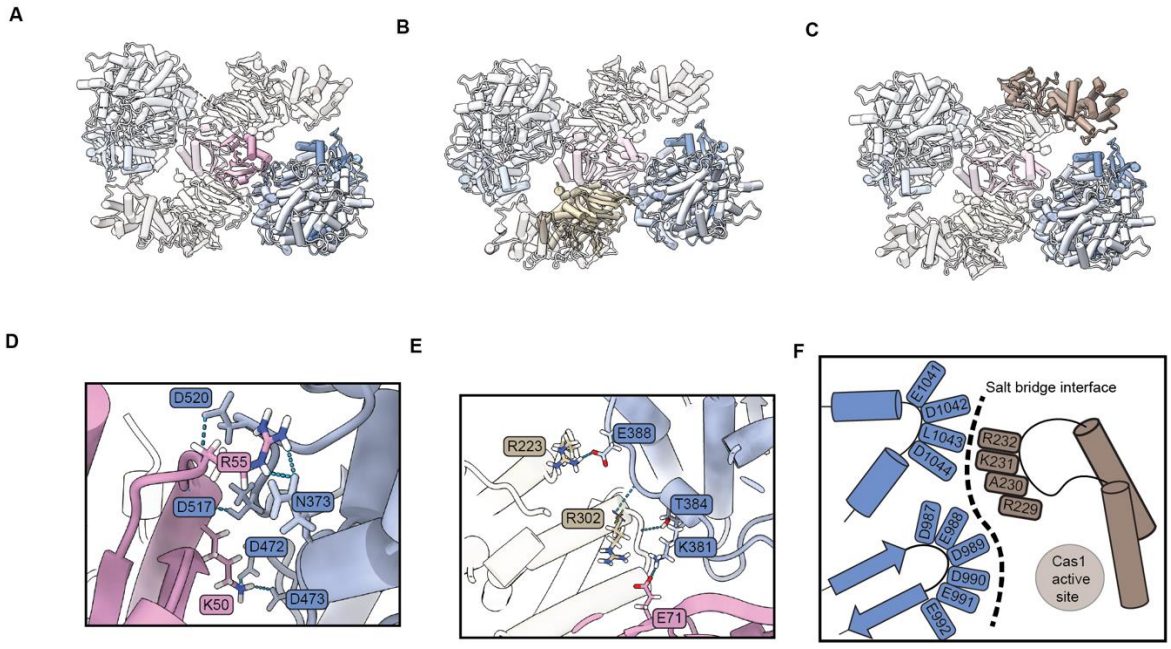

Supplementary Figure 8.4. The Cas2 C-terminus determines the extent of complex interactions A, F, K, P. The variable Cas2 C-terminus determines the identity and extent of partner interaction. Structures of experimentally determined Cas1-2 complexes from the PDB shown at 50% transparency, with one subunit of the Cas2 dimer colored to show the core ferredoxin-like fold (pink) and the variable C-terminus (purple). B,G,L,Q. Cas2-containing subunit only, rotated 90 degrees to visualize the dimer interface. C,H,M,R) Ribbon diagrams for each Cas2 structure in panels D, I, N, and S. Depiction of the residues involved at each interface for a single Cas2 within the dimer. E, J, O, T) Interaction diagrams depicting the subunit interactions for a single Cas2 dimer, with the number of residues and buried surface area of each interface depicted.



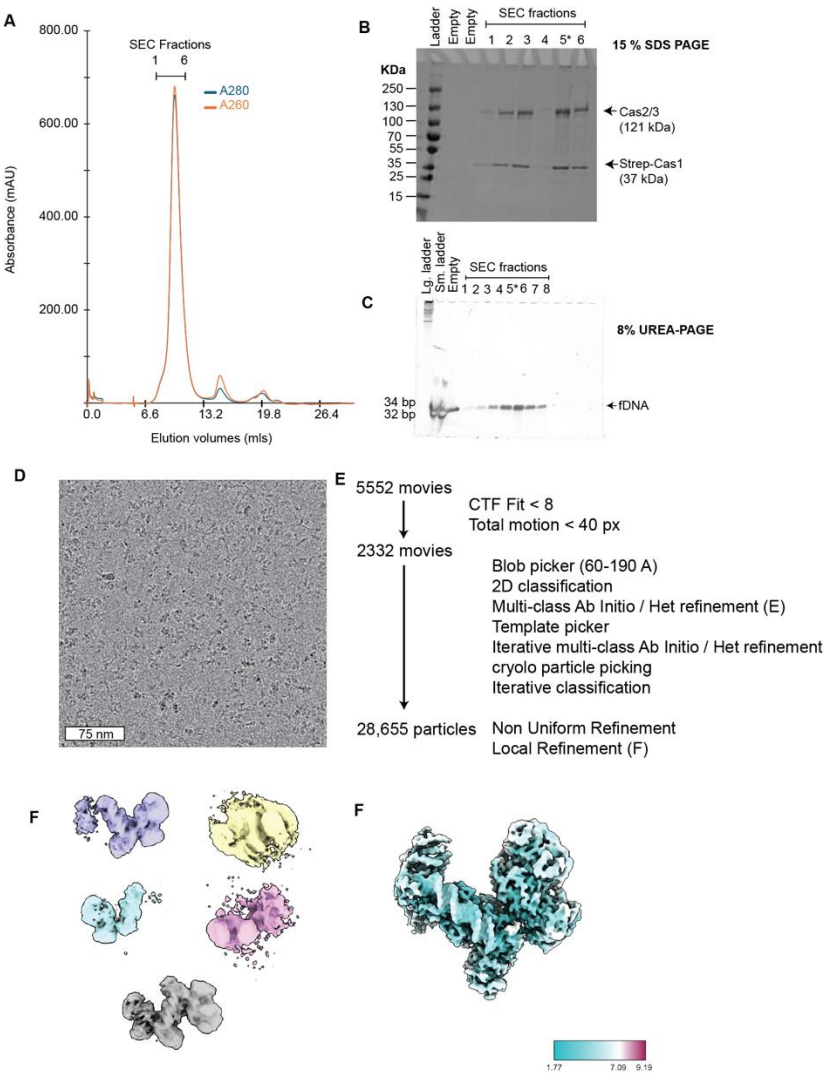

Supplementary Figure 8.5. Positively charged foreign DNA binding residues are conserved in the Cas1-2/3 complex. A. Multiple sequence alignments of Cas1 (n=679) and Cas2/3 (n=745) from type I-F systems were used to calculate Consurf conservation scores, and mapped back to the atomic model of the complex (1 – least conserved, 9 – most conserved) B,D. Conserved positively charged residues line a channel formed from the fDNA binding face to the Cas1 active site. C,E DNA binding residues on the fDNA binding face of the complex sit in conserved patches of positive charge.



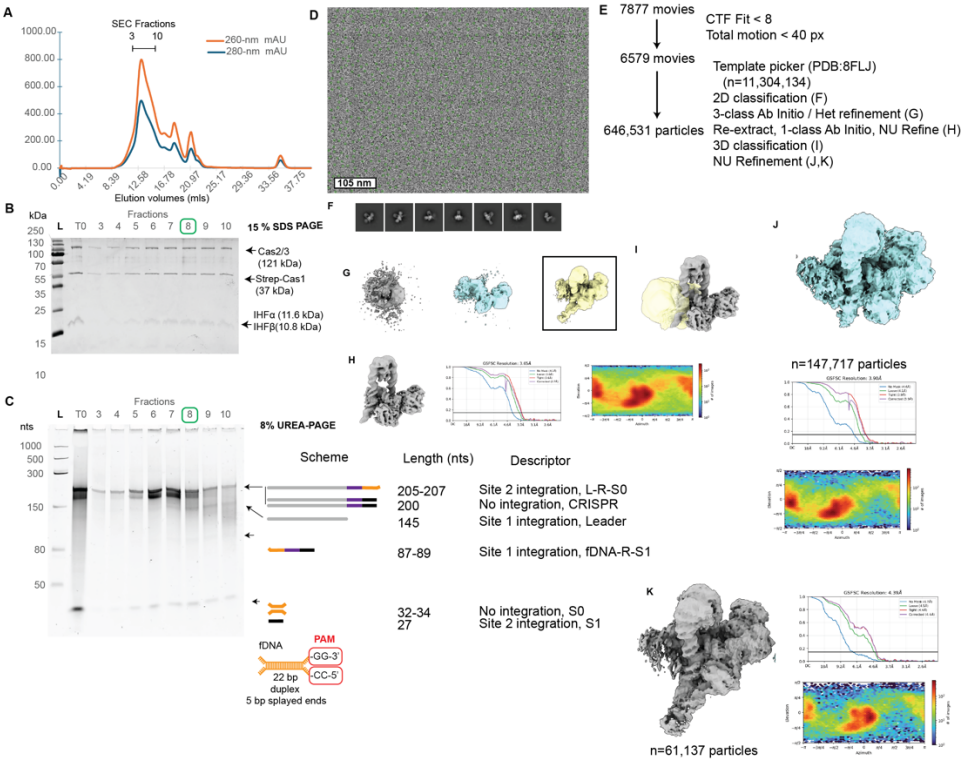

Supplementary Figure 8.6. A conserved domain inserted into the HD active site explains nuclease inhibition. A. label cas2 in the fig Schematic of Cas2/3 architecture depicting the location of the inhibitory domain and WebLogo built from an amino acid sequence alignment of 685 Cas2/3s. B. Atomic model of Cas1-2/3, colored by domain B-C. The well-structured helix-loop-helix formed by residues 478-513 obstructs the path of ssDNA by inserting into the HD active site. D. The inhibitory loop sits in a positively charged channel, the path of ssDNA to the HD domain E-F. Previously determined 3.7 Å cryoEM structure of the V. cholerae phage ICP1 CRISPR interference complex (PDB: 8k24) binding both a short fragment of foreign DNA on the foreign DNA face and the target sequence (the R-loop strand of fDNA colored orange). Bound to Csy, Cas3 is degrading a fragment of ssDNA. The inhibitory domain is unstructured and flipped out and down from the HD active site. G. The positively charged channel is empty, ready to accommodate ssDNA from threaded up from the helicase wedge. H. Previously determined crystal structure of T. fusca Cas3 bound to ssDNA (orange) (PDB: 4QQX). The inhibitory domain is partially unstructured, but the resolved main chain adopts the same orientation as the active Cas3 in panel C, not the inhibited Cas3 in panel B.



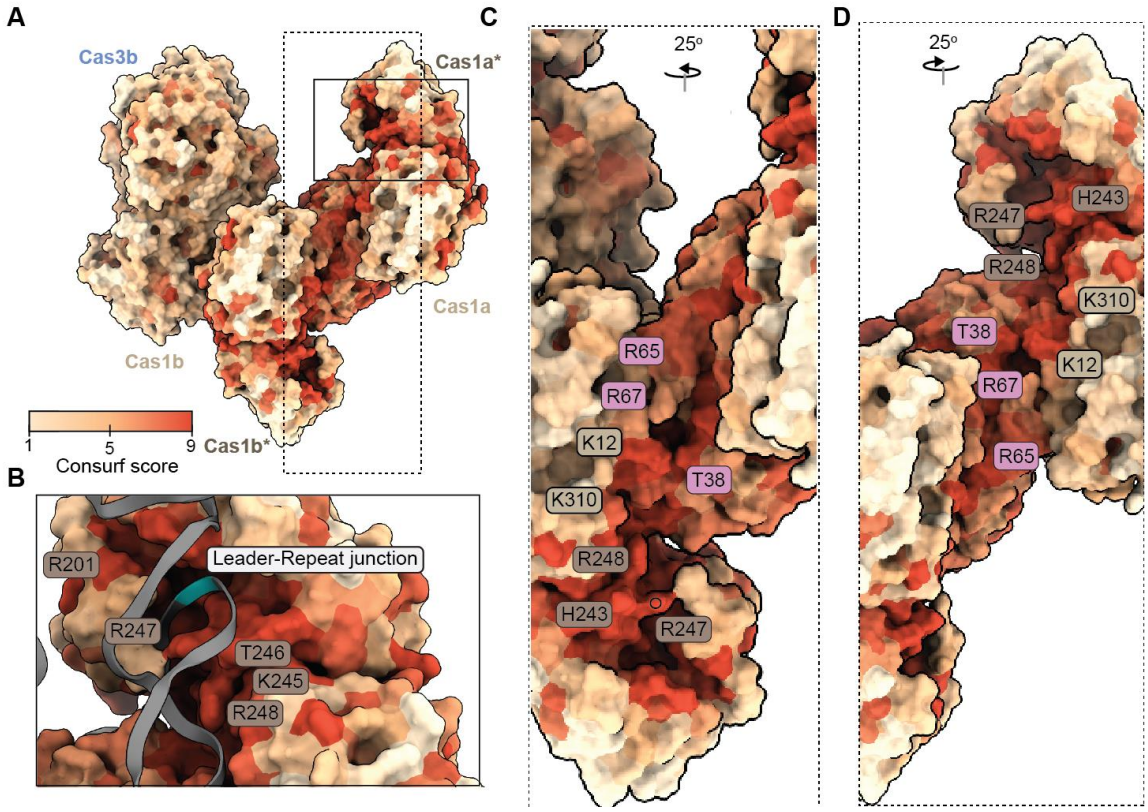

Supplementary Figure 8.7. Three interfaces clamp Cas3 in place, inhibiting both interference and integration interactions. A. Global view of the Cas1-2/3 complex, with the Cas2/3 interface responsible for blocking genome-interacting residues (R55) and locking the RecA1 domain in place shown for a single Cas2/3 subunit, all other subunits displayed at 80% transparency. B. Global view of the Cas1-2/3 complex, with the Cas1-Cas3 interface responsible for locking the RecA1 domain in place shown for a single Cas2/3 subunit, all other subunits displayed at 80% transparency. C. Global view of the Cas1-2/3 complex, with the Cas1-CTD interface responsible for locking the RecA1 domain in place shown for a single Cas1-Cas3 interaction, all other subunits displayed at 80% transparency. D. Salt bridge interactions between Cas2 (R55, K50) and Cas3 (D472, D473, N373, D520) reveal that the genome interaction residues are sequested in Cas3's RecA1 domain. E. Salt bridge interactions between the inactive Cas1 and the RecA1 domain of Cas3 hold Cas3 in place. F. Poorly resolved electrostatic interactions between the active Cas1 and the CTD pin Cas3 in place.




Supplementary Figure 8.8. Purification and data processing of the Cas1-2/3 capture complex. A, The Cas1-2/3 complex bound to a fragment of dsDNA elutes as an ~415 kDa complex in size exclusion chromatography. B. SDS-PAGE analysis confirms the main peak contains both Cas1 and Cas2/3 (* indicates fraction on which data was collected) C.
A UREA-PAGE denaturing gel confirms that the main peak also contains the fragment of dsDNA. D. Sample micrograph at a nominal magnification of 96000 E) Data processing summary E) Voluems from a 5-class Ab initio reconstruction and heterogenous refinement after preliminary particle picking. Thre grey volume was used subsequently as a template. F. Local refinement of 28,655 particles colored by local resolution at estimation (FSC 0.5) reveals most of the complex is resolved below 3 Angstroms, but the edges of the complex are low resolution.




Supplementary Figure 8.9. Purification and cryoEM data analysis of PAM-containing integration complex. A. Size exclusion chromatogram of the integration reaction after 10 minutes at 25ºC. B. 15% SDS PAGE gel of the main peak fractions reveals all of the protein components of the integration complex are present. Fraction 8 (green box) was frozen for cryoEM. C. 8% UREA-PAGE gel of main peak fractions reveals that all of the DNA components of a partial integration reaction are present in fraction 8. D. Sample micrograph from data collection. E. Data analysis workflow. F. Representative 2D class averages selected for further processing. G. 3-class Ab Initio and Heterogeneous refinement identifies a single integration complex class (black box) H. Consensus Non-Uniform refinement of particles from G. I. Focus mask used for 3D classification of particles from H. J. A class of particles contains the U-bend, but no loop. K. A smaller class of particles contains the loop and U-bend, with the repeat stretched partially through the repeat channel.




Supplementary Figure 8.10. Positively charged residues in the repeat channel are conserved in type I-F CRISPR systems. **A.** Consurf analysis of the integration complex with DNA removed to show the conserved residues of the repeat channel. B. K245, T246, R247, and R248 are conserved and interact with the B-form portion of the CRISPR repeat in the channel just after the leader-repeat junction. C,D Conserved basic and polar residues line the region of the repeat channel that distorts the repeat DNA.

Supplementary Tables

| **Dataset** | Cas1-2/3 Capture |
|---|---|
| Complex | Consensus |
| Source refinement | J411 Refinement 23 |
| **Data collection and processing** | |
| Magnification | 81000 |
| Voltage (kV) | 300 |
| Electron exposure (e-/A^2 | 52.51 |
| Defocus range (um) | 0.8-2.5 |
| Pixel size (A) | 1.0691 |
| Number of frames | 40 frames |
| Total/Accepted micrographs | 3944/3293 |
| Symmetry imposed | C2 |
| Initial particle images (no.) | 1716883 (Template picker) |
| Final particle images (no.) | 303,250 (J411) |
| Map resolution (A) | 3.3 |
| FSC Threshold | 0.143 |
| Map resolution range (A) | 2.8-5.1 |
| **Refinement** | |
| Initial model used (PDB code) | Alphafold 3 |
| Model resolution (A) | 3.3 |
| Fsc treshold | 0.143 |
| Map sharpening B factor (A2) | NA |
| Model composition | |
| Non-hydrogen atoms | 26088 |
| Protein residues | 3380 |
| Nucleotides | 0 |
| B factors (A^2) (min/max/mean) | |
| Protein | 41.05/262.82/132.76 |
| Nucleotides | NA |
| R.M.S deviations | |
| Bond lengths | 0.003 (0) |
| Bond angles | 0.517 (3) |
| Validation | |
| MolProbity Score | 1.46 |
| Clashscore | 6.59 |
| Poor rotamers | 0.12% |
| Ramachandran plot | |
| Favored | 97.53 |
| Allowed | 2.41 |
| Disallowed | 0.06 |

Supplementary Table 8.1. Refinement statistics for the Cas1-2/3 complex

| **Dataset** | Cas1-2/3 Delivery |
|---|---|
| Complex | Consensus |
| Source refinement | J1775 Refine4 |
| **Data collection and processing** | |
| Magnification | 96000 |
| Voltage (kV) | 300 |
| Electron exposure (e-/A^2 | 51.78 |
| Defocus range (um) | 0.9-2.8 |
| Pixel size (A) | 0.833 |
| Number of frames | 40 frames |
| Total/Accepted micrographs | 5403/5552 |
| Symmetry imposed | C1 |
| Initial particle images (no.) | 1,083,100 (crYolo) |
| Final particle images (no.) | 28,665 |
| Map resolution (A) | 3.49 |
| FSC Threshold | 0.143 |
| Map resolution range (A) | 1.77-41.78 |
| **Refinement** | |
| Initial model used (PDB code) | 8FLJ + Alphafold 2 (for Cas3) |
| Model resolution (A) | 3.5 |
| Fsc treshold | 0.143 |
| Map sharpening B factor (A2) | NA |
| Model composition | |
| Non-hydrogen atoms | 39,223 (19025 Hyrdogens) |
| Protein residues | 2426 |
| Nucleotides | 66 |
| B factors (A^2) (min/max/mean) | |
| Protein | 78.01/323.72/143.26 |
| Nucleotides | 88.99/771.23/213.09 |
| R.M.S deviations | |
| Bond lengths | 0.008 (7) |
| Bond angles | 0.999 (18) |
| Validation | |
| MolProbity Score | 1.48 |
| Clashscore | 8.93 |
| Poor rotamers | 0.16% |
| Ramachandran plot | |
| Favored | 98.47 |
| Allowed | 1.53 |
| Disallowed | 0 |

Supplementary Table 8.2. Refinement statistics for Cas1-2/3 bound to dsDNA fragment

# CHAPTER NINE

# CONCLUSION

Viruses reprogram cells across the domains of life. The genomic invasion of viral-like selfish genetic elements restructures genomes and provides raw genetic substrate for evolutionary innovation in the form of new genes and regulatory sequences. In addition to raw substrate, the molecular arms race between cellular hosts and viral parasites continuously spawns sophisticated defense systems to stop viral predation. Between gene domestication and the conflict-driven proliferation of defense systems, the struggle between virus and host has given rise to a significant amount of biological diversity. This diversity has been a gold mine for creative minds, who have repurposed restriction modification systems for molecular cloning[1], domesticated viral capsids for therapeutic delivery vehicles[2,3], and CRISPR systems for genome engineering[4], transcriptome engineering[5], molecular recording[6], and diagnostic assays[7–9].

While this list of conflict-based biotechnologies is not exhaustive, it highlights a key principle: to repurpose a molecular machine first requires mechanistic understanding of molecular function. Three recent technological advances have accelerated a mechanistic, atomic level understanding of the evolutionary outcomes of genetic conflict: accurate long-read sequencing, reliable structural prediction *in silico*, and rapid, high-resolution structure determination by cryogenic electron microscopy (cryoEM).

The development of long-read sequencing technology transformed metagenomics by enabling the assembly of full genomes from uncultured environmental samples[10]. Mining of this metagenomic data has already revealed a dizzying array of both viruses[11] and previously undescribed CRISPR subtypes[12], suggesting that current taxonomies of both viruses and viral

defense systems only touch the tip of the iceberg. Long-read sequencing has also transformed genomics by enabling the assembly of complete, telomere-to-telomere genomes[13,14] While repetitive regions of the genome historically presented a significant barrier to genome assembly, long-read sequencing enables read through of these long repetitive regions. This allows the accurate assembly of repetitive telomeric and centromeric regions, in addition to accurate localization of multi-copy genome parasites[13].

However, a bottleneck in the functional annotation of genomic and metagenomic data is the enormous diversity at the primary sequence level. This sequence diversity obscures the structural homology of evolutionarily related proteins. Many genes encode for proteins of a single evolutionary origin but low overall sequence homology, such as the PNMA family of genes and retroviral capsid genes[15,16]. However, the lack of sequence homology frustrated functional annotation attempts of both genomic and metagenomic data. Prior to 2021, the best strategy for predicting the function of a protein was through comparison to known protein structures on homology prediction servers like Phyre[17]. However, in 2021 there were only 100,000 protein structures in the Protein Data Bank, compared to over a billion protein sequences[18]. Alphafold, first released in 2021, addressed the desperate need for fast and accurate protein structure prediction, first predicting 365,000 protein structures, and subsequently predicting over 200 million structures from sequences found in UniProt[19,20]. Subsequent releases of Alphafold predict not only monomers, but multimers, and even multimers in the presence of nucleic acids and other non-protein ligands[21,22].

This remarkable development in the realm of structural prediction has paralleled an explosion in experimentally determined structures, largely due to the “Resolution Revolution” in

cryogenic electron microscopy (cryoEM)[23]. This year or next, cryoEM is likely to displace X-ray crystallography as the technique of choice in structural biology (as determined by the number of structures deposited annually to the PDB[24]). With the advent of direct electron detectors, sophisticated data-processing algorithms and GPU-accelerated data-processing pipelines, it is now possible to pull a plasmid from the freezer and obtain an atomic-resolution structure of the protein the plasmid encodes in days to weeks with standard protein purification and data collection workflows. And while Alphafold provides reliable monomer and multimer predictions, it is not yet capable of capturing a protein complex in distinct and biologically relevant conformational states. cryoEM, on the other hand, can be used to capture time-resolved structures and conformational changes triggered by ligand binding or disassociation.

## From Sequence to Structure

My doctoral work has leveraged long read sequencing combined with Alphafold structural predictions to update pre-existing gene annotations and create *de novo* functional predictions. We used long-read sequencing to sequence the genome of an algae with promise as a biofuel producer (Chapter 2) and Alphafold to enhance the functional prediction of the genome content. We annotated the selfish genetic elements within the algal genome, and future work will need to investigate preliminary unpublished results suggesting that stressful growth conditions activate algal retrotransposons. We have also leveraged the new Telomere-to-Telomere human genome to refute claims of human CRISPR systems (Chapter 3), and performed a comprehensive evolutionary analysis of ancient domesticated viral-like genes in placental mammals. We combined this evolutionary analysis with Alphafold structure predictions to determine that, in addition to viral capsid genes, an RNA-binding domain of viral origin is found in the PNMA

family of domesticated genes (Chapter 4). Knockout studies in mice of two of these genes, PNMA1 and PNMA4, demonstrate that they play a critical role in placental reproduction. Preliminary structural evidence suggests that these two proteins have retained their viral-like function of forming capsids while performing host functions in reproductive tissues (Chapter 5). Ongoing and future work will be needed to characterize the precise cargo contents and packaging mechanism for loading the cargo into these domesticated viral capsid genes, and there is exciting potential to develop these capsids into reproductive therapeutics. Taken together, the work presented in this dissertation highlights how Alphafold predictions can be combined with long-read sequencing data to generate functional predictions for proteins with no known function.

## Structure Determines Function

However, Alphafold cannot yet capture the conformational dynamics of a protein in solution with interaction partners, nor does it always accurately model nucleic acid folding against protein binding partners. cryoEM provides experimental validation of Alphafold-predicted structures and captures the conformational landscape of a protein complex in solution. We have used cryoEM to capture an ATP-driven bacterial immune system that recognizes viral triggers and subsequently releases a tRNA nuclease (Chapter 6). We have also used cryoEM to determine the structure of the CRISPR integration complex (Chapter 7). Follow-up work has revealed that CRISPR adaptation is not a single step reaction, but rather a series of structural checkpoints that license integration of foreign DNA fragments into the genome (Chapter 8).

However, a major barrier in rapid cryoEM structure determination remains the time required for expression and purification of a sample of interest. We have used *in vitro* transcription and translation to determine a high resolution structure of ferritin in under 24 hours,

significantly reducing the time required to determine an atomic structure by cryoEM (Appendix A). Ongoing work will attempt to optimize this technique for more complex biological samples.

## Concluding Remarks

This is an exciting time to be a structural biologist interested in genetic conflict. Recent advances in sequencing, structure prediction, and structure determination have opened the door to a vast and unexplored landscape of molecular machines. Many of these machines have been shaped by genetic conflict, either as host response to the threat of viral predation, or as viral genes repurposed for novel host function. My work has contributed to the fields of bacterial defense by revealing how an ATP-driven immune system assembles into a multimeric complex that detects diverse viral proteins and initiates abortive infection through tRNA cleavage and by demonstrating the specific and directional integration of foreign DNA into CRISPR arrays is driven by genomic folding, and structural checkpoints. My work has also contributed to the blossoming field of viral domestication by providing a deep evolutionary analysis of domesticated viral genes present in the human genome that are conserved across placental mammals. While this work advances several fields, it also opens up more questions than it answers. One persistent question in the CRISPR field remains: do other adaptation modules besides Cas1-Cas2 exist for integrating fragments of foreign DNA precisely into the genome? A related open question in eukaryotes is: what drives the genesis of piRNA clusters, which store fragments of actively replicating transposons? Another persistent question remains: what is the cellular and organismal function of domesticated metaviral genes in placental mammals? Despite decades of work, the cellular function of these genes, and the role they played in placental evolution remains elusive.

References


1. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences* **70**, 3240–3244 (1973).

2. Madigan, V. *et al.* Human paraneoplastic antigen Ma2 (PNMA2) forms icosahedral capsids that can be engineered for mRNA delivery. *Proceedings of the National Academy of Sciences* **121**, e2307812120 (2024).

3. Segel, M. *et al.* Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. *Science* **373**, 882–889 (2021).

4. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816–821 (2012).

5. Nemudraia, A., Nemudryi, A. & Wiedenheft, B. Repair of CRISPR-guided RNA breaks enables site-specific RNA excision in human cells. *Science* **384**, 808–814 (2024).

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

7. Broughton, J. P. *et al.* CRISPR–Cas12-based detection of SARS-CoV-2. *Nat Biotechnol* **38**, 870–874 (2020).

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

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

10. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* **37**, 937–944 (2019).

11. Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* **15**, 161–168 (2017).

12. Hu, Y. *et al.* Metagenomic discovery of novel CRISPR-Cas13 systems. *Cell Discov* **8**, 1–4 (2022).

13. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

14. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

15. Krupovic, M. & Koonin, E. V. Homologous Capsid Proteins Testify to the Common Ancestry of Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses. *Journal of Virology* **91**, e00210-17 (2017).

16. Schüller, M., Jenne, D. & Voltz, R. The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. *J. Neuroimmunol.* **169**, 172–176 (2005).

17. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845–858 (2015).

18. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* **48**, D570–D578 (2020).

19. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

20. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).

21. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13**, 1265 (2022).

22. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

23. Kühlbrandt, W. The Resolution Revolution. *Science* **343**, 1443–1444 (2014).

24. Bank, R. P. D. PDB Statistics: Number of Released PDB Structures per Year. https://www.rcsb.org/stats/all-released-structures.

# CUMULATIVE REFERENCES CITED

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105. https://doi.org/10.1093/bioinformatics/bti263

Abby, S.S., Néron, B., Ménager, H., Touchon, M., Rocha, E.P.C., 2014a. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLOS ONE 9, e110726. https://doi.org/10.1371/journal.pone.0110726

Abe, S., Tanaka, J., Kojima, M., Kanamaru, S., Hirata, K., Yamashita, K., Kobayashi, A., Ueno, T., 2022. Cell-free protein crystallization for nanocrystal structure determination. Sci Rep 12, 16031. https://doi.org/10.1038/s41598-022-19681-9

Abed, M., Verschueren, E., Budayeva, H., Liu, P., Kirkpatrick, D.S., Reja, R., Kummerfeld, S.K., Webster, J.D., Gierke, S., Reichelt, M., Anderson, K.R., Newman, R.J., Roose-Girma, M., Modrusan, Z., Pektas, H., Maltepe, E., Newton, K., Dixit, V.M., 2019. The Gag protein PEG10 binds to RNA and regulates trophoblast stem cell lineage specification. PLOS ONE 14, e0214110. https://doi.org/10.1371/journal.pone.0214110

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., Bodenstein, S.W., Evans, D.A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A.I., Cowie, A., Figurnov, M., Fuchs, F.B., Gladman, H., Jain, R., Khan, Y.A., Low, C.M.R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E.D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., Jumper, J.M., 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500. https://doi.org/10.1038/s41586-024-07487-w

Acton, O., Grant, T., Nicastro, G., Ball, N.J., Goldstone, D.C., Robertson, L.E., Sader, K., Nans, A., Ramos, A., Stoye, J.P., Taylor, I.A., Rosenthal, P.B., 2019. Structural basis for Fullerene geometry in a human endogenous retrovirus capsid. Nature Communications 10, 5822. https://doi.org/10.1038/s41467-019-13786-y

Aid, F., 2020. Plant Lipid Metabolism. https://doi.org/10.5772/intechopen.81355

Albers, S., Czech, A., 2016. Exploiting tRNAs to Boost Virulence. Life (Basel) 6, 4. https://doi.org/10.3390/life6010004

Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S., Ishihama, A., 1999. Growth Phase-Dependent Variation in Protein Composition of the *Escherichia coli* Nucleoid. J Bacteriol 181, 6361–6370. https://doi.org/10.1128/JB.181.20.6361-6370.1999

Alkhnbashi, O.S., Meier, T., Mitrofanov, A., Backofen, R., Voß, B., 2020. CRISPR-Cas bioinformatics. Methods 172, 3–11. https://doi.org/10.1016/j.ymeth.2019.07.013

Alonso, J.C., Weise, F., Rojo, F., 1995. The Bacillus subtilis Histone-like Protein Hbsu Is Required for DNA Resolution and DNA Inversion Mediated by the β Recombinase of

Plasmid pSM19035. Journal of Biological Chemistry 270, 2938–2945. https://doi.org/10.1074/jbc.270.7.2938

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Álvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Molecular Phylogenetics and Evolution. https://doi.org/10.1016/S1055-7903(03)00208-2

Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., Gouil, Q., 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biology. https://doi.org/10.1186/s13059-020-1935-5

Andersen, K.R., Leksa, N.C., Schwartz, T.U., 2013. Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. Proteins: Structure, Function, and Bioinformatics 81, 1857–1861. https://doi.org/10.1002/prot.24364

Anisimova, M., Nielsen, R., Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164, 1229–1236. https://doi.org/10.1093/genetics/164.3.1229

Antine, S.P., Johnson, A.G., Mooney, S.E., Leavitt, A., Mayer, M.L., Yirmiya, E., Amitai, G., Sorek, R., Kranzusch, P.J., 2024. Structural basis of Gabija anti-phage defence and viral immune evasion. Nature 625, 360–365. https://doi.org/10.1038/s41586-023-06855-2

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., Ogata, H., 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz859

Aravind, L., Leipe, D.D., Koonin, E.V., 1998. Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic Acids Research 26, 4205–4213. https://doi.org/10.1093/nar/26.18.4205

Arriola, M.B., Velmurugan, N., Zhang, Y., Plunkett, M.H., Hondzo, H., Barney, B.M., 2018. Genome sequences of Chlorella sorokiniana UTEX 1602 and Micractinium conductrix SAG 241.80: implications to maltose excretion by a green alga. The Plant Journal. https://doi.org/10.1111/tpj.13789

Ashkenazi, A., Faingold, O., Shai, Y., 2013. HIV-1 fusion protein exerts complex immunosuppressive effects. Trends in Biochemical Sciences 38, 345–349. https://doi.org/10.1016/j.tibs.2013.04.003

Ashley, J., Cordy, B., Lucia, D., Fradkin, L.G., Budnik, V., Thomson, T., 2018. Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. Cell 172, 262-274.e11. https://doi.org/10.1016/j.cell.2017.12.022

Avcilar-Kucukgoze, I., Gamper, H., Hou, Y.-M., Kashina, A., 2020. Purification and Use of tRNA for Enzymatic Post-translational Addition of Amino Acids to Proteins. STAR Protocols 1, 100207. https://doi.org/10.1016/j.xpro.2020.100207

Avidan, O., Pick, U., 2015. Acetyl-CoA synthetase is activated as part of the PDH-bypass in the oleaginous green algaChlorella desiccata. Journal of Experimental Botany. https://doi.org/10.1093/jxb/erv424

Azam, A.H., Chihara, K., Kondo, K., Nakamura, T., Ojima, S., Tamura, A., Yamashita, W., Cui, L., Takahashi, Y., Watashi, K., Kiga, K., 2023. Viruses encode tRNA and anti-retron to evade bacterial immunity. https://doi.org/10.1101/2023.03.15.532788

Aziz, R.K., Breitbart, M., Edwards, R.A., 2010. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Research 38, 4207–4217. https://doi.org/10.1093/nar/gkq140

Baik, M., Jeong, J.Y., Park, S.J., Yoo, S.P., Lee, J.Oh., Lee, J.S., Haque, M.N., Lee, H.-J., 2020. Testosterone deficiency caused by castration increases adiposity in male rats in a tissue-specific and diet-dependent manner. Genes & Nutrition 15, 14. https://doi.org/10.1186/s12263-020-00673-1

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., Talbot, R.T., Gustincich, S., Freeman, T.C., Mattick, J.S., Hume, D.A., Heutink, P., Carninci, P., Jeddeloh, J.A., Faulkner, G.J., 2011. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537. https://doi.org/10.1038/nature10531

Ball, N.J., Nicastro, G., Dutta, M., Pollard, D.J., Goldstone, D.C., Sanz-Ramos, M., Ramos, A., Müllers, E., Stirnnagel, K., Stanke, N., Lindemann, D., Stoye, J.P., Taylor, W.R., Rosenthal, P.B., Taylor, I.A., 2016. Structure of a Spumaretrovirus Gag Central Domain Reveals an Ancient Retroviral Capsid. PLOS Pathogens 12, e1005981. https://doi.org/10.1371/journal.ppat.1005981

Bank, R.P.D., n.d. PDB Statistics: Number of Released PDB Structures per Year [WWW Document]. URL https://www.rcsb.org/stats/all-released-structures (accessed 8.12.24).

Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., James, J.A., Salto-Tellez, M., Hamilton, P.W., 2017. QuPath: Open source software for digital pathology image analysis. Sci Rep 7, 16878. https://doi.org/10.1038/s41598-017-17204-5

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11. https://doi.org/10.1186/s13100-015-0041-9

Barton, A.R., Sherman, M.A., Mukamel, R.E., Loh, P.-R., 2021. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. Nat Genet 53, 1260–1269. https://doi.org/10.1038/s41588-021-00892-1

Bauer, D.E., Hatzivassiliou, G., Zhao, F., Andreadis, C., Thompson, C.B., 2005. ATP citrate lyase is an important component of cell growth and transformation. Oncogene. https://doi.org/10.1038/sj.onc.1208773

Béguin, P., Chekli, Y., Sezonov, G., Forterre, P., Krupovic, M., 2019. Sequence motifs recognized by the casposon integrase of Aciduliprofundum boonei. Nucleic acids research 47, 6386–6395. https://doi.org/10.1093/nar/gkz447

Behrenfeld, M.J., 2001. Biospheric Primary Production During an ENSO Transition. Science. https://doi.org/10.1126/science.1055071

Beijerinck, M.W., 1898. On a Contagium vivum fluidum causing the Spotdisease of the Tobacco-leaves. Koninklijke Nederlandse Akademie van Wetenschappen Proceedings Series B Physical Sciences 1, 170–176.

Bellou, S., Aggelis, G., 2013. Biochemical activities in Chlorella sp. and Nannochloropsis salina during lipid and sugar synthesis in a lab-scale open pond simulating reactor. Journal of Biotechnology. https://doi.org/10.1016/j.jbiotec.2013.01.010

Bénit, L., De Parseval, N., Casella, J.F., Callebaut, I., Cordonnier, A., Heidmann, T., 1997. Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J Virol 71, 5652–5657.

Berchowitz, L.E., Kabachinski, G., Walker, M.R., Carlile, T.M., Gilbert, W.V., Schwartz, T.U., Amon, A., 2015. Regulated Formation of an Amyloid-like Translational Repressor Governs Gametogenesis. Cell 163, 406–418. https://doi.org/10.1016/j.cell.2015.08.060

Bernheim, A., Sorek, R., 2020. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol 18, 113–119. https://doi.org/10.1038/s41579-019-0278-2

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A.H.Q., Kumar, M.S., Li, C., Dvornicic, M., Soldo, J.P., Koh, J.Y., Tong, C., Ng, O.T., Barkham, T., Young, B., Marimuthu, K., Chng, K.R., Sikic, M., Nagarajan, N., 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol 37, 937–944. https://doi.org/10.1038/s41587-019-0191-2

Bingham, R., Ekunwe, S.I.N., Falk, S., Snyder, L., Kleanthous, C., 2000. The Major Head Protein of Bacteriophage T4 Binds Specifically to Elongation Factor Tu*. Journal of Biological Chemistry 275, 23219–23226. https://doi.org/10.1074/jbc.M002546200

Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C., Brown, C.M., 2016a. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics 17, 356. https://doi.org/10.1186/s12864-016-2627-0

Blanc-Mathieu, R., Verhelst, B., Derelle, E., Rombauts, S., Bouget, F.-Y., Carré, I., Château, A., Eyre-Walker, A., Grimsley, N., Moreau, H., Piégu, B., Rivals, E., Schackwitz, W., Van De Peer, Y., Piganeau, G., 2014. An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. BMC Genomics. https://doi.org/10.1186/1471-2164-15-1103

Blikstad, V., Benachenhou, F., Sperber, G.O., Blomberg, J., 2008. Endogenous retroviruses. Cell. Mol. Life Sci. 65, 3348–3365. https://doi.org/10.1007/s00018-008-8495-2

Boissinot, S., Davis, J., Entezam, A., Petrov, D., Furano, A.V., 2006. Fitness cost of LINE-1 (L1) activity in humans. Proceedings of the National Academy of Sciences 103, 9590–9594. https://doi.org/10.1073/pnas.0603334103

Bold, H.C., 1949. The Morphology of Chlamydomonas chlamydogama, Sp. Nov. Bulletin of the Torrey Botanical Club. https://doi.org/10.2307/2482218

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

Bou-Nader, C., Muecksch, F., Brown, J.B., Gordon, J.M., York, A., Peng, C., Ghirlando, R., Summers, M.F., Bieniasz, P.D., Zhang, J., 2021. HIV-1 matrix-tRNA complex structure reveals basis for host control of Gag localization. Cell Host & Microbe 29, 1421-1436.e7. https://doi.org/10.1016/j.chom.2021.07.006

Brandt, J., Schrauth, S., Veith, A.-M., Froschauer, A., Haneke, T., Schultheis, C., Gessler, M., Leimeister, C., Volff, J.-N., 2005a. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345, 101–111. https://doi.org/10.1016/j.gene.2004.11.022

Brandt, J., Veith, A.-M., Volff, J.-N., 2005b. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. Cytogenet. Genome Res. 110, 307–317. https://doi.org/10.1159/000084963

Brazelton, W.J., Baross, J.A., 2009. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. The ISME Journal 3, 1420–1424. https://doi.org/10.1038/ismej.2009.79

Brookfield, J.F.Y., 2005. The ecology of the genome — mobile DNA elements and their hosts. Nat Rev Genet 6, 128–136. https://doi.org/10.1038/nrg1524

Brosius, J., 1991. Retroposons—Seeds of Evolution. Science 251, 753–753. https://doi.org/10.1126/science.1990437

Broughton, J.P., Deng, X., Yu, G., Fasching, C.L., Servellita, V., Singh, J., Miao, X., Streithorst, J.A., Granados, A., Sotomayor-Gonzalez, A., Zorn, K., Gopez, A., Hsu, E., Gu, W., Miller, S., Pan, C.-Y., Guevara, H., Wadford, D.A., Chen, J.S., Chiu, C.Y., 2020. CRISPR–Cas12-based detection of SARS-CoV-2. Nat Biotechnol 38, 870–874. https://doi.org/10.1038/s41587-020-0513-4

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., van der Oost, J., 2008. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science 321, 960–964. https://doi.org/10.1126/science.1159689

Brown, M.W., Dillard, K.E., Xiao, Y., Dolan, A.E., Hernandez, E.T., Dahlhauser, S., Kim, Y., Myler, L.R., Anslyn, E., Ke, A., Finkelstein, I., 2017. Assembly and translocation of a CRISPR-Cas primed acquisition complex. bioRxiv 41, 1–11. https://doi.org/10.1101/208058

Bryant, P., Pozzati, G., Elofsson, A., 2022. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun 13, 1265. https://doi.org/10.1038/s41467-022-28865-w

Bulletti, C., Bulletti, F.M., Sciorio, R., Guido, M., 2022. Progesterone: The Key Factor of the Beginning of Life. International Journal of Molecular Sciences 23, 14138. https://doi.org/10.3390/ijms232214138

Burns, K.H., 2020. Our Conflict with Transposable Elements and Its Implications for Human Disease. Annu Rev Pathol 15, 51–70. https://doi.org/10.1146/annurev-pathmechdis-012419-032633

Burns, K.H., 2017. Transposable elements in cancer. Nat Rev Cancer 17, 415–424. https://doi.org/10.1038/nrc.2017.35

Burns, K.H., Boeke, J.D., 2012. Human Transposon Tectonics. Cell 149, 740–752. https://doi.org/10.1016/j.cell.2012.04.019

Butler, M., Goodwin, T., Simpson, M., Singh, M., Poulter, R., 2001. Vertebrate LTR Retrotransposons of the Tf1/Sushi Group. J Mol Evol 52, 260–274. https://doi.org/10.1007/s002390010154

Cael, B.B., Heathcote, A.J., Seekell, D.A., 2017. The volume and mean depth of Earth's lakes. Geophysical Research Letters. https://doi.org/10.1002/2016gl071378

Campbell, M.S., Holt, C., Moore, B., Yandell, M., 2014. Genome Annotation and Curation Using MAKER and MAKER-P. Current Protocols in Bioinformatics. https://doi.org/10.1002/0471250953.bi0411s48

Campillos, M., Doerks, T., Shah, P.K., Bork, P., 2006. Computational characterization of multiple Gag-like human proteins. Trends in Genetics 22, 585–589. https://doi.org/10.1016/j.tig.2006.09.006

Campodonico, W., Black, H.H., Lau, C.I., Whiteley, A.M., 2023. The gag-like gene RTL8 antagonizes PEG10-mediated virus like particles in humans. https://doi.org/10.1101/2023.02.03.527044

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Müller, K.M., Pande, N., Shang, Z., Yu, N., Gutell, R.R., 2002. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-3-2

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J., 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msab293

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., Yandell, M., 2007. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research. https://doi.org/10.1101/gr.6743907

Cao, Q., Padmanabhan, K., Richter, J.D., 2009. Pumilio 2 controls translation by competing with eIF4E for 7-methyl guanosine cap recognition. RNA. https://doi.org/10.1261/rna.1884610

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carmichael, R.E., Henley, J.M., 2018. Transcriptional and post-translational regulation of Arc in synaptic plasticity. Semin Cell Dev Biol 77, 3–9. https://doi.org/10.1016/j.semcdb.2017.09.007

Cavazzana-Calvo, M., Hacein-Bey, S., Basile, G. de S., Gross, F., Yvon, E., Nusbaum, P., Selz, F., Hue, C., Certain, S., Casanova, J.-L., Bousso, P., Deist, F.L., Fischer, A., 2000. Gene Therapy of Human Severe Combined Immunodeficiency (SCID)-X1 Disease. Science 288, 669–672. https://doi.org/10.1126/science.288.5466.669

CeS, C., 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282, 2012–8. https://doi.org/10.1126/science.282.5396.2012

Chalmers, R., Guhathakurta, A., Benjamin, H., Kleckner, N., 1998. IHF Modulation of Tn10 Transposition: Sensory Transduction of Supercoiling Status via a Proposed Protein/DNA Molecular Spring. Cell 93, 897–908. https://doi.org/10.1016/S0092-8674(00)81449-X

Charif, D., Lobry, J.R., 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. https://doi.org/10.1007/978-3-540-35306-5_10

Charlier, C., Segers, K., Wagenaar, D., Karim, L., Berghmans, S., Jaillon, O., Shay, T., Weissenbach, J., Cockett, N., Gyapay, G., Georges, M., 2001. Human–Ovine Comparative Sequencing of a 250-kb Imprinted Domain Encompassing the Callipyge (clpg) Locus and Identification of Six Imprinted Transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. Genome Res 11, 850–862. https://doi.org/10.1101/gr.172701

Cheng, R., Huang, F., Wu, H., Lu, X., Yan, Y., Yu, B., Wang, X., Zhu, B., 2021. A nucleotide-sensing endonuclease from the Gabija bacterial defense system. Nucleic Acids Research 49, 5216–5229. https://doi.org/10.1093/nar/gkab277

Cheng, Y., 2018. Single-particle cryo-EM—How did it get here and where will it go. Science 361, 876–880. https://doi.org/10.1126/science.aat4346

Chevallereau, A., Westra, E.R., 2022. Bacterial immunity: Mobile genetic elements are hotspots for defence systems. Current Biology 32, R923–R926. https://doi.org/10.1016/j.cub.2022.07.075

Cho, G., Lim, Y., Zand, D., Golden, J.A., 2008. Sizn1 Is a Novel Protein That Functions as a Transcriptional Coactivator of Bone Morphogenic Protein Signaling. Molecular and Cellular Biology 28, 1565–1572. https://doi.org/10.1128/MCB.01038-07

Chou, M.-Y., Hu, M.-C., Chen, P.-Y., Hsu, C.-L., Lin, T.-Y., Tan, M.-J., Lee, C.-Y., Kuo, M.-F., Huang, P.-H., Wu, V.-C., Yang, S.-H., Fan, P.-C., Huang, H.-Y., Akbarian, S., Loo, T.-H., Stewart, C.L., Huang, H.-P., Gau, S.S.-F., Huang, H.-S., 2022. RTL1/PEG11 imprinted in human and mouse brain mediates anxiety-like and social behaviors and regulates neuronal excitability in the locus coeruleus. Human Molecular Genetics 31, 3161–3180. https://doi.org/10.1093/hmg/ddac110

Clark, M.B., Jänicke, M., Gottesbühren, U., Kleffmann, T., Legge, M., Poole, E.S., Tate, W.P., 2007. Mammalian gene PEG10 expresses two reading frames by high efficiency -1 frameshifting in embryonic-associated tissues. J. Biol. Chem. 282, 37359–37369. https://doi.org/10.1074/jbc.M705676200

Clarke, H.J., 2018. Regulation of germ cell development by intercellular signaling in the mammalian ovarian follicle. WIREs Developmental Biology 7, e294. https://doi.org/10.1002/wdev.294

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cockram, C., Thierry, A., Koszul, R., 2021. Generation of gene-level resolution chromosome contact maps in bacteria and archaea. STAR Protocols 2, 100512. https://doi.org/10.1016/j.xpro.2021.100512

Cohen, S.N., Chang, A.C.Y., Boyer, H.W., Helling, R.B., 1973. Construction of Biologically Functional Bacterial Plasmids In Vitro. Proceedings of the National Academy of Sciences 70, 3240–3244. https://doi.org/10.1073/pnas.70.11.3240

Conine, C.C., Rando, O.J., 2022. Soma-to-germline RNA communication. Nat Rev Genet 23, 73–88. https://doi.org/10.1038/s41576-021-00412-1

Cordaux, R., Batzer, M.A., 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet 10, 691–703. https://doi.org/10.1038/nrg2640

Cornelis, G., Funk, M., Vernochet, C., Leal, F., Tarazona, O.A., Meurice, G., Heidmann, O., Dupressoir, A., Miralles, A., Ramirez-Pinilla, M.P., Heidmann, T., 2017. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. Proceedings of the National Academy of Sciences 114, E10991–E11000. https://doi.org/10.1073/pnas.1714590114

Coticchio, G., Dal Canto, M., Mignini Renzini, M., Guglielmo, M.C., Brambillasca, F., Turchi, D., Novara, P.V., Fadini, R., 2015. Oocyte maturation: gamete-somatic cells interactions, meiotic resumption, cytoskeletal dynamics and cytoplasmic reorganization. Human Reproduction Update 21, 427–454. https://doi.org/10.1093/humupd/dmv011

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., Mozziconacci, J., 2012. Normalization of a chromosomal contact map. BMC Genomics 13, 436. https://doi.org/10.1186/1471-2164-13-436

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., Pourcel, C., 2018b. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research 46, W246–W251. https://doi.org/10.1093/nar/gky425

Crichton, J.H., Dunican, D.S., MacLennan, M., Meehan, R.R., Adams, I.R., 2013. Defending the genome from the enemy within: mechanisms of retrotransposon suppression in the mouse germline. Cell Mol Life Sci 71, 1581–1605. https://doi.org/10.1007/s00018-013-1468-0

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. https://doi.org/10.1107/S2059798318002425

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome research 14, 1188–90. https://doi.org/10.1101/gr.849004

Culbertson, E.M., Levin, T.C., 2023. Eukaryotic CD-NTase, STING, and viperin proteins evolved via domain shuffling, horizontal transfer, and ancient inheritance from prokaryotes. PLOS Biology 21, e3002436. https://doi.org/10.1371/journal.pbio.3002436

d'Herelle, F., 1917. Sur un microbe invisible antagoniste des bacilles dysentétriques. Comptes rendus de l'Académie des sciences 165, 373–375.

Dalmau, J., Gultekin, S.H., Voltz, R., Hoard, R., DesChamps, T., Balmaceda, C., Batchelor, T., Gerstner, E., Eichen, J., Frennier, J., Posner, J.B., Rosenfeld, M.R., 1999. Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. Brain 122, 27–39. https://doi.org/10.1093/brain/122.1.27

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience. https://doi.org/10.1093/gigascience/giab008

Danilovich, N., Babu, P.S., Xing, W., Gerdes, M., Krishnamurthy, H., Sairam, M.R., 2000. Estrogen Deficiency, Obesity, and Skeletal Abnormalities in Follicle-Stimulating Hormone Receptor Knockout (FORKO) Female Mice**This investigation was supported in part by the Canadian Institutes of Health Research. Endocrinology 141, 4295–4308. https://doi.org/10.1210/endo.141.11.7765

Datsenko, K.A., Pougach, K., Tikhonov, A., Wanner, B.L., Severinov, K., Semenova, E., 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun 3, 945. https://doi.org/10.1038/ncomms1937

Daugherty, M.D., Schaller, A.M., Geballe, A.P., Malik, H.S., 2016. Evolution-guided functional analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. eLife 5, e14228. https://doi.org/10.7554/eLife.14228

Dawkins, R., 1976. The Selfish Gene. Oxford University Press.

De Bigault Du Granrut, A., Cacas, J.-L., 2016. How Very-Long-Chain Fatty Acids Could Signal Stressful Conditions in Plants? Frontiers in plant science. https://doi.org/10.3389/fpls.2016.01490

de Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., Pollock, D.D., 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLOS Genetics 7, e1002384. https://doi.org/10.1371/journal.pgen.1002384

de Mattos, K., Pierre, K.J., Tremblay, J.J., 2023. Hormones and Signaling Pathways Involved in the Stimulation of Leydig Cell Steroidogenesis. Endocrines 4, 573–594. https://doi.org/10.3390/endocrines4030041

De Rosier, D.J., Klug, A., 1968. Reconstruction of Three Dimensional Structures from Electron Micrographs. Nature 217, 130–134. https://doi.org/10.1038/217130a0

Deamer, D., Akeson, M., Branton, D., 2016. Three decades of nanopore sequencing. Nature Biotechnology. https://doi.org/10.1038/nbt.3423

Deep, A., Gu, Y., Gao, Y.-Q., Ego, K.M., Herzik, M.A., Zhou, H., Corbett, K.D., 2022. The SMC-family Wadjet complex protects bacteria from plasmid transformation by recognition and cleavage of closed-circular DNA. Molecular Cell 82, 4145-4159.e7. https://doi.org/10.1016/j.molcel.2022.09.008

Deininger, P.L., Batzer, M.A., 2002. Mammalian Retroelements. Genome Res. 12, 1455–1465. https://doi.org/10.1101/gr.282402

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics. https://doi.org/10.1038/nrg1603

Deo, V.K., Kato, T., Park, E.Y., 2015. Chimeric Virus-Like Particles Made Using GAG and M1 Capsid Proteins Providing Dual Drug Delivery and Vaccination Platform. Mol. Pharmaceutics 12, 839–845. https://doi.org/10.1021/mp500860x

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., Moineau, S., 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. Journal of Bacteriology 190, 1390–1400. https://doi.org/10.1128/JB.01412-07

Devine, M.F., Kothapalli, N., Elkhooly, M., Dubey, D., 2021. Paraneoplastic neurological syndromes: clinical presentations and management. Ther Adv Neurol Disord 14, 1756286420985323. https://doi.org/10.1177/1756286420985323

Dhingra, Y., Suresh, S.K., Juneja, P., Sashital, D.G., 2022. PAM binding ensures orientational integration during Cas4-Cas1-Cas2-mediated CRISPR adaptation. Molecular Cell 82, 4353-4367.e6. https://doi.org/10.1016/j.molcel.2022.09.030

Dimitriu, T., Szczelkun, M.D., Westra, E.R., 2020. Evolutionary Ecology and Interplay of Prokaryotic Innate and Adaptive Immune Systems. Current Biology 30, R1189–R1202. https://doi.org/10.1016/j.cub.2020.08.028

Dodonova, S.O., Prinz, S., Bilanchone, V., Sandmeyer, S., Briggs, J.A.G., 2019. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. PNAS 116, 10048–10057. https://doi.org/10.1073/pnas.1900931116

Dohm, J.C., Peters, P., Stralis-Pavese, N., Himmelbauer, H., 2020. Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics. https://doi.org/10.1093/nargab/lqaa037

Doolittle, R.F., Johnson, M.S., McClure, M.A., 1989. Origins and Evolutionary Relationships of Retroviruses. The Quarterly Review of Biology 64, 1–30. https://doi.org/10.1086/416128

Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601–603. https://doi.org/10.1038/284601a0

Dopkins, N., Nixon, D.F., 2024. Activation of human endogenous retroviruses and its physiological consequences. Nat Rev Mol Cell Biol 25, 212–222. https://doi.org/10.1038/s41580-023-00674-z

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., Sorek, R., 2018. Systematic discovery of antiphage defense systems in the microbial pangenome. Science 359, eaar4120. https://doi.org/10.1126/science.aar4120

Dot, E.W., Thomason, L.C., Chappie, J.S., 2023. Everything OLD is new again: How structural, functional, and bioinformatic advances have redefined a neglected nuclease family. Molecular Microbiology 120, 122–140. https://doi.org/10.1111/mmi.15074

Drost, H.-G., Gabel, A., Grosse, I., Quint, M., 2015. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msv012

Dupressoir, A., Lavialle, C., Heidmann, T., 2012. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta 33, 663–671. https://doi.org/10.1016/j.placenta.2012.05.005

Durnaoglu, S., Lee, S.-K., Ahnn, J., 2021. Human Endogenous Retroviruses as Gene Expression Regulators: Insights from Animal Models into Human Diseases. Mol Cells 44, 861–878. https://doi.org/10.14348/molcells.2021.5016

Eddy, S.R., 2015. Homology searches for structural RNAs: from proof of principle to practical use. RNA 21, 605–607. https://doi.org/10.1261/rna.050484.115

Eddy, S.R., 2011. Accelerated Profile HMM Searches. PLOS Computational Biology 7, e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Edwards, C.A., Mungall, A.J., Matthews, L., Ryder, E., Gray, D.J., Pask, A.J., Shaw, G., Graves, J.A.M., Rogers, J., SAVOIR consortium, Dunham, I., Renfree, M.B., Ferguson-Smith, A.C., 2008. The evolution of the DLK1-DIO3 imprinted domain in mammals. PLoS Biol 6, e135. https://doi.org/10.1371/journal.pbio.0060135

Elde, N.C., Child, S.J., Geballe, A.P., Malik, H.S., 2009. Protein kinase R reveals an evolutionary model for defeating viral mimicry. Nature 457, 485–489. https://doi.org/10.1038/nature07529

Elde, N.C., Malik, H.S., 2009. The evolutionary conundrum of pathogen mimicry. Nature Reviews Microbiology 7, 787–797. https://doi.org/10.1038/nrmicro2222

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., Finn, R.D., 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47, D427–D432. https://doi.org/10.1093/nar/gky995

Elguindy, M.M., Mendell, J.T., 2021. NORAD-induced Pumilio phase separation is required for genome stability. Nature 595, 303–308. https://doi.org/10.1038/s41586-021-03633-w

Ellinghaus, D., Kurtz, S., Willhoeft, U., 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 18. https://doi.org/10.1186/1471-2105-9-18

Emerson, R.O., Thomas, J.H., 2011. Gypsy and the birth of the SCAN domain. J Virol 85, 12043–12052. https://doi.org/10.1128/JVI.00867-11

Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., 2010a. Features and development of Coot. Acta Cryst D 66, 486–501. https://doi.org/10.1107/S0907444910007493

Epstein, I., Finkbeiner, S., 2018. The Arc of cognition: Signaling cascades regulating Arc and implications for cognitive function and disease. Semin Cell Dev Biol 77, 63–72. https://doi.org/10.1016/j.semcdb.2017.09.023

Eriksen, M.S., Nikolaienko, O., Hallin, E.I., Grødem, S., Bustad, H.J., Flydal, M.I., Merski, I., Hosokawa, T., Lascu, D., Akerkar, S., Cuéllar, J., Chambers, J.J., O'Connell, R., Muruganandam, G., Loris, R., Touma, C., Kanhema, T., Hayashi, Y., Stratton, M.M., Valpuesta, J.M., Kursula, P., Martinez, A., Bramham, C.R., 2021. Arc self-association and formation of virus-like capsids are mediated by an N-terminal helical coil motif. The FEBS Journal 288, 2930–2955. https://doi.org/10.1111/febs.15618

Erlendsson, S., Morado, D.R., Cullen, H.B., Feschotte, C., Shepherd, J.D., Briggs, J.A.G., 2020. Structures of virus-like capsids formed by the Drosophila neuronal Arc proteins. Nat Neurosci 1–4. https://doi.org/10.1038/s41593-019-0569-y

Ernits, K., Saha, C.K., Brodiazhenko, T., Chouhan, B., Shenoy, A., Buttress, J.A., Duque-Pedraza, J.J., Bojar, V., Nakamoto, J.A., Kurata, T., Egorov, A.A., Shyrokova, L., Johansson, M.J.O., Mets, T., Rustamova, A., Džigurski, J., Tenson, T., Garcia-Pino, A., Strahl, H., Elofsson, A., Hauryliuk, V., Atkinson, G.C., 2023. The structural basis of hyperpromiscuity in a core combinatorial network of type II toxin–antitoxin and related

phage defense systems. Proceedings of the National Academy of Sciences 120, e2305393120. https://doi.org/10.1073/pnas.2305393120

Evans, J.H., 1958. The Survival of Freshwater Algae During Dry Periods: Part I. An Investigation of the Algae of Five Small Ponds. The Journal of Ecology. https://doi.org/10.2307/2256910

Evans, K.M., Wortley, A.H., Mann, D.G., 2007. An Assessment of Potential Diatom "Barcode" Genes (cox1, rbcL, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in Sellaphora (Bacillariophyta). Protist. https://doi.org/10.1016/j.protis.2007.04.001

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., Hassabis, D., 2021. Protein complex prediction with AlphaFold-Multimer. https://doi.org/10.1101/2021.10.04.463034

Fagerlund, Robert D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., Krause, K.L., Bostina, M., Scheltema, R.A., Staals, R.H.J., Fineran, P.C., 2017. Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proceedings of the National Academy of Sciences 114, E5122–E5128. https://doi.org/10.1073/pnas.1618421114

Fan, J., Ning, K., Zeng, X., Luo, Y., Wang, D., Hu, J., Li, J., Xu, H., Huang, J., Wan, M., Wang, W., Zhang, D., Shen, G., Run, C., Liao, J., Fang, L., Huang, S., Jing, X., Su, X., Wang, A., Bai, L., Hu, Z.M., Xu, J., Li, Y., 2015. Genomic Foundation of Starch to Lipid Switch in Oleaginous Chlorella. Plant Physiology. https://doi.org/10.1104/pp.15.01174

Fan, X., Wang, J., Zhang, X., Yang, Z., Zhang, J.-C., Zhao, L., Peng, H.-L., Lei, J., Wang, H.-W., 2019. Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution. Nat Commun 10, 2386. https://doi.org/10.1038/s41467-019-10368-w

Field, C.B., 1998. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. Science. https://doi.org/10.1126/science.281.5374.237

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., Ysebaert, M., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507. https://doi.org/10.1038/260500a0

Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Kuñne, T., Boekhorst, J., Van Hijum, S.A.F.T., Staals, R.H.J., Brouns, S.J.J., 2014. Degenerate target sites mediate rapid primed CRISPR adaptation. Proceedings of the National Academy of Sciences of the United States of America 111. https://doi.org/10.1073/pnas.1400071111

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39, W29–W37. https://doi.org/10.1093/nar/gkr367

Finnegan, D.J., 2012. Retrotransposons. Curr. Biol. 22, R432-437. https://doi.org/10.1016/j.cub.2012.04.025

Flajnik, M.F., Kasahara, M., 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat Rev Genet 11, 47–59. https://doi.org/10.1038/nrg2703

Flamand, M.C., Goblet, J.P., Duc, G., Briquet, M., Boutry, M., 1992. Sequence and transcription analysis of mitochondrial plasmids isolated from cytoplasmic male-sterile lines of Vicia faba. Plant Mol Biol 19, 913–923. https://doi.org/10.1007/BF00040524

FlowJo Software, 2023.

Foley, C.J., Freedman, H., Choo, S.L., Onyskiw, C., Fu, N.Y., Yu, V.C., Tuszynski, J., Pratt, J.C., Baksh, S., 2008. Dynamics of RASSF1A/MOAP-1 association with death receptors. Mol Cell Biol 28, 4520–4535. https://doi.org/10.1128/MCB.02011-07

Frank, J.A., Singh, M., Cullen, H.B., Kirou, R.A., Benkaddour-Boumzaouad, M., Cortes, J.L., Garcia Pérez, J., Coyne, C.B., Feschotte, C., 2022. Evolution and antiviral activity of a human protein of retroviral origin. Science 378, 422–428. https://doi.org/10.1126/science.abq7871

Fu, N.Y., Sukumaran, S.K., Yu, V.C., 2007. Inhibition of ubiquitin-mediated degradation of MOAP-1 by apoptotic stimuli promotes Bax function in mitochondria. Proceedings of the National Academy of Sciences 104, 10051–10056. https://doi.org/10.1073/pnas.0700007104

Fueyo, R., Judd, J., Feschotte, C., Wysocka, J., 2022. Roles of transposable elements in the regulation of mammalian transcription. Nat Rev Mol Cell Biol 23, 481–497. https://doi.org/10.1038/s41580-022-00457-y

Fujioka, Y., Shiura, H., Ishii, M., Ono, R., Endo, T., Kiyonari, H., Hirate, Y., Ito, H., Kanai-Azuma, M., Kohda, T., Kaneko-Ishino, T., Ishino, F., 2023. Targeting retrovirus-derived Rtl8a and 8b causes late onset obesity and neurodevelopmental defects. https://doi.org/10.1101/2023.05.28.542606

Fulnečková, J., Hasíková, T., Fajkus, J., Lukešová, A., Eliáš, M., Sýkorová, E., 2012. Dynamic Evolution of Telomeric Sequences in the Green Algal Order Chlamydomonadales. Genome Biology and Evolution. https://doi.org/10.1093/gbe/evs007

Gallagher, J.R., Kim, A.J., Gulati, N.M., Harris, A.K., 2019. Negative-Stain Transmission Electron Microscopy of Molecular Complexes for Image Analysis by 2D Class Averaging. Curr Protoc Microbiol 54, e90. https://doi.org/10.1002/cpmc.90

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., Burdett, T., Hayhurst, J., Baker, J., Ferrer, J., Gonzalez-Uriarte, A., Jupp, S., Karim, M.A., Koscielny, G., Machlitt-Northen, S., Malangone, C., Pendlington, Z.M., Roncaglia, P., Suveges, D., Wright, D., Vrousgou, O., Papa, E., Parkinson, H., MacArthur, J.A.L., Todd, J.A., Barrett, J.C., Schwartzentruber, J., Hulcoop, D.G., Ochoa, D., McDonagh, E.M., Dunham, I., 2021. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res 49, D1311–D1320. https://doi.org/10.1093/nar/gkaa840

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6, 343–345. https://doi.org/10.1038/nmeth.1318

Gifford, R.J., Blomberg, J., Coffin, J.M., Fan, H., Heidmann, T., Mayer, J., Stoye, J., Tristem, M., Johnson, W.E., 2018. Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology 15, 59. https://doi.org/10.1186/s12977-018-0442-1

Glukhov, A.S., Krutilina, A.I., Kaliman, A.V., Shlyapnikov, M.G., Ksenzenko, V.N., 2018. Bacteriophage T5 Mutants Carrying Deletions in tRNA Gene Region. Mol Biol 52, 1–6. https://doi.org/10.1134/S0026893318010065

Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., Ferrin, T.E., 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Science 27, 14–25. https://doi.org/10.1002/pro.3235

Goemann, C.L.C., Wilkinson, R., Henriques, W., Bui, H., Goemann, H.M., Carlson, R.P., Viamajala, S., Gerlach, R., Wiedenheft, B., 2023. Genome sequence, phylogenetic analysis, and structure-based annotation reveal metabolic potential of Chlorella sp. SLA-04. Algal Research 69, 102943. https://doi.org/10.1016/j.algal.2022.102943

Goemann, H.M., Gay, J.D., Mueller, R.C., Brookshire, E.N.J., Miller, P., Poulter, B., Peyton, B.M., 2021. Aboveground and belowground responses to cyanobacterial biofertilizer supplement in a semi-arid, perennial bioenergy cropping system. Global Change Biology. Bioenergy. https://doi.org/10.1111/gcbb.12892

Goncalves, E.C., Wilkie, A.C., Kirst, M., Rathinasabapathi, B., 2016. Metabolic regulation of triacylglycerol accumulation in the green algae: identification of potential targets for engineering to improve oil yield. Plant Biotechnology Journal. https://doi.org/10.1111/pbi.12523

Gong, S., Gaccioli, F., Dopierala, J., Sovio, U., Cook, E., Volders, P.-J., Martens, L., Kirk, P.D.W., Richardson, S., Smith, G.C.S., Charnock-Jones, D.S., 2021. The RNA landscape of the human placenta in health and disease. Nat Commun 12, 2639. https://doi.org/10.1038/s41467-021-22695-y

Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., Myhrvold, C., Bhattacharyya, R.P., Livny, J., Regev, A., Koonin, E.V., Hung, D.T., Sabeti, P.C., Collins, J.J., Zhang, F., 2017. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science 356, 438–442. https://doi.org/10.1126/science.aam9321

Goren, M.G., Doron, S., Globus, R., Amitai, G., Sorek, R., Qimron, U., 2016. Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Reports 16, 2811–2818. https://doi.org/10.1016/j.celrep.2016.08.043

Gouveia-Oliveira, R., Sackett, P.W., Pedersen, A.G., 2007. MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics 8, 312. https://doi.org/10.1186/1471-2105-8-312

Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Greenbaum, M.P., Iwamori, N., Agno, J.E., Matzuk, M.M., 2009. Mouse TEX14 Is Required for Embryonic Germ Cell Intercellular Bridges but Not Female Fertility1. Biology of Reproduction 80, 449–457. https://doi.org/10.1095/biolreprod.108.070649

Greenbaum, M.P., Iwamori, T., Buchold, G.M., Matzuk, M.M., 2011. Germ Cell Intercellular Bridges. Cold Spring Harb Perspect Biol 3, a005850. https://doi.org/10.1101/cshperspect.a005850

Greenbaum, M.P., Yan, W., Wu, M.-H., Lin, Y.-N., Agno, J.E., Sharma, M., Braun, R.E., Rajkovic, A., Matzuk, M.M., 2006. TEX14 is essential for intercellular bridges and fertility in male mice. Proceedings of the National Academy of Sciences 103, 4982–4987. https://doi.org/10.1073/pnas.0505123103

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Gut, F., Käshammer, L., Lammens, K., Bartho, J.D., Boggusch, A.-M., van de Logt, E., Kessler, B., Hopfner, K.-P., 2022. Structural mechanism of endonucleolytic processing of blocked DNA ends and hairpins by Mre11-Rad50. Mol Cell 82, 3513-3522.e6. https://doi.org/10.1016/j.molcel.2022.07.019

Hadley, W., 2016. Ggplot2: Elegrant graphics for data analysis.

Haft, D.H., Selengut, J., Mongodin, E.F., Nelson, K.E., 2005. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. PLOS Computational Biology 1, e60. https://doi.org/10.1371/journal.pcbi.0010060

Haggård-Ljungquist, E., Barreiro, V., Calendar, R., Kurnit, D.M., Hans, C., 1989. The P2 phage *old* gene: sequence, transcription and translational control. Gene 85, 25–33. https://doi.org/10.1016/0378-1119(89)90460-5

Haig, D., 2016. Transposable elements: Self-seekers of the germline, team-players of the soma. BioEssays 38, 1158–1166. https://doi.org/10.1002/bies.201600125

Hallin, E.I., Eriksen, M.S., Baryshnikov, S., Nikolaienko, O., Grødem, S., Hosokawa, T., Hayashi, Y., Bramham, C.R., Kursula, P., 2018. Structure of monomeric full-length ARC sheds light on molecular flexibility, protein interactions, and functional modalities. J. Neurochem. 147, 323–343. https://doi.org/10.1111/jnc.14556

Hamard-Peron, E., Muriaux, D., 2011. Retroviral matrix and lipids, the intimate interaction. Retrovirology 8, 15. https://doi.org/10.1186/1742-4690-8-15

Haniford, D.B., 2006. Transpososome Dynamics and Regulation in Tn10 Transposition. Critical Reviews in Biochemistry and Molecular Biology 41, 407–424. https://doi.org/10.1080/10409230600987415

Hanifzadeh, M., Garcia, E.C., Viamajala, S., 2018. Production of lipid and carbohydrate from microalgae without compromising biomass productivities: Role of Ca and Mg. Renewable Energy. https://doi.org/10.1016/j.renene.2018.05.012

Hanifzadeh, M.-M., 2018. Approaches for Sustainable Production of Microalgae with High Productivity and Flexible Composition (Thesis).

Hanschen, E.R., Hovde, B.T., Starkenburg, S.R., 2020. An evaluation of methodology to determine algal genome completeness. Algal Research. https://doi.org/10.1016/j.algal.2020.102019

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Hayward, A., 2017. Origin of the retroviruses: when, where, and how? Current Opinion in Virology, Animal models for viral diseases • Paleovirology 25, 23–27. https://doi.org/10.1016/j.coviro.2017.06.006

He, Y., Wang, Y., Liu, B., Helmling, C., Sušac, L., Cheng, R., Zhou, Z.H., Feigon, J., 2021. Structures of telomerase at several steps of telomere repeat synthesis. Nature 593, 454–459. https://doi.org/10.1038/s41586-021-03529-9

Heine, P.A., Taylor, J.A., Iwamoto, G.A., Lubahn, D.B., Cooke, P.S., 2000. Increased adipose tissue in male and female estrogen receptor-α knockout mice. Proceedings of the National Academy of Sciences 97, 12729–12734. https://doi.org/10.1073/pnas.97.23.12729

Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., Downing, K.H., 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. Journal of Molecular Biology 213, 899–929. https://doi.org/10.1016/S0022-2836(05)80271-2

Henderson, R., Unwin, P.N.T., 1975. Three-dimensional model of purple membrane obtained by electron microscopy. Nature 257, 28–32. https://doi.org/10.1038/257028a0

Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E., Hatfull, G.F., 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proceedings of the National Academy of Sciences 96, 2192–2197. https://doi.org/10.1073/pnas.96.5.2192

Henriques, W.S., Young, J.M., Nemudryi, A., Nemudraia, A., Wiedenheft, B., Malik, H.S., 2024. The Diverse Evolutionary Histories of Domesticated Metaviral Capsid Genes in Mammals. Molecular Biology and Evolution 41, msae061. https://doi.org/10.1093/molbev/msae061

Henry, K., 2019. Paraneoplastic syndromes: Definitions, classification, pathophysiology and principles of treatment. Seminars in Diagnostic Pathology, Paraneoplastic syndromes 36, 204–210. https://doi.org/10.1053/j.semdp.2019.01.002

Henzy, J.E., Gifford, R.J., Kenaley, C.P., Johnson, W.E., 2017. An Intact Retroviral Gene Conserved in Spiny-Rayed Fishes for over 100 My. Molecular Biology and Evolution 34, 634–639. https://doi.org/10.1093/molbev/msw262

Herzik, M.A., Wu, M., Lander, G.C., 2019. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. Nature Communications 10, 1–9. https://doi.org/10.1038/s41467-019-08991-8

Hickman, A.B., Dyda, F., 2015. Mechanisms of DNA transposition. Mobile DNA III 529–553. https://doi.org/10.1128/9781555819217.ch25

Higashiyama, T., Maki, S., Yamada, T., 1995. Molecular organization of Chlorella vulgaris chromosome I: presence of telomeric repeats that are conserved in higher plants. Molecular and General Genetics MGG. https://doi.org/10.1007/bf00290130

Higgins, C.F., Linton, K.J., 2004. The ATP switch model for ABC transporters. Nat Struct Mol Biol 11, 918–926. https://doi.org/10.1038/nsmb836

Hollis, B., Day, F.R., Busch, A.S., Thompson, D.J., Soares, A.L.G., Timmers, P.R.H.J., Kwong, A., Easton, D.F., Joshi, P.K., Timpson, N.J., Ong, K.K., Perry, J.R.B., 2020. Genomic

analysis of male puberty timing highlights shared genetic basis with hair colour and lifespan. Nat Commun 11, 1536. https://doi.org/10.1038/s41467-020-14451-5

Holm, L., 2020. DALI and the persistence of protein shape. Protein Science 29, 128–140. https://doi.org/10.1002/pro.3749

Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-12-491

Hopfner, K.-P., Karcher, A., Shin, D.S., Craig, L., Arthur, L.M., Carney, J.P., Tainer, J.A., 2000. Structural Biology of Rad50 ATPase: ATP-Driven Conformational Control in DNA Double-Strand Break Repair and the ABC-ATPase Superfamily. Cell 101, 789–800. https://doi.org/10.1016/S0092-8674(00)80890-9

Hopfner, K.-P., Tainer, J.A., 2003. Rad50/SMC proteins and ABC transporters: unifying concepts from high-resolution structures. Current Opinion in Structural Biology 13, 249–255. https://doi.org/10.1016/S0959-440X(03)00037-X

Hovde, B.T., Hanschen, E.R., Steadman Tyler, C.R., Lo, C.-C., Kunde, Y., Davenport, K., Daligault, H., Msanne, J., Canny, S., Eyun, S.-I., Riethoven, J.-J.M., Polle, J., Starkenburg, S.R., 2018. Genomic characterization reveals significant divergence within Chlorella sorokiniana (Chlorellales, Trebouxiophyceae). Algal Research. https://doi.org/10.1016/j.algal.2018.09.012

Hu, C., Almendros, C., Nam, K.H., Costa, A.R., Vink, J.N.A., Haagsma, A.C., Bagde, S.R., Brouns, S.J.J., Ke, A., 2021. Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. Nature 598, 515–520. https://doi.org/10.1038/s41586-021-03951-z

Hu, Y., Chen, Y., Xu, J., Wang, X., Luo, S., Mao, B., Zhou, Q., Li, W., 2022. Metagenomic discovery of novel CRISPR-Cas13 systems. Cell Discov 8, 1–4. https://doi.org/10.1038/s41421-022-00464-5

Huang, N.-J., Zhang, L., Tang, W., Chen, C., Yang, C.-S., Kornbluth, S., 2012. The Trim39 ubiquitin ligase inhibits APC/CCdh1-mediated degradation of the Bax activator MOAP-1. Journal of Cell Biology 197, 361–367. https://doi.org/10.1083/jcb.201111141

Hübel, C., Gaspar, H.A., Coleman, J.R.I., Finucane, H., Purves, K.L., Hanscombe, K.B., Prokopenko, I., Investigators, M., Graff, M., Ngwa, J.S., Workalemahu, T., Consortium, E.D.W.G. of the P.G., Consortium, M.D.D.W.G. of the P.G., Consortium, S.W.G. of the P.G., Consortium, T.S.-C.D.W.G. of the P.G., O'Reilly, P.F., Bulik, C.M., Breen, G., 2019. Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 180, 428–438. https://doi.org/10.1002/ajmg.b.32709

Hubley, R., Finn, R.D., Clements, J., 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res 44. https://doi.org/10.1093/nar/gkv1272

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A.I., Marand, A.P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M.K., Piri, R.D., Wang, N., Kim, D. won, Zeng, Y., O'Connor, C.H., Li, X., Gilbert, A.M., Baggs, E., Krasileva, K.V., Portwood, J.L., Cannon, E.K.S., Andorf, C.M., Manchanda, N., Snodgrass, S.J., Hufnagel, D.E., Jiang, Q., Pedersen, S., Syring, M.L., Kudrna, D.A., Llaca, V., Fengler, K., Schmitz, R.J., Ross-Ibarra, J., Yu, J., Gent, J.I., Hirsch, C.N., Ware, D., Dawe, R.K., 2021. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science 373, 655–662. https://doi.org/10.1126/science.abg5289

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Hynes, M.J., Murray, S.L., 2010. ATP-Citrate Lyase Is Required for Production of Cytosolic Acetyl Coenzyme A and Development in Aspergillus nidulans. Eukaryotic Cell. https://doi.org/10.1128/ec.00080-10

Ikeda, H., Laigret, F., Martin, M.A., Repaske, R., 1985. Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. J Virol 55, 768–777. https://doi.org/10.1128/JVI.55.3.768-777.1985

Illergård, K., Ardell, D.H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. Proteins. https://doi.org/10.1002/prot.22458

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., Mirny, L.A., 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9, 999–1003. https://doi.org/10.1038/nmeth.2148

Irie, M., Itoh, J., Matsuzawa, A., Ikawa, M., Kiyonari, H., Kihara, M., Suzuki, T., Hiraoka, Y., Ishino, F., Kaneko-Ishino, T., 2022. Retrovirus-derived RTL5 and RTL6 genes are novel constituents of the innate immune system in the eutherian brain. Development 149, dev200976. https://doi.org/10.1242/dev.200976

Irie, M., Yoshikawa, M., Ono, R., Iwafune, H., Furuse, T., Yamada, I., Wakana, S., Yamashita, Y., Abe, T., Ishino, F., Kaneko-Ishino, T., 2015. Cognitive Function Related to the Sirh11/Zcchc16 Gene Acquired from an LTR Retrotransposon in Eutherians. PLOS Genetics 11, e1005521. https://doi.org/10.1371/journal.pgen.1005521

Isaev, A., Drobiazko, A., Sierro, N., Gordeeva, J., Yosef, I., Qimron, U., Ivanov, N.V., Severinov, K., 2020. Phage T7 DNA mimic protein Ocr is a potent inhibitor of BREX defence. Nucleic Acids Research 48, 5397–5406. https://doi.org/10.1093/nar/gkaa290

Ishino, F., Itoh, J., Irie, M., Matsuzawa, A., Naruse, M., Suzuki, T., Hiraoka, Y., Kaneko-Ishino, T., 2023. Retrovirus-Derived RTL9 Plays an Important Role in Innate Antifungal Immunity in the Eutherian Brain. International Journal of Molecular Sciences 24, 14884. https://doi.org/10.3390/ijms241914884

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

Isola, J.V.V., Ocañas, S.R., Hubbart, C.R., Ko, S., Mondal, S.A., Hense, J.D., Carter, H.N.C., Schneider, A., Kovats, S., Alberola-Ila, J., Freeman, W.M., Stout, M.B., 2024. A single-cell atlas of the aging mouse ovary. Nat Aging 4, 145–162. https://doi.org/10.1038/s43587-023-00552-5

Ito, J., Baba, T., Kawasaki, J., Nishigaki, K., 2016. Ancestral Mutations Acquired in Refrex-1, a Restriction Factor against Feline Retroviruses, during its Cooption and Domestication. Journal of Virology 90, 1470–1485. https://doi.org/10.1128/JVI.01904-15

Ito, J., Watanabe, S., Hiratsuka, T., Kuse, K., Odahara, Y., Ochi, H., Kawamura, M., Nishigaki, K., 2013. Refrex-1, a Soluble Restriction Factor against Feline Endogenous and Exogenous Retroviruses. Journal of Virology 87, 12029–12040. https://doi.org/10.1128/JVI.01267-13

Ivanov, D., Stone, J.R., Maki, J.L., Collins, T., Wagner, G., 2005. Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain. Mol. Cell 17, 137–143. https://doi.org/10.1016/j.molcel.2004.12.015

Iwasaki, S., Suzuki, S., Pelekanos, M., Clark, H., Ono, R., Shaw, G., Renfree, M.B., Kaneko-Ishino, T., Ishino, F., 2013. Identification of a novel PNMA-MS1 gene in marsupials suggests the LTR retrotransposon-derived PNMA genes evolved differently in marsupials and eutherians. DNA Res. 20, 425–436. https://doi.org/10.1093/dnares/dst020

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., Brouns, S.J.J., 2017a. CRISPR-Cas: Adapting to change. Science 356, eaal5056. https://doi.org/10.1126/science.aal5056

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., Brouns, S.J.J., 2017b. CRISPR-Cas: Adapting to change. Science 356, eaal5056. https://doi.org/10.1126/science.aal5056

Jain, I., Kolesnik, M., Minakhin, L., Morozova, N., Shiriaeva, A., Kirillov, A., Medvedeva, S., Kuznedelov, K., Borukhov, S., Makarova, K.S., Koonin, E.V., Severinov, K., Semenova, E., 2021. tRNA anticodon cleavage by target-activated CRISPR-Cas13a effector. https://doi.org/10.1101/2021.11.10.468108

Jangam, D., Feschotte, C., Betrán, E., 2017. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends in Genetics 33, 817–831. https://doi.org/10.1016/j.tig.2017.07.011

Jansen, Ruud., Embden, Jan.D.A. van, Gaastra, Wim., Schouls, Leo.M., 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Molecular Microbiology 43, 1565–1575. https://doi.org/10.1046/j.1365-2958.2002.02839.x

Jiang, X., Jia, X., Sun, J., Qi, C., Lu, L., Wang, Y., Zhang, L., Wei, M., 2020. Overexpressed coiled-coil domain containing protein 8 (CCDC8) mediates newly synthesized HIV-1 Gag lysosomal degradation. Scientific Reports 10, 11416. https://doi.org/10.1038/s41598-020-68341-3

Jiang, Y., Meidler, R., Amitsur, M., Kaufmann, G., 2001. Specific interaction between anticodon nuclease and the tRNALys wobble base1. Journal of Molecular Biology 305, 377–388. https://doi.org/10.1006/jmbi.2000.4282

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., Charpentier, E., 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (New York, N.Y.) 337, 816–821. https://doi.org/10.1126/science.1225829

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaneko-Ishino, T., Ishino, F., 2015. Mammalian-specific genomic functions: Newly acquired traits generated by genomic imprinting and LTR retrotransposon-derived genes in mammals. Proc Jpn Acad Ser B Phys Biol Sci 91, 511–538. https://doi.org/10.2183/pjab.91.511

Kaneko-Ishino, T., Ishino, F., 2012. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. Front Microbiol 3, 262. https://doi.org/10.3389/fmicb.2012.00262

Karlsson, T., Rask-Andersen, M., Pan, G., Höglund, J., Wadelius, C., Ek, W.E., Johansson, Å., 2019. Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. Nat Med 25, 1390–1395. https://doi.org/10.1038/s41591-019-0563-7

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/mst010

Katoh, Kazutaka, Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30, 772–780. https://doi.org/10.1093/molbev/mst010

Ke, H.Z., Paralkar, V.M., Grasser, W.A., Crawford, D.T., Qi, H., Simmons, H.A., Pirie, C.M., Chidsey-Frink, K.L., Owen, T.A., Smock, S.L., Chen, H.K., Jee, W.S.S., Cameron, K.O., Rosati, R.L., Brown, T.A., Dasilva-Jardine, P., Thompson, D.D., 1998. Effects of CP-336,156, a New, Nonsteroidal Estrogen Agonist/Antagonist, on Bone, Serum Cholesterol, Uterus, and Body Composition in Rat Models. Endocrinology 139, 2068–2076. https://doi.org/10.1210/endo.139.4.5902

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.E., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10, 845–858. https://doi.org/10.1038/nprot.2015.053

Kelly, D.M., Jones, T.H., 2015. Testosterone and obesity. Obesity Reviews 16, 581–606. https://doi.org/10.1111/obr.12282

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., 1958. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature. https://doi.org/10.1038/181662a0

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., Price, A.L., 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. The American Journal of Human Genetics 104, 65–75. https://doi.org/10.1016/j.ajhg.2018.11.008

Kieper, S.N., Almendros, C., Brouns, S.J.J., 2019. Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type I-D CRISPR-Cas systems. FEMS Microbiology Letters 366, 2016–2020. https://doi.org/10.1093/femsle/fnz129

Kim, F.J., Battini, J.-L., Manel, N., Sitbon, M., 2004. Emergence of vertebrate retroviruses and envelope capture. Virology 318, 183–191. https://doi.org/10.1016/j.virol.2003.09.026

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., Voytas, D.F., 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res 8, 464–478. https://doi.org/10.1101/gr.8.5.464

Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S.J.J., Joo, C., 2020. Selective loading and processing of prespacers for precise CRISPR adaptation. Nature 579, 141–145. https://doi.org/10.1038/s41586-020-2018-1

Kingsford, C., Schatz, M.C., Pop, M., 2010. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-11-21

Kitazawa, M., Hayashi, S., Imamura, M., Takeda, S., Oishi, Y., Kaneko-Ishino, T., Ishino, F., 2020. Deficiency and overexpression of Rtl1 in the mouse cause distinct muscle abnormalities related to Temple and Kagami-Ogata syndromes. Development 147, dev185918. https://doi.org/10.1242/dev.185918

Kitazawa, M., Sutani, A., Kaneko-Ishino, T., Ishino, F., 2021. The role of eutherian-specific RTL1 in the nervous system and its implications for the Kagami-Ogata and Temple syndromes. Genes Cells 26, 165–179. https://doi.org/10.1111/gtc.12830

Kitazawa, M., Tamura, M., Kaneko-Ishino, T., Ishino, F., 2017. Severe damage to the placental fetal capillary network causes mid- to late fetal lethality and reduction in placental size in Peg11/Rtl1 KO mice. Genes Cells 22, 174–188. https://doi.org/10.1111/gtc.12465

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

Kokošar, J., Kordiš, D., 2013. Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. Mol Biol Evol 30, 1015–1031. https://doi.org/10.1093/molbev/mst014

Kolde, R., 2019. pheatmap: Pretty Heatmaps.

Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A., 2019. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology. https://doi.org/10.1038/s41587-019-0072-8

Kong, J.N., Hardin, K., Dinkins, M., Wang, G., He, Q., Mujadzic, T., Zhu, G., Bielawski, J., Spassieva, S., Bieberich, E., 2015. Regulation of Chlamydomonas flagella and ependymal cell motile cilia by ceramide-mediated translocation of GSK3. Molecular Biology of the Cell. https://doi.org/10.1091/mbc.e15-06-0371

Koonin, E.V., 2017. Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. Biol Direct 12, 5. https://doi.org/10.1186/s13062-017-0177-2

Koonin, E.V., Krupovic, M., 2015a. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet 16, 184–192. https://doi.org/10.1038/nrg3859

Koonin, E.V., Krupovic, M., 2015b. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nature Reviews Genetics 16, 184–192. https://doi.org/10.1038/nrg3859

Koonin, E.V., Makarova, K.S., 2019. Origins and evolution of CRISPR-Cas systems. Philos Trans R Soc Lond B Biol Sci 374, 20180087. https://doi.org/10.1098/rstb.2018.0087

Koonin, E.V., Makarova, K.S., Wolf, Y.I., Krupovic, M., 2020. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nature Reviews Genetics 21, 119–131. https://doi.org/10.1038/s41576-019-0172-9

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Phillippy, A.M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology. https://doi.org/10.1038/nbt.2280

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. Genome Research. https://doi.org/10.1101/gr.215087.116

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D.W., 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22, 3096–3098. https://doi.org/10.1093/bioinformatics/btl474

Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T.F., Morgan, H., Atienza-Herrero, J., Blake, A., Chen, C.-K., Easty, R., Di Fenza, A., Fiegel, T., Grifiths, M., Horne, A., Karp, N.A., Kurbatova, N., Mason, J.C., Matthews, P., Oakley, D.J., Qazi, A., Regnart, J., Retha, A., Santos, L.A., Sneddon, D.J., Warren, J., Westerberg, H., Wilson, R.J., Melvin, D.G., Smedley, D., Brown, S.D.M., Flicek, P., Skarnes, W.C., Mallon, A.-M., Parkinson, H., 2014. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Research 42, D802–D809. https://doi.org/10.1093/nar/gkt977

Krishnan, A., Burroughs, A.M., Iyer, L.M., Aravind, L., 2020. Comprehensive classification of ABC ATPases and their functional radiation in nucleoprotein dynamics and biological conflict systems. Nucleic Acids Research 48, 10045–10075. https://doi.org/10.1093/nar/gkaa726

Krüger, D.H., Reuter, M., Hansen, S., Schroeder, C., 1982. Influence of phage T3 and T7 gene functions on a type III (EcoP1) DNA restriction-modification system in vivo. Mol Gen Genet 185, 457–461. https://doi.org/10.1007/BF00334140

Krupovic, M., Béguin, P., Koonin, E.V., 2017. Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Current Opinion in Microbiology, Mobile genetic elements and HGT in prokaryotes * Microbiota 38, 36–43. https://doi.org/10.1016/j.mib.2017.04.004

Krupovic, M., Blomberg, J., Coffin, J.M., Dasgupta, I., Fan, H., Geering, A.D., Gifford, R., Harrach, B., Hull, R., Johnson, W., Kreuze, J.F., Lindemann, D., Llorens, C., Lockhart, B., Mayer, J., Muller, E., Olszewski, N.E., Pappu, H.R., Pooggin, M.M., Richert-Pöggeler,

K.R., Sabanadzovic, S., Sanfaçon, H., Schoelz, J.E., Seal, S., Stavolone, L., Stoye, J.P., Teycheney, P.-Y., Tristem, M., Koonin, E.V., Kuhn, J.H., 2018. Ortervirales: New Virus Order Unifying Five Families of Reverse-Transcribing Viruses. Journal of Virology 92, e00515-18. https://doi.org/10.1128/JVI.00515-18

Krupovic, M., Koonin, E.V., 2017a. Homologous Capsid Proteins Testify to the Common Ancestry of Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses. Journal of Virology 91, e00210-17. https://doi.org/10.1128/JVI.00210-17

Krupovic, M., Koonin, E.V., 2017b. Multiple origins of viral capsid proteins from cellular ancestors. PNAS 114, E2401–E2410. https://doi.org/10.1073/pnas.1621061114

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., Koonin, E.V., 2014. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biology 12, 36. https://doi.org/10.1186/1741-7007-12-36

Kühlbrandt, W., 2014. The Resolution Revolution. Science 343, 1443–1444. https://doi.org/10.1126/science.1251652

Kumar, S., Suleski, M., Craig, J.M., Kasprowicz, A.E., Sanderford, M., Li, M., Stecher, G., Hedges, S.B., 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. Mol Biol Evol 39, msac174. https://doi.org/10.1093/molbev/msac174

Kunin, V., Sorek, R., Hugenholtz, P., 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8, R61. https://doi.org/10.1186/gb-2007-8-4-r61

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., Brouns, S.J.J., 2016. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Molecular Cell 63, 852–864. https://doi.org/10.1016/j.molcel.2016.07.011

Kurata, T., Saha, C.K., Buttress, J.A., Mets, T., Brodiazhenko, T., Turnbull, K.J., Awoyomi, O.F., Oliveira, S.R.A., Jimmy, S., Ernits, K., Delannoy, M., Persson, K., Tenson, T., Strahl, H., Hauryliuk, V., Atkinson, G.C., 2022. A hyperpromiscuous antitoxin protein domain for the neutralization of diverse toxin domains. Proceedings of the National Academy of Sciences 119, e2102212119. https://doi.org/10.1073/pnas.2102212119

Kuypers, M.M.M., Marchant, H.K., Kartal, B., 2018. The microbial nitrogen-cycling network. Nature Reviews Microbiology. https://doi.org/10.1038/nrmicro.2018.9

Lammens, A., Hopfner, K.-P., 2010. Structural Basis for Adenylate Kinase Activity in ABC ATPases. Journal of Molecular Biology 401, 265–273. https://doi.org/10.1016/j.jmb.2010.06.029

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, Christina, Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, Christopher, Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M.J., International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, C. for G.R., The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Department of Genome Analysis, I. of M.B., GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, Multimegabase Sequencing Center, T.I. for S.B., Stanford Genome Technology Center:, University of Oklahoma's Advanced

Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Cold Spring Harbor Laboratory, L.A.H.G.C., GBF—German Research Centre for Biotechnology:, *Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):, Scientific management: National Human Genome Research Institute, U.N.I. of H., Stanford Human Genome Center:, University of Washington Genome Center:, Department of Molecular Biology, K.U.S. of M., University of Texas Southwestern Medical Center at Dallas:, Office of Science, U.D. of E., The Wellcome Trust:, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062

Lane, N., 2015. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals.' Philos Trans R Soc Lond B Biol Sci 370, 20140344. https://doi.org/10.1098/rstb.2014.0344

Laureau, R., Dyatel, A., Dursuk, G., Brown, S., Adeoye, H., Yue, J.-X., De Chiara, M., Harris, A., Ünal, E., Liti, G., Adams, I.R., Berchowitz, L.E., 2021. Meiotic Cells Counteract Programmed Retrotransposon Activation via RNA-Binding Translational Repressor Assemblies. Developmental Cell 56, 22-35.e7. https://doi.org/10.1016/j.devcel.2020.11.008

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., Heidmann, T., 2013. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philosophical Transactions of the Royal Society B: Biological Sciences 368, 20120507. https://doi.org/10.1098/rstb.2012.0507

Lavoie, B.D., Chaconas, G., 1993. Site-specific HU binding in the Mu transpososome: conversion of a sequence-independent DNA-binding protein into a chemical nuclease. Genes & Development 7, 2510–2519. https://doi.org/10.1101/gad.7.12b.2510

Law, J., Salla, M., Zare, A., Wong, Y., Luong, L., Volodko, N., Svystun, O., Flood, K., Lim, J., Sung, M., Dyck, J.R.B., Tan, C.T., Su, Y.-C., Yu, V.C., Mackey, J., Baksh, S., 2015. Modulator of apoptosis 1 (MOAP-1) is a tumor suppressor protein linked to the RASSF1A protein. J Biol Chem 290, 24100–24118. https://doi.org/10.1074/jbc.M115.648345

Laxmikanthan, G., Xu, C., Brilot, A.F., Warren, D., Steele, L., Seah, N., Tong, W., Grigorieff, N., Landy, A., Van Duyne, G.D., 2016. Structure of a holliday junction complex reveals mechanisms governing a highly regulated DNA transaction. eLife 5, 1–23. https://doi.org/10.7554/eLife.14313

Ledvina, H.E., Whiteley, A.T., 2024. Conservation and similarity of bacterial and eukaryotic innate immunity. Nat Rev Microbiol 22, 420–434. https://doi.org/10.1038/s41579-024-01017-1

Lee, H., Sashital, D.G., 2022a. Creating memories: molecular mechanisms of CRISPR adaptation. Trends in Biochemical Sciences 47, 464–476. https://doi.org/10.1016/j.tibs.2022.02.004

Lee, H., Sashital, D.G., 2022b. Creating memories: molecular mechanisms of CRISPR adaptation. Trends in Biochemical Sciences 1–13. https://doi.org/10.1016/j.tibs.2022.02.004

Lee, Y.H., Pang, S.W., Poh, C.L., Tan, K.O., 2016. Distinct functional domains of PNMA5 mediate protein-protein interaction, nuclear localization, and apoptosis signaling in human cancer cells. J. Cancer Res. Clin. Oncol. 142, 1967–1977. https://doi.org/10.1007/s00432-016-2205-5

Lesch, B.J., Silber, S.J., McCarrey, J.R., Page, D.C., 2016. Parallel evolution of male germline epigenetic poising and somatic development in animals. Nat Genet 48, 888–894. https://doi.org/10.1038/ng.3591

Levasseur, A., Bekliz, M., Chabrière, E., Pontarotti, P., La Scola, B., Raoult, D., 2016. MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. Nature 531, 249–252. https://doi.org/10.1038/nature17146

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty191

Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw152

Li, M., Wang, R., Zhao, D., Xiang, H., 2014. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Research 42, 2483–2492. https://doi.org/10.1093/nar/gkt1154

Li, Q., Pangas, S.A., Jorgez, C.J., Graff, J.M., Weinstein, M., Matzuk, M.M., 2008. Redundant Roles of SMAD2 and SMAD3 in Ovarian Granulosa Cells In Vivo. Molecular and Cellular Biology 28, 7001–7011. https://doi.org/10.1128/MCB.00732-08

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

Li, W., Lee, M.-H., Henderson, L., Tyagi, R., Bachani, M., Steiner, J., Campanac, E., Hoffman, D.A., von Geldern, G., Johnson, K., Maric, D., Morris, H.D., Lentz, M., Pak, K., Mammen, A., Ostrow, L., Rothstein, J., Nath, A., 2015. Human endogenous retrovirus-K contributes to motor neuron disease. Science Translational Medicine 7, 307ra153-307ra153. https://doi.org/10.1126/scitranslmed.aac8201

Li, W.-H., Gu, Z., Wang, H., Nekrutenko, A., 2001. Evolutionary analyses of the human genome. Nature 409, 847–849. https://doi.org/10.1038/35057039

Li, Y., Lou, Y., Mu, T., Ke, A., Ran, Z., Xu, J., Chen, J., Zhou, C., Yan, X., Xu, Q., Tan, Y., 2017. Sphingolipids in marine microalgae: Development and application of a mass

spectrometric method for global structural characterization of ceramides and glycosphingolipids in three major phyla. Analytica Chimica Acta. https://doi.org/10.1016/j.aca.2017.07.039

Liang, M.-H., Jiang, J.-G., 2016. Characterization and nitrogen deficiency response of ATP-citrate lyase from unicellular alga Dunaliella tertiolecta. Algal Research. https://doi.org/10.1016/j.algal.2016.09.025

Liao, C., Sharma, S., Svensson, S.L., Kibe, A., Weinberg, Z., Alkhnbashi, O.S., Bischler, T., Backofen, R., Caliskan, N., Sharma, C.M., Beisel, C.L., 2022. Spacer prioritization in CRISPR–Cas9 immunity is enabled by the leader RNA. Nature Microbiology. https://doi.org/10.1038/s41564-022-01074-3

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., Buonaiuto, S., Chang, X.H., Cheng, H., Chu, J., Colonna, V., Eizenga, J.M., Feng, X., Fischer, C., Fulton, R.S., Garg, S., Groza, C., Guarracino, A., Harvey, W.T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F.J., Mitchell, M.W., Munson, K.M., Mwaniki, M.N., Novak, A.M., Olsen, H.E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J.A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M.R., Antonacci-Fulton, L.L., Baid, G., Baker, C.A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D.E., Cook-Deegan, R.M., Cornejo, O.E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A.L., Formenti, G., Frankish, A., Gao, Y., Garrison, N.A., Giron, C.G., Green, R.E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H.P., Kenny, E.E., Koenig, B.A., Kolesnikov, A., Korbel, J.O., Kordosky, J., Koren, S., Lee, H., Lewis, A.P., Magalhães, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N.D., Popejoy, A.B., Puiu, D., Rautiainen, M., Regier, A.A., Rhie, A., Sacco, S., Sanders, A.D., Schneider, V.A., Schultz, B.I., Shafin, K., Smith, M.W., Sofia, H.J., Abou Tayoun, A.N., Thibaud-Nissen, F., Tricomi, F.F., Wagner, J., Walenz, B., Wood, J.M.D., Zimin, A.V., Bourque, G., Chaisson, M.J.P., Flicek, P., Phillippy, A.M., Zook, J.M., Eichler, E.E., Haussler, D., Wang, T., Jarvis, E.D., Miga, K.H., Garrison, E., Marschall, T., Hall, I.M., Li, H., Paten, B., 2023. A draft human pangenome reference. Nature 617, 312–324. https://doi.org/10.1038/s41586-023-05896-x

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., Moriarty, N.W., Oeffner, R.D., Poon, B.K., Prisant, M.G., Read, R.J., Richardson, J.S., Richardson, D.C., Sammito, M.D., Sobolev, O.V., Stockwell, D.H., Terwilliger, T.C., Urzhumtsev, A.G., Videau, L.L., Williams, C.J., Adams, P.D., 2019. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Cryst D 75, 861–877. https://doi.org/10.1107/S2059798319011471

Liebschner, Dorothee, Afonine, P.V., Baker, M.L., Bunkóczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A.J., Moriarty, N.W., Oeffner, R.D., Poon, B.K., Prisant, M.G., Read, R.J., Richardson, J.S., Richardson, D.C., Sammito, M.D., Sobolev, O.V., Stockwell, D.H., Terwilliger, T.C., Urzhumtsev, A.G., Videau, L.L., Williams, C.J.,

Adams, P.D., 2019. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. Acta Crystallogr D Struct Biol 75, 861–877. https://doi.org/10.1107/S2059798319011471

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130. https://doi.org/10.1126/science.ade2574

Linheiro, R.S., Bergman, C.M., 2008. Testing the palindromic target site model for DNA transposon insertion using the Drosophila melanogaster P-element. Nucleic Acids Research 36, 6199–6208. https://doi.org/10.1093/nar/gkn563

Lioy, V.S., Cournac, A., Marbouty, M., Duigou, S., Mozziconacci, J., Espéli, O., Boccard, F., Koszul, R., 2018. Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. Cell 172, 771-783.e18. https://doi.org/10.1016/j.cell.2017.12.027

Little, A.J., Corbett, E., Ortega, F., Schatz, D.G., 2013. Cooperative recruitment of HMGB1 during V(D)J recombination through interactions with RAG1 and DNA. Nucleic Acids Research 41, 3289–3301. https://doi.org/10.1093/nar/gks1461

Liu, C., Zhang, Y., Liu, C.C., Schatz, D.G., 2022. Structural insights into the evolution of the RAG recombinase. Nat Rev Immunol 22, 353–370. https://doi.org/10.1038/s41577-021-00628-6

Liu, D., Haniford, D.B., Chalmers, R.M., 2011. H-NS mediates the dissociation of a refractory protein-DNA complex during Tn10/IS10 transposition. Nucleic Acids Research 39, 6660–6668. https://doi.org/10.1093/nar/gkr309

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., Munoz-Pomer, A., Sempere, J.M., Latorre, A., Moya, A., 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39, D70–D74. https://doi.org/10.1093/nar/gkq1061

Loenen, W.A.M., Dryden, D.T.F., Raleigh, E.A., Wilson, G.G., Murray, N.E., 2014. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Research 42, 3–19. https://doi.org/10.1093/nar/gkt990

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S.,

Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F., 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580–585. https://doi.org/10.1038/ng.2653

Luria, S.E., Anderson, T.F., 1942. The Identification and Characterization of Bacteriophages with the Electron Microscope. Proc Natl Acad Sci U S A 28, 127-130.1. https://doi.org/10.1073/pnas.28.4.127

Lynch, C., Tristem, M., 2003. A Co-opted gypsy-type LTR-Retrotransposon Is Conserved in the Genomes of Humans, Sheep, Mice, and Rats. Current Biology 13, 1518–1523. https://doi.org/10.1016/S0960-9822(03)00618-3

Mabb, A.M., Ehlers, M.D., 2018. Arc ubiquitination in synaptic plasticity. Semin Cell Dev Biol 77, 10–16. https://doi.org/10.1016/j.semcdb.2017.09.009

Macaulay, A.D., Gilbert, I., Caballero, J., Barreto, R., Fournier, E., Tossou, P., Sirard, M.-A., Clarke, H.J., Khandjian, É.W., Richard, F.J., Hyttel, P., Robert, C., 2014. The Gametic Synapse: RNA Transfer to the Bovine Oocyte1. Biology of Reproduction 91, 90, 1–12. https://doi.org/10.1095/biolreprod.114.119867

Macaulay, A.D., Gilbert, I., Scantland, S., Fournier, E., Ashkar, F., Bastien, A., Saadi, H.A.S., Gagné, D., Sirard, M.-A., Khandjian, É.W., Richard, F.J., Hyttel, P., Robert, C., 2016. Cumulus Cell Transcripts Transit to the Bovine Oocyte in Preparation for Maturation1. Biology of Reproduction 94, 16, 1–11. https://doi.org/10.1095/biolreprod.114.127571

Madigan, V., Zhang, Y., Raghavan, R., Wilkinson, M.E., Faure, G., Puccio, E., Segel, M., Lash, B., Macrae, R.K., Zhang, F., 2024. Human paraneoplastic antigen Ma2 (PNMA2) forms icosahedral capsids that can be engineered for mRNA delivery. Proceedings of the National Academy of Sciences 121, e2307812120. https://doi.org/10.1073/pnas.2307812120

Maertens, G.N., Hare, S., Cherepanov, P., 2010. The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature 468, 326–329. https://doi.org/10.1038/nature09517

Mager, D.L., Stoye, J.P., 2015. Mammalian Endogenous Retroviruses. Microbiology Spectrum 3, 3.1.09. https://doi.org/10.1128/microbiolspec.MDNA3-0009-2014

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., Moineau, S., Mojica, F.J.M., Scott, D., Shah, S.A., Siksnys, V., Terns, M.P., Venclovas, Č., White, M.F., Yakunin, A.F., Yan, W., Zhang, F., Garrett, R.A., Backofen, R., van der Oost, J., Barrangou, R., Koonin, E.V., 2020a. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18, 67–83. https://doi.org/10.1038/s41579-019-0299-x

Makarova, K.S., Wolf, Y.I., Koonin, E.V., 2018. Classification and Nomenclature of CRISPR-Cas Systems: Where from Here? The CRISPR Journal 1, 325–336. https://doi.org/10.1089/crispr.2018.0033

Makarova, K.S., Wolf, Y.I., Snir, S., Koonin, E.V., 2011. Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. Journal of Bacteriology 193, 6039–6056. https://doi.org/10.1128/JB.05535-11

Manago, F., Mereu, M., Mastwal, S., Mastrogiacomo, R., Scheggia, D., Emanuele, M., De Luca, M.A., Weinberger, D.R., Wang, K.H., Papaleo, F., 2016. Genetic Disruption of Arc/Arg3.1 in Mice Causes Alterations in Dopamine and Neurobehavioral Phenotypes Related to Schizophrenia. Cell Rep 16, 2116–2128. https://doi.org/10.1016/j.celrep.2016.07.044

Manni, M. e, Matthew, Seppey, M., Felipe, Evgeny, 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv pre-print server. https://doi.org/None arxiv:2106.11799

Martin, S., Cule, M., Basty, N., Tyrrell, J., Beaumont, R.N., Wood, A.R., Frayling, T.M., Sorokin, E., Whitcher, B., Liu, Y., Bell, J.D., Thomas, E.L., Yaghootkar, H., 2021. Genetic Evidence for Different Adiposity Phenotypes and Their Opposing Influences on Ectopic Fat and Risk of Cardiometabolic Disease. Diabetes 70, 1843–1856. https://doi.org/10.2337/db21-0129

Mastronarde, D.N., 2005a. Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol 152, 36–51. https://doi.org/10.1016/j.jsb.2005.07.007

Matthews, B.J., Dudchenko, O., Kingan, S.B., Koren, S., Antoshechkin, I., Crawford, J.E., Glassford, W.J., Herre, M., Redmond, S.N., Rose, N.H., Weedall, G.D., Wu, Y., Batra, S.S., Brito-Sierra, C.A., Buckingham, S.D., Campbell, C.L., Chan, S., Cox, E., Evans, B.R., Fansiri, T., Filipović, I., Fontaine, A., Gloria-Soria, A., Hall, R., Joardar, V.S., Jones, A.K., Kay, R.G.G., Kodali, V.K., Lee, J., Lycett, G.J., Mitchell, S.N., Muehling, J., Murphy, M.R., Omer, A.D., Partridge, F.A., Peluso, P., Aiden, A.P., Ramasamy, V., Rašić, G., Roy, S., Saavedra-Rodriguez, K., Sharan, S., Sharma, A., Smith, M.L., Turner, J., Weakley, A.M., Zhao, Z., Akbari, O.S., Black, W.C., Cao, H., Darby, A.C., Hill, C.A., Johnston, J.S., Murphy, T.D., Raikhel, A.S., Sattelle, D.B., Sharakhov, I.V., White, B.J., Zhao, L., Aiden,

E.L., Mann, R.S., Lambrechts, L., Powell, J.R., Sharakhova, M.V., Tu, Z., Robertson, H.M., McBride, C.S., Hastie, A.R., Korlach, J., Neafsey, D.E., Phillippy, A.M., Vosshall, L.B., 2018. Improved reference genome of Aedes aegypti informs arbovirus vector control. Nature 563, 501–507. https://doi.org/10.1038/s41586-018-0692-z

McBee, R.M., Rozmiarek, S.A., Meyerson, N.R., Rowley, P.A., Sawyer, S.L., 2015. The effect of species representation on the detection of positive selection in primate gene data sets. Mol Biol Evol 32, 1091–1096. https://doi.org/10.1093/molbev/msu399

McCafferty, C.L., Verbeke, E.J., Marcotte, E.M., Taylor, D.W., 2020. Structural Biology in the Multi-Omics Era. J Chem Inf Model 60, 2424–2429. https://doi.org/10.1021/acs.jcim.9b01164

McCLINTOCK, B., 1950. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences of the United States of America 36, 344–355. https://doi.org/10.1073/pnas.36.6.344

McGinn, J., Marraffini, L.A., 2016. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Molecular Cell 64, 616–623. https://doi.org/10.1016/j.molcel.2016.08.038

McLaughlin, R.N., Malik, H.S., 2017. Genetic conflicts: the usual suspects and beyond. J Exp Biol 220, 6–17. https://doi.org/10.1242/jeb.148148

Mehrotra, S., Goyal, V., 2014. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genomics, proteomics & bioinformatics. https://doi.org/10.1016/j.gpb.2014.07.003

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Marechal-Drouard, L., Marshall, W.F., Qu, L.H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fernandez, E., Fukuzawa, H., Gonzalez-Ballester, D., Gonzalez-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.A., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S., Ral, J.P., Riano-Pachon, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., others, 2007. The Chlamydomonas Genome Reveals the

Evolution of Key Animal and Plant Functions. Science. https://doi.org/10.1126/science.1143609

Mestre, M.R., González-Delgado, A., Gutiérrez-Rus, L.I., Martínez-Abarca, F., Toro, N., 2020. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. Nucleic Acids Research 48, 12632–12647. https://doi.org/10.1093/nar/gkaa1149

Michel, A.M., Fox, G., M. Kiran, A., De Bo, C., O'Connor, P.B.F., Heaphy, S.M., Mullan, J.P.A., Donohue, C.A., Higgins, D.G., Baranov, P.V., 2014. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Research 42, D859–D864. https://doi.org/10.1093/nar/gkt1035

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M., 2022a. ColabFold: making protein folding accessible to all. Nature Methods. https://doi.org/10.1038/s41592-022-01488-1

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M., 2022b. ColabFold: making protein folding accessible to all. Nat Methods 19, 679–682. https://doi.org/10.1038/s41592-022-01488-1

Mirdita, M., Steinegger, M., Söding, J., 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 35, 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Research 49, D412–D419. https://doi.org/10.1093/nar/gkaa913

Mitchell, A.L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., Finn, R.D., 2020. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research 48, D570–D578. https://doi.org/10.1093/nar/gkz1035

Modzelewski, A.J., Gan Chong, J., Wang, T., He, L., 2022. Mammalian genome innovation through transposon domestication. Nat Cell Biol 24, 1332–1340. https://doi.org/10.1038/s41556-022-00970-4

Mogila, I., Tamulaitiene, G., Keda, K., Timinskas, A., Ruksenaite, A., Sasnauskas, G., Venclovas, Č., Siksnys, V., Tamulaitis, G., 2023. Ribosomal stalk-captured CARF-RelE ribonuclease inhibits translation following CRISPR signaling. Science 382, 1036–1041. https://doi.org/10.1126/science.adj2107

Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E.V., van der Oost, J., 2016a. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science 353, aad5147. https://doi.org/10.1126/science.aad5147

Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E.V., van der Oost, J., 2016b. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science 353, aad5147. https://doi.org/10.1126/science.aad5147

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., Almendros, C., 2009a. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733–740. https://doi.org/10.1099/mic.0.023960-0

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., Almendros, C., 2009b. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733–740. https://doi.org/10.1099/mic.0.023960-0

Mojica, F.J.M., Díez-Villaseñor, C., Soria, E., Juez, G., 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Molecular Microbiology 36, 244–246. https://doi.org/10.1046/j.1365-2958.2000.01838.x

Molaro, A., Young, J.M., Malik, H.S., 2018. Evolutionary origins and diversification of testis-specific short histone H2A variants in mammals. Genome Res 28, 460–473. https://doi.org/10.1101/gr.229799.117

Montaño, S.P., Pigli, Y.Z., Rice, P.A., 2012. The Mu transpososome structure sheds light on DDE recombinase evolution. Nature 491, 413–417. https://doi.org/10.1038/nature11602

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. https://doi.org/10.1093/nar/gkm321

Mountjoy, E., Schmidt, E.M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M.A., Wright, D., Hercules, A., Papa, E., Fauman, E.B., Barrett, J.C., Todd, J.A., Ochoa, D., Dunham, I., Ghoussaini, M., 2021. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet 53, 1527–1533. https://doi.org/10.1038/s41588-021-00945-5

Muriaux, D., Darlix, J.-L., 2010. Properties and functions of the nucleocapsid protein in virus assembly. RNA Biol 7, 744–753. https://doi.org/10.4161/rna.7.6.14065

Murray, P.S., Li, Z., Wang, J., Tang, C.L., Honig, B., Murray, D., 2005. Retroviral Matrix Domains Share Electrostatic Homology: Models for Membrane Binding Function throughout the Viral Life Cycle. Structure 13, 1521–1531. https://doi.org/10.1016/j.str.2005.07.010

Musharova, O., Medvedeva, S., Klimuk, E., Guzman, N.M., Titova, D., Zgoda, V., Shiriaeva, A., Semenova, E., Severinov, K., Savitskaya, E., 2021. Prespacers formed during primed adaptation associate with the Cas1–Cas2 adaptation complex and the Cas3 interference nuclease–helicase. Proceedings of the National Academy of Sciences 118, e2021291118. https://doi.org/10.1073/pnas.2021291118

Mustapa, M., Sallehudin, N., Mohamed, M., Mohammad Noor, N., Raus, R., 2016. Decontamination of Chlorella sp. Culture using antibiotics and antifungal cocktail treatment. ARPN Journal of Engineering and Applied Sciences.

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P.M.G.E., Grigoras, I.T., Malinauskaite, L., Malinauskas, T., Miehling, J., Uchański, T., Yu, L., Karia, D., Pechnikova, E.V., de Jong, E., Keizer, J., Bischoff, M., McCormack, J., Tiemeijer, P., Hardwick, S.W., Chirgadze, D.Y., Murshudov, G., Aricescu, A.R., Scheres, S.H.W., 2020. Single-particle cryo-EM at atomic resolution. Nature 587, 152–156. https://doi.org/10.1038/s41586-020-2829-0

Naruse, M., Ono, R., Irie, M., Nakamura, K., Furuse, T., Hino, T., Oda, K., Kashimura, M., Yamada, I., Wakana, S., Yokoyama, M., Ishino, F., Kaneko-Ishino, T., 2014. Sirh7/Ldoc1 knockout mice exhibit placental P4 overproduction and delayed parturition. Development 141, 4763–4771. https://doi.org/10.1242/dev.114520

Nash, H.A., Robertson, C.A., 1981. Purification and properties of the Escherichia coli protein factor required for lambda integrative recombination. Journal of Biological Chemistry 256, 9246–9253. https://doi.org/10.1016/S0021-9258(19)52537-6

Naville, M., Warren, I.A., Haftek-Terreau, Z., Chalopin, D., Brunet, F., Levin, P., Galiana, D., Volff, J.-N., 2016. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin. Microbiol. Infect 22, 312–323. https://doi.org/10.1016/j.cmi.2016.02.001.

Nemudraia, A., Nemudryi, A., Wiedenheft, B., 2024. Repair of CRISPR-guided RNA breaks enables site-specific RNA excision in human cells. Science 384, 808–814. https://doi.org/10.1126/science.adk5518

Nethery, M.A., Korvink, M., Makarova, K.S., Wolf, Y.I., Koonin, E.V., Barrangou, R., 2021. CRISPRclassify: Repeat-Based Classification of CRISPR Loci. The CRISPR Journal 4, 558–574. https://doi.org/10.1089/crispr.2021.0021

Newpher, T.M., Harris, S., Pringle, J., Hamilton, C., Soderling, S., 2018. Regulation of spine structural plasticity by Arc/Arg3.1. Semin Cell Dev Biol 77, 25–32. https://doi.org/10.1016/j.semcdb.2017.09.022

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msu300

Nicholls, R.A., 2011. Conformation-independent comparison of protein structures.

Nicholson, T.J., Jackson, S.A., Croft, B.I., Staals, R.H.J., Fineran, P.C., Brown, C.M., 2019. Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biology 16, 566–576. https://doi.org/10.1080/15476286.2018.1509662

Nielsen, L.D., Pedersen, C.P., Erlendsson, S., Teilum, K., 2019. The Capsid Domain of Arc Changes Its Oligomerization Propensity through Direct Interaction with the NMDA Receptor. Structure 27, 1071-1081.e5. https://doi.org/10.1016/j.str.2019.04.001

Nikolaienko, O., Patil, S., Eriksen, M.S., Bramham, C.R., 2018. Arc protein: a flexible hub for synaptic plasticity and cognition. Seminars in Cell & Developmental Biology, Arc/ARg3.1 77, 33–42. https://doi.org/10.1016/j.semcdb.2017.09.006

Novikova, I.V., Sharma, N., Moser, T., Sontag, R., Liu, Y., Collazo, M.J., Cascio, D., Shokuhfar, T., Hellmann, H., Knoblauch, M., Evans, J.E., 2018. Protein structural biology using cell-free platform from wheat germ. Adv Struct Chem Imaging 4, 13. https://doi.org/10.1186/s40679-018-0062-9

Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., Doudna, J.A., 2016. CRISPR Immunological Memory Requires a Host Factor for Specificity. Molecular Cell 62, 824–833. https://doi.org/10.1016/j.molcel.2016.04.027

Nuñez, J.K., Harrington, L.B., Kranzusch, P.J., Engelman, A.N., Doudna, J.A., 2015a. Foreign DNA capture during CRISPR-Cas adaptive immunity. Nature 527, 535–538. https://doi.org/10.1038/nature15760

Nuñez, J.K., Lee, A.S.Y., Engelman, A., Doudna, J.A., 2015b. Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature 519, 193–198. https://doi.org/10.1038/nature14237

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S.J., Diekhans, M., Logsdon, G.A., Alonge, M., Antonarakis, S.E., Borchers, M., Bouffard, G.G., Brooks, S.Y., Caldas, G.V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L.G., Dishuck, P.C., Durbin, R., Dvorkina, T., Fiddes, I.T., Formenti, G., Fulton, R.S., Fungtammasan, A., Garrison, E., Grady, P.G.S., Graves-Lindsay, T.A., Hall, I.M., Hansen, N.F., Hartley, G.A., Haukness, M., Howe, K., Hunkapiller, M.W., Jain, C., Jain, M., Jarvis, E.D., Kerpedjiev,

P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V.V., Marschall, T., McCartney, A.M., McDaniel, J., Miller, D.E., Mullikin, J.C., Myers, E.W., Olson, N.D., Paten, B., Peluso, P., Pevzner, P.A., Porubsky, D., Potapova, T., Rogaev, E.I., Rosenfeld, J.A., Salzberg, S.L., Schneider, V.A., Sedlazeck, F.J., Shafin, K., Shew, C.J., Shumate, A., Sims, Y., Smit, A.F.A., Soto, D.C., Sović, I., Storer, J.M., Streets, A., Sullivan, B.A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B.P., Wenger, A., Wood, J.M.D., Xiao, C., Yan, S.M., Young, A.C., Zarate, S., Surti, U., McCoy, R.C., Dennis, M.Y., Alexandrov, I.A., Gerton, J.L., O'Neill, R.J., Timp, W., Zook, J.M., Schatz, M.C., Eichler, E.E., Miga, K.H., Phillippy, A.M., 2022. The complete sequence of a human genome. Science 376, 44–53. https://doi.org/10.1126/science.abj6987

Nuzhat, N., Ikami, K., Abbott, H., Tanner, H., Spradling, A.C., Lei, L., 2019. Altered germline cyst and oocyte differentiation in Tex14 mutant mice reveal a new mechanism underlying female reproductive life-span. https://doi.org/10.1101/814384

Okuno, H., Minatohara, K., Bito, H., 2018. Inverse synaptic tagging: An inactive synapse-specific mechanism to capture activity-induced Arc/arg3.1 and to locally regulate spatial distribution of synaptic weights. Semin Cell Dev Biol 77, 43–50. https://doi.org/10.1016/j.semcdb.2017.09.025

Olson, E.D., Musier-Forsyth, K., 2019. Retroviral Gag protein - RNA interactions: Implications for specific genomic RNA packaging and virion assembly. Semin Cell Dev Biol 86, 129–139. https://doi.org/10.1016/j.semcdb.2018.03.015

Ono, R., Kobayashi, S., Wagatsuma, H., Aisaka, K., Kohda, T., Kaneko-Ishino, T., Ishino, F., 2001. A Retrotransposon-Derived Gene, PEG10, Is a Novel Imprinted Gene Located on Human Chromosome 7q21. Genomics 73, 232–237. https://doi.org/10.1006/geno.2001.6494

Ono, R., Kuroki, Y., Naruse, M., Ishii, M., Iwasaki, S., Toyoda, A., Fujiyama, A., Shaw, G., Renfree, M.B., Kaneko-Ishino, T., Ishino, F., 2011. Identification of tammar wallaby SIRH12, derived from a marsupial-specific retrotransposition event. DNA Res 18, 211–219. https://doi.org/10.1093/dnares/dsr012

Ono, R., Nakamura, K., Inoue, K., Naruse, M., Usami, T., Wakisaka-Saito, N., Hino, T., Suzuki-Migishima, R., Ogonuki, N., Miki, H., Kohda, T., Ogura, A., Yokoyama, M., Kaneko-Ishino, T., Ishino, F., 2006. Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38, 101–106. https://doi.org/10.1038/ng1699

Orgel, L.E., Crick, F.H.C., 1980. Selfish DNA: the ultimate parasite. Nature 284, 604–607. https://doi.org/10.1038/284604a0

Pang, S.W., Lahiri, C., Poh, C.L., Tan, K.O., 2018. PNMA family: Protein interaction network and cell signalling pathways implicated in cancer and apoptosis. Cell Signal 45, 54–62. https://doi.org/10.1016/j.cellsig.2018.01.022

Pastuzyn, E.D., Day, C.E., Kearns, R.B., Kyrke-Smith, M., Taibi, A.V., McCormick, J., Yoder, N., Belnap, D.M., Erlendsson, S., Morado, D.R., Briggs, J.A.G., Feschotte, C., Shepherd, J.D., 2018. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell 172, 275-288.e18. https://doi.org/10.1016/j.cell.2017.12.024

Payne, A., Holmes, N., Rakyan, V., Loose, M., 2018. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty841

Pertea, G., Pertea, M., 2020. GFF Utilities: GffRead and GffCompare. F1000Research. https://doi.org/10.12688/f1000research.23297.2

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. https://doi.org/10.1038/nbt.3122

Petit, M.-A., Ehrlich, D., Jannière, L., 1995. pAMβ1 resolvase has an atypical recombination site and requires a histone-like protein HU. Molecular Microbiology 18, 271–282. https://doi.org/10.1111/j.1365-2958.1995.mmi_18020271.x

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E., 2021a. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30, 70–82. https://doi.org/10.1002/pro.3943

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E., 2021b. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Science 30, 70–82. https://doi.org/10.1002/pro.3943

Peyret, H., 2015. A protocol for the gentle purification of virus-like particles produced in plants. Journal of Virological Methods 225, 59–63. https://doi.org/10.1016/j.jviromet.2015.09.005

Plath, N., Ohana, O., Dammermann, B., Errington, M.L., Schmitz, D., Gross, C., Mao, X., Engelsberg, A., Mahlke, C., Welzl, H., Kobalz, U., Stawrakakis, A., Fernandez, E., Waltereit, R., Bick-Sander, A., Therstappen, E., Cooke, S.F., Blanquet, V., Wurst, W., Salmen, B., Bösl, M.R., Lipp, H.-P., Grant, S.G.N., Bliss, T.V.P., Wolfer, D.P., Kuhl, D., 2006. Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. Neuron 52, 437–444. https://doi.org/10.1016/j.neuron.2006.08.024

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.-P., Couvin, D., Toffano-Nioche, C., Vergnaud, G., 2020. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Research 48, D535–D544. https://doi.org/10.1093/nar/gkz915

Prakash, A., Jeffryes, M., Bateman, A., Finn, R.D., 2017. The HMMER Web Server for Protein Sequence Similarity Search. Current Protocols in Bioinformatics 60, 3.15.1-3.15.23. https://doi.org/10.1002/cpbi.40

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5, e9490. https://doi.org/10.1371/journal.pone.0009490

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 35, D61–D65. https://doi.org/10.1093/nar/gkl842

Pucker, B., Holtgräwe, D., Stadermann, K.B., Frey, K., Huettel, B., Reinhardt, R., Weisshaar, B., 2019. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLOS ONE 14, e0216233. https://doi.org/10.1371/journal.pone.0216233

Punjani, A., Rubinstein, J.L., Fleet, D.J., Brubaker, M.A., 2017a. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Methods 14, 290–296. https://doi.org/10.1038/nmeth.4169

Punjani, A., Rubinstein, J.L., Fleet, D.J., Brubaker, M.A., 2017b. CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. Nature Methods 14, 290–296. https://doi.org/10.1038/nmeth.4169

Punjani, A., Zhang, H., Fleet, D.J., 2020. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. Nat Methods 17, 1214–1221. https://doi.org/10.1038/s41592-020-00990-8

Qu, K., Glass, B., Doležal, M., Schur, F.K.M., Murciano, B., Rein, A., Rumlová, M., Ruml, T., Kräusslich, H.-G., Briggs, J.A.G., 2018. Structure and architecture of immature and mature murine leukemia virus capsids. PNAS 115, E11751–E11760. https://doi.org/10.1073/pnas.1811580115

Ramachandran, A., Summerville, L., Learn, B.A., DeBell, L., Bailey, S., 2020. Processing and integration of functionally oriented prespacers in the Escherichia coli CRISPR system depends on bacterial host exonucleases. Journal of Biological Chemistry 295, 3403–3414. https://doi.org/10.1074/jbc.RA119.012196

Ranwez, V., Douzery, E.J.P., Cambon, C., Chantret, N., Delsuc, F., 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Molecular Biology and Evolution 35, 2582–2584. https://doi.org/10.1093/molbev/msy159

Razmjou, S., Abdulnour, J., Bastard, J.-P., Fellahi, S., Doucet, É., Brochu, M., Lavoie, J.-M., Rabasa-Lhoret, R., Prud'homme, D., 2018. Body composition, cardiometabolic risk factors, physical activity, and inflammatory markers in premenopausal women after a 10-year

follow-up: a MONET study. Menopause 25, 89. https://doi.org/10.1097/GME.0000000000000951

Resemann, H.C., Herrfurth, C., Feussner, K., Hornung, E., Ostendorf, A.K., Gömann, J., Mittag, J., Van Gessel, N., Vries, J.D., Ludwig-Müller, J., Markham, J., Reski, R., Feussner, I., 2021. Convergence of sphingolipid desaturation across over 500 million years of plant evolution. Nature Plants. https://doi.org/10.1038/s41477-020-00844-3

Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., Allen, J., Asri, M., Bzikadze, A.V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., Garcia Giron, C., Garrison, E., Gershman, A., Gerton, J.L., Grady, P.G.S., Guarracino, A., Haggerty, L., Halabian, R., Hansen, N.F., Harris, R., Hartley, G.A., Harvey, W.T., Haukness, M., Heinz, J., Hourlier, T., Hubley, R.M., Hunt, S.E., Hwang, S., Jain, M., Kesharwani, R.K., Lewis, A.P., Li, H., Logsdon, G.A., Lucas, J.K., Makalowski, W., Markovic, C., Martin, F.J., Mc Cartney, A.M., McCoy, R.C., McDaniel, J., McNulty, B.M., Medvedev, P., Mikheenko, A., Munson, K.M., Murphy, T.D., Olsen, H.E., Olson, N.D., Paulin, L.F., Porubsky, D., Potapova, T., Ryabov, F., Salzberg, S.L., Sauria, M.E.G., Sedlazeck, F.J., Shafin, K., Shepelev, V.A., Shumate, A., Storer, J.M., Surapaneni, L., Taravella Oill, A.M., Thibaud-Nissen, F., Timp, W., Tomaszkiewicz, M., Vollger, M.R., Walenz, B.P., Watwood, A.C., Weissensteiner, M.H., Wenger, A.M., Wilson, M.A., Zarate, S., Zhu, Y., Zook, J.M., Eichler, E.E., O'Neill, R.J., Schatz, M.C., Miga, K.H., Makova, K.D., Phillippy, A.M., 2023. The complete sequence of a human Y chromosome. Nature 621, 344–354. https://doi.org/10.1038/s41586-023-06457-y

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Rice, P.A., Yang, S., Mizuuchi, K., Nash, H.A., 1996. Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn. Cell 87, 1295–1306. https://doi.org/10.1016/S0092-8674(00)81824-3

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N.J., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H.J., Fineran, P.C., 2014. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Research 42, 8516–8526. https://doi.org/10.1093/nar/gku527

Robinson, H.L., Astrin, S.M., Senior, A.M., Salazar, F.H., 1981. Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. Journal of Virology 40, 745–751. https://doi.org/10.1128/jvi.40.3.745-751.1981

Rocha, E.P.C., Bikard, D., 2022. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? PLOS Biology 20, e3001514. https://doi.org/10.1371/journal.pbio.3001514

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., Mann, R.S., 2010. Origins of specificity in protein-DNA recognition. Annual Review of Biochemistry 79, 233–269. https://doi.org/10.1146/annurev-biochem-060408-091030

Rohwer, F., Azam, F., 2000. Detection of DNA Damage in Prokaryotes by Terminal Deoxyribonucleotide Transferase-Mediated dUTP Nick End Labeling. Applied and Environmental Microbiology 66, 1001–1006. https://doi.org/10.1128/AEM.66.3.1001-1006.2000

Rojo, F., Alonso, J.C., 1995. The β recombinase of plasmid pSM19035 binds to two adjacent sites, making different contacts at each of them. Nucl Acids Res 23, 3181–3188. https://doi.org/10.1093/nar/23.16.3181

Rollie, C., Graham, S., Rouillon, C., White, M.F., 2018. Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Research 46, 1007–1020. https://doi.org/10.1093/nar/gkx1232

Rollie, C., Schneider, S., Brinkmann, A.S., Bolt, E.L., White, M.F., 2015. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. eLife 4, 1–19. https://doi.org/10.7554/eLife.08716

Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C., Wiedenheft, B., 2017. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proceedings of the National Academy of Sciences 114, 201616395. https://doi.org/10.1073/pnas.1616395114

Rosario, R., Smith, R.W.P., Adams, I.R., Anderson, R.A., 2017. RNA immunoprecipitation identifies novel targets of DAZL in human foetal ovary. Molecular Human Reproduction 23, 177–186. https://doi.org/10.1093/molehr/gax004

Rousset, F., Depardieu, F., Miele, S., Dowding, J., Laval, A.-L., Lieberman, E., Garry, D., Rocha, E.P.C., Bernheim, A., Bikard, D., 2022. Phages and their satellites encode hotspots of antiviral systems. Cell Host & Microbe 30, 740-753.e5. https://doi.org/10.1016/j.chom.2022.02.018

Rowland, S.-J., Stark, W.M., Boocock, M.R., 2002. Sin recombinase from Staphylococcus aureus: synaptic complex architecture and transposon targeting: Sin recombinase. Molecular Microbiology 44, 607–619. https://doi.org/10.1046/j.1365-2958.2002.02897.x

Rozewicki, J., Li, S., Amada, K.M., Standley, D.M., Katoh, K., 2019. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res 47, W5–W10. https://doi.org/10.1093/nar/gkz342

Ruperti, F., Papadopoulos, N., Musser, J., Arendt, D., 2022. Beyond sequence similarity: cross-phyla protein annotation by structural prediction and alignment. https://doi.org/10.1101/2022.07.05.498892

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S.A., Sørensen, S.J., 2020. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. The CRISPR Journal 3, 462–469. https://doi.org/10.1089/crispr.2020.0059

Ruth, K.S., Day, F.R., Tyrrell, J., Thompson, D.J., Wood, A.R., Mahajan, A., Beaumont, R.N., Wittemans, L., Martin, S., Busch, A.S., Erzurumluoglu, A.M., Hollis, B., O'Mara, T.A., McCarthy, M.I., Langenberg, C., Easton, D.F., Wareham, N.J., Burgess, S., Murray, A., Ong, K.K., Frayling, T.M., Perry, J.R.B., 2020. Using human genetics to understand the disease impacts of testosterone in men and women. Nat Med 26, 252–258. https://doi.org/10.1038/s41591-020-0751-5

Sagendorf, J.M., Markarian, N., Berman, H.M., Rohs, R., 2019. DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. Nucleic Acids Research gkz889. https://doi.org/10.1093/nar/gkz889

Salehnia, M., Zavareh, S., 2013. The Effects of Progesterone on Oocyte Maturation and Embryo Development. Int J Fertil Steril 7, 74–81.

Sambrook, J., 1989. Molecular Cloning: A Laboratory Manual, 2nd edition. ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanchez-Garcia, R., Gomez-Blanco, J., Cuervo, A., Carazo, J.M., Sorzano, C.O.S., Vargas, J., 2021. DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. Commun Biol 4, 1–8. https://doi.org/10.1038/s42003-021-02399-1

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74, 5463–5467.

Santiago-Frangos, A., Buyukyoruk, M., Wiegand, T., Krishna, P., Wiedenheft, B., 2021a. Distribution and phasing of sequence motifs that facilitate CRISPR adaptation. Current Biology 31, 3515-3524.e6. https://doi.org/10.1016/j.cub.2021.05.068

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

Santiago-Frangos, A., Hall, L.N., Nemudraia, A., Nemudryi, A., Krishna, P., Wiegand, T., Wilkinson, R.A., Snyder, D.T., Hedges, J.F., Cicha, C., Lee, H.H., Graham, A., Jutila, M.A., Taylor, M.P., Wiedenheft, B., 2021c. Intrinsic signal amplification by type III CRISPR-Cas systems provides a sequence-specific SARS-CoV-2 diagnostic. CR Med 2. https://doi.org/10.1016/j.xcrm.2021.100319

Scheffler, K., Giannini, F., Lemonnier, T., Mogessie, B., 2022. The prophase oocyte nucleus is a homeostatic G-actin buffer. J. Cell Sci 135, 259807. https://doi.org/10.1242/jcs.259807.

Scheres, S.H.W., 2012. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. Journal of Structural Biology 180, 519–530. https://doi.org/10.1016/j.jsb.2012.09.006

Scheres, S.H.W., Chen, S., 2012. Prevention of overfitting in cryo-EM structure determination. Nat Methods 9, 853–854. https://doi.org/10.1038/nmeth.2115

Schiltz, C.J., Adams, M.C., Chappie, J.S., 2020. The full-length structure of Thermus scotoductus OLD defines the ATP hydrolysis properties and catalytic mechanism of Class 1 OLD family nucleases. Nucleic Acids Research 48, 2762–2776. https://doi.org/10.1093/nar/gkaa059

Schiltz, C.J., Lee, A., Partlow, E.A., Hosford, C.J., Chappie, J.S., 2019. Structural characterization of Class 2 OLD family nucleases supports a two-metal catalysis mechanism for cleavage. Nucleic Acids Research 47, 9448–9463. https://doi.org/10.1093/nar/gkz703

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

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: A new way to display consensus sequences. Nucleic Acids Research 18, 6097–6100. https://doi.org/10.1093/nar/18.20.6097

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., Fulton, R.S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K.M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J.T., Threadgold, G., Torrance, J., Wood, J.M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A.M., Durbin, R., Wilson, R.K., Flicek, P., Eichler, E.E., Church, D.M., 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 27, 849–864. https://doi.org/10.1101/gr.213611.116

Schomberg, D.W., Couse, J.F., Mukherjee, A., Lubahn, D.B., Sar, M., Mayo, K.E., Korach, K.S., 1999. Targeted Disruption of the Estrogen Receptor-α Gene in Female Mice: Characterization of Ovarian Responses and Phenotype in the Adult*. Endocrinology 140, 2733–2744. https://doi.org/10.1210/endo.140.6.6823

Schüller, M., Jenne, D., Voltz, R., 2005. The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J. Neuroimmunol. 169, 172–176. https://doi.org/10.1016/j.jneuroim.2005.08.019

Segel, M., Lash, B., Song, J., Ladha, A., Liu, C.C., Jin, X., Mekhedov, S.L., Macrae, R.K., Koonin, E.V., Zhang, F., 2021. Mammalian retrovirus-like protein PEG10 packages its own mRNA and can be pseudotyped for mRNA delivery. Science 373, 882–889. https://doi.org/10.1126/science.abg6155

Seifert, F.U., Lammens, K., Stoehr, G., Kessler, B., Hopfner, K., 2016. Structural mechanism of ATP-dependent DNA binding and DNA end bridging by eukaryotic Rad50. The EMBO Journal 35, 759–772. https://doi.org/10.15252/embj.201592934

Sekita, Y., Wagatsuma, H., Nakamura, K., Ono, R., Kagami, M., Wakisaka, N., Hino, T., Suzuki-Migishima, R., Kohda, T., Ogura, A., Ogata, T., Yokoyama, M., Kaneko-Ishino, T., Ishino, F., 2008. Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. Nat Genet 40, 243–248. https://doi.org/10.1038/ng.2007.51

Semenova, E., Savitskaya, E., Musharova, O., Strotskaya, A., Vorontsova, D., Datsenko, K.A., Logacheva, M.D., Severinov, K., 2016. Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. Proceedings of the National Academy of Sciences 113, 7626–7631. https://doi.org/10.1073/pnas.1602639113

Shao, C., Sun, S., Liu, K., Wang, Jiahao, Li, S., Liu, Q., Deagle, B.E., Seim, I., Biscontin, A., Wang, Q., Liu, X., Kawaguchi, S., Liu, Yalin, Jarman, S., Wang, Yue, Wang, H.-Y., Huang, G., Hu, J., Feng, B., De Pittà, C., Liu, Shanshan, Wang, R., Ma, K., Ying, Y., Sales, G., Sun, T., Wang, X., Zhang, Y., Zhao, Y., Pan, S., Hao, X., Wang, Yang, Xu, J., Yue, B., Sun, Y., Zhang, H., Xu, M., Liu, Yuyan, Jia, X., Zhu, J., Liu, Shufang, Ruan, J., Zhang, G., Yang, H., Xu, X., Wang, Jun, Zhao, X., Meyer, B., Fan, G., 2023. The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. Cell 186, 1279-1294.e19. https://doi.org/10.1016/j.cell.2023.02.005

Shigemoto, K., Brennan, J., Walls, E., Watson, C.J., Stott, D., Rigby, P.W., Reith, A.D., 2001. Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting. Nucleic Acids Res 29, 4079–4088. https://doi.org/10.1093/nar/29.19.4079

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., Ueda, T., 2001. Cell-free translation reconstituted with purified components. Nat Biotechnol 19, 751–755. https://doi.org/10.1038/90802

Shipman, S.L., Nivala, J., Macklis, J.D., Church, G.M., 2016a. Molecular recordings by directed CRISPR spacer acquisition. Science 353, aaf1175. https://doi.org/10.1126/science.aaf1175

Shipman, S.L., Nivala, J., Macklis, J.D., Church, G.M., 2016b. Molecular recordings by directed CRISPR spacer acquisition. Science 353, aaf1175. https://doi.org/10.1126/science.aaf1175

Shiura, H., Ono, R., Tachibana, S., Kohda, T., Kaneko-Ishino, T., Ishino, F., 2021. PEG10 viral aspartic protease domain is essential for the maintenance of fetal capillary structure in the mouse placenta. Development 148, dev199564. https://doi.org/10.1242/dev.199564

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv351

Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M.Q., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15, 161–168. https://doi.org/10.1038/nrmicro.2016.177

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., Havulinna, A.S., Pirruccello, J.P., Qian, J., Shcherbina, A., Rodriguez, F., Assimes, T.L., Agarwala, V., Tibshirani, R., Hastie, T., Ripatti, S., Pritchard, J.K., Daly, M.J., Rivas, M.A., 2021. Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat Genet 53, 185–194. https://doi.org/10.1038/s41588-020-00757-z

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., Huang, J., Du, Y., Lin, L., Dong, Z., Yang, L., Liu, X., Jiang, H., Xu, X., Wang, J., Yang, H., Bolund, L., Mardinoglu, A., Zhang, C., von Feilitzen, K., Lindskog, C., Pontén, F., Luo, Y., Hökfelt, T., Uhlén, M., Mulder, J., 2020. An atlas of the protein-coding genes in the human, pig, and mouse brain. Science 367, eaay5947. https://doi.org/10.1126/science.aay5947

Smit, A., Hubley, R., Green, P., 2013. RepeatMasker Open-4.0.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9, 657–663. https://doi.org/10.1016/s0959-437x(99)00031-3

Söding, J., Biegert, A., Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Research 33, W244–W248. https://doi.org/10.1093/nar/gki408

Sorek, R., Lawrence, C.M., Wiedenheft, B., 2013. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annual Review of Biochemistry 82, 237–266. https://doi.org/10.1146/annurev-biochem-072911-172315

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., Dym, M., de Massy, B., Mikkelsen, T.S., Kaessmann, H., 2013. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Reports 3, 2179–2190. https://doi.org/10.1016/j.celrep.2013.05.031

Sousa Martins, J.P., Liu, X., Oke, A., Arora, R., Franciosi, F., Viville, S., Laird, D.J., Fung, J.C., Conti, M., 2016. DAZL and CPEB1 regulate mRNA translation synergistically during oocyte maturation. J Cell Sci 129, 1271–1282. https://doi.org/10.1242/jcs.179218

Sperber, G.O., Airola, T., Jern, P., Blomberg, J., 2007. Automated recognition of retroviral sequences in genomic data—RetroTector©. Nucleic Acids Research 35, 4964–4976. https://doi.org/10.1093/nar/gkm515

Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F., 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12, 2. https://doi.org/10.1186/s13100-020-00230-y

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J.L., Makarova, K.S., Koonin, E.V., Zhang, F., 2019. RNA-guided DNA insertion with CRISPR-associated transposases. Science 365, 48–53. https://doi.org/10.1126/science.aax9181

Sugiura, K., Su, Y.-Q., Eppig, J.J., 2010. Does Bone Morphogenetic Protein 6 (BMP6) Affect Female Fertility in the Mouse?1. Biology of Reproduction 83, 997–1004. https://doi.org/10.1095/biolreprod.110.086777

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., Carragher, B., 2005. Automated molecular microscopy: The new Leginon system. Journal of Structural Biology 151, 41–60. https://doi.org/10.1016/j.jsb.2005.03.010

Suttle, C.A., 2005. Viruses in the sea. Nature 437, 356–361. https://doi.org/10.1038/nature04160

Suttle, C.A., 1994. The significance of viruses to mortality in aquatic microbial communities. Microb Ecol 28, 237–243. https://doi.org/10.1007/BF00166813

Sutton, J.M., Millwood, J.D., Case Mccormack, A., Fierst, J.L., 2021. Optimizing experimental design for genome sequencing and assembly with Oxford Nanopore Technologies. Gigabyte. https://doi.org/10.46471/gigabyte.27

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. https://doi.org/10.1093/nar/gkl315

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., the UniProt Consortium, 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932. https://doi.org/10.1093/bioinformatics/btu739

Suzuki, A., Yanagisawa, M., Greene, R.W., 2020. Loss of Arc attenuates the behavioral and molecular responses for sleep homeostasis in mice. Proceedings of the National Academy of Sciences 117, 10547–10553. https://doi.org/10.1073/pnas.1906840117

Suzuki, S., Ono, R., Narita, T., Pask, A.J., Shaw, G., Wang, C., Kohda, T., Alsop, A.E., Graves, J.A.M., Kohara, Y., Ishino, F., Renfree, M.B., Kaneko-Ishino, T., 2007. Retrotransposon Silencing by DNA Methylation Can Drive Mammalian Genomic Imprinting. PLOS Genetics 3, e55. https://doi.org/10.1371/journal.pgen.0030055

Takaji, M., Komatsu, Y., Watakabe, A., Hashikawa, T., Yamamori, T., 2009. Paraneoplastic Antigen-Like 5 Gene (PNMA5) Is Preferentially Expressed in the Association Areas in a Primate Specific Manner. Cerebral Cortex 19, 2865–2879. https://doi.org/10.1093/cercor/bhp062

Tan, C.T., Zhou, Q.-L., Su, Y.-C., Fu, N.Y., Chang, H.-C., Tao, R.N., Sukumaran, S.K., Baksh, S., Tan, Y.-J., Sabapathy, K., Yu, C.-D., Yu, V.C., 2016. MOAP-1 Mediates Fas-Induced Apoptosis in Liver by Facilitating tBid Recruitment to Mitochondria. Cell Reports 16, 174–185. https://doi.org/10.1016/j.celrep.2016.05.068

Tan, K.O., Fu, N.Y., Sukumaran, S.K., Chan, S.-L., Kang, J.H., Poon, K.L., Chen, B.S., Yu, V.C., 2005. MAP-1 is a mitochondrial effector of Bax. PNAS 102, 14623–14628. https://doi.org/10.1073/pnas.0503524102

Tan, K.O., Tan, K.M., Chan, S.L., Yee, K.S., Bevort, M., Ang, K.C., Yu, V.C., 2001. MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. J Biol Chem 276, 2802–2807. https://doi.org/10.1074/jbc.M008955200

Tan, K.W.M., Lee, Y.K., 2016. The dilemma for lipid productivity in green microalgae: importance of substrate provision in improving oil yield without sacrificing growth. Biotechnology for Biofuels. https://doi.org/10.1186/s13068-016-0671-2

Tan, Y.Z., Baldwin, P.R., Davis, J.H., Williamson, J.R., Potter, C.S., Carragher, B., Lyumkis, D., 2017. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nat Methods 14, 793–796. https://doi.org/10.1038/nmeth.4347

Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. https://doi.org/10.1002/0471250953.bi0410s25

Taylor, G.M., Gao, Y., Sanders, D.A., 2001. Fv-4: identification of the defect in Env and the mechanism of resistance to ecotropic murine leukemia virus. J Virol 75, 11244–11248. https://doi.org/10.1128/JVI.75.22.11244-11248.2001

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., Bernheim, A., 2022. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun 13, 2561. https://doi.org/10.1038/s41467-022-30269-9

Torsvik, T., Dundas, I.D., 1974. Bacteriophage of Halobacterium salinarium. Nature 248, 680–681. https://doi.org/10.1038/248680a0

Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13, 36–46. https://doi.org/10.1038/nrg3117

Udrițe, A.-A., Iordachescu, M., Ciceoi, R., Bădulescu, L., 2022. Next-Generation Sequencing of Local Romanian Tomato Varieties and Bioinformatics Analysis of the Ve Locus. International Journal of Molecular Sciences 23, 9750. https://doi.org/10.3390/ijms23179750

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C.A.-K., Odeberg, J., Djureinovic, D., Takanen, J.O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., Pontén, F., 2015. Tissue-based map of the human proteome. Science 347, 1260419. https://doi.org/10.1126/science.1260419

Vadlamani, A., Pendyala, B., Viamajala, S., Varanasi, S., 2019. High Productivity Cultivation of Microalgae without Concentrated CO2 Input. ACS Sustainable Chemistry & Engineering. https://doi.org/10.1021/acssuschemeng.8b04094

Vadlamani, A., Viamajala, S., Pendyala, B., Varanasi, S., 2017. Cultivation of Microalgae at Extreme Alkaline pH Conditions: A Novel Approach for Biofuel Production. ACS Sustainable Chemistry & Engineering. https://doi.org/10.1021/acssuschemeng.7b01534

van den Berg, D.F., van der Steen, B.A., Costa, A.R., Brouns, S.J., 2023. Phage tRNAs evade tRNA-targeting host defenses through anticodon loop mutations. eLife 12, e85183. https://doi.org/10.7554/eLife.85183

van Gent, D.C., Hiom, K., Paull, T.T., Gellert, M., 1997. Stimulation of V(D)J cleavage by high mobility group proteins. The EMBO Journal 16, 2665–2670. https://doi.org/10.1093/emboj/16.10.2665

van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., Steinegger, M., 2023. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 1–4. https://doi.org/10.1038/s41587-023-01773-0

van Riet, J., Saha, C., Strepis, N., Brouwer, R.W.W., Martens-Uzunova, E.S., van de Geer, W.S., Swagemakers, S.M.A., Stubbs, A., Halimi, Y., Voogd, S., Tanmoy, A.M., Komor, M.A., Hoogstrate, Y., Janssen, B., Fijneman, R.J.A., Niknafs, Y.S., Chinnaiyan, A.M., van IJcken, W.F.J., van der Spek, P.J., Jenster, G., Louwen, R., 2022. CRISPRs in the human genome are differentially expressed between malignant and normal adjacent to tumor tissue. Commun Biol 5, 1–13. https://doi.org/10.1038/s42003-022-03249-4

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space

with high-accuracy models. Nucleic Acids Research 50, D439–D444. https://doi.org/10.1093/nar/gkab1061

Varela, M., Spencer, T.E., Palmarini, M., Arnaud, F., 2009. Friendly Viruses. Annals of the New York Academy of Sciences 1178, 157–172. https://doi.org/10.1111/j.1749-6632.2009.05002.x

Vargiu, L., Rodriguez-Tomé, P., Sperber, G.O., Cadeddu, M., Grandi, N., Blikstad, V., Tramontano, E., Blomberg, J., 2016. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology 13, 7. https://doi.org/10.1186/s12977-015-0232-y

Varshney, P., Mikulic, P., Vonshak, A., Beardall, J., Wangikar, P.P., 2015. Extremophilic micro-algae and their potential contribution in biotechnology. Bioresour Technol. https://doi.org/10.1016/j.biortech.2014.11.040

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Miklos, G.L.G., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V.D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.-R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z.Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S.C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph,

S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A.D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The Sequence of the Human Genome. Science 291, 1304–1351. https://doi.org/10.1126/science.1058040

Villanueva, R.A.M., Chen, Z.J., 2019. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). Measurement: Interdisciplinary Research and Perspectives 17, 160–167. https://doi.org/10.1080/15366367.2019.1565254

Vinarov, D.A., Lytle, B.L., Peterson, F.C., Tyler, E.M., Volkman, B.F., Markley, J.L., 2004. Cell-free protein production and labeling protocol for NMR-based structural proteomics. Nat Methods 1, 149–153. https://doi.org/10.1038/nmeth716

Volff, J.-N., 2009. Cellular genes derived from Gypsy/Ty3 retrotransposons in mammalian genomes. Annals of the New York Academy of Sciences 1178, 233–243. https://doi.org/10.1111/j.1749-6632.2009.05005.x.

Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, K., Semenova, E., 2015. Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. Nucleic Acids Research 43, 10848–10860. https://doi.org/10.1093/nar/gkv1261

Vos, M.D., Dallol, A., Eckfeld, K., Allen, N.P.C., Donninger, H., Hesson, L.B., Calvisi, D., Latif, F., Clark, G.J., 2006. The RASSF1A tumor suppressor activates Bax via MOAP-1. J Biol Chem 281, 4557–4563. https://doi.org/10.1074/jbc.M512128200

Walker, J.C.G., 1980. The Oxygen Cycle. https://doi.org/10.1007/978-3-662-24940-6_5

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

Walkinshaw, M.D., Taylor, P., Sturrock, S.S., Atanasiu, C., Berge, T., Henderson, R.M., Edwardson, J.M., Dryden, D.T.F., 2002. Structure of Ocr from Bacteriophage T7, a Protein that Mimics B-Form DNA. Molecular Cell 9, 187–194. https://doi.org/10.1016/S1097-2765(02)00435-5

Wang, H., Huang, Z., Lee, Y., Song, X., Shen, X., Shu, C., Wu, L.H., Pakkiri, L.S., Lim, P.L., Zhang, X., Drum, C.L., Zhu, J., Li, R., 2024. Rejuvenation of aged oocyte through exposure to young follicular microenvironment. https://doi.org/10.1101/2024.03.11.584343

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., Wang, Y., 2015. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell 163, 840–853. https://doi.org/10.1016/j.cell.2015.10.008

Wang, Jianbin, Jiang, Y., Vincent, M., Sun, Y., Yu, H., Wang, Jing, Bao, Q., Kong, H., Hu, S., 2005. Complete genome sequence of bacteriophage T5. Virology 332, 45–65. https://doi.org/10.1016/j.virol.2004.10.049

Wang, R., Li, M., Gong, L., Hu, S., Xiang, H., 2016. DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. Nucleic Acids Research 44, 4266–4277. https://doi.org/10.1093/nar/gkw260

Wang, X., Yao, D., Xu, J.G., Li, A.R., Xu, J., Fu, P., Zhou, Y., Zhu, Y., 2016. Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. Nature Structural and Molecular Biology 23, 868–870. https://doi.org/10.1038/nsmb.3269

Wang, Yunhao, Zhao, Y., Bollas, A., Wang, Yuru, Au, K.F., 2021. Nanopore sequencing technology, bioinformatics and applications. Nature Biotechnology. https://doi.org/10.1038/s41587-021-01108-x

Wei, M., Zhao, X., Liu, M., Huang, Z., Xiao, Y., Niu, M., Shao, Y., Kleiman, L., 2015. Inhibition of HIV-1 assembly by coiled-coil domain containing protein 8 in human cells. Sci Rep 5, 14724. https://doi.org/10.1038/srep14724

Wei, Y., Chesne, M.T., Terns, R.M., Terns, M.P., 2015. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Research 43, 1749–1758. https://doi.org/10.1093/nar/gku1407

Wein, T., Sorek, R., 2022. Bacterial origins of human cell-autonomous innate immune mechanisms. Nat Rev Immunol 22, 629–638. https://doi.org/10.1038/s41577-022-00705-4

Weiß, M., Giacomelli, G., Assaya, M.B., Grundt, F., Haouz, A., Peng, F., Petrella, S., Wehenkel, A.M., Bramkamp, M., 2023. The MksG nuclease is the executing part of the bacterial plasmid defense system MksBEFG. Nucleic Acids Research 51, 3288–3306. https://doi.org/10.1093/nar/gkad130

Weissenberger, G., Henderikx, R.J.M., Peters, P.J., 2021. Understanding the invisible hands of sample preparation for cryo-EM. Nat Methods 18, 463–471. https://doi.org/10.1038/s41592-021-01130-6

Wells, J.N., Feschotte, C., 2020. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet 54, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Westra, E.R., van Houte, S., Oyesiku-Blakemore, S., Makin, B., Broniewski, J.M., Best, A., Bondy-Denomy, J., Davidson, A., Boots, M., Buckling, A., 2015. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. Current Biology 25, 1043–1049. https://doi.org/10.1016/j.cub.2015.01.065

Weyrich, L.S., Farrer, A.G., Eisenhofer, R., Arriola, L.A., Young, J., Selway, C.A., Handsley-Davis, M., Adler, C.J., Breen, J., Cooper, A., 2019. Laboratory contamination over time during low-biomass sample analysis. Molecular Ecology Resources. https://doi.org/10.1111/1755-0998.13011

Whitfield, C.R., Wardle, S.J., Haniford, D.B., 2009. The global bacterial regulator H-NS promotes transpososome formation and transposition in the Tn5 system. Nucleic Acids Research 37, 309–321. https://doi.org/10.1093/nar/gkn935

Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E., 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Computational Biology 13, e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., Doudna, J.A., 2009. Structural Basis for DNase Activity of a Conserved Protein Implicated in CRISPR-Mediated Genome Defense. Structure 17, 904–912. https://doi.org/10.1016/j.str.2009.03.019

Wiegand, T., Semenova, E., Shiriaeva, A., Fedorov, I., Datsenko, K., Severinov, K., Wiedenheft, B., 2020a. Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. The CRISPR Journal 3, 378–387. https://doi.org/10.1089/crispr.2020.0069

Wiegand, T., Semenova, E., Shiriaeva, A., Fedorov, I., Datsenko, K., Severinov, K., Wiedenheft, B., 2020b. Reproducible Antigen Recognition by the Type I-F CRISPR-Cas System. The CRISPR Journal 3, 378–387. https://doi.org/10.1089/crispr.2020.0069

Wiegand, T., Wilkinson, R., Santiago-Frangos, A., Lynes, M., Hatzenpichler, R., Wiedenheft, B., 2023. Functional and Phylogenetic Diversity of Cas10 Proteins. The CRISPR Journal 6, 152–162. https://doi.org/10.1089/crispr.2022.0085

Wilkinson, M., Troman, L.A., Wan Nur Ismah, W.A., Chaban, Y., Avison, M.B., Dillingham, M.S., Wigley, D.B., 2016. Structural basis for the inhibition of RecBCD by Gam and its synergistic antibacterial effect with quinolones. eLife 5, e22963. https://doi.org/10.7554/eLife.22963

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., Jain, S., Lewis, S.M., Arendall, W.B., Snoeyink, J., Adams, P.D., Lovell, S.C., Richardson, J.S., Richardson, D.C., 2018. MolProbity: More and better reference data for improved all-atom structure validation: PROTEIN SCIENCE.ORG. Protein Science 27, 293–315. https://doi.org/10.1002/pro.3330

Wills, N.M., Moore, B., Hammer, A., Gesteland, R.F., Atkins, J.F., 2006. A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. J Biol Chem 281, 7082–7088. https://doi.org/10.1074/jbc.M511629200

Wright, A.V., Doudna, J.A., 2016. Protecting genome integrity during CRISPR immune adaptation. Nature Structural & Molecular Biology 23, 876–883. https://doi.org/10.1038/nsmb.3289

Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., Doudna, J.A., 2017a. Structures of the CRISPR genome integration complex. Science 357, 1113–1118. https://doi.org/10.1126/science.aao0679

Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., Doudna, J.A., 2017b. Structures of the CRISPR genome integration complex. Science 357, 1113–1118. https://doi.org/10.1126/science.aao0679

Wu, M., Lander, G.C., Herzik, M.A., 2020. Sub-2 Angstrom resolution structure determination using single-particle cryo-EM at 200 keV. Journal of Structural Biology: X 4, 100020. https://doi.org/10.1016/j.yjsbx.2020.100020

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

Xiao, Y, Xiao, Yibei, Luo, M., Dolan, A.E., Liao, M., Ke, A., 2018. Structure basis for RNA-guided DNA degradation by Cascade and Cas3 0839, 1–12. https://doi.org/10.1126/science.aat0839

Xie, T., Zhang, Z., Fang, Q., Du, B., Gong, X., 2021. Structural basis of substrate recognition and translocation by human ABCA4. Nat Commun 12, 3853. https://doi.org/10.1038/s41467-021-24194-6

Xu, J., Erlendsson, S., Singh, M., Holling, G.A., Regier, M., Ibiricu, I., Einstein, J., Hantak, M.P., Day, G.S., Piquet, A.L., Smith, T.L., Clardy, S.L., Whiteley, A.M., Feschotte, C., Briggs, J.A.G., Shepherd, J.D., 2024. PNMA2 forms immunogenic non-enveloped virus-like capsids associated with paraneoplastic neurological syndrome. Cell 187, 831-845.e19. https://doi.org/10.1016/j.cell.2024.01.009

Yan, C., Wang, P., DeMayo, J., DeMayo, F.J., Elvin, J.A., Carino, C., Prasad, S.V., Skinner, S.S., Dunbar, B.S., Dube, J.L., Celeste, A.J., Matzuk, M.M., 2001. Synergistic Roles of Bone Morphogenetic Protein 15 and Growth Differentiation Factor 9 in Ovarian Function. Molecular Endocrinology 15, 854–866. https://doi.org/10.1210/mend.15.6.0662

Yang, C.-R., Rajkovic, G., Daldello, E.M., Luong, X.G., Chen, J., Conti, M., 2020. The RNA-binding protein DAZL functions as repressor and activator of mRNA translation during oocyte maturation. Nat Commun 11, 1399. https://doi.org/10.1038/s41467-020-15209-9

Yang, F., Wang, P.J., 2016. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. Seminars in Cell & Developmental Biology, Molecular Mechanisms in Spermatogenesis 59, 118–125. https://doi.org/10.1016/j.semcdb.2016.03.001

Yang, J.Y., Fang, W., Miranda-Sanchez, F., Brown, J.M., Kauffman, K.M., Acevero, C.M., Bartel, D.P., Polz, M.F., Kelly, L., 2021. Degradation of host translational machinery drives tRNA acquisition in viruses. cels 12, 771-779.e5. https://doi.org/10.1016/j.cels.2021.05.019

Yang, M., Bozdagi, O., Scattoni, M.L., Wöhr, M., Roullet, F.I., Katz, A.M., Abrams, D.N., Kalikhman, D., Simon, H., Woldeyohannes, L., Zhang, J.Y., Harris, M.J., Saxena, R., Silverman, J.L., Buxbaum, J.D., Crawley, J.N., 2012. Reduced Excitatory Neurotransmission and Mild Autism-Relevant Phenotypes in Adolescent Shank3 Null Mutant Mice. J. Neurosci. 32, 6525–6541. https://doi.org/10.1523/JNEUROSCI.6107-11.2012

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24, 1586–1591. https://doi.org/10.1093/molbev/msm088

Yap, M.W., Colbeck, E., Ellis, S.A., Stoye, J.P., 2014. Evolution of the Retroviral Restriction Gene Fv1: Inhibition of Non-MLV Retroviruses. PLOS Pathogens 10, e1003968. https://doi.org/10.1371/journal.ppat.1003968

Yip, K.M., Fischer, N., Paknia, E., Chari, A., Stark, H., 2020. Atomic-resolution protein structure determination by cryo-EM. Nature 587, 157–161. https://doi.org/10.1038/s41586-020-2833-4

Yirmiya, E., Leavitt, A., Lu, A., Ragucci, A.E., Avraham, C., Osterman, I., Garb, J., Antine, S.P., Mooney, S.E., Hobbs, S.J., Kranzusch, P.J., Amitai, G., Sorek, R., 2024. Phages overcome bacterial immunity via diverse anti-defence proteins. Nature 625, 352–359. https://doi.org/10.1038/s41586-023-06869-w

Yosef, I., Goren, M.G., Qimron, U., 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Research 40, 5569–5576. https://doi.org/10.1093/nar/gks216

Young, G.R., Yap, M.W., Michaux, J.R., Steppan, S.J., Stoye, J.P., 2018. Evolutionary journey of the retroviral restriction gene Fv1. PNAS 115, 10130–10135. https://doi.org/10.1073/pnas.1808516115

Youngson, N.A., Kocialkowski, S., Peel, N., Ferguson-Smith, A.C., 2005. A Small Family of Sushi-Class Retrotransposon-Derived Genes in Mammals and Their Relation to Genomic Imprinting. J Mol Evol 61, 481–490. https://doi.org/10.1007/s00239-004-0332-0

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y., 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated

data. Methods in Ecology and Evolution 8, 28–36. https://doi.org/10.1111/2041-210X.12628

Yuan, L., Yin, P., Yan, H., Zhong, X., Ren, C., Li, K., Chin Heng, B., Zhang, W., Tong, G., 2021. Single-cell transcriptome analysis of human oocyte ageing. Journal of Cellular and Molecular Medicine 25, 6289–6303. https://doi.org/10.1111/jcmm.16594

Zagore, L.L., Sweet, T.J., Hannigan, M.M., Weyn-Vanhentenryck, S.M., Jobava, R., Hatzoglou, M., Zhang, C., Licatalosi, D.D., 2018. DAZL Regulates Germ Cell Survival through a Network of PolyA-Proximal mRNA Interactions. Cell Reports 25, 1225-1240.e6. https://doi.org/10.1016/j.celrep.2018.10.012

Zayed, H., 2003. The DNA-bending protein HMGB1 is a cellular cofactor of Sleeping Beauty transposition. Nucleic Acids Research 31, 2313–2322. https://doi.org/10.1093/nar/gkg341

Zhang, F., Ji, Q., Chaturvedi, J., Morales, M., Mao, Y., Meng, X., Dong, L., Deng, J., Qian, S.-B., Xiang, Y., 2023. Human SAMD9 is a poxvirus-activatable anticodon nuclease inhibiting codon-specific protein synthesis. Science Advances 9, eadh8502. https://doi.org/10.1126/sciadv.adh8502

Zhang, T., Tamman, H., Coppieters ’t Wallant, K., Kurata, T., LeRoux, M., Srikant, S., Brodiazhenko, T., Cepauskas, A., Talavera, A., Martens, C., Atkinson, G.C., Hauryliuk, V., Garcia-Pino, A., Laub, M.T., 2022. Direct activation of a bacterial innate immune system by a viral capsid protein. Nature 612, 132–140. https://doi.org/10.1038/s41586-022-05444-z

Zhang, X.-L., Liu, P., Yang, Z.-X., Zhao, J.-J., Gao, L.-L., Yuan, B., Shi, L.-Y., Zhou, C.-X., Qiao, H.-F., Liu, Y.-H., Ying, X.-Y., Zhang, J.-Q., Ling, X.-F., Zhang, D., 2017. Pnma5 is essential to the progression of meiosis in mouse oocytes through a chain of phosphorylation. Oncotarget 8, 96809–96825. https://doi.org/10.18632/oncotarget.18425

Zhang, Y., Yan, Z., Qin, Q., Nisenblat, V., Chang, H.-M., Yu, Y., Wang, T., Lu, C., Yang, M., Yang, S., Yao, Y., Zhu, X., Xia, X., Dang, Y., Ren, Y., Yuan, P., Li, R., Liu, P., Guo, H., Han, J., He, H., Zhang, K., Wang, Y., Wu, Y., Li, M., Qiao, J., Yan, J., Yan, L., 2018. Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. Molecular Cell 72, 1021-1034.e4. https://doi.org/10.1016/j.molcel.2018.10.029

Zimmerman, S.B., 2002. Toroidal nucleoids in *Escherichia coli* exposed to chloramphenicol. Journal of Structural Biology 138, 199–206. https://doi.org/10.1016/S1047-8477(02)00036-9

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology, Computation

Resources for Molecular Biology 430, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007

Zirkin, B.R., Papadopoulos, V., 2018. Leydig cells: formation, function, and regulation†. Biology of Reproduction 99, 101–111. https://doi.org/10.1093/biolre/ioy059

# APPENDIX

## RAPID CRYOEM STRUCTURE DETERMINATION FROM CELL-FREE TRANSCRIPTION AND TRANSLATION

## Abstract

CryoEM has transformed structural biology, but a major bottleneck remains in the time required prepare purified samples. Here we show that small-volume cell-free transcription and translation yields adequate protein quantities for high resolution structure determination using cryoEM. We express a tagless mouse ferritin and obtain a high resolution structure with under 24 hours of hands-on time in the lab and determine the structure of a heterohexameric CRISPR complex.


## Main Text

Electron microscopy transformed biology by rendering invisible molecular structures visible[1–3]. Electron cryo-microscopy (cryoEM) enables structural determination of diverse macromolecules from small quantities of pure sample[4–6]. Advances in grid preparation, electron detectors, and image processing pipelines made it possible to reconstruct atomic models of macromolecular structures[7–10]. However, due to the constraints of producing proteins *in vivo,* sample preparation remains a major bottleneck in the cryoEM workflow[11]. Amplifying, digesting, ligating, and transforming genes followed by expression and purification of target proteins and nucleic acids in cellular systems requires at least a week of work in the lab, often significantly longer. Shotgun cryoEM removes the purification step by coupling mass spectrometry with cryoEM to identify and determine the structures of complexes found in crude fractions of cell cultures, but this technique does not address the cloning bottleneck[12]. *In vitro* transcription and translation using a wheat-germ cell-free transcription and translation system have been used for structure determination by nuclear magnetic resonance (NMR)[13], x-ray

crystallography[14] and for the low resolution structure of the *Arabidopsis thaliana* protein PDX1 by cryoEM[15]. However, to our knowledge, no methods exist that couples the rapid synthesis of a structural target using *in vitro* transcription and translation and minimal purification with rapid, high-resolution structure determination by cryoEM.

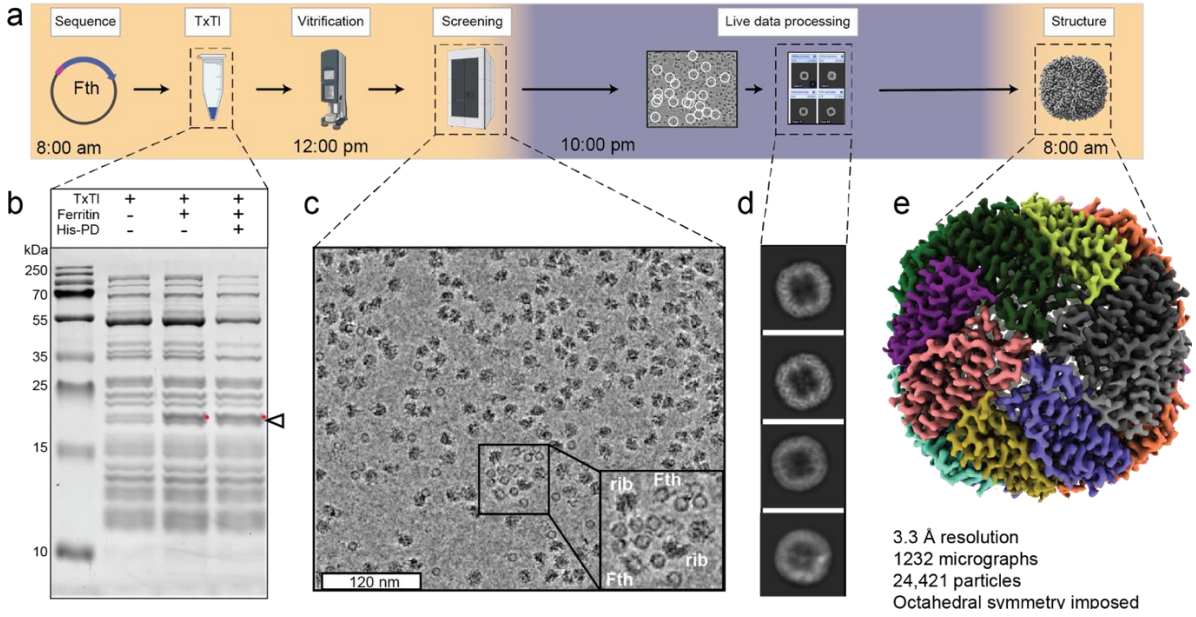



Figure A.1. Sequence to structure in 24 hours. a. Schematic and timeline of cell-free single-day cryoEM sample preparation workflow b. SDS-PAGE of cell free transcription and translation reactions shows 21 kDa heavy chain ferritin band at the expected molecular weight (Txtl = transcription and translation reaction, His-PD=His-tag pull-down with magnetic beads, open left-facing arrow indicate untagged ferritin band) c. Dose-weighted and aligned micrograph shows visible 12 nm ferritin particles mixed in with ribosomes in 1:3 dilution of original TxTl reaction at a nominal magnification of 36000 (rib = ribosome, Fth = ferritin) d. Four 2D classes containing centered, single particles, were selected in cryoSPARC live. By 8 am these classes contained 24,421 particles. e. ~3.3 Å cryoEM density map of ferritin at 8 am the morning after TxTl reaction, from 24,421 particles with octahedral (O) symmetry imposed in a streaming Homogenous Refinement Job in cryoSPARC live.

We hypothesized that protein produced from a commercially available *in vitro* transcription and translation reaction with minimal subsequent purification would be sufficient to obtain a high resolution cryoEM structure. To test this hypothesis, we expressed tagless heavy chain ferritin, a

gold-standard for benchmarking cryoEM techniques[16], using New England Biolabs PureExpress kit (referred to as TxTl)[17]. After letting the TxTl reaction run for three hours at 37º C, we performed a His-pulldown with magnetic beads to remove transcription components from the reaction. (**Figure A.2**). SDS-PAGE confirmed successful expression of ferritin protein, and ferritin was retained in the reaction mix after His-pull-down (**Figure A.2b**). To test if the quantity of ferritin was sufficient to visualize by EM, we spotted the undiluted sample on grids, and visualized by 2% uranyl acetate negative stain in a 120 keV electron microscope. Though quite concentrated, we discerned multiple clusters of 12 nm ferritin particles in all regions of the grid examined (**Figure A.2**). We subsequently performed the same TxTl reaction and spotted 3 uL of a 1:3 dilution of the TxTl reaction on Cu 300 R1.2/1.3 grids at 4 C and 95% humidity before plunge freezing in liquid ethane. Frozen grids were clipped and loaded for screening into a 200 keV Talos Arctic equipped with a Gatan K3 direct electron detector. The screened grids had sufficiently thin ice and particle distribution (**Figure A.1c**) to proceed with an overnight data collection automated using SerialEM. We aligned frames and performed CTF correction in cryoSPARC Live during data acquisition, and used a ferritin structure (PDB: 6z6u) for template picking[8,9]. We rejected any micrographs with CTF fits < 8 Å. During the first hour of data collection, we selected 2D classes with centered ferritin particles from the streaming 2D classification job for ab inito reconstruction and streaming homogeneous refinement in cryoSPARC Live with octahedral (“O”) symmetry imposed (**Figure A.1d).** We let the data collection and live processing run overnight. By 8 am the following morning, we had accumulated 24,421 particles from the template picker and an initial reconstruction with a resolution of ~3.3 Å (**Figure A.2**). We used cryoSPARC Live for all subsequent motion

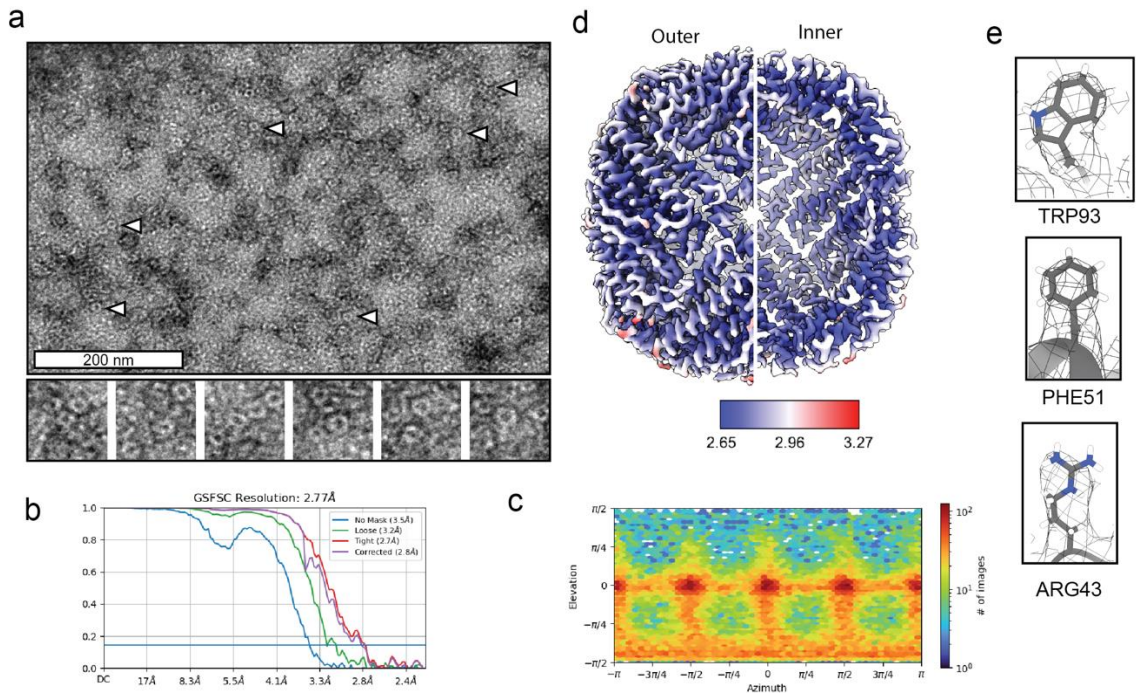



Figure A.2. TxTl cryoEM produces high resolution structures a. Negative stain reveals visible ferritin structures b. Global resolution as calculated by the global fourier shell correlation at the standard 0.143 cutoff. c. Txtl samples suffer from orientation bias, partially overcome by imposing octahedral (O) symmetry in the homogeneous refinement. d. Local resolution estimate shows that most of the density map has a sub-3 angstrom resolution. e. Atomic model fit into the electron density map generated from a homogenous refinement.

correction, CTF correction, and particle picking and then exported particles and corrected micrographs into the main instance. We collected a total of 3726 movies (50 frames with a total exposure dose of 56.37 e/Å$^2$**)** at 36000x nominal magnification (super resolution pixel size: 0.552 Å/px). Subsequent refinement of the full dataset of 3726 movies yielded a 2.77 Å electron density map of ferritin in which a previously built atomic model could be unambiguously fit (**Figure A.2**). In the density map, the main chain is clearly defined and side chain densities can be effectively placed (**Figure A.2**).

In summary, we show here that cell-free transcription and translation can be used to determine high resolution structures with minimal purification using a 200 keV electron

microscope. We anticipate that this technique will be useful for structural studies of translational inhibitors and toxins previously untenable for single particle analysis because of cell toxicity. We also anticipate that the time savings in this single-day workflow will provide benefits in the structural determination of symmetric and large particles. We anticipate that protocol optimization will allow for further enhancement of protocol for rapid high resolution structure determination.

## References


1. De Rosier, D. J. & Klug, A. Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature* **217**, 130–134 (1968).
2. Henderson, R. & Unwin, P. N. T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**, 28–32 (1975).
3. Henderson, R. *et al.* Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology* **213**, 899–929 (1990).
4. Kühlbrandt, W. The Resolution Revolution. *Science* **343**, 1443–1444 (2014).
5. Cheng, Y. Single-particle cryo-EM—How did it get here and where will it go. *Science* **361**, 876–880 (2018).
6. Fan, X. *et al.* Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution. *Nat Commun* **10**, 2386 (2019).
7. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).
8. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290–296 (2017).
9. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).
10. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180**, 519–530 (2012).
11. Weissenberger, G., Henderikx, R. J. M. & Peters, P. J. Understanding the invisible hands of sample preparation for cryo-EM. *Nat Methods* **18**, 463–471 (2021).

12. McCafferty, C. L., Verbeke, E. J., Marcotte, E. M. & Taylor, D. W. Structural Biology in the Multi-Omics Era. *J Chem Inf Model* **60**, 2424–2429 (2020).

13. Vinarov, D. A. *et al.* Cell-free protein production and labeling protocol for NMR-based structural proteomics. *Nat Methods* **1**, 149–153 (2004).

14. Abe, S. *et al.* Cell-free protein crystallization for nanocrystal structure determination. *Sci Rep* **12**, 16031 (2022).

15. Novikova, I. V. *et al.* Protein structural biology using cell-free platform from wheat germ. *Adv Struct Chem Imaging* **4**, 13 (2018).

16. Wu, M., Lander, G. C. & Herzik, M. A. Sub-2 Angstrom resolution structure determination using single-particle cryo-EM at 200 keV. *Journal of Structural Biology: X* **4**, 100020 (2020).

17. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat Biotechnol* **19**, 751–755 (2001).

18. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519–530 (2018).

19. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).

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

21. Santiago-Frangos, A. *et al.* Structure reveals why genome folding is necessary for site-specific integration of foreign DNA into CRISPR arrays. *Nat Struct Mol Biol* 1–11 (2023) doi:10.1038/s41594-023-01097-2.