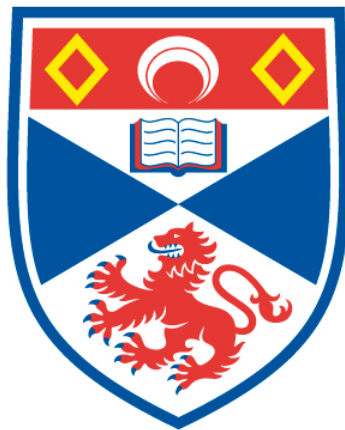


MEMORIES ARE MADE OF THIS: INVESTIGATING THE
CRISPR-CAS ADAPTATION MECHANISM

Clare Rollie

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2016

Full metadata for this thesis is available in
St Andrews Research Repository
at:

<http://research-repository.st-andrews.ac.uk/>

Identifiers to use to cite or link to this thesis:

DOI: <https://doi.org/10.17630/10023-10814>

<http://hdl.handle.net/10023/10814>

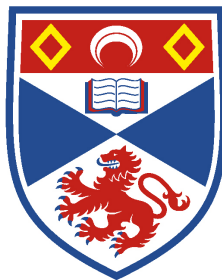
This item is protected by original copyright

This item is licensed under a
Creative Commons License

<https://creativecommons.org/licenses/by-nc-nd/4.0>

Memories are made of this: investigating the
CRISPR-Cas adaptation mechanism

Clare Rollie



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of PhD
at the
University of St Andrews

26th September 2016

Table of Contents

Table of Contents	iii
Figures, tables and equations	ix
Abbreviations	xiii
Declarations	xvii
Acknowledgements	xix
Abstract	xxi
Chapter 1: Introduction	1
1.1 Prokaryotic defence mechanisms against invading genetic elements	1
1.1.1 Surface modification	1
1.1.2 Restriction/modification (R-M) systems	2
1.1.3 Abortive infection	3
1.1.4 Argonaute	4
1.2 CRISPR-Cas discovery	5
1.2.1 Repeat arrays in prokaryotic genomes	5
1.2.2 Understanding the function of CRISPR-Cas	6
1.3 Structure of a CRISPR-Cas immune system.....	8
1.3.1 Repeats.....	8
1.3.2 Spacers.....	10
1.3.3 Protospacer adjacent motif	10
1.3.4 Leader sequence	12
1.3.5 <i>cas</i> genes	12
1.4 Diversity and classification of the CRISPR-Cas systems	12
1.4.1 Class 1	14
1.4.2 Class 2.....	15
1.5 Casposons and the origins of CRISPR-Cas.....	17
1.6 The three stages of CRISPR-Cas immunity	19
1.6.1 Stage 1: Adaptation	19
1.6.1.1 <i>cas1</i> and <i>cas2</i> are sufficient for naïve adaptation	19
1.6.1.2 Protospacer capture.....	22
1.6.1.3 Primed adaptation	24
1.6.1.4 Mechanism of integration	25
1.6.2 Stage 2: Transcription and processing of the CRISPR array.....	28
1.6.3 Stage 3: Interference	31
1.6.3.1 Type I interference	31
1.6.3.2 Type III interference	33
1.6.3.3 Transcription-dependent DNA targeting.....	36
1.7 Autoimmunity.....	38
1.8 Viral escape.....	39
1.8.1 CRISPR-Cas resistance through mutation	39
1.8.2 Anti-CRISPRs	40
1.9 Regulation of CRISPR-Cas activity	41
1.9.1 H-NS	41
1.9.2 LeuO	42
1.9.3 CRP	43
1.9.4 Envelope stress	44
1.10 Scope of this thesis	44

1.10.1 CRISPR-Cas systems of <i>S. solfataricus</i>	44
1.10.2 Aims of this thesis.....	46
Chapter 2: Materials and methods.....	47
2.1 Materials.....	47
2.1.1 Oligonucleotides.....	47
2.1.2 Restriction enzymes.....	51
2.1.3 Vectors for recombinant protein expression.....	52
2.1.4 Vectors used as substrates.....	52
2.1.5 Strains.....	54
2.2 Methods.....	54
2.2.1 Cloning and protein expression.....	54
2.2.1.1 Site-directed mutagenesis.....	54
2.2.1.2 SDS-PAGE.....	55
2.2.1.3 Restriction digests and DNA ligation.....	55
2.2.1.4 Protein over-expression and purification.....	56
2.2.1.5 Protein concentration determination.....	57
2.2.2 Substrate preparation.....	57
2.2.2.1 Gel purification.....	57
2.2.2.2 Ethanol precipitation.....	58
2.2.2.3 Nucleic acid concentration determination.....	58
2.2.2.4 Preparation of substrates with multiple DNA strands.....	58
2.2.2.5 Plasmid DNA preparation.....	58
2.2.2.6 5' end-labelling with [γ - ³² P] ATP.....	59
2.2.3 Substrate ladders.....	59
2.2.3.1 A + G Maxam-Gilbert DNA Ladder.....	59
2.2.3.2 RNA alkaline hydrolysis ladder.....	60
2.2.4 Binding assays.....	60
2.2.4.1 Electrophoretic mobility shift assay (EMSA).....	60
2.2.4.1.1 EMSA with small molecule extract.....	60
2.2.4.2 Fluorescence anisotropy.....	61
2.2.4.3 DNaseI footprinting.....	61
2.2.5 <i>In vitro</i> transcription.....	61
2.2.6 Assaying changes in gene expression.....	62
2.2.6.1 RNA extraction from <i>S. solfataricus</i> cell pellets.....	62
2.2.6.2 Reverse transcription-quantitative PCR (RT-qPCR).....	63
2.2.6.2.1 Primer efficiency determination.....	63
2.2.6.2.2 One-step RT-qPCR.....	63
2.2.6.3 Western Blot.....	64
2.2.6.3.1 Antibody generation.....	64
2.2.6.3.2 Blot method and scanning.....	64
2.2.7 Assessing complex formation.....	64
2.2.7.1 Isothermal titration calorimetry (ITC).....	64
2.2.7.2 Gel filtration.....	65
2.2.8 Activity assays.....	65
2.2.8.1 Nuclease assays.....	65
2.2.8.1.1 Cas1 nuclease assay.....	65
2.2.8.1.2 Cas2 nuclease assay.....	66
2.2.8.2 Disintegration reactions.....	66
2.2.8.2.1 Standard disintegration reaction.....	66
2.2.8.2.2 Disintegration-coupled <i>SacI</i> digest.....	66
2.2.8.2.3 Time course disintegration reactions.....	67
2.2.8.3 Integration reactions.....	67

2.2.8.3.1	Integration assay with radiolabelled protospacer	67
2.2.8.3.2	Integration time course	67
2.2.8.3.3	<i>Bst</i> UI digest of integration reaction products	68
2.2.8.4	PCR amplification of integration sites	68
2.2.8.4.1	Integration assays with <i>S. solfataricus</i> lysate	69
Chapter 3: Regulation of the CRISPR-Cas system in response to infection 71		
3.1	Introduction.....	71
3.1.1	Changes in <i>cas</i> gene expression during viral infection.....	71
3.1.2	CRISPR-Cas regulation in archaea	71
3.1.3	Transcriptional regulators in archaea	72
3.1.3.1	Cbp1.....	72
3.1.3.2	Csa3 proteins	73
3.2	Results	75
3.2.1	Infection time course of <i>S. solfataricus</i> with SMV1 and pMGB.....	75
3.2.2	Changes in Cas protein levels in response to infection	77
3.2.3	Change in <i>cas</i> gene transcript levels during infection.....	79
3.2.4	Calculating primer efficiency	80
3.2.5	Relative gene transcript levels from RT-qPCR	81
3.2.6	A putative transcriptional regulator	83
3.2.7	A DNaseI footprint of Csa3 _{CD} binding region	87
3.2.8	Searching for a small molecule ligand for Csa3 _{CD}	87
3.2.9	<i>In vitro</i> transcription	90
3.2.10	Effect of Csa3 _{CD} on transcription <i>in vitro</i>	92
3.3	Discussion	94
3.3.1	CRISPR-Cas upregulation in response to infection	94
3.3.2	Strong upregulation of Cas1 in response to infection	94
3.3.3	Differential regulation of CRISPR-Cas components	95
3.3.4	A Csa3 binding site upstream of the adaptation genes	96
3.3.5	Reduced Csa3 _{CD} during early infection.....	96
3.3.6	Csa3 _{CD} does not affect transcription <i>in vitro</i>	97
Chapter 4: Characterisation of Cas1 and Cas2 from <i>S. solfataricus</i>..... 99		
4.1	Introduction.....	99
4.1.1	Structure of the Cas1 protein	99
4.1.2	Biochemical activity of Cas1	100
4.1.3	Structure of the Cas2 protein	100
4.1.4	Biochemical activity	101
4.2	Results	102
4.2.1	Expression and purification of a Cas1 and Cas2 from <i>S. solfataricus</i>	102
4.2.2	Substrate preference of the Cas1 _{CD} protein.....	103
4.2.3	Cas1 _{CD} does not bind CRISPR sequences preferentially.....	105
4.2.4	Cas1 _{CD} is not a nuclease of ssDNA or Holliday junctions.....	107
4.2.5	Cas2 _{CD} does not bind strongly to nucleic acids	109
4.2.6	Cas2 proteins possess ribonuclease activity	110
4.2.7	Cas1 _{CD} and Cas2 _{CD} do not interact <i>in vitro</i>	112
4.3	Discussion	114
4.3.1	ssDNA binding	115
4.3.2	Cas1 is not a non-specific nuclease	116
4.3.3	Cas2 _{CD} does not stably bind nucleic acid	117
4.3.4	The role of ribonuclease activity of Cas2	118
4.3.5	Lack of Cas1-Cas2 complex formation in <i>S. solfataricus</i>	118

Chapter 5: Cas1 performs a sequence-specific disintegration reaction	121
5.1 Introduction.....	121
5.1.1 Protospacer capture.....	121
5.1.2 Spacer integration.....	123
5.1.3 Repeat duplication and repair	123
5.1.4 Similarities with viral integrases	124
5.1.5 Disintegration	124
5.2 Results	125
5.2.1 Cas1 has a high affinity for branched DNA.....	125
5.2.2 Cas1 performs disintegration on branched DNA	127
5.2.3 Disintegration occurs precisely at the branch point	131
5.2.4 Metal dependence	132
5.2.5 Identifying the nucleophile in disintegration	133
5.2.6 Concentration-dependent disintegration	134
5.2.7 Clues to the nature of incoming DNA during integration.....	135
5.2.8 Length of incoming DNA	137
5.2.9 Disintegration by Cas1 can be used to form DNA dumbbells	138
5.2.10 Residues required for disintegration	140
5.2.11 Sequence specificity of Cas1 in the disintegration reaction.....	142
5.2.11.1 +1 position.....	143
5.2.11.2 -1 position.....	146
5.2.11.3 -2 position.....	148
5.2.11.4 Incoming nucleotide	150
5.2.11.5 Disintegration reactions with substrates matching site 1 and 2 .	152
5.3 Conclusions.....	154
5.3.1 Incoming DNA.....	154
5.3.2 Cas2 is not required for disintegration	155
5.3.3 Cas1 has an intrinsic sequence specificity	155
Chapter 6: An <i>in vitro</i> reconstitution of integration	159
6.1 Introduction.....	159
6.1.1 Integration of spacers is not strictly sequence-specific <i>in vitro</i>	159
6.1.2 Supercoiled plasmid DNA is important for integration	160
6.2 Results	161
6.2.1 Cas1-Cas2 integrates short oligonucleotides into supercoiled DNA...	161
6.2.2 ssDNA is a substrate for integration	163
6.2.3 Integration is not specific for a CRISPR array <i>in vitro</i>	165
6.2.4 Protospacer end structure influences integration.....	169
6.2.5 Modifying 3' overhang length.....	171
6.2.6 No PAM processing during <i>in vitro</i> integration.....	173
6.2.7 The effect of protospacer duplex length on integration.....	175
6.2.8 Mutation of conserved residues outwith the Cas1 _{CD} active site	176
6.2.9 Sequencing of the integration sites of <i>S. solfataricus</i> Cas1-Cas2	179
6.2.10 Searching for a host factor.....	182
6.3 Discussion	184
6.3.1 Substrate structure is important for integration	185
6.3.2 Cas2 enhances integration <i>in vitro</i>	186
6.3.3 Intrinsic specificity of Cas1 guides integration	186
6.3.4 Integration host factor	187
6.3.5 A host factor in <i>S. solfataricus</i>	188
Chapter 7: Conclusions and future directions.....	191
7.1 Summary	191
7.2 CRISPR-Cas regulation.....	192

Table of contents

7.3	Characterisation of Cas1 _{CD} and Cas2 _{CD}	193
7.4	Disintegration	195
7.5	Integration	195
7.6	Capture.....	198
7.7	Conclusion.....	199
References		201
Appendices		217
	Appendix A: Triplicate Ct values from RT-qPCR.....	217
	Appendix B: Published work from this thesis	219

Figures, tables and equations

Figure 1.1 Diverse defence mechanisms of prokaryotes.....	3
Figure 1.2 The structure and function of a CRISPR-Cas system.....	9
Figure 1.3 PAM determines orientation of protospacer insertion.....	11
Figure 1.4 An up-to-date classification of the CRISPR-Cas systems	13
Figure 1.5 Conserved architecture of Class 1 interference complexes	16
Figure 1.6 The multi-partite origin of the CRISPR-Cas system.....	18
Figure 1.7 Cas1 protein sequence and structure.....	21
Figure 1.8 A model for self versus non-self discrimination during spacer uptake....	23
Figure 1.9 The last nucleotide of repeat 1 is determined by PAM	26
Figure 1.10 Model for the integration of a new spacer.....	27
Figure 1.11 pre-CRISPR RNA processing pathways in type I and III systems.....	30
Figure 1.12 Target DNA binding and interference in type I systems	32
Figure 1.13 Type III-B Cmr complex cleaves RNA with a 6 nt periodicity.....	35
Figure 1.14 Co-transcriptional DNA interference in Type III systems	37
Figure 1.15 Self-protection from CRISPR immunity.....	39
Figure 1.16 Transcriptional regulation in bacteria.....	43
Table 2.1 Sequence of oligonucleotides used in this thesis	47
Table 2.2 Complex substrates and disintegration substrate junction sequences	51
Table 2.3 Vectors made in this study.....	52
Figure 3.1 The structure of a Csa3 _{CD} protein from <i>S. solfataricus</i>	74
Figure 3.2 Infection time course of <i>S. solfataricus</i> with SMV1 and pMGB.....	76
Figure 3.3 Changes in Cas protein levels in response to infection	78
Equation 1. Primer efficiency calculation from a standard curve	80
Figure 3.4 Efficiency of primer sets used in RT-qPCR	81
Equation 2. The Pfaffl equation.....	81
Figure 3.5 Fold changes in <i>cas</i> transcripts identified by RT-qPCR	82
Figure 3.6 Two putative operator sequences in <i>S. solfataricus</i>	84
Equation 3. Binding isotherm assuming 1:1 binding of protein:nucleic acid	85
Figure 3.7 Csa3 _{CD} binds the semi-palindromic 1451 operator sequence	86
Figure 3.8 DNaseI footprint analysis of Csa3 _{CD} binding to operator DNA.....	88
Figure 3.9 Csa3 _{CD} binding in the presence of potential regulatory ligands.....	89
Figure 3.10 <i>In vitro</i> transcription from the 1451 promoter.....	91
Figure 3.11 The effect of Csa3 _{CD} on transcription efficiency.....	93
Figure 4.1 Cas1 and Cas2 proteins from <i>S. solfataricus</i>	103
Figure 4.2 Cas1 _{CD} binds preferentially to single-stranded DNA.....	104

Figure 4.3 DNA binding by Cas1 _{CD} is not sequence dependent.....	106
Figure 4.4 Cas1 does not cut Holliday junction or single-strand DNA sequences.	108
Figure 4.5 Cas2 _{CD} does not show nucleic acid binding	110
Figure 4.6 <i>Sulfolobus solfataricus</i> Cas2 proteins have ribonuclease activity	111
Figure 4.7 Cas1 _{CD} and Cas2 _{CD} do not interact <i>in vitro</i>	113
Figure 4.8 Cas1-Cas2 form a complex in <i>E. coli</i>	117
Figure 5.1 Model of the adaptation stage of CRISPR-Cas immunity	122
Figure 5.2 Cas1 binds branched DNA structures	126
Figure 5.3 SsoCas1 activity on branched DNA.....	128
Figure 5.4 SsoCas1 performs a disintegration reaction on branched substrates ..	129
Figure 5.5 Disintegration occurs at the branch point	131
Figure 5.6 Metal-dependent disintegration by Sso- and EcoCas1.....	133
Figure 5.7 Requirements for disintegration.....	134
Figure 5.8 Concentration-dependent disintegration activity.....	135
Figure 5.9 A double-stranded 5' flap supports disintegration by SsoCas1	136
Figure 5.10 5' flap structure influences disintegration by Cas1	138
Figure 5.11 Disintegration by Cas1 produces DNA 'dumbbells' efficiently	139
Figure 5.12 Conserved residues required for disintegration	141
Figure 5.13 The +1 position influences disintegration activity by SsoCas1	143
Equation 4 Single exponential for reaction rate calculation	144
Figure 5.14 Sequence specificity for the +1 position of EcoCas1.....	145
Figure 5.15 Specificity of Cas1 for the -1 position during disintegration	147
Figure 5.16 Specificity at the -2 position during disintegration by Cas1	149
Figure 5.17 Specificity of Cas1 for the 'incoming' nucleotide during disintegration	151
Figure 5.18 Disintegration of a site 1 versus site 2 substrate	153
Figure 5.19 Nucleotide sequence at site 1 is key to integration and disintegration	156
Figure 6.1 Cas1 _{CD} and Cas2 _{CD} joins short oligonucleotides to plasmid DNA	162
Figure 6.2 Single- and double-stranded DNA integrated by Cas1 _{CD}	165
Figure 6.3 <i>in vitro</i> integration does not require a CRISPR array.....	166
Figure 6.4 CRISPR C secondary structure and variants.....	167
Figure 6.5 Restriction digest of Cas1 _{CD} -Cas2 _{CD} integration products	168
Figure 6.6 Protospacer end structure affects integration efficiency	170
Figure 6.7 Modifying 3' overhang length affects integration	172
Figure 6.8 PAM sequences in protospacer ends are not processed	174
Figure 6.9 Effect of protospacer duplex length on integration	176
Figure 6.10 Cas1 _{CD} residues important for integration <i>in vitro</i>	178
Figure 6.11 A PCR assay to amplify integration sites.....	180
Figure 6.12 Sequence motifs at Cas1-Cas2 integration sites.....	182

Figure 6.13 A component of *S. solfataricus* lysate needed for specific integration 184

Abbreviations

3'	3 prime DNA end
5'	5 prime DNA end
A	Adenine
Å	Ångström
Abi	Abortive infection
ADP	Adenosine diphosphate
eAgo	Eukaryotic argonaute
pAgo	Prokaryotic argonaute
AHL	N-acyl-L-homoserine lactone
Asp	Aspartic acid
ATP	Adenosine triphosphate
ATP [γ - ³² P]	Adenosine triphosphate with a 32-phosphate radioactive isotope in the gamma phosphate position
bp	Base pair
BSA	Bovine serum albumin
C	Cytosine
cAMP	Cyclic adenosine monophosphate
c-di-AMP	Cyclic diadenosine monophosphate
cDNA	Complementary DNA
C-term	C-terminal domain
Cas	CRISPR-associated
Cascade	CRISPR-associated complex for antiviral defence
CasX _{AB}	Indicates the Cas protein is located between CRISPR loci A and B
CasX _{CD}	Indicates the Cas protein is located between CRISPR loci C and D
Cbp1	CRISPR DNA repeat binding protein 1
CMR	CRISPR RAMP module
COG	Cluster of orthologous groups
CRISPR	Clustered regularly interspaced short palindromic repeats
CRP	cAMP receptor protein
crRNA	CRISPR RNA
Csm	Type III-A or -D interference complex
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate (A, T, C and G)

Abbreviations

dpi	Days post-infection
ds	Double-stranded
DTT	1,4 – dithiothreitol
Eco	<i>Escherichia coli</i>
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
EMDB	Electron Microscopy Data Bank
EMSA	Electrophoretic mobility shift assay
FAM	Carboxyfluorescein
g	Gram
G	Guanine
gp5.5	product of T7 phage gene 5.5
H-NS	Heat-stable nucleoid-structuring protein
HD	Histidine-aspartate
HGT	Horizontal gene transfer
HJ	Holliday junction
IHF	Integration host factor
IP	(Phosphor)Imaging plate
IPTG	Isopropyl β -D-thiogalactopyranoside
ITC	Isothermal titration calorimetry
kb	Kilobase
K _D	Equilibrium Dissociation constant
kDa	Kilodalton
l	Litre
LB	Luria Bertani medium
Lit	Late inhibitor of T4
LRP	Leucine-responsive regulatory protein
LSU	Large subunit
M	Molar
mA	Milliampere
MES	2-(N-morpholino)ethanesulfonic acid
mg	Milligram
MID	Middle
min	Minute/minutes
ml	Millilitre
mM	Millimolar
mRNA	Messenger RNA
MTase	Methyltransferase

Abbreviations

MW	Molecular weight
nm	Nanometre
nt	Nucleotide
NTD	N-terminal domain
OH	Hydroxyl
OD	Optical density
OD ₆₀₀	Absorbance at 600 nm
ON	Overnight
Orc	Overcome classical resistance
P1, P2	<i>Sulfolobus solfataricus</i> strains, P1 and P2
PAM	Protospacer adjacent motif
PAZ	PIWI-Argonaute-Zwille
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
<i>Pfu</i>	<i>Pyrococcus furiosus</i>
Phage	Bacteriophage
PIWI	P-element-induced wimpy testis
PNK	Polynucleotide kinase
Pol	Polymerase
pre-crRNA	Precursor CRISPR RNA
R-M	Restriction-modification system
RAMP	Repeat-associated mysterious protein
REase	Restriction endonuclease
RNA	Ribonucleic acid
RNA pol/RNAP	RNA polymerase
RNase	Ribonuclease
rNTP	Ribonucleotide triphosphate (A, U, C and G)
rpm	Revolutions per minute
RRM	RNA-recognition motif
RT	Room temperature
RT-qPCR	Reverse transcriptase polymerase chain reaction
SDM	Site directed mutagenesis
SDS	Sodium dodecyl sulphate
SDS-PAGE	SDS polyacrylamide gel electrophoresis
sec	Second/seconds
SIRV2	<i>Sulfolobus islandicus</i> rod-shaped virus 2

Abbreviations

Site 1	The junction between the CRISPR leader and 5' end of repeat 1
Site 2	The junction between the 5' end of repeat 1 and the first spacer
SM	Small molecule
SMV1	<i>Sulfolobus monocaudavirus 1</i>
ss	Single-stranded
Sso	<i>Sulfolobus solfataricus</i>
SSU	Small subunit
SSV	<i>Sulfolobus shibatae</i> virus
STIV	<i>Sulfolobus</i> turreted icosahedral virus
STSV	Single-tailed fusiform <i>Sulfolobus</i> virus
T	Thymine
TBE	Tris-borate EDTA
TBP	TATA-binding protein
TFB	Transcription factor β
TE	Tris-EDTA
TEV	Tobacco etch virus
Ter	Termination sites
TIR	Inverted terminal repeats
tracrRNA	Trans-activating RNA
U	Uracil
UV	Ultraviolet
wHTH	Winged helix-turn-helix
WT	Wild type

Declarations

Candidate's declarations:

I, Clare Rollie, hereby certify that this thesis, which is approximately 65,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student and as a candidate for the degree of PhD in September 2011; the higher study for which this is a record was carried out in the University of St Andrews between 2011 and 2016.

Date: _____ Signature of candidate: _____

Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: _____ Signature of supervisor: _____

Permission for publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any *bona fide* library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any

Declarations

third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY

b) Embargo on all of printed copy for a period of 1 year on the following ground:
Publication would preclude future publication.

ELECTRONIC COPY

b) Embargo on all of electronic copy for a period of 1 year on the following ground:
Publication would preclude future publication.

Date: _____

Signature of candidate: _____

Date: _____

Signature of supervisor: _____

Acknowledgements

Firstly and foremost, I would like to thank my supervisor Professor Malcolm White for his support, guidance and encouragement throughout my studies. I also wish to express my gratitude to members of the White lab, both past and present for their help and advice. Specifically, I acknowledge Dr Shirley Graham who directly contributed to this work by preparing *S. solfataricus* cell lysate and by carrying out the PCR assays detailed in Chapter 6. Dr Jing Zhang and Dr Christophe Rouillon taught me the basics of many of the assays used in this work. I also thank Mike and Richard for making the PhD experience so much more enjoyable!

I acknowledge the Garrett lab (University of Copenhagen), particularly previous member Dr Susanne Erdmann, for providing cell pellets used in Chapter 3. I also thank Dr Ed Bolt (University of Nottingham) for providing *E. coli* proteins used in Chapter 5. Project students Kotryna Temcinaite and James Robson contributed by cloning and purifying variant Cas1 proteins used in this work. I really appreciate the time Tom, Christophe, Shirley and Malcolm gave up to read and comment on parts of this thesis.

I am very grateful to the University of St Andrews for funding my PhD and giving me the opportunity to live and study in such a great wee town.

My heartfelt thanks go to my friends and family for their love and encouragement, and for helping me regain perspective when I needed to. Finally, thanks to Tom who supports me in everything I do.

Abstract

CRISPR-Cas is an adaptive immune system unique to prokaryotes, which prevents infection by foreign genetic elements. Key to the function of CRISPR-Cas immunity is the ability to adapt to new threats by incorporating short segments, termed spacers, of invading DNA into the clustered regularly interspaced short palindromic repeat (CRISPR) array of the host. Spacers constitute *immunological memories*, used by CRISPR-associated (Cas) proteins to mount a sequence-specific attack on subsequent infections. This immunisation of the host is called CRISPR adaptation.

Adaptation requires the integration of new spacers at a precise site in the CRISPR array. Two proteins, Cas1 and Cas2, are essential for adaptation; however, the mechanisms of spacer integration remain poorly understood. The work described here focused on understanding adaptation in *Sulfolobus solfataricus*. Using biochemical assays, I aimed to characterise the activity of the Cas1 and Cas2 proteins in this organism in order to understand their role in the insertion of new spacers. Additionally, I aimed to investigate how the expression of CRISPR-Cas components is regulated in this organism in response to viral infection.

The results presented here show that expression of Cas1 was strongly upregulated in response to infection. A Csa3 protein from *S. solfataricus* was found to bind to the promoter for transcription of *cas1*, implying a role in the regulation observed. I reconstituted *in vitro* both the integration reaction performed by the Cas1 and Cas2 proteins of *S. solfataricus* and the reverse of this reaction, disintegration. Cas1 was shown to impose sequence specificity on these reactions, selecting sites similar to the leader-repeat junction of the CRISPR locus. Finally, I demonstrated that, in addition to the intrinsic specificity of Cas1, there was a requirement for an additional host factor for site-specific integration in *S. solfataricus*.

Chapter 1: Introduction

1.1 Prokaryotic defence mechanisms against invading genetic elements

Prokaryotic life is under constant threat from foreign genetic elements, with viruses in some ecosystems outnumbering their prokaryote hosts by at least 15:1 (Suttle, 2007). The presence of foreign genetic elements can provide prokaryotes with fitness-enhancing genes through horizontal gene transfer (HGT). However, the ultimate goal of these selfish elements is to propagate, which may be costly for the host. Therefore, prokaryotes must be equipped to protect themselves from invasion in order to thrive. The formidable threat posed by viruses and other genetic elements has led to the expansion and diversification of host defence systems. These systems are often present in large 'defence islands', which can account for up to 10% of the total genome size (Makarova et al., 2013a). Some of the key defence mechanisms employed by prokaryotes to prevent infection are introduced below and a summary is provided in Figure 1.1.

1.1.1 Surface modification

Viruses often gain entry to host cells by docking on cell surface components such as proteins, lipopolysaccharides and motility structures. Selective pressure to overcome viral infection frequently leads to the loss, variation or masking of these docking sites in prokaryotic populations. For example, bacteriophage (phage) λ relies on the *Escherichia coli* sugar-transporting protein, LamB, for docking and invasion of the host. An infection experiment carried out with *E. coli* and phage λ in glucose-limited media led to selection for and emergence of a host population with greatly reduced LamB expression. The emergence of this resistant population coincided with a massive reduction in phage densities (Meyer et al., 2012). The loss or mutation of surface receptors often results in a heavy fitness cost, explaining why such drastic changes are not often fixed in natural host populations (Meaden et al., 2015). Instead transient phenotypic changes, called phase variations, are common and often occur in response to environmental cues. Phase variation can lead to members of a bacterial population displaying different surface receptors, thus

preventing the rapid adaptation of viruses to infect the host (Veening et al., 2008). Interestingly, bacteria seem to be able to estimate risk of infection and cooperate through quorum sensing to counteract virus docking (Hoyland-Kroghsbo et al., 2013). At high *E. coli* population density, when risk of infection is greatest, the extracellular concentration of the excreted small molecule signal N-acyl-L-homoserine lactone (AHL) accumulates. *E. coli* respond to high AHL by reducing the expression of surface receptors used by phage λ to enter cells. Quorum sensing in this way was found to reduce levels of phage adsorption and greatly increase the number of cells surviving a phage infection (Hoyland-Kroghsbo et al., 2013).

Masking of receptors is also employed by bacteria to avoid docking of phage. This masking can occur by direct protein-protein interaction, as found for the blocking of the phage-targeted surface receptor OmpA in *E. coli*, which is bound and masked by the plasmid-encoded lipoprotein TraT. Bacterial strains containing the TraT-encoding plasmid were found to be much more resistant to phage infection compared to strains missing the conjugative plasmid (Riede & Eschbach, 1986). Surface receptors can also be masked by excretion of exopolysaccharides, with the expression of an hydrophilic exopolysaccharide in *Lactococcus* shown to interfere with the adsorption of phage (Forde & Fitzgerald, 1999).

1.1.2 Restriction/modification (R-M) systems

If a foreign genetic element succeeds in gaining access to the cell, the next line of host defence involves the recognition and destruction of non-self DNA. Restriction/modification systems represent an extremely widespread and diverse form of immunity, active in around 90% of sequenced prokaryotic genomes (Roberts et al., 2010). This defence system relies on prokaryotic genomes coding for a series of restriction endonucleases (REases) with strict sequence specificity for recognition sites of 4-8 nucleotides (nts). When foreign DNA enters the cell, REases target and cleave at their recognition motif, leading to the neutralisation of the infectious element (Took & Dryden, 2005). These restriction sites are also very common in host genomes, but the second component of the restriction modification systems, usually a methyltransferase (MTase), modifies these sites so that self-DNA avoids detection by REases.

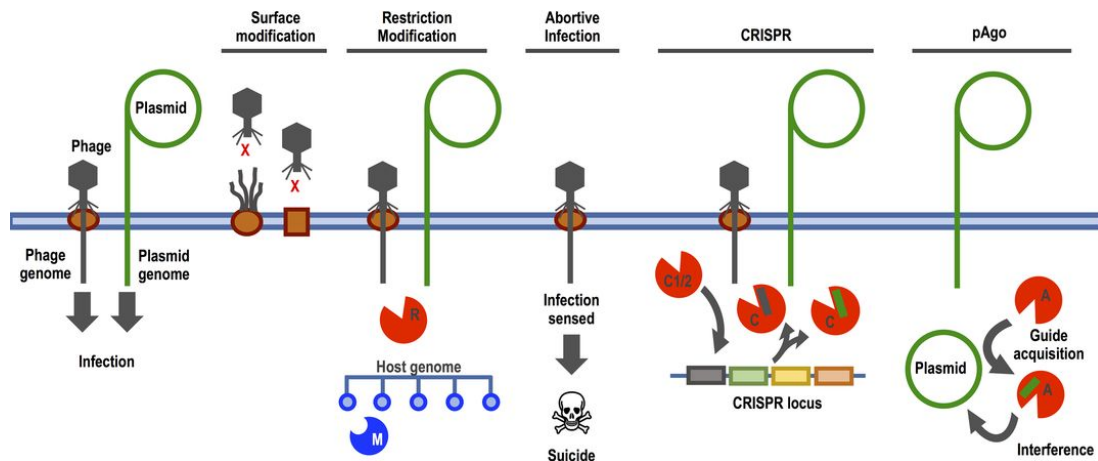


Figure 1.1 Diverse defence mechanisms of prokaryotes

A summary of the array of defence mechanisms prokaryotes use to protect themselves from foreign genetic elements is shown. Innate defence mechanisms function by: blocking entry of viruses (surface modification), recognising and degrading non-self DNA (restriction-modification and argonaute), or limiting the spread of viral particles by triggering cellular suicide (abortive infection). The CRISPR system is an adaptive immune system that provides sequence-specific immunity against foreign genetic elements. Specific proteins involved in the different systems are labelled: M, methylase; R, restriction enzyme; A, prokaryote argonaute; C, Cas proteins. Adapted from Houte et al., 2016.

Viruses and plasmids have evolved ways to circumvent these defences, facilitating their entry and propagation in prokaryote hosts. Anti-restriction mechanisms include acquiring a MTase, or hijacking the host MTase, in order to methylate and disguise restriction sites from the host REase (Krüger & Bickle, 1983). The loss of restriction sites or incorporation of unusual bases can also block the recognition of foreign DNA by the host. Furthermore, the T3 and T7 phage have been demonstrated to express the Ocr (overcome classical restriction) protein, which binds and directly inhibits *E. coli* REases by blocking their interaction with restriction sequences (Krüger & Bickle, 1983).

1.1.3 Abortive infection

Programmed cell death or growth arrest is a last line of defence in the battle against foreign genetic elements. This involves the self-sacrifice, or induced-dormancy, of a host cell after the failure of other immune mechanisms, in order to prevent further infection of the host population (Makarova et al., 2012). Abortive infection (Abi) systems are commonly encoded on acquired elements, such as prophage and plasmids, and are activated by the entry of phage DNA (Georgiou et al., 1998). Expression of a single gene is often sufficient to lead to death of the host, through inhibition of replication or transcription (Chopin et al., 2005). A well-understood example is that of the T4 phage-induced abortive infection in *E. coli* activated by the

Lit (late inhibitor of T4) protein (Georgiou et al., 1998). The expression of the phage major head coat protein late in the infection cycle triggers the Lit protein, encoded from a prophage region of the *E. coli* K12 genome, to specifically cleave the translation elongation factor Tu. This induced cleavage inhibits the translation of viral and host proteins, limiting phage multiplication and also leading to host cell death (Georgiou et al., 1998).

Abortive infection can also be brought about by toxin/antitoxin systems of bacteria and archaea (Makarova et al., 2011b; Fineran et al., 2009). Toxin genes code for proteins that cause cell death by the destruction of host cell membrane, or the degradation of RNA transcripts. These activities are neutralised by the anti-toxin gene, which codes for an RNA or protein that interferes with the production or activity of the associated toxin. Under stress or viral infection, the production of the anti-toxin is inhibited and the unstable protein rapidly degraded. This, in turn, releases the destructive activities of the toxin to bring about cell death. In the Gram-negative bacterium *Pectobacterium atrosepticum* viral infection results in reduced transcription of the *toxI* anti-toxin gene. Freed from inhibition the associated toxin (ToxN) causes growth retardation of the host, thus preventing rapid spread of phage (Fineran et al., 2009).

1.1.4 Argonaute

Argonaute proteins make up part of the RNAi system in eukaryotes and have recently been found to contribute to host defence against foreign genetic elements in prokaryotes (Makarova et al., 2009). The argonaute proteins in eukaryotes (eAgos) invariably contain 3 domains: the PAZ (PIWI-Argonaute-Zwille) nucleic-acid-binding domain, the PIWI (P-element-induced wimpy testis) nuclease domain, and a MID (middle) domain (Makarova et al., 2009). These proteins use short single-stranded (ss) RNA guides to silence complementary RNAs through binding or cleavage.

Argonautes are present in 32% of archaeal and 9% of bacterial genomes (Swarts et al., 2014b). These prokaryotic argonautes (pAgos) are more structurally diverse than their eukaryotic counterparts and a subset lack the PAZ domain (Makarova et al., 2009). The truncated group have no nuclease activity and are thought to rely on nucleases encoded by adjacent genes for nucleic acid processing (Makarova et al.,

2009). Furthermore, unlike eAgos, pAgos often have a much higher affinity for short 15-30 nt DNA guides, compared to RNA guides, and are primarily DNA targeting systems (Yuan et al., 2005; Swarts et al., 2014b). The short guides are used to target pAgo proteins to complementary sequences which are degraded directly by pAgo, or bound by pAgo and destroyed by an as yet unidentified partner nuclease (Swarts et al., 2014a; Olovnikov et al., 2013). This DNA- or RNA-guided DNA interference reduces transformation efficiency and leads to silencing of gene expression from plasmid DNA, protecting the host from foreign DNA elements (Swarts et al., 2014a; Olovnikov et al., 2013). Guides are preferentially selected by pAgo from plasmid and foreign genetic elements, leading to their subsequent destruction, while host genetic material is protected (Swarts et al., 2014a; Olovnikov et al., 2013). The reason for the preferential selection and targeting of extrachromosomal elements is still not understood, but the answer will be key to advancing our understanding of how argonaute functions in immune defence.

1.2 CRISPR-Cas discovery

1.2.1 Repeat arrays in prokaryotic genomes

The discovery of an adaptive immune system present in ~40% of bacteria and ~90% of archaea (Grissa et al., 2007) revolutionised our understanding of prokaryote immunity. Unlike other immune mechanisms described above, this system can adapt throughout the lifetime of a bacterium to provide immunity to a constantly evolving viral threat – a feature previously thought to be unique to the immune systems of higher eukaryotes.

The first step in the discovery of CRISPR-Cas came in 1987 when tandem arrays of 29 base pair (bp) repeat sequences were identified in the genome of *E. coli* by Ishino and colleagues (Ishino et al., 1987). These repeats contained palindromic sequences and were separated by 32 bp non-repetitive intervening sequences, later called spacers. At the time the significance of these repeats was unknown, however shortly after this first report, similar arrays were identified in other bacteria and archaeal genomes, with studies of these repeats in *Mycobacterium tuberculosis* and *Haloferax mediterranei* being among the first (Hermans et al., 1991; Mojica et al., 1993).

These arrays were later recognised as a distinct and widely distributed family of prokaryotic repeats (Mojica et al., 2000). Following several iterations the arrays were named after one of their components, the clustered regularly interspaced short palindromic repeat (CRISPR) sequences (Jansen et al., 2002). AT-rich sequences often several hundred base pairs in length were found to be commonly associated with CRISPR arrays (Jansen et al., 2002). These sequences, called leaders, were similar between different CRISPR arrays of the same species, but varied between species (Jansen et al., 2002). Sequencing of a cDNA library from *Archaeoglobus fulgidus* showed that CRISPR arrays were actively transcribed from a putative promoter in the leader sequence to form a long CRISPR transcript (Tang et al., 2002). The identification of a ladder of shorter CRISPR RNA signals on northern blots implied that the long tandem repeat transcript underwent processing following transcription (Tang et al., 2002).

The presence of a CRISPR array on a plasmid in *H. mediterranei* led to faulty separation of genetic material during cell division, prompting the initial suggestion that CRISPR arrays were involved in replicon partitioning (Mojica et al., 1995). A greater understanding of the function of these repeats came when a set of protein-coding genes were found to be associated with the repeat-spacer arrays (Jansen et al., 2002). These CRISPR-associated (*cas*) genes coded for proteins with predicted nuclease and helicase functions, leading to suggested roles in DNA metabolism and maintenance of the CRISPR arrays (Jansen et al., 2002). Makarova and colleagues went on to identify many other *cas* genes in the genomes of hyperthermophiles (Makarova et al., 2002). Given the putative functions of the encoded proteins and elevated levels of DNA damage at high temperature, the authors predicted that Cas proteins might form part of a novel DNA repair system (Makarova et al., 2002).

1.2.2 Understanding the function of CRISPR-Cas

A key step in understanding the function of the CRISPR and *cas* elements came when a subset of spacer sequences located between CRISPR repeats were found to match segments of viral or plasmid DNA (Mojica et al., 2005). Following a review of the literature, the authors noted that hosts carrying a spacer that matched a particular genetic element seemed to be protected from infection by this element. In contrast, closely related strains lacking this spacer remained susceptible to infection. They concluded that incorporation of a foreign sequence into the CRISPR array

imparted sequence-specific immunity (Mojica et al., 2005). Another study matched 75% of identifiable *Streptococcus thermophilus* spacers to phage genomes and a further 20% to conjugative plasmids. The authors also identified a correlation between the number of spacers in an array and resistance to infection (Bolotin et al., 2005). Around the same time, Pourcel and colleagues reported that the spacer complement of three closely related strains of *Yersinia pestis* differed at the leader-proximal end, whereas the leader-distal spacers were conserved (Pourcel et al., 2005). They concluded that the spacer content from a common ancestor had been modified over time by the addition of new spacers in a polarized fashion, directed by the leader (Pourcel et al., 2005). Taking these results together, a hypothesis was generated that suggested that the CRISPR array and *cas* genes make up a prokaryote immune system, capable of acquiring sequence-specific immunity to invading viruses and plasmids by incorporating a short segment of foreign DNA (Makarova et al., 2006). The mechanism of targeting was suggested to involve the inhibition of invader gene expression by anti-sense CRISPR RNA binding, similar to RNAi interference in eukaryotes (Mojica et al., 2005; Makarova et al., 2006).

The first experimental proof that CRISPR-Cas acted as an immune system came when *S. thermophilus* was shown to acquire resistance to a viral challenge following the incorporation of spacers matching the infecting phage into its CRISPR array (Barrangou et al., 2007). The authors also demonstrated that the deletion of spacers or *cas* genes reversed the acquired resistance. It was concluded that CRISPR spacer content defines resistance to phage, and Cas proteins are essential in mediating immunity (Barrangou et al., 2007). Following this initial proof, the CRISPR-Cas system was subsequently demonstrated to prevent horizontal gene transfer of conjugative plasmids in *Staphylococcus epidermidis* (Marraffini & Sontheimer, 2008). Spacers with a perfect match to the *nickase* gene of the conjugative plasmid, which is crucial for transfer, provided the host with immunity. In contrast, mismatches between spacer and targeted sequence rendered the host susceptible to infection once more (Marraffini & Sontheimer, 2008). It was not until 2011 that experimental proof of immunity mediated by CRISPR-Cas was demonstrated in archaea. The CRISPR-Cas system of the hyperthermophile *S. solfataricus* was shown to protect the host from viral infection and, unlike the targeting described previously, targeting in this system was more promiscuous, with spacers containing up to three mismatches to viral sequences still providing immunity (Manica et al., 2011).

1.3 Structure of a CRISPR-Cas immune system

The CRISPR-Cas immune response can be divided into three stages (shown in Figure 1.2, B), each involving a different set of Cas proteins. First is the adaptation stage, when a piece of foreign DNA, called a protospacer, is inserted into the CRISPR array. The second step requires the transcription of the CRISPR array and the processing of the long transcript into units called CRISPR RNAs (crRNAs), which contain individual spacer transcripts and short repeat arms. Finally, in the interference stage, these crRNAs are loaded into Cas protein complexes and are used to guide these complexes to target and degrade foreign genetic elements. To bring about this sophisticated immune response CRISPR-Cas systems of prokaryotes are equipped with several key components, which are shown in Figure 1.2 (A) and introduced below.

1.3.1 Repeats

The number and length of CRISPR repeat-spacer arrays vary between bacteria and archaea. Archaeal CRISPR-Cas systems are often made up of between two and eight arrays, but can contain up to 20, representing up to 1% of the host genome (Lillestøl et al., 2006, 2009). Within a CRISPR array the sequence of the repeat is usually constant, with the exception of some degenerate sequences found at the leader-distal end of the CRISPR array. Repeat sequences differ between organisms and their length ranges from 24-48 bp. However, the 3' end of repeats from different CRISPR-Cas systems were found to have a common sequence motif that was predicted to be important for Cas protein binding (Kunin et al., 2007). The repeats have been grouped into 12 clusters based on their secondary structure and consensus sequence (Kunin et al., 2007). There is a clear relationship between particular repeat clusters and *cas* gene complements, implying coevolution of these elements. However, similarity in sequence and structure does not depend on the hosts being closely related, with some bacteria and archaea sharing the same repeat cluster (Kunin et al., 2007). This supports the theory that CRISPR arrays and *cas* genes have been distributed widely between prokaryotes by HGT (Godde & Bickerton, 2006).

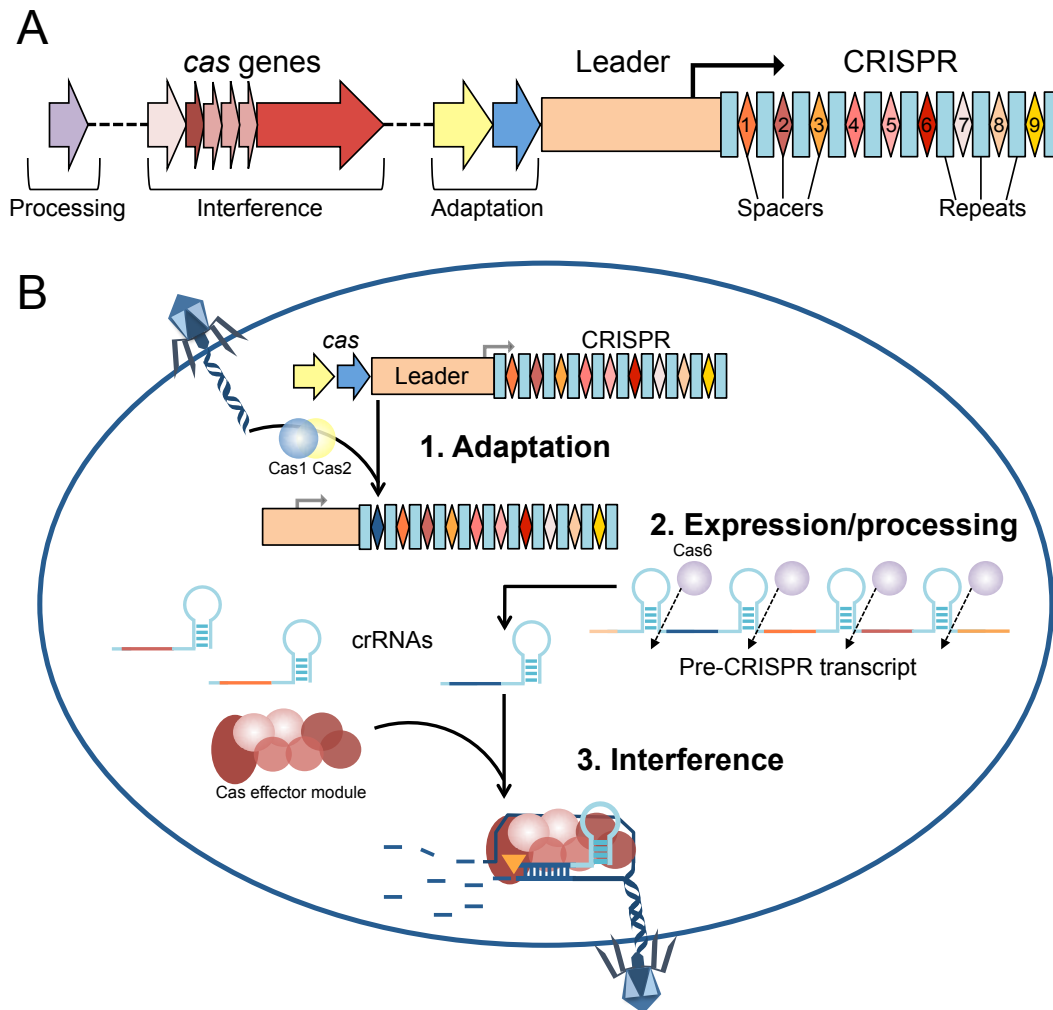


Figure 1.2 The structure and function of a CRISPR-Cas system

A. A functional CRISPR locus requires an AT-rich leader sequence containing the promoter for the array. A cassette of CRISPR-associated (*cas*) genes is also required to code for proteins involved in mediating CRISPR-Cas immunity. The CRISPR array itself consists of short, often palindromic, tandem repeats (blue rectangles) separated by spacer sequences (multi-coloured diamonds). Spacer sequences often match phage or plasmid DNA. **B.** A schematic showing an overview of CRISPR-Cas immunity. The immune response is mediated in 3 stages: adaptation, expression/processing and interference. **1.** During the first stage a segment of foreign DNA called a protospacer is excised and integrated at the leader-proximal end of the CRISPR locus (blue diamond). Universally conserved proteins Cas1 and Cas2 are involved in this step. **2.** The CRISPR array is transcribed to form a long pre-CRISPR RNA transcript (pre-crRNA) containing the sequence of the added spacer. This transcript is then processed into short mature crRNAs containing one spacer sequence and repeat arms. This process is carried out by an endoribonuclease, usually Cas6. **3.** The short crRNAs are loaded into a large interference protein/complex and are used to guide this effector to target and destroy matching foreign sequences.

The division of CRISPR repeat sequences into clusters also revealed clear differences in secondary structure. The first repeats identified contained inverted sequences, leading to the term 'palindromic' being added to the CRISPR acronym (Jansen et al., 2002). This dyad symmetry was predicted to lead to crRNAs having

a hairpin structure, thought to be crucial for binding and processing by Cas proteins. However, the grouping of repeats into clusters revealed that while many bacterial groups contain a hairpin structure, those of the archaea are often unstructured (Kunin et al., 2007). The abundance of unstructured repeats gave rise to the argument that the 'P' in the CRISPR acronym should stand for 'prokaryotic', rather than 'palindromic' (Lawrence & White, 2011).

1.3.2 Spacers

Spacers are generally between 30 and 45 bp in length and do not contain any conserved secondary structure (Kunin et al., 2007). These spacers derive from viral and plasmid sequences and act to immunise the host against these invaders (Mojica et al., 2005). CRISPR spacer sequences were found to match both sense and anti-sense, and coding and intergenic regions of foreign genetic material, suggesting that spacers are captured directly from double-stranded DNA (dsDNA) and not mRNA (Shah et al., 2009). When the host is exposed to an infection, new spacers are added at the leader-proximal end of the CRISPR array, which is often one of the most variable regions of the prokaryotic genome, differing even between closely related strains (Pourcel et al., 2005). This polarized addition of spacers results in the CRISPR array becoming a chronological record of past infections encountered by the host. Interestingly, Horvath and colleagues reported that spacers at the leader-distal end of the array, acquired during ancient infections, were often deleted (Horvath et al., 2008). Another study found that under growth in the absence of infection, a large section of the *S. solfataricus* P2 genome containing the CRISPR arrays was deleted (Lillestøl et al., 2006). These findings suggest that the maintenance of a CRISPR array can be a fitness cost to the host, which can be reduced by the shortening or loss of the array in the absence of infection.

1.3.3 Protospacer adjacent motif

The segments of viral or plasmid DNA matching CRISPR spacers are known as protospacers. Protospacers were found to be selected non-specifically from both strands of plasmid or viral DNA and from both coding and non-coding sequences (Makarova et al., 2006). However, examination of the sequences flanking protospacers in *S. thermophilus* led to the identification of a short conserved motif,

implying a degree of sequence-specific protospacer selection (Deveau et al., 2008; Bolotin et al., 2005).

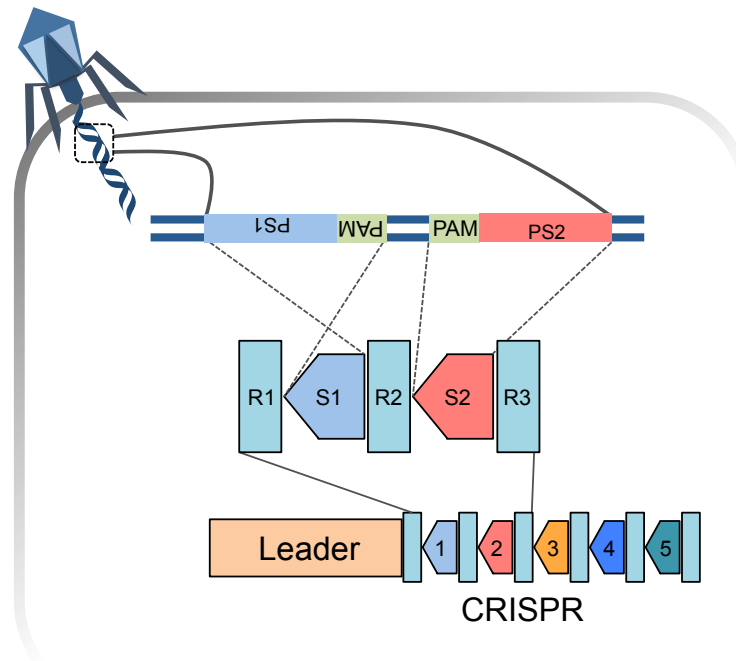


Figure 1.3 PAM determines orientation of protospacer insertion

Protospacers can come from both strands of viral DNA and their selection requires the presence of a short recognition sequence called a protospacer adjacent motif (PAM). Protospacers 1 and 2 (PS1/PS2) are located on different strands of incoming DNA, but are both inserted as spacers (S1/2) with the end of the spacer that was previously PAM-adjacent oriented towards the leader sequence. Adapted from Mojica et al., 2009.

These short sequence motifs were later found to be a universal feature of CRISPR-Cas variants and were called protospacer adjacent motifs (PAMs) (Díez-Villaseñor et al., 2009). The PAM is located immediately 5' or 3' of the protospacer (dependent on the CRISPR-Cas system) and is between 2 and 5 nt in length (Díez-Villaseñor et al., 2009). The sequence differs between CRISPR-Cas systems and is linked to the type of repeat present in the array as defined by Kunin and colleagues (Kunin et al., 2007). This short sequence motif is crucial both during adaptation of the array by the addition of new spacers and during targeting of foreign DNA for destruction by CRISPR-Cas. The PAM sequence also governs the orientation of insertion of new protospacers with the PAM-proximal end of the protospacer always being inserted toward the leader as shown in Figure 1.3 (Díez-Villaseñor et al., 2009; Deveau et al., 2008).

1.3.4 Leader sequence

A functional CRISPR array requires an intact AT-rich leader sequence, which is usually between 100 and 550 bp in length (Jansen et al., 2002). The leader sequence is often longer in archaea and a correlation between longer leader lengths and increasing growth temperature of the host has been identified (Lillestøl et al., 2006). One of the six CRISPR arrays (CRISPR E) of *S. solfataricus* P2, which does not acquire new spacers during infection, was found to lack a leader sequence, leading the authors to conclude that the leader is an important docking site for Cas proteins required for the insertion of new spacers (Lillestøl et al., 2006). This theory is supported by the polarized insertion and orientation of new spacers with respect to the leader sequence (Pourcel et al., 2005; Lillestøl et al., 2006). A second important feature of the leader sequence is that it often contains a promoter region required to initiate transcription of the CRISPR array (Pul et al., 2010; Lillestøl et al., 2006).

1.3.5 *cas* genes

A set of *cas* genes located close to the CRISPR array is needed to code for proteins required to mediate each of the three stages in CRISPR immunity (Jansen et al., 2002). CRISPR-Cas systems across species have a diverse complement of *cas* genes with only *cas1* and *cas2*, required for adaptation of the array, being universally conserved (Makarova et al., 2006). The *cas* gene complement can often be predicted from the repeat type of the CRISPR array, testament to the close functional relationship and coevolution of these elements. However, in organisms with several CRISPR arrays, one set of *cas* genes may code for proteins that serve multiple arrays. The classification of CRISPR-Cas system based on the associated *cas* genes is explored in detail below.

1.4 Diversity and classification of the CRISPR-Cas systems

Since the discovery of CRISPR elements there have been several attempts to classify the systems based on host organism, repeat structure and *cas* gene complement. However, a unifying classification has proved difficult to implement as

the transfer of systems between organisms by horizontal gene transfer, the rearrangement of CRISPR elements within a genome and the rapid evolution of *cas* genes complicates the identification of evolutionarily related elements (Makarova et al., 2015). The most up-to-date and comprehensive classification groups CRISPR-Cas systems into 2 classes, 6 types and 19 subtypes (Makarova et al., 2015; Shmakov et al., 2015) (Figure 1.4). This classification takes into account the locus architecture, *cas* gene complement, and the presence of certain signature *cas* genes. The most conserved feature of the different types is the presence of the *cas1*, *cas2*, and often *cas4*, genes involved in the adaptation stage of CRISPR-Cas immunity. The second key component of the CRISPR-Cas systems is the presence of a gene, or genes, coding for proteins involved in CRISPR interference. These interference-associated genes are thought to have been acquired and exchanged as modules and to have evolved independently, resulting in an incredible diversity in interference machinery between CRISPR-Cas subtypes (Makarova et al., 2015).

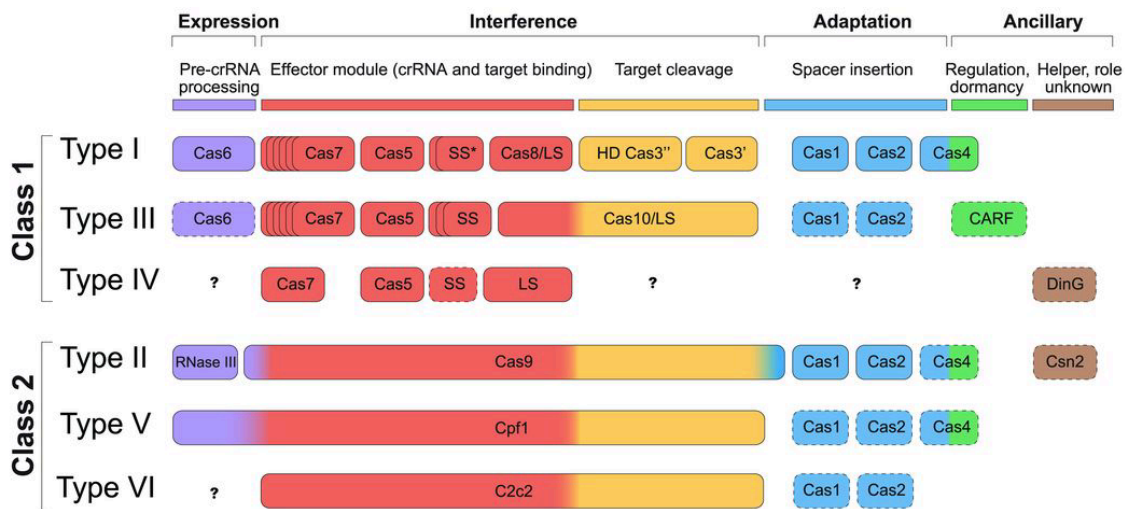


Figure 1.4 An up-to-date classification of the CRISPR-Cas systems

CRISPR-Cas systems are divided into two classes based on the interference module. Class 1 systems have a multi-subunit interference complex, whereas in class 2 systems interference is carried out by a single protein. Classes are further divided into 6 types based on their *cas* gene complement. The *cas* components are coloured dependent on the role they play in immunity; multiple colours imply a role in more than one stage of CRISPR-Cas defence. Dashed lines indicate that these components are missing in some subtypes and question marks indicate as yet unidentified genes, which complete the system. Interference complexes in class 1 systems contain a large and small subunit (LS and SS), which can be present as a fused construct in some subtypes (indicated by *). Adapted from Mohanraju et al., 2016.

The first level of grouping of CRISPR-Cas systems is based on the structure and composition of the interference machinery used to target and degrade foreign DNA. The class 1 systems have a multi-subunit interference complex, whereas those of

class 2 have a single, large protein complex that carries out a similar role (Makarova et al., 2015).

1.4.1 Class 1

Class 1 includes systems of types I, III and IV. These subdivisions are based on the presence of a signature protein: Cas3 in type I, Cas10 in type III, and the large Csf1 subunit in the recently-identified type IV system. All have multisubunit interference complexes, which are responsible for the interference stage in CRISPR defence.

The type I system is the most prevalent in both archaea and bacteria and is further divided into subtypes I-A to I-F and I-U (Makarova et al., 2011). Interference in this type is carried out by the CRISPR-associated complex for anti-viral defence (Cascade) (Brouns et al., 2008). The Cascade complex has been said to resemble a seahorse; with a head and a tail domain linked by a backbone made up of multiple subunits of the Cas7 family (Jore et al., 2011; Wiedenheft et al., 2011) (Figure 1.5, A). Cas5 is positioned at the base of the Cas7 backbone and the large subunit (often Cas8) makes up the tail of the Cascade complex. Both Cas5 and Cas7 are of the RAMP (repeat-associated mysterious protein) family and contain a characteristic RNA recognition motif (RRM), which is involved in crRNA binding. The Cas6 protein, required for CRISPR transcript processing, forms the head of the seahorse in some subtypes and two copies of the small subunit make up the 'belly' of the complex. The signature protein of the type I systems is Cas3, a helicase/nuclease protein often with a N-terminal HD-nuclease domain (Makarova et al., 2011). This protein is recruited to foreign nucleic acid targets by the Cascade complex and degrades DNA during the interference stage of CRISPR immunity (Beloglazova et al., 2011).

The type III systems are found in 30% of bacteria and 70% of archaea and are equipped with one of two multisubunit interference complexes (Shah & Garrett, 2011). The type III-A and III-D systems use the Csm complex, whereas the type III-B and III-C systems have a complex referred to as Cmr. A Cas6 endoribonuclease also processes CRISPR transcripts in type III systems, but is not stably associated with either Cmr or Csm. Cas6 is often encoded in a separate *cas* gene operon from the type III systems and is thought to only weakly associate with the complexes to deliver crRNAs, with one Cas6 potentially feeding multiple interference machines (Sokolowski et al., 2014). The type III complexes are said to resemble 'seaworms',

with narrow elongated topologies formed by two backbone filaments spiralling around one another (Staals et al., 2013, 2014) (Figure 1.5, A). These filaments are made up of Cas7-family proteins and multiple copies of the small subunit. The foot, or tail, of the complex is made up of the signature protein of this system type, the Cas10 large subunit (Staals et al., 2013, 2014).

Type I and III complex subunits have a low sequence similarity; however, the overall composition and topology of the complexes is strikingly similar (Figure 1.5, A and B). This is thought to provide evidence of their shared roles and suggests that the complexes have diverged from a common ancestor (Makarova et al., 2015).

The type IV systems have a minimal multisubunit effector complex resembling a degenerate Cascade. The large subunit, Csf1, is unique to this subtype and acts as the signature protein. Whether this system is a true CRISPR-Cas subtype is still unclear, as the known examples lack several key cas genes and are often located far from a CRISPR array (Makarova et al., 2015).

1.4.2 Class 2

The class 2 CRISPR-Cas systems (types II, V and VI) are almost always found in bacteria, with only one putative class 2 system in the archaea (Shmakov et al., 2015). This class of CRISPR-Cas system contains a single, multidomain protein involved in interference. In the type II systems, interference is carried out by endonuclease Cas9. This large protein has HNH and RuvC-like nuclease domains and is required for all three stages of the CRISPR response (Jinek et al., 2012).

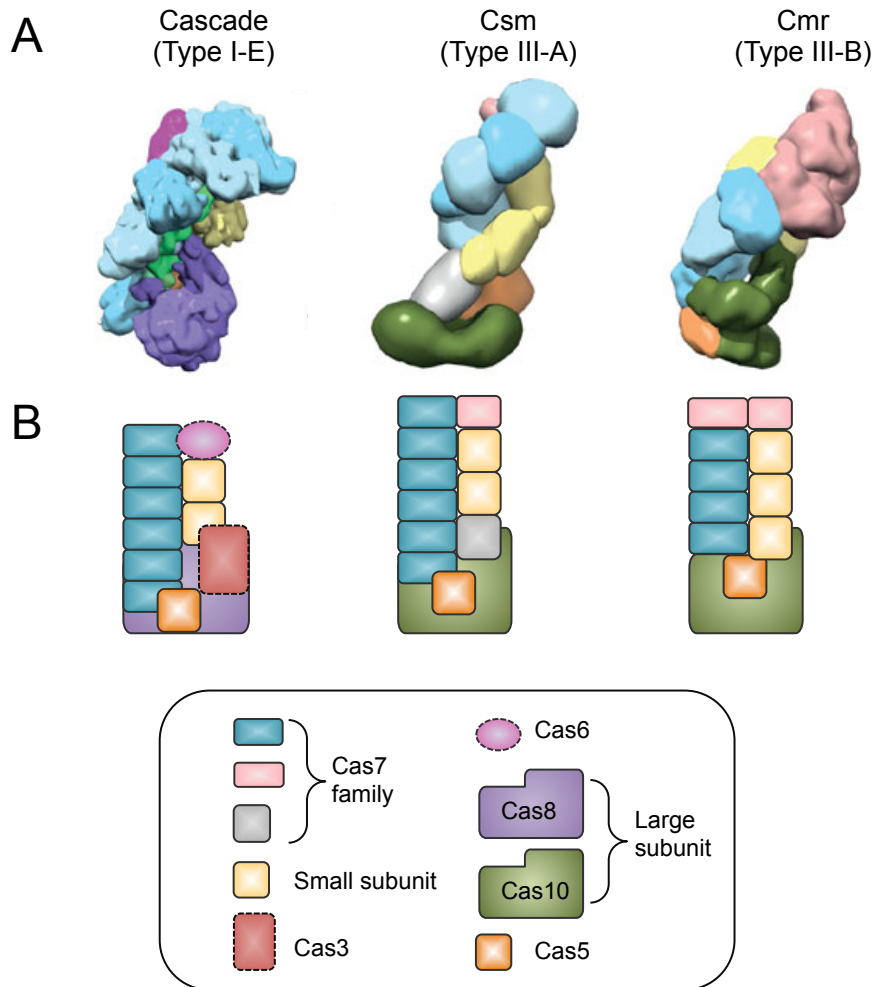


Figure 1.5 Conserved architecture of Class 1 interference complexes

A. Electron microscopy structures of Class 1 complexes. The left-hand structure is that of the *E. coli* Cascade complex (type I-E) without Cas3 (Electron Microscopy Data Bank (EMDB) accession 5314) (Wiedenheft et al., 2011). The middle structure is that of a type III-A Csm complex from *S. solfataricus* (EMDB accession 2420) (Rouillon et al., 2013) and the right-hand structure is of a Cmr complex from *Pyrococcus furiosus* (EMDB accession 5740) (Spilman et al., 2013). A key to the colours of the subunits is shown at the bottom of the figure. The architecture of all three complexes shares a large base subunit (Cas8/10), and helical filaments made up of Cas7 family proteins. **B.** Schematic showing subunit composition of the Cascade (I-E), Csm (III-A) and Cmr (III-B) complexes. Colours represent the same subunits as in **A**. The schematic shows the Cas3 helicase/nuclease docked on the large subunit of Cascade. It is recruited to the complex following binding of a complementary DNA and is required for target degradation. Dashed lines indicate subunits that are only transiently associated or non-essential. Figure adapted from van der Oost et al., 2014.

Type II loci also encode a trans-activating RNA (tracrRNA) that is partially complementary to the guide crRNA. tracrRNA and a cellular RNase III are required for CRISPR transcript processing in this system (Jinek et al., 2012). A partial duplex of tracrRNA and crRNA is used to guide the Cas9 protein to foreign sequences matching the crRNA guide. On recognition of a complementary target and correct PAM, both strands of the target are cut by one of the two Cas9 nuclease domains at

specific positions (Jinek et al., 2012). Due to the precise DNA cleavage, single protein component and the ability to modify targeted DNA by introducing a new crRNA, the Cas9 protein has become the tool of choice for a plethora of gene-editing projects, many of which are predicted to lead to significant therapeutic outcomes.

Other Class 2 systems (V and VI) also possess a single protein interference module. These large proteins have been shown to process CRISPR transcripts into short guide crRNAs and also to mediate sequence-specific DNA targeting and cleavage (Shmakov et al., 2015; Zetsche et al., 2015; Fonfara et al., 2016). These activities take place in separate active sites of the Class 2 complexes and while the type IV interference proteins require both crRNA and tracrRNA for activity (Shmakov et al., 2015), in the type V effectors there is no requirement for tracrRNA (Zetsche et al., 2015; Fonfara et al., 2016).

1.5 Casposons and the origins of CRISPR-Cas

Two groups of *cas1* genes have been identified that are not associated with a CRISPR array (Makarova et al., 2013b). The first set of *cas1* genes is not located in a conserved genomic location and their function remains unclear. However, solo *cas1* genes of the second group are consistently found close to genomic islands containing a gene coding for a divergent DNA polymerase B (PolB), flanked by inverted repeats and located between direct repeats. This arrangement is typical of mobile elements, called transposons, and led to the name 'casposons' being given to these *cas1* loci (Krupovic et al., 2014).

Casposons are the first examples of prokaryotic transposable elements similar to the eukaryotic Polinton/Maverick type II transposons (Krupovic et al., 2014). These elements use the associated PolB to replicate and also often contain genes coding for proteins required for processing and insertion of the transposon (Kapitonov & Jurka, 2006; Krupovic et al., 2014). The *cas1* gene is universally conserved in CRISPR-Cas systems and is required for the addition of new spacers during the adaptation stage of CRISPR-Cas immunity (Makarova et al., 2011). Given the predicted role of Cas1 in integrating new protospacers, the solo *cas1* genes were hypothesised to code for proteins that act as the integrases required for insertion of casposons (Krupovic et al., 2014). This theory was confirmed when the casposon-

encoded Cas1 from *Aciduliprofundum boonei* was shown to integrate short oligonucleotides and larger casposon elements into target DNA (Hickman & Dyda, 2015). Integration required the substrate to have inverted terminal repeats (TIR) and insertion led to the duplication of the target site.

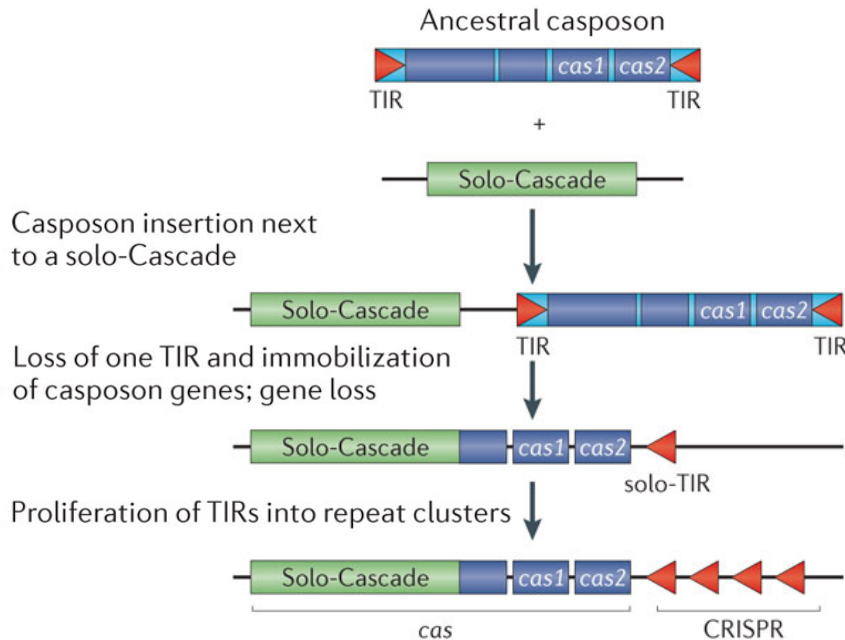


Figure 1.6 The multi-partite origin of the CRISPR-Cas system

A proposed model explaining the origins of CRISPR-Cas systems. It shows a transposable element containing an ancestral *cas1* gene, called a casposon, being inserted next to an innate immune system. The loss of casposon genes and degradation of one terminal inverted repeat (TIR) is then thought to have produced a fully-formed and functional CRISPR-Cas system. Adapted from Koonin & Krupovic 2015.

It has been hypothesised that the casposon *cas1* genes are the ancestors of the CRISPR *cas1*, which were hijacked by the cell to form an integral part of an adaptive immune system (Figure 1.6). A crucial step in the formation of a CRISPR-Cas system is thought to have taken place when a casposon was inserted next to a stand-alone RNA-guided interference complex (Koonin & Krupovic, 2015). In support of this hypothesis, the type IV systems are composed of a primitive *cascade*-like effector complex, but do not seem to have a *cas1* or *cas2* gene (Makarova et al., 2015). The degeneration of one TIR of the casposon is thought to have trapped the ancestral *cas1* and inverted repeats in the host genome next to the primitive interference complex – forming a functional CRISPR-Cas system. The integrase activity of the casposon Cas1 protein is then thought to have led to the insertion of new pieces of DNA between repeat sequences, forming the CRISPR array (Koonin & Krupovic, 2015). As Cas2 proteins are homologous to the VapD toxins, it is

suggested that *cas2* genes were acquired by the CRISPR-Cas systems following the integration of a mobile toxin/antitoxin system (Makarova et al., 2006). This hypothesis highlights the contribution of transposable elements to the evolution of prokaryotes, and to the CRISPR-Cas immune system in particular.

1.6 The three stages of CRISPR-Cas immunity

The three stages required for CRISPR-Cas immunity will be described in detail in this section. This thesis will focus on the mechanism of adaptation in organisms containing type I and type III CRISPR-Cas systems, belonging to class I. Therefore, the remainder of this introduction will focus on the function of class 1 systems, while immunity mediated by class 2 systems will not be discussed in detail.

1.6.1 Stage 1: Adaptation

The first stage in CRISPR-Cas-mediated immunity is known as adaptation. It involves the capture of a foreign segment of DNA and its incorporation into the CRISPR array. When the experimental work for this project began, very little was known about this step in CRISPR-Cas immunity. However, some recent breakthroughs have significantly advanced our understanding of adaptation. A general introduction to what was known about adaptation prior to the work described in this thesis will be given here, while recent key breakthroughs will be introduced and discussed in individual results chapters.

1.6.1.1 *cas1* and *cas2* are sufficient for naïve adaptation

Naïve adaptation requires CRISPR-Cas systems to detect the entry of a previously unencountered foreign genetic element and to immunise the host against this threat by capturing and incorporating a new spacer. The universally conserved Cas1 and Cas2 proteins were first implicated in the capture and integration of a protospacer when they were demonstrated to play no role in the expression or interference stages of CRISPR-Cas immunity in *E. coli* (Brouns et al., 2008). The overexpression of Cas1 and Cas2 alone was later shown to be sufficient for the integration of new spacers into a minimal CRISPR array, containing a leader and one repeat (Yosef et al., 2012). This indicated that, in the type I-E system of *E. coli* at least, *cas1* and *cas2* were the only *cas* genes required for adaptation.

Additionally the protospacers selected for incorporation in this system were excised primarily from regions flanked by a AWG (W = A or T) PAM motif, suggesting that Cas1 and Cas2 were capable of sequence-specific spacer selection (Yosef et al., 2012).

The sequences of Cas1 proteins differ considerably, with the few conserved residues being clustered around a metal-binding active site (Figure 1.7, A). However, structural studies have revealed that Cas1 proteins with little sequence similarity share a very similar structure. Cas1 proteins exist as dimers that have a conserved novel fold and common domain organisation, with the overall protein architecture said to resemble a butterfly (Wiedenheft et al., 2009) (Figure 1.7, B). The N-terminal β -strand domain makes up the lower wings of the butterfly, while the helix-rich C-terminal domain makes up the large upper wings. A group of conserved residues (E141, H208 and D221 in *E. coli*) in the C-terminal domain co-ordinate a divalent metal ion required for activity.

Various biochemical studies of Cas1 proteins have identified a non-specific, metal-dependent nuclease activity on various nucleic acid substrates. The Cas1 from *E. coli* was found to cleave linear and branched DNA and RNA substrates and to associate with proteins involved in DNA repair pathways (Babu et al., 2011). The Cas1 from *Pseudomonas aeruginosa* was also found to cut DNA substrates and to generate DNA fragments of roughly 80 bp (Wiedenheft et al., 2009). These activities are not easily reconciled with the predicted role of Cas1 in the integration of new spacers. Firstly, the nonspecific nature of the nucleic acid cleavage reported does not fit with the role of Cas1 in integrating spacers in a highly specific manner only between the leader and repeat 1 of the CRISPR. Secondly, the 80 bp DNA fragments reported to be produced by Cas1 are much larger than the expected (32/33 bp) spacer size in *E. coli*. Furthermore, the Cas1 protein from *S. solfataricus* was shown to possess no nuclease activity, but rather to promote the annealing of complementary single-stranded DNA (Han et al., 2009).

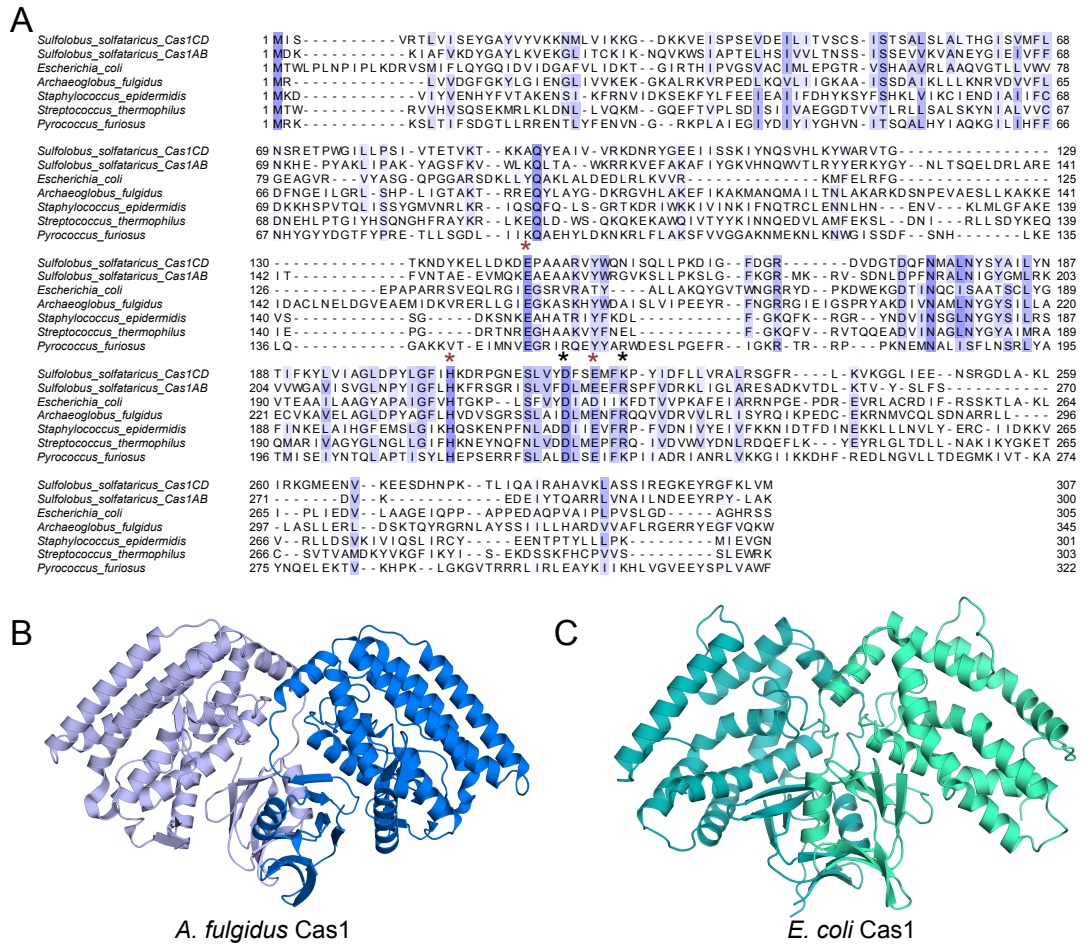


Figure 1.7 Cas1 protein sequence and structure
A. A global alignment of Cas1 proteins from across CRISPR-Cas subtypes. Two Cas1 proteins from *S. solfataricus* were included in the alignment. Other protein sequences come from the Cas1 of: *E. coli* (Type I-E); *A. fulgidus* (I-A); *S. epidermidis* (III-A); *P. furiosus* (III-B); *S. thermophilus* (II-A). Asterisks indicate residues in the Cas1 active site required for nuclease activity (residues E141, H208, D218, D221 and K224 in *E. coli* Cas1) and red asterisks indicate residues that co-ordinate a divalent metal ion. Residues are coloured based on their conservation, with the lighter colours showing weakly conserved residues and darker colours indicating more conservation at this position. The alignment was performed using Jalview (Waterhouse et al., 2009) and the BLOSUMSCORE conservation threshold was set to 30%. **B.** The crystal structure of the *A. fulgidus* Cas1 showing a dimeric structure with α -helix-rich C-terminal and a β -sheet-rich N-terminal domain (PDB code 4N06) (Kim et al., 2013). **C.** The crystal structure of *E. coli* Cas1 (PDB 3NKD) (Babu et al., 2011).

When the work in this thesis began, the role of Cas2 in the integration of new spacers was also poorly understood. The crystal structure revealed Cas2 to be a small dimeric protein with a ferredoxin-like fold (Beloglazova et al., 2008). Biochemical investigations identified both metal-dependent RNase and DNase activities for Cas2 proteins (Beloglazova et al., 2008; Nam et al., 2012), while another study failed to identify any nuclease activity for the Cas2 of *Desulfovibrio vulgaris* (Samai et al., 2010). Cas1 and Cas2 are always coded for by genes located adjacent, or close to one another, in the same *cas* operon (Makarova et al.,

2006) and are present as a fused construct in *Trichomonas tenax* (Plagens et al., 2012). Given the genetic association and shared role in adaptation, Cas1 and Cas2 have long been predicted to form a functional complex *in vivo*. However, this association was only recently demonstrated for the proteins of *E. coli* (Nuñez et al., 2014) (discussed in Chapter 3). Genes coding for the Cas4 exonuclease (Zhang et al., 2012a) and the uncharacterised Csa1 protein are also often located close to, or present as a fusion construct with *cas1* and *cas2* genes, indicating that these proteins may also play a role in adaptation in some CRISPR-Cas systems. In summary, the activities identified for Cas1 and Cas2 proteins at the beginning of this project were conflicting and did not link to their predicted roles *in vivo*. Therefore, a primary aim of the work described in this thesis was to characterise biochemically the activity of Cas1 and Cas2 proteins from *S. solfataricus* in an attempt to learn more about the role they play in adaptation.

1.6.1.2 Protospacer capture

Initiation of adaptation requires the host CRISPR-Cas system to detect the presence of a foreign genetic element and capture a spacer from this invading DNA. This process relies on a level of self versus non-self discrimination to avoid incorporation of spacers matching the host, which may lead to targeting of the host genome and autoimmunity. Evidence of this substrate discrimination was observed during a study of adaptation in *E. coli*, where protospacers were found to be 200 times more likely to come from an invading plasmid compared to the host genome, when the length of each element was taken into account (Yosef et al., 2012). However, the mechanism by which this selection bias was maintained remained a mystery until recently.

Levy and colleagues discovered that hotspots for protospacer uptake from plasmids coincided with replication termination sites (Ter sites), where stalled replication forks often lead to double-strand breaks (Levy et al., 2015). The authors also showed that the helicase/nuclease RecBCD was crucial for the biased uptake from these sites and that protospacer uptake hotspots were delineated at one extreme by a Ter site and on the other by an octameric Chi motif (Levy et al., 2015). The RecBCD nuclease/helicase is known to be recruited to DNA ends caused by DNA breaks or the injection of a linear viral genome and to catalyse DNA unwinding and

degradation from these sites until a Chi motif is reached (Dillingham & Kowalczykowski, 2008).

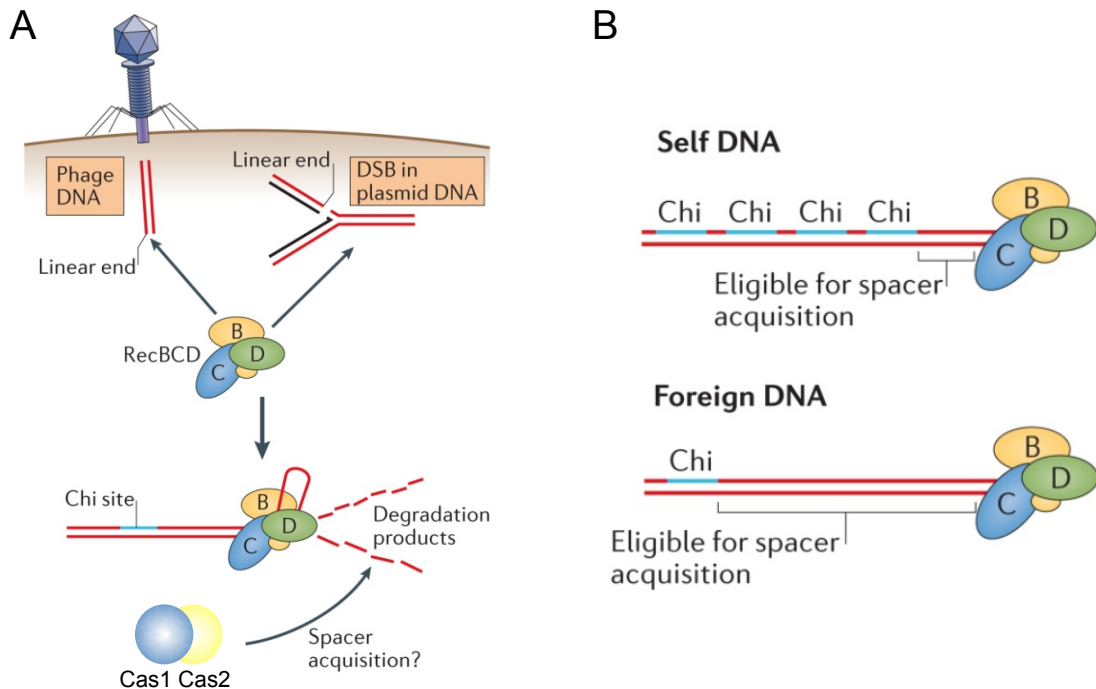


Figure 1.8 A model for self versus non-self discrimination during spacer uptake

A. Shows a schematic explaining the role of linear ends in protospacer generation. RecBCD recognises and degrades linear DNA ends produced from double-strand breaks or invading viral DNA until a Chi site is reached. The degradation products of RecBCD are then thought to be used as protospacer precursors in some CRISPR-Cas systems. **B.** The *E. coli* host genome is highly enriched in Chi sites compared to foreign DNA. As Chi sites act to pause RecBCD DNA degradation, this results in far fewer DNA fragments that can be used as spacer substrates being produced from host DNA, compared to viral DNA. Adapted from Amitai & Sorek 2016.

Taking these results together the authors formulated a model in which, firstly, RecBCD is recruited to linear DNA ends created either by the collapse of a stalled replication fork at a Ter site or the injection of a linear viral genome. Secondly, the helicase/nuclease activity of RecBCD degrades DNA until a stop signal, in the form of a Chi site, is reached. Finally, the authors predicted that Cas1 and Cas2 use DNA fragments produced by RecBCD activity as the substrates for integration during adaptation (Levy et al., 2015) (Figure 1.8, A). This model explains the propensity for spacer uptake from between Chi and Ter sites in plasmid DNA. In addition, it also elegantly explains the preference for incorporation of exogenous DNA, as Chi sites are around 14-fold enriched in *E. coli* DNA compared to exogenous DNA, meaning considerably less host DNA is eligible for RecBCD degradation compared to viral or plasmid DNA (Levy et al., 2015) (Figure 1.8, B).

As RecBCD is only present in Gram-negative bacteria (Dillingham & Kowalczykowski, 2008), its role in producing spacer precursors must be fulfilled by alternative proteins in other CRISPR systems. The Cas4 protein contains a RecB-like nuclease motif and has been shown to exhibit DNA unwinding activity on double-stranded DNA and endo- and exonuclease activity on single-stranded DNA (Zhang et al., 2012a; Lemak et al., 2013). Therefore, this protein might have the potential to substitute for RecBCD in some CRISPR-Cas systems. The protospacer precursors produced by both of these potential routes are likely to be single-stranded in nature and may reanneal with a complementary strand or undergo replication before integration (Levy et al., 2015). Furthermore, protospacers resulting from Cas4 or RecBCD degradation would vary greatly in length, indicating that the adaptation machinery may act as a molecular ruler to specify a particular spacer length before insertion.

1.6.1.3 Primed adaptation

Adaptation during infection often results in the incorporation of multiple spacers into the CRISPR array. Swartz and colleagues found that there was a bias towards multiple spacers acquired against an invader originating from the same DNA strand. They suggested that a positive feedback loop existed, where the original addition of a spacer and subsequent targeting of one strand of the invading DNA, promoted further spacer uptake specifically from the targeted strand (Swartz et al., 2012).

This theory was expanded when mismatches between the CRISPR guide RNA and phage DNA were found to trigger hyperactive spacer uptake into the CRISPR array (Datsenko et al., 2012). Adaptation under these circumstances, referred to as 'primed adaptation', required Cas1 and Cas2 as well as the Cascade interference complex and the associated Cas3 helicase/nuclease (Datsenko et al., 2012). Primed adaptation is thought to be a type of failsafe mechanism employed by prokaryotes to retain robust immunity to rapidly evolving invaders, by ensuring that point mutations in phage do not lead to their escape from the CRISPR-Cas immune system of the host (Datsenko et al., 2012).

Due to the bias for spacers acquired during primed adaptation to come from the same strand as the original protospacer, a scanning mechanism was suggested for priming in type I-E systems (Datsenko et al., 2012). It is hypothesised that a mismatch between the crRNA, used to guide the Cascade complex to a target, and

viral DNA leads to a conformational change in the interference complex, which instead of recruiting Cas3 for target degradation, recruits Cas1 and Cas2 for further spacer uptake. The priming complex is then thought to move away from the mismatched protospacer, scanning the same DNA strand for PAM sequences. PAM recognition is then thought to trigger the uptake of new protospacers through an unknown mechanism (Datsenko et al., 2012). Primed protospacer uptake has also been identified in the type I-F systems. However, protospacers captured by priming in this system come from both strands of the foreign DNA, suggesting that the unidirectional sliding mechanism, thought to exist in *E. coli*, may not be a widespread method for primed protospacer selection (Richter et al., 2014).

1.6.1.4 Mechanism of integration

This insertion of new spacers is polarized and always occurs between leader and repeat 1 (Pourcel et al., 2005; Lillestøl et al., 2006), which suggests that these elements contain important motifs that guide docking of the Cas1 and Cas2 proteins. In support of this hypothesis, the last 60 bp of the leader and the first repeat in *E. coli* were shown to be essential and sufficient for integration of new spacers (Yosef et al., 2012). However, the precise nature of the motifs at the leader-repeat 1 junction that guide integration remains unknown, and their identification will be key to understanding the mechanism of adaptation.

The integration of a new spacer in *E. coli* leads to the expansion of the CRISPR array by ~61 bp, which corresponds to the addition of a new protospacer (32/33 bp) and the duplication of the repeat sequence (29 bp) (Yosef et al., 2012). Mutations introduced into the existing leader-proximal repeat were found to be replicated with the newly added repeat 1 following integration of a spacer (Yosef et al., 2012). This finding suggested that one strand of the first repeat acts as a template for synthesis of a new repeat during adaptation (Yosef et al., 2012). PAM sequences are crucial for the selection and correct orientation of protospacers, with the protospacer end that was previously proximal to the PAM always being inserted towards the leader sequence (Díez-Villaseñor et al., 2009). Swarts and colleagues reported that in *E. coli* the uptake of spacers with divergent PAM sequences led to a change in the last nucleotide of the duplicated repeat during adaptation (Swarts et al., 2012). This nucleotide was found to always match the last nucleotide of the PAM associated with the most recently added spacer (Swarts et al., 2012; Goren et al., 2012) (Figure

1.9, A). As the consensus PAM motif for spacer selection in *E. coli* is AWG, the last nucleotide of the first repeat is usually a G (Goren et al., 2012). These data indicated that the last nucleotide of the PAM is excised with the protospacer and is inserted into the *E. coli* CRISPR as the final nucleotide of the duplicated repeat (Swarts et al., 2012; Goren et al., 2012). This process seems to guide the orientation in which new spacers are inserted during adaptation (Figure 1.9, B).

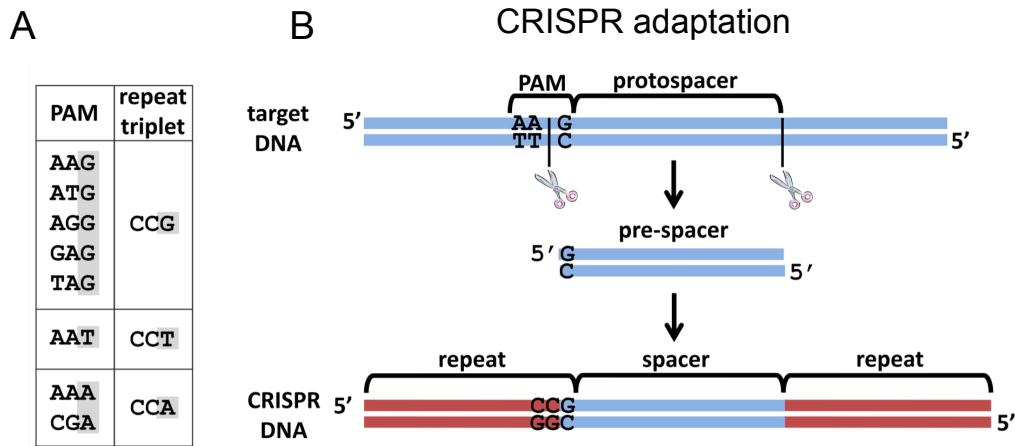


Figure 1.9 The last nucleotide of repeat 1 is determined by PAM

A. In the type I-E CRISPR-Cas system of *E. coli*, the last nucleotide of the PAM sequence associated with spacer 1 always matches the last nucleotide of repeat 1. The consensus PAM is AWG, where W is an A or T. The selection of a consensus PAM results in the last nucleotide of the repeat being a G. Selection of a PAM differing at the third position led to a change in the last nucleotide of the duplicated repeat during adaptation. **B.** Shows the model proposed to explain the link between the PAM sequence and the last nucleotide of repeat 1. Cas1 and Cas2 proteins select a protospacer with a recognized PAM and excise the sequence as well as the last nucleotide of the PAM. The protospacer end containing the residue derived from PAM is always orientated towards the leader during adaptation. The last nucleotide of repeat 1 is not copied from the previous repeat, but instead is made up by the incorporated, PAM-derived, nucleotide. Adapted from Swarts et al., 2012.

The exact mechanism of spacer integration is still not understood. However, it has been reported that an erroneous insertion, two nucleotides upstream of the correct insertion site at the leader-repeat 1 junction, was accompanied by the duplicated repeat being shortened at the leader-distal end by two nucleotides (Díez-Villaseñor et al., 2013). The authors concluded that adaptation happens through the nicking of the CRISPR locus, usually at the leader-repeat 1 junction, which is followed by a second locus nicking a defined distance from the first, selected through a molecular-ruler rather than sequence-specific mechanism (Díez-Villaseñor et al., 2013). The spacer 3' ends were then hypothesised to be joined to these nicked ends, forming a gapped intermediate, which is repaired by host factors to complete adaptation (Díez-Villaseñor et al., 2013).

More recently experimental evidence of adaptation intermediates has strengthened the staggered nicking and insertion model of adaptation (Arslan et al., 2014). Following expression of Cas1 and Cas2 for 18 hours in the presence of a plasmid containing a CRISPR locus, Southern blots were carried out with probes against sequences upstream and downstream of the expected insertion site (Arslan et al., 2014). As well as a major product corresponding to the entire CRISPR array, minor products of shorter lengths were obtained, indicating cleavage of the CRISPR array at specific points (Arslan et al., 2014). The expression of Cas1 and Cas2 as well as an intact repeat 1 sequence was crucial for the detection of adaptation intermediates. Further analysis of these minor products revealed that they were produced by the staggered nicking of the CRISPR locus at the 5' ends of the first repeat and the coordinated joining of a protospacer to the repeat ends (Arslan et al., 2014) (Figure 1.10).

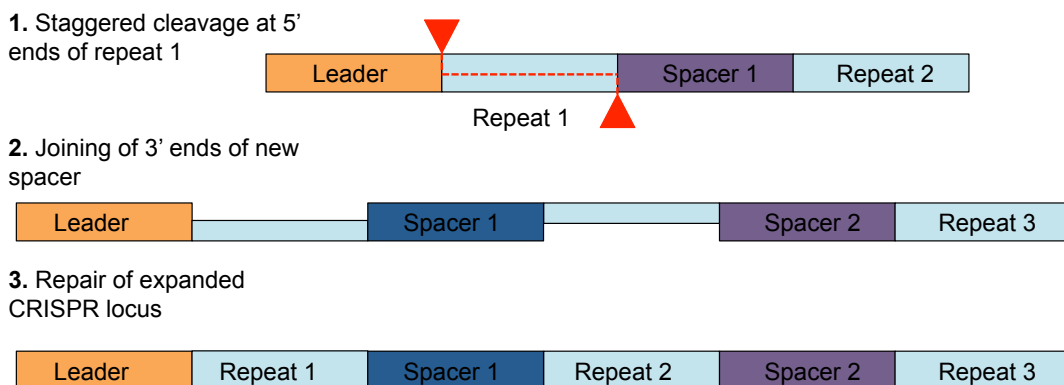


Figure 1.10 Model for the integration of a new spacer

The integration of a new spacer into the CRISPR locus is thought to happen through a concerted cleavage and ligation reaction. The putative mechanism begins with staggered nicks being made at the 5' ends of the first repeat (1). The 3' ends of the incoming protospacer are joined to the repeat, forming a gapped intermediate (2). Cas1 and Cas2 are thought to catalyse this activity; however, their exact role is unclear. The final step in the integration is the repair of the CRISPR locus (3). A host polymerase and ligase are thought to be required for this step. Figure based on data from Arslan et al., 2014.

As no nicking of the locus occurred without the joining of a protospacer, the authors predicted that spacer end joining occurred through a one-step transesterification reaction mediated by Cas1 and Cas2, in which the 3' hydroxyl residues (3' OH) of the incoming protospacer are used to attack the host locus (Arslan et al., 2014). This suggested concerted cleavage-ligation mechanism of adaptation had clear similarities to the integration of target DNA by viral integrases (Chow et al., 1992). However, as no transesterification activity had been identified for the Cas1 or Cas2 proteins and the precise DNA manipulation steps required for adaptation remained

enigmatic, further research was required to understand the process of spacer insertion during adaptation.

The final step in adaptation is thought to require the repair of the gapped intermediates created by the staggered nicking of the host genome and joining of protospacer ends during spacer integration. The gap-filling and ligation roles required for this step are predicted to be provided by a host polymerase and ligase. A recent report of DNA polymerase I being essential for both naïve and primed adaptation in *E. coli* supports this hypothesis (Ivančić-Baće et al., 2015). However, further work is required to confirm the exact mechanism and requirements of this final stage in adaptation.

1.6.2 Stage 2: Transcription and processing of the CRISPR array

Following insertion of a new protospacer during adaptation the CRISPR array is transcribed and processed to generate short CRISPR RNAs (crRNAs), crucial for guiding sequence-specific degradation of matching invader sequences (Brouns et al., 2008). The promoter for transcription of the CRISPR is usually located in the corresponding leader sequence (Lillestøl et al., 2009; Pul et al., 2010). Initiation in the leader leads to the production of a long transcript RNA containing repeat and spacer units, which is referred to as precursor CRISPR RNA (pre-crRNA) (Lillestøl et al., 2009; Tang et al., 2002; Lillestøl et al., 2006). The pre-crRNA transcript then undergoes processing into mature crRNAs containing a single spacer sequence flanked by short repeat arms (Hale et al., 2008; Carte et al., 2008). Interestingly, Hale and colleagues identified a gradient in the abundance of crRNA, with spacer sequences closest to the promoter being the most abundant and those at the end of the array being rare (Hale et al., 2008). This gradient was predicted to have functional significance, as transcripts guiding interference to recently encountered invaders are plentiful, facilitating a robust immunity against these elements. In contrast, crRNAs matching ancient infections, which may no longer threaten the organism, are only minimally transcribed, thus saving energy and resources (Hale et al., 2008).

In type I and III CRISPR-Cas systems, primary processing of pre-crRNA is carried out by Cas6 proteins (Brouns et al., 2008). Cas6 proteins are members of the RAMP family and usually contain two ferredoxin-like folds (Makarova et al., 2006;

Carte et al., 2008). Often Cas6 proteins have evolved with a particular repeat cluster and specifically target and process only this cluster (Sokolowski et al., 2014). A conserved histidine residue in the active site of Cas6 proteins is required for processing in most systems. However, an atypical cleavage mode was identified for the unusual, dimeric Cas6 of *S. solfataricus*, which lacks the canonical active site residue (Reeks et al., 2013). Additionally, none of the predicted catalytic residues of Cas6 were found to be absolutely required for processing and instead Cas6 binding was suggested to promote RNA auto-catalysis (Reeks et al., 2013).

Pre-crRNA is cleaved 8 nucleotides upstream of the 3' end of repeat sequences, forming unit-length crRNAs containing an intact spacer sequence, flanked by defined 5' and 3' repeat handles. This cleavage is guided by different motifs depending on the CRISPR-Cas type. In subtypes with structured repeat clusters (Kunin et al., 2007), processing by Cas6 occurs at the base of the stem loop and both sequence and structure of the pre-crRNA is crucial (Brouns et al., 2008; Haurwitz et al., 2010) (Figure 1.11). A conserved sequence motif and a structured pre-crRNA are also crucial for processing in the type I-C CRISPR-Cas of *Bacillus halodurans*. However, in this system the cleavage is carried out by the endoribonuclease Cas5 (Nam et al., 2012b).

In systems where the pre-crRNA is unstructured, the sequence of the repeat defines the binding site of Cas6 and a ruler mechanism is thought to determine the length of the crRNA product (Carte et al., 2008). In the type III-B system of *P. furiosus*, Wang and colleagues demonstrated that the 5' end of the unstructured repeat was tethered between the ferredoxin folds on one side of Cas6, while the rest of the repeat was wrapped around the protein. The length of the processed crRNA was defined by the distance between the tethered 5' end of the RNA and putative active site on the other side of the enzyme (Wang et al., 2011).

In some type I systems the Cas6 protein is a crucial component of the Cascade complex (Brouns et al., 2008). These enzymes process structured pre-crRNA into crRNA units and remain bound to the 5' handle (Haurwitz et al., 2010; Sternberg et al., 2012). No further crRNA processing is required for interference in these systems and crRNA is directly fed from Cas6 to the Cascade complex (Figure 1.11). In type III and some type I systems Cas6 is not strongly associated with the interference complexes (Carte et al., 2008). Instead, it is thought to process pre-crRNA and transiently associate with an interference complex to deliver the mature

guide (Lintner et al., 2011b). This mode of processing allows multiple turnover of substrate by the Cas6 enzyme and in the type III systems also seems to facilitate Cas6 enzymes supplying multiple interference complexes with mature RNA (Sokolowski et al., 2014).

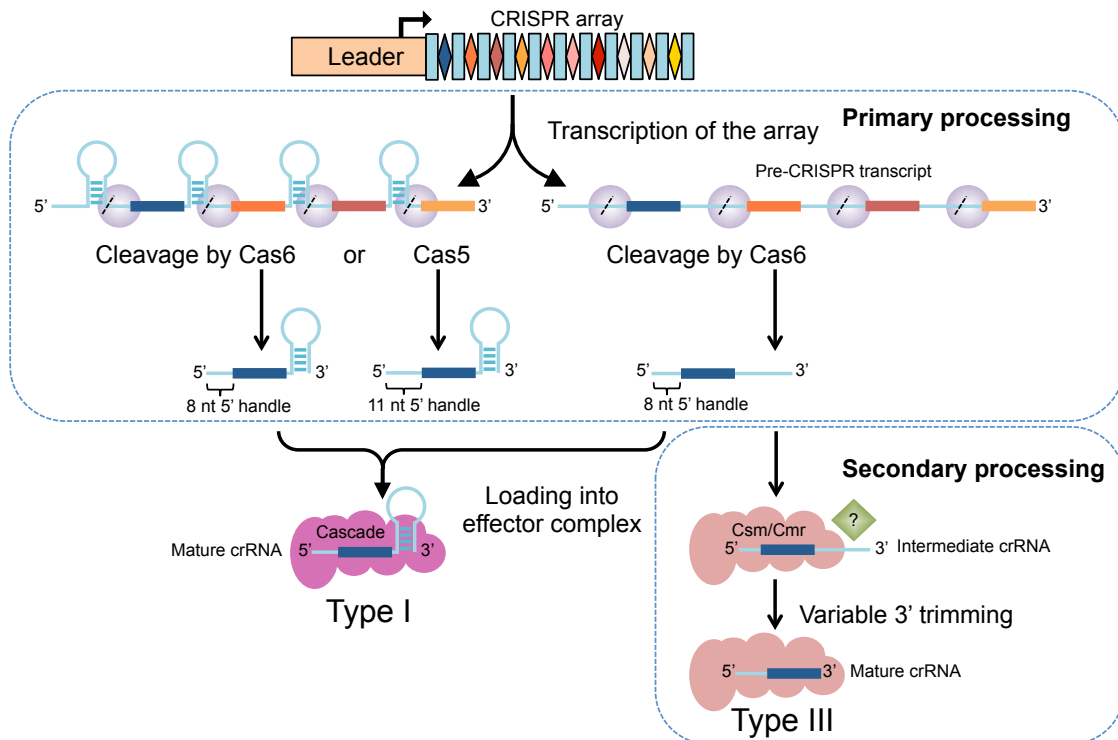


Figure 1.11 pre-CRISPR RNA processing pathways in type I and III systems

Following the addition of a new spacer the entire CRISPR locus is transcribed to produce a long pre-CRISPR RNA. Depending on the repeat sequence this RNA may contain hairpins or be unstructured. The transcript undergoes primary processing by Cas6 or Cas5 (type I-C) endoribonucleases. A single cut is made in each repeat unit as indicated by dashed lines. In type I systems these processed units are loaded into Cascade complexes as the mature crRNA form, containing an 8/11 nt 5' handle and a 3' repeat end. In type III systems these units are the intermediate form of crRNA and are further processed at the 3' end by an unknown nuclease following loading into Cmr or Csm complexes. Adapted from Charpentier et al., 2015 and Hochstrasser & Doudna, 2014.

In type III systems, following cleavage by Cas6, the unstructured, repeat-derived 3' handle of the intermediate crRNA is removed through a poorly understood mechanism (Carte et al., 2008; Hale et al., 2009) (Figure 1.11). This secondary processing was shown to be independent of the structure or sequence of the crRNA in *S. epidermidis*, but required the Csm3 backbone subunit of the type III-A complex (Hatoum-Aslan et al., 2013). Each Csm3 subunit was demonstrated to bind along the length of crRNAs, protecting 6 nt segments. The 3' ends of intermediate crRNAs not bound by the Csm complex are exposed to varying degrees of trimming

by unidentified host nucleases to produce mature guide crRNAs of multiple lengths (Hatoum-Aslan et al., 2013).

1.6.3 Stage 3: Interference

In the final stage of CRISPR immunity, called the interference stage, Cas proteins use the mature guide crRNAs to target and destroy complementary foreign genetic material. The elucidation of the function of CRISPR-Cas as a prokaryotic immune system prompted initial comparisons with RNAi-mediated gene silencing in eukaryotes, and led to suggestions that mRNA may be the target of CRISPR-Cas interference (Makarova et al., 2006). However, subsequent studies demonstrated that spacers originating from both sense and anti-sense strands and non-coding regions conferred CRISPR-Cas-mediated immunity (Shah et al., 2009), implying that dsDNA rather than RNA was the target of CRISPR interference. Furthermore, the introduction of a self-splicing intron into a targeted protospacer led to the loss of CRISPR-mediated immunity. As the RNA sequence was unchanged in this experiment and retained complementarity to the guide crRNA, the loss of targeting confirmed that CRISPR-Cas immunity relied on DNA interference (Marraffini & Sontheimer, 2008). The first direct evidence of CRISPR-Cas DNA targeting was obtained in *S. thermophilus*, where acquired CRISPR spacers against a phage or plasmid were found to lead to Cas9-dependent cleavage within the protospacer sequence of the foreign DNA (Garneau et al., 2010).

1.6.3.1 Type I interference

In type I systems, dsDNA targeting is carried out by the Cas3 nuclease recruited to a target-bound Cascade complex (Brouns et al., 2008; Jore et al., 2011). As the type I-E interference mechanism of *E. coli* is the most well studied, this summary will focus on the structure and mechanism of interference of the Cascade complex present in this subtype. Following processing by Cas6, the Cascade complex assembles on the mature guide crRNA, with the fully formed complex loosely resembling a seahorse (Jore et al., 2011). The spacer region of the crRNA is tightly bound in a groove formed by six Cas7 subunits, which make up the backbone of Cascade (Wiedenheft et al., 2011) (Figure 1.12, A). The 5' end of the crRNA is coordinated in a pocket formed between the large subunit (Cas8e/Cse1), Cas7 and Cas5 subunits at the tail of the complex. The 3' end of the crRNA, containing the repeat-derived hairpin, is tightly bound by Cas6 and protrudes from the head of the

complex. A dimer of Cse2, the small subunit, spans the ‘belly’ of the complex, connecting head and tail domains (Wiedenheft et al., 2011).

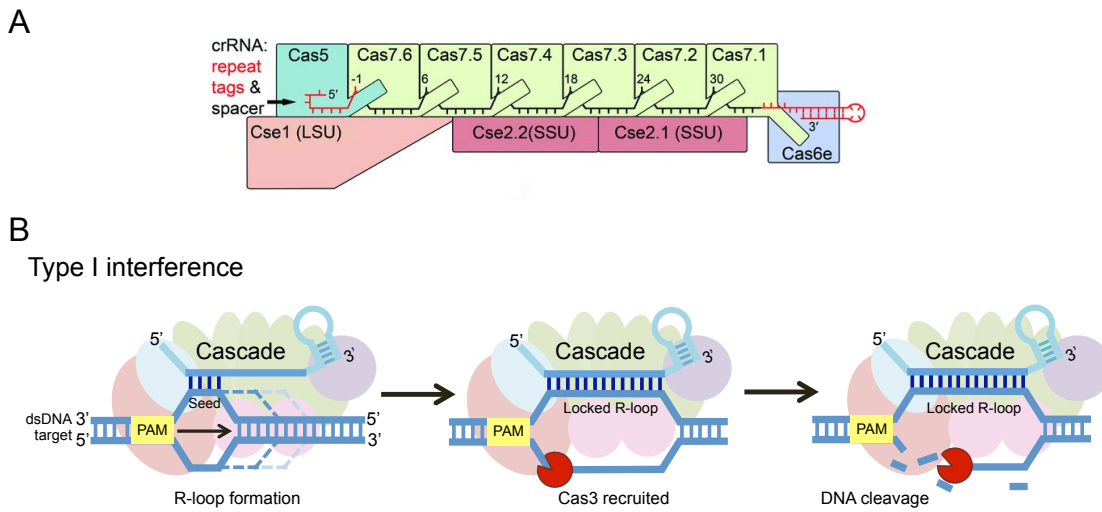


Figure 1.12 Target DNA binding and interference in type I systems

A. Cartoon of Cascade (type I-E) bound to a guide crRNA. A thumb domain of the multiple Cas7-family backbone subunits flips every 6th base of the crRNA strand, preventing base pairing at these sites with the target. Cas6 binds the 3' crRNA hairpin and forms the ‘head’ of the Cascade complex in type I-E. The large subunit (LSU) and Cas5 make up the base and the small subunit (SSU) makes up the belly of the Cascade complex. **B.** Model of DNA interference mediated by Cascade. Colours match those used in **A**. In the first step, dsDNA is scanned for a viable PAM sequence; this scanning is carried out by the large subunit. Once a PAM is reached, the target duplex is destabilised, allowing base-pairing of a single DNA strand with the crRNA guide. Complementarity at the first 7 nucleotides of the crRNA spacer sequence, called the seed, is crucial for R-loop formation. Base pairing continues along the crRNA and the displaced strand is bound by the small subunits of the belly. Complete base pairing between target and guide leads to R-loop locking and the recruitment of helicase/nuclease Cas3, which docks on the large subunit. Cas3 nicks the non-target strand before unwinding and degrading the target in a 3' - 5' direction. Adapted from Plagens et al., 2015 and Rutkauskas et al., 2015.

The Cas7 subunit closest to the base of the complex adopts a different conformation to the other backbone subunits. This opens a gap in the backbone, which exposes a short region of crRNA for base pairing and also allows dsDNA to pass through this space to be scanned for potential targets (Hochstrasser et al., 2014). Initial recognition of a target relies on the large subunit detecting a PAM sequence during this scanning. An *in vitro* study showed that supercoiled DNA is required as a target for Cascade binding (Westra et al., 2012). A flexible loop of the large subunit makes sequence-specific contacts with a cognate PAMs, resulting in the opening of the duplex DNA (Sashital et al., 2012). The next step in validation of the target involves the exposed 5' end of the crRNA base pairing with a seven nucleotide seed sequence in the target adjacent to the identified PAM (Semenova et al., 2011).

Mutations in either the PAM or positions 1-5 and 7 of the seed abolish interference mediated by Cascade (Semenova et al., 2011).

Seed complementarity triggers binding along the rest of the crRNA sequence in a 5' to 3' direction to the potential target (Figure 1.12, B). The crRNA:target duplex does not adopt a helical arrangement and instead has a ribbon-like structure composed of short 5 bp stretches of complementarity, with every 6th base of the crRNA flipped out by a 'thumb' region of the Cas7 backbone subunits (Wiedenheft et al., 2011; Mulepati et al., 2014) (Figure 1.12, A). Base pairing of crRNA and the target strand displaces the non-target DNA strand, forming an R-loop (Jore et al., 2011) (Figure 1.12, B). The non-target strand is co-ordinated by the large subunit of Cascade and is thought to feed around the dimeric small subunit that makes up the 'belly' of the complex (Mulepati et al., 2014). Once R-loop formation is complete, it is locked in place and causes a conformational change in the structure of the complex, which is thought to open up a platform for Cas3 docking on the large subunit Cas8 (Hochstrasser et al., 2014) (Figure 1.12, B).

Cas3 proteins possess a metal-dependent HD-nuclease domain and an ATP-dependent helicase domain, which together open and degrade ss- and branched-DNA structures (Makarova et al., 2006; Beloglazova et al., 2011). Following recruitment to the Cascade complex, Cas3 nicks the displaced non-target strand near the PAM motif, then unwinds and degrades DNA in a 3' - 5' direction (Westra et al., 2012; Hochstrasser et al., 2014) (Figure 1.12, B).

1.6.3.2 Type III interference

There are two subtypes of type III interference systems, defined by their differing *cas* gene complements. The type III-A complexes are known as Csm and the type III-B complexes are known as Cmr (Makarova et al., 2011). In contrast to Cascade, which requires recruitment of Cas3 for interference, the type III complexes have intrinsic nuclease activity. Functional interference depends on the complex being loaded with a crRNA containing a conserved eight nucleotide 5' tag and a variable 3' end, which undergoes secondary processing by an unknown host nuclease (Carte et al., 2008; Hatoum-Aslan et al., 2013).

The complex architecture of type III effectors is similar to that of Cascade. The conserved 5' crRNA tag is bound at the base of the complex, with the spacer

sequence being coordinated by multiple Cas7 family subunits (Cmr4 or Csm3) forming the complex backbone (Staals et al., 2013, 2014). The crRNA is coordinated in a ribbon-like formation in a groove formed by the backbone RRM proteins, with every 6th base of the guide crRNA being flipped by 90° away from the target-binding cleft (Osawa et al., 2015) (Figure 1.13). The type III complexes have a second filament, made up of multiple copies of the small subunit (Cmr5 or Csm2), which twists around the crRNA binding backbone. The shape of type III complexes has been described as 'sea-worm'-like, with a broad foot and elongated helical body (Staals et al., 2013). The foot of the complex is formed by a Cas10-like large subunit (Cmr2 or Csm1) and a Cas5-like subunit (Cmr3 or Csm4), and the head of the complex by a Cas7 family subunit (Cmr6 or Csm3) (Staals et al., 2013, 2014) (Figure 1.13).

The first interference activity identified for type III-B Cmr complexes was *in vitro* RNA cleavage. Hale and colleagues found that the Cmr complex of *P. furiosus* cleaved ssRNAs that were complementary to the bound crRNA guide at a fixed distance from the 3' end (Hale et al., 2009). A Cmr protein from *S. solfataricus* was also shown to cleave crRNA-targeted ssRNA at UA dinucleotides (Zhang et al., 2012b). Both of these activities required divalent metal ions and the crRNA guide to have an intact and sequence-specific 5' tag. More recently, Cmr complexes have been found to cleave ssRNA targets with a 6-nucleotide periodicity (Staals et al., 2013). The cleavage positions were selected by a molecular-ruler mechanism, anchored at the 5' tag of the crRNA (Figure 1.13). The first cleavage occurred at the 3' end of the target RNA and progressed in 6-nt increments towards the 5' end of the substrate (Staals et al., 2013). Zhang and colleagues also identified the 6 nt cleavage pattern for the *S. solfataricus* Cmr protein, previously shown to cut at UA dinucleotides (Zhang et al., 2012b, 2016). The authors reported that the cleavage mode adopted by this Cmr complex was determined by the ratio of target RNA to protein complex. When RNA was in excess the UA cleavage mode presided, whereas at low RNA concentrations the 6-nt periodicity of cleavage was evident (Zhang et al., 2016).

Type III-B interference

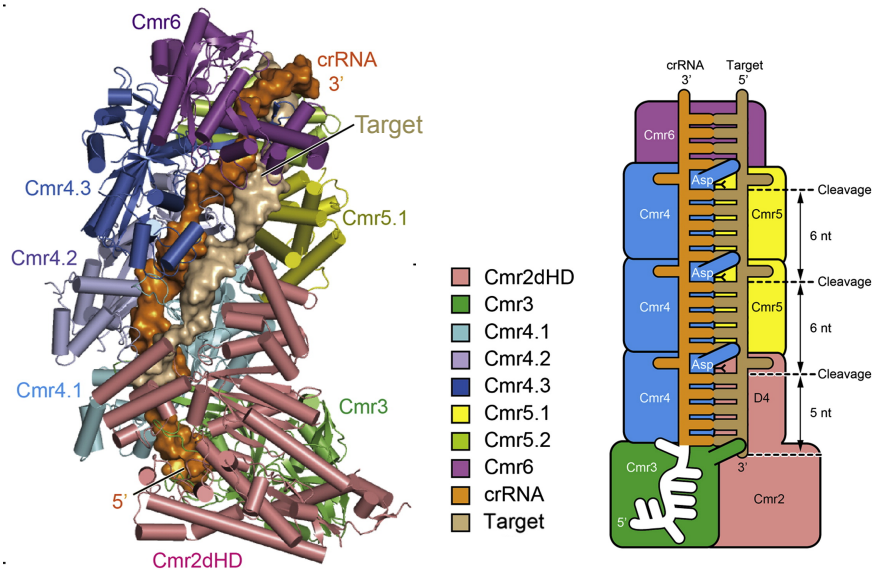


Figure 1.13 Type III-B Cmr complex cleaves RNA with a 6 nt periodicity

The left-hand structure is the crystal structure of a chimeric Cmr complex with bound crRNA and target analogue. On the right is a schematic depicting the binding and co-ordination of a crRNA guide and target strand by the Cmr complex. The colours used for different subunits are shown in the central key. The thumb domains of backbone Cas7-family subunits (Cmr4) disrupt base pairing every 6 nucleotides, positioning a scissile phosphate bond of the target close to an Asp residue thought to catalyse cleavage of the target RNA at these positions. Adapted from Osawa et al., 2015.

The crystallisation of a chimeric Cmr complex bound to a crRNA:target duplex provided a structural explanation for the 6 nt cleavage pattern observed (Osawa et al., 2015). A stable Cmr complex was attained by mixing the large subunit (Cmr2) and Cmr3 protein from *P. furiosus*, which together make up the base of the complex, with the small subunit (Cmr5) and Cmr4 proteins from *A. fulgidus*, required to form the helical filaments of the body (Osawa et al., 2015). The target-bound crystal structure of this complex elucidated how Cas7 family (Cmr4) subunits in the Cmr backbone disrupt base pairing of the guide and target every 6 nt. A β -hairpin ‘thumb’ domain of Cmr4 was found to flip guide and target bases by 90° from the duplex at these positions, bringing scissile phosphate bonds close to a catalytic aspartate (Asp) residue of the Cmr4 subunits (Osawa et al., 2015). The crystal structure also revealed that the 5’ end of the crRNA was bound in a pocket at the base of the Cmr complex and Cmr3, the Cas5-like subunit, specifically recognised the second residue of the 5’ tag. This specific interaction explained the required sequence conservation of the 5’ tag and also the 5’ anchoring of the molecular ruler during RNA cleavage (Osawa et al., 2015).

The type III-A interference machinery was initially demonstrated to target DNA and block horizontal gene transfer *in vivo* (Marraffini & Sontheimer, 2008). The Palm polymerase/cyclase domain of the large subunit of Csm was found to be essential for plasmid targeting *in vivo* (Hatoum-Aslan et al., 2014). However, attempts to reproduce this DNA targeting by Csm *in vitro* failed (Rouillon et al., 2013; Staals et al., 2014). Instead, Csm complexes were found to cut RNA at 6 nt intervals (Tamulaitis et al., 2014; Staals et al., 2014). This activity is thought to be orchestrated in a very similar way to the RNA cleavage observed for the Cmr complex. In Csm a putative thumb domain of the Csm3 backbone subunits is thought to disrupt base pairing of crRNA and target at 6 nt intervals (Osawa et al., 2015; Tamulaitis et al., 2014), while a conserved Csm3 Asp residue is essential for RNA cleavage (Tamulaitis et al., 2014). The coordination of crRNA and disruption of guide and target duplex by Cas7 backbone subunits is also observed in Cascade (Mulepati et al., 2014) and therefore seems to be a general feature of Class I interference complexes, supporting the prediction that these complexes share a common ancestor (Makarova et al., 2015).

1.6.3.3 Transcription-dependent DNA targeting

The *in vitro* RNA targeting activity of the Csm complex was difficult to reconcile with the plasmid interference mediated by this complex *in vivo* (Marraffini & Sontheimer, 2008). A link between the two activities was made when *in vivo* DNA interference by a Csm complex was shown to require transcription of the targeted protospacer (Goldberg et al., 2014). Samai and colleagues subsequently demonstrated that transcription through a target sequence led to cleavage of both the RNA transcript and the non-template DNA strand by the Csm complex from *S. epidermidis* (Samai et al., 2015). The protospacer transcript was cut within the crRNA complementary region by Csm3 subunits with the expected 6 nt periodicity. However, DNA processing took place on the non-template strand outwith the crRNA complementary region. DNA interference was independent of RNA cleavage and was mediated by the Palm domain of the large subunit (Samai et al., 2015). The authors speculated that the DNA targeting might be triggered by a transcription bubble passing through the targeted site and opening the DNA, allowing targeting by Csm (Samai et al., 2015).

Transcription-dependent DNA cleavage was also identified for a type III-B Cmr complex in *Sulfolobus islandicus* (Deng et al., 2013). Another study showed that the addition of RNA complementary to the Cmr guide activated cleavage of a DNA substrate, without the requirement for active transcription (Estrella et al., 2016). Both RNA and single-stranded DNAs or the single-stranded region of DNA bubbles were found to be cut by Cmr in this study and the cleavage was mediated by the HD-domain of the large subunit (Estrella et al., 2016).

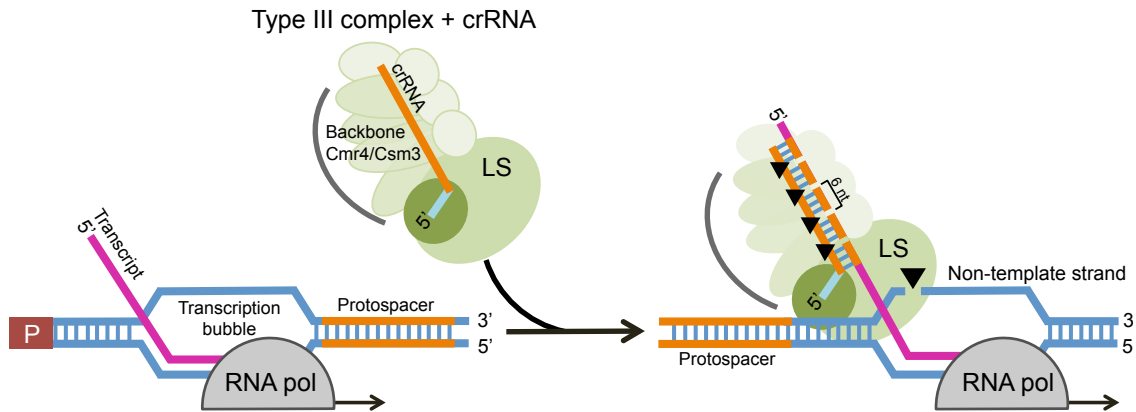


Figure 1.14 Co-transcriptional DNA interference in Type III systems

Model for co-transcriptional targeting by type III systems. In the first step, transcription is activated at a promoter site (P) and a transcription bubble passes through a target sequence. Type III complexes then bind the crRNA-complementary region of the nascent transcript and backbone Cas7-family subunits cleave the target sequence every 6 nt. The localisation of the type III complex to the transcription bubble also facilitates cleavage of the non-template strand by the large subunit (LS). Adapted from Estrella et al., 2016.

Taking these results together it seems that transcription of a protospacer likely activates DNA interference in type III systems. A model has been proposed in which a nascent protospacer transcript is bound by a crRNA-guided type III complex. This binding is thought to direct the complex to a transcription bubble and to trigger DNA interference to the non-template strand of the bubble (Estrella et al., 2016) (Figure 1.14). This model predicts spatially- and temporally-controlled type III interference only on transcriptionally active elements, which has been speculated to provide an advantage to the host by allowing silent, potentially fitness-enhancing transposable elements to be tolerated (Goldberg & Marraffini, 2015). The role of RNA interference of these complexes remains enigmatic; however, RNA targeting by a type III-A complex has been shown to confer resistance to an RNA phage *in vivo* (Tamulaitis et al., 2014), implying that the dual cleavage may provide a robust CRISPR-Cas immunity against both DNA and RNA threats.

1.7 Autoimmunity

Given the potency of CRISPR interference, a mechanism to avoid autoimmunity caused by targeting of host spacers is essential. This self-targeting is avoided in Type I systems due to the strict requirement for a PAM motif for interference by the Cascade complex (Westra et al., 2013). This mechanism is referred to as non-self activation (van der Oost et al., 2014). Westra and colleagues identified four different PAM sequences, which were essential for DNA targeting by the type I-E system in *E. coli* (Westra et al., 2012). As these sequences are removed during spacer capture and not found adjacent to the spacer in the CRISPR array, they also act to protect the host from self-targeting (Westra et al., 2013) (Figure 1.15, A).

In the type III-A system a different mechanism to avoid autoimmunity exists, called self-inactivation (van der Oost et al., 2014). Mismatches between the conserved 8-nt 5' tag of crRNA and the region upstream of the target sequence were found to license interference and prevent transformation of a protospacer-containing plasmid in *S. epidermidis* (Marraffini & Sontheimer, 2010). In contrast, full complementarity of the repeat-derived 5' tag with target inactivated cleavage and allowed transformation (Marraffini & Sontheimer, 2010) (Figure 1.15, B). This finding explained why spacers in the CRISPR locus were not targeted, as the upstream repeat sequence was a full match to the 8 nt crRNA tag. Positions -2 to -4 upstream of the targeted protospacer were found to be particularly important for self versus non-self discrimination, with two consecutive mismatches needed for interference (Marraffini & Sontheimer, 2010). This result was corroborated in *S. solfataricus*, as three consecutive matches between the crRNA 5' handle and the protospacer adjacent sequence were shown to be sufficient to abolish interference (Manica et al., 2013).

The requirement for non-complementarity between the crRNA 5' handle and target DNA in type III-A systems extends to co-transcriptional DNA interference (Samai et al., 2015). However, RNA transcript cleavage by the Csm complex occurred even when the 5' tag was fully complementary to the target transcript (Samai et al., 2015). RNA cleavage by the *S. thermophilus* Csm complex was also shown to be independent of a PAM sequence or complementarity in the 5' crRNA tag (Tamulaitis et al., 2014). The flexible RNA targeting may reflect the lower autoimmune cost of targeting self-RNA compared to self-DNA.

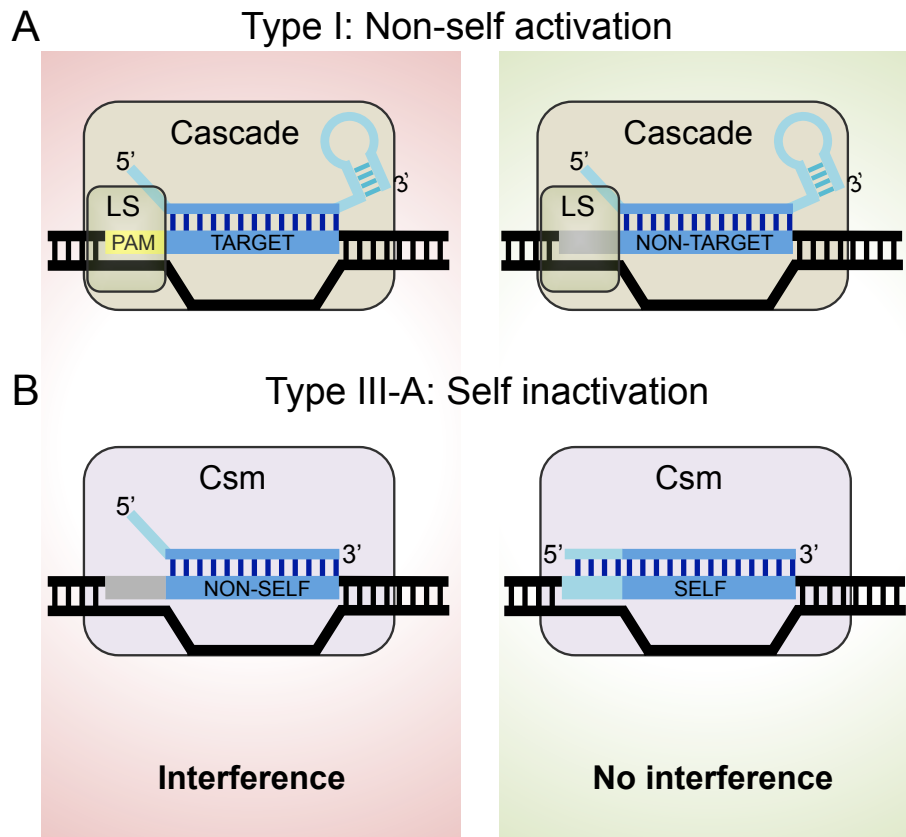


Figure 1.15 Self-protection from CRISPR immunity

A. In type I CRISPR-Cas systems, only crRNA-matching targets with an adjacent PAM motif will be degraded during interference. As the crRNA-matching spacer in the CRISPR array is always flanked by non-PAM repeats, it is not targeted and autoimmunity is prevented. The large subunit (LS) of Cascade is responsible for PAM detection during interference. PAM is represented by a yellow box and a non-PAM sequence by a grey box. **B.** In type III-A systems complementarity between the 5' handle of the crRNA and the target abolishes interference. As the repeat-derived 5' handle of the crRNA will always have full complementarity to the repeats surrounding host spacers, self-cleavage is avoided. Adapted from Westra et al., 2013; Marraffini & Sontheimer 2010.

1.8 Viral escape

1.8.1 CRISPR-Cas resistance through mutation

There exists a rapid co-evolutionary 'arms race' between viruses and prokaryotes, with both predator and prey rapidly adapting to outwit the defensive or offensive strategies of the other (Stern & Sorek, 2012). Viruses have a high turnover rate and error-prone replication, which means that mutations that circumvent host defences are acquired rapidly and spread throughout a viral population. Phage that infect *S. thermophilus* were found to rapidly evolve to escape targeting by a matching spacer of the CRISPR-Cas system (Deveau et al., 2008). Escape phage were often found

to possess a single mutation in the protospacer region, or in the short PAM sequence, which led to a renewed ability to infect the host (Deveau et al., 2008). Furthermore, in the type I-E system of *E. coli*, a point mutation in the PAM or 5' seed sequence of the targeted protospacer was sufficient to lead to phage escape (Semenova et al., 2011). In order to minimise resistance, CRISPR-Cas immune systems often incorporate multiple spacers during an infection (Deveau et al., 2008). In addition, it is now clear that the sequence specificity of targeting also often has a degree of flexibility. For example, for Cascade interference in *E. coli*, multiple mutations outside of the PAM or 5' seed sequence were tolerated by the interference machinery (Semenova et al., 2011). The type III systems also exhibit promiscuous targeting, with protospacers containing up to 15 mismatches to the crRNA, outside of the 5' seed sequence, still leading to effective interference (Manica et al., 2013). In some type I systems, mismatches that lead to ineffective interference promote hyperactive uptake of new spacers by primed adaptation, allowing the host to reacquire immunity to the invader (Datsenko et al., 2012; Richter et al., 2014).

1.8.2 Anti-CRISPRs

The recent discovery of anti-CRISPR proteins has provided a fascinating insight into the diverse methods used by phage to evade the CRISPR-Cas system. The first evidence of an anti-CRISPR mechanism was identified when several lysogenic *P. aeruginosa* strains containing prophage DNA were found to be much less resistant to viral infection than a wildtype strain with an active CRISPR-Cas system (Bondy-Denomy et al., 2012). A group of genes in the prophage DNA coding for small proteins of unknown function were found to be responsible for this effect. Overexpression of members of this group had no effect on the production of Cas proteins or crRNA, indicating that the small proteins must inhibit the interference stage of CRISPR immunity (Bondy-Denomy et al., 2012).

It was subsequently shown that these anti-CRISPR proteins inhibit, in a highly specific manner, the activity of interference complexes (Bondy-Denomy et al., 2015). No common mode of action was observed for this inhibition as some anti-CRISPR proteins directly interacted with the type I-F Cascade, while another bound Cas3 and prevented its recruitment to the complex (Bondy-Denomy et al., 2015). Anti-CRISPRs have also been identified in other mobile genetic elements in the *P.*

aeruginosa genome, where they are predicted to provide these regions with protection against CRISPR-Cas during their transfer between hosts (Pawluk et al., 2014). Many anti-CRISPRs have now been identified across the proteobacteria and it is hypothesised that their existence may provide an explanation for the surprisingly widespread influence of HGT on prokaryote evolution even in species possessing active CRISPR-Cas systems (Gophna et al., 2015; Pawluk et al., 2016).

1.9 Regulation of CRISPR-Cas activity

There are autoimmune consequences of having a continuously active CRISPR-Cas immune system in the absence of invaders. These include the incorporation of self-spacers and aberrant targeting of self-DNA. In general, for every 250 spacer insertions across CRISPR arrays, 1 self-spacer is inserted (Stern et al., 2010). These self-spacers are often the most recently incorporated spacer in an array and are frequently associated with degraded or deleted CRISPR-Cas loci, which implies that these incorporations result in severe autoimmune penalties for the host (Stern et al., 2010). Studies of the CRISPR-Cas system in *E. coli* have revealed that these consequences are avoided by the transcriptional silencing of the immune system under control conditions (Pul et al., 2010). In archaea, much less is known about the regulation of the CRISPR-Cas system, with some early studies suggesting that Cas proteins and pre-CRISPR transcripts were expressed constitutively in archaea (Lillestøl et al., 2006; Hale et al., 2009). A key aim of this thesis was to learn more about transcription regulation of the CRISPR-Cas system in the hyperthermophilic archaeon *S. solfataricus*. A detailed introduction to archaeal CRISPR-Cas regulation is provided with the results of this investigation in Chapter 3.

Recent studies have helped to partially elucidate the mechanism of CRISPR-Cas regulation in some bacterial systems. The components involved in this regulation are introduced below and summarised in Figure 1.16.

1.9.1 H-NS

An initial investigation found that that under control conditions the *P_{cas}* promoter controlling *cas* gene expression in the *E. coli* CRISPR-Cas system was transcriptionally silent (Pul et al., 2010). The global repressor, heat-stable nucleoid-structuring (H-NS) protein was found to be responsible for this silencing (Pul et al.,

2010). The H-NS protein repressed transcription from *Pcas*, and to a lesser extent the promoter of the CRISPR array (*Pcrispr*), by binding to upstream AT-rich intergenic sequences. The initial H-NS nucleation was followed by co-operative binding of other H-NS proteins along the DNA strand, preventing transcription by blocking RNA polymerase access to the promoters (Pul et al., 2010). The authors demonstrated *in vitro* and *in vivo* that the disruption or removal of the H-NS protein lifted the repression of transcription from *Pcas* and *Pcrispr*. A study in *Salmonella enterica* serovar Typhi also identified H-NS, as well as the leucine-responsive regulatory protein (LRP), as a repressor of transcription of CRISPR-cas elements (Medina-Aparicio et al., 2011).

1.9.2 LeuO

A further breakthrough in our understanding of the regulation of CRISPR-related elements in *E. coli* came when it was reported that elevated levels of the H-NS antagonist LeuO act to de-repress transcription from the *Pcas* promoters (Westra et al., 2010). LeuO was found to bind to an intergenic region upstream of *Pcas* and to block repression by preventing co-operative binding and polymerisation of H-NS across the promoter (Westra et al., 2010). The authors speculated that H-NS and LeuO work antagonistically to control the transcription of the CRISPR-Cas system in *E. coli*.

One factor that complicates formation of a model for CRISPR-Cas regulation in *E. coli* is the fact that transcription of LeuO is itself under repression by H-NS (Klauck et al., 1997). However, it has been hypothesised that on viral infection the entry of AT-rich foreign DNA will provoke H-NS to leave its repressor sites and bind these incoming sequences (Westra et al., 2010; Dillon et al., 2010). This titration of H-NS is thought to allow LeuO production and kick-start a positive feedback loop, leading to further LeuO expression and its binding and activation of *Pcas* and *Pcrispr* promoters. However, as the intracellular levels of LeuO even in the presence of phage infection are not sufficient to lift *Pcas* repression, it seems that further factors produced by either the host or invader are required to activate CRISPR-Cas in *E. coli* (Westra et al., 2010). One such factor may be the T7 phage protein gp5.5, which has been shown to bind to H-NS and release transcriptional inhibition of H-NS-regulated genes (Ali et al., 2011).

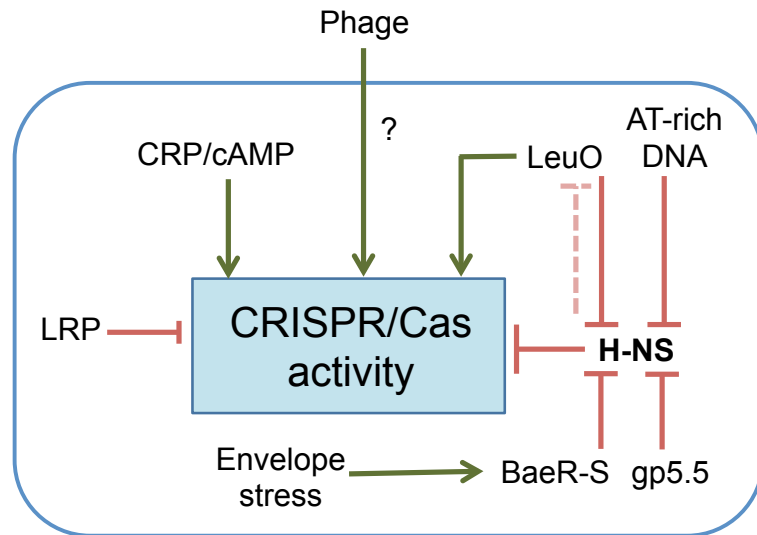


Figure 1.16 Transcriptional regulation in bacteria

Network model describing regulation of CRISPR-Cas activity in bacteria. Green triangular arrows represent factors that enhance CRISPR-Cas activity and red flat arrows those that repress the system. See text for more details of the regulation mechanisms. The regulation of CRISPR-Cas is not accounted for by the mechanisms identified up to now, implying that other host or viral factors are involved (represented by a '?') (Adapted from Richter, Chang, et al., 2012).

1.9.3 CRP

Another protein thought to be involved in the regulation of bacterial CRISPR-Cas systems is the cAMP-receptor protein (CRP). Disruption of the gene coding for CRP in *T. thermophilus* led to a reduction in transcription from several promoters, including those controlling expression of *cas* operons (Shinkai et al., 2007). Furthermore, upregulation of *cas* gene expression during phage infection was shown to be severely impaired in a *T. thermophilus* CRP knockout strain (Agari et al., 2010). These studies suggest that CRP may contribute to the induction of expression of the *cas* genes of *T. thermophilus* during infection. The type I-F *cas* operon of *P. atrosepticum* was also recently shown to be upregulated by a cAMP-CRP complex (Patterson et al., 2015). The authors suggested that as CRP and cAMP are required for phage λ to enter the lytic lifecycle stage, the evolution of an antiviral system that responds to this same signal is a considerable advantage to the host (Patterson et al., 2015).

1.9.4 Envelope stress

Finally, envelope and heat stress have been documented to induce the CRISPR-Cas response in *E. coli*. Envelope stress is thought to be triggered by phage entry or accumulation in the host, which in turn activates the two component regulatory system BaeSR (Perez-Rodriguez et al., 2011). A BaeR binding site was identified downstream of the H-NS binding site responsible for the repression of the *cas* operon. Following phosphorylation by BaeS, BaeR is thought to target this site and trigger, by an unknown mechanism, the expression of the *cas* operon in *E. coli* (Perez-Rodriguez et al., 2011).

In conclusion, transcription of the CRISPR-Cas systems of bacteria is controlled by a diverse range of environmental and infection-induced factors, both within and between organisms. In general, these CRISPR-Cas systems seemed to be silenced in the absence of invaders in order to avoid autoimmunity and are turned on, or upregulated, on the entry of foreign DNA to provide protection of the host.

1.10 Scope of this thesis

1.10.1 CRISPR-Cas systems of *S. solfataricus*

This thesis will mainly investigate aspects of the CRISPR-Cas system of the hyperthermophilic creanarchaeon *Sulfolobus solfataricus*, which lives at 80°C and pH 2 – 3. The complete genome sequence is available (She et al., 2001) and two related strains (P1 and P2) have been identified with different CRISPR-Cas complements. The CRISPR-Cas system of *S. solfataricus* is complex and makes up more than 1% of the total genome length (Lillestøl et al., 2006). It includes 6 CRISPR arrays, denoted A – F, and three different CRISPR-Cas types (type I-A, III-D and III-B). Loci E and F are inactive, lacking a functional leader sequence or associated *cas* genes (Lillestøl et al., 2009). These loci cannot incorporate new spacers and are identical in P1 and P2. In contrast, loci A - D are complete and differ at the leader-proximal end between strains, indicating unique spacer acquisition after the strains diverged (Lillestøl et al., 2009). The *S. solfataricus* genome encodes multiple Cas6 proteins responsible for processing two families of unstructured CRISPR transcripts. The repeats of CRISPR loci A and B constitute one family, and those of loci C – F the second (Sokolowski et al., 2014). Processed

crRNAs are then loaded into type I or III interference complexes to guide CRISPR immunity.

This thesis will focus on the two adaptation cassettes, one located between CRISPR A and B and one between CRISPR C and D, of *S. solfataricus* and the role the Cas1 and Cas2 proteins encoded by these operons play in acquiring spacers to immunise the host against infection. Before this work began very little was known about the process of adaptation in *S. solfataricus*. An early study of adaptation in this archaeon found that new spacers were only added if the CRISPR array was associated with a set of *cas* genes and preceded by a long leader sequence (Lillestøl et al., 2009). Further investigation has shown that activation of adaptation *in vitro* is not trivial and the mechanism differs amongst the six CRISPR arrays of *S. solfataricus* (Erdmann & Garrett, 2012). Infection of a *S. solfataricus* P2 culture with an environmental sample of viruses activated adaptation of the CRISPR loci C, D and E (Erdmann & Garrett, 2012). However, infection with isolated single viruses failed to provoke spacer uptake (Erdmann & Garrett, 2012). Spacers added to CRISPR E were found to be integrated at every repeat, not at leader repeat junction as for other arrays, perhaps indicating a different mechanism of adaptation. New spacers were not added to CRISPR arrays A and B during this infection study, instead adaptation of these arrays was only observed following a freeze-thaw cycle (Erdmann & Garrett, 2012). This led the authors to conclude that adaptation of CRISPR A and B occurs in response to environmental stress.

Biochemical studies of the Cas1 and Cas2 proteins, known to be required for adaptation in the *E. coli* CRISPR system (see section 1.6.1.1) (Yosef et al., 2012), have yielded some insights into the activities of these proteins, but few clues to the role Cas1 and Cas2 play in adaptation. The crystal structure of the Cas2 protein associated with CRISPR arrays A and B (Cas2_{AB}) has been solved, revealing a dimeric protein with a ferredoxin-like fold (Beloglazova et al., 2008). Biochemical assays showed that this protein degrades RNA in a metal-dependent manner (Beloglazova et al., 2008). An initial characterisation of the Cas1 protein located between loci C and D (Cas1_{CD}) failed to identify an enzymatic activity, but did report that this Cas1 protein had a high affinity for nucleic acids and enhanced DNA strand annealing (Han et al., 2009).

The adaptation cassettes in *S. solfataricus* also code for Cas4-family proteins, suggesting a role for these proteins in adaptation. The *S. solfataricus* Cas4 protein

Sso0001 has been shown to form a toroidal ring and possess both exo- and endonuclease activities on single-stranded DNA (Zhang et al., 2012a; Lemak et al., 2013). It is possible the exonuclease activity of these proteins may be essential for the creation and processing of short protospacer duplexes for integration by the Cas1 and Cas2 proteins.

1.10.2 Aims of this thesis

Very little is known about transcriptional regulation of the CRISPR-Cas systems of archaea. Therefore, a primary aim of this thesis was to investigate how the expression of Cas proteins in *S. solfataricus* P2 changes during viral infection. Chapter 3 describes the results of this investigation and also identifies a putative transcriptional regulator in *S. solfataricus*. Secondly, this thesis aimed to characterise the activity of Cas1 and Cas2 proteins from *S. solfataricus* P2. The few reports available on the activity of these proteins revealed diverse activities *in vitro* that did not fit well with the predicted roles of Cas1 and Cas2 *in vivo*. Chapter 4 describes the *in vitro* activity of a Cas1 and Cas2 protein from *S. solfataricus*. The biochemical characterisation of the Cas1 protein led to the identification of a transesterification activity carried out by this protein as well as by the Cas1 protein from the type I-E system of *E. coli*. Chapter 5 describes the study of this reaction, its link to adaptation and how the strict sequence specificity imposed by the Cas1 *in vitro* helps to define the site of integration *in vivo*. A final aim of this project was to reconstitute the integration of new spacers by Cas1 and Cas2 *in vitro*, to try to better understand the substrate requirements and DNA manipulation steps involved. Chapter 6 presents data on the *in vitro* reconstitution of protospacer integration into supercoiled substrates by Cas1 and Cas2 from *S. solfataricus*.

Overall, the work contained in this thesis aimed to enhance the understanding of how new CRISPR memories are stored during the adaptation stage of CRISPR-Cas immunity.

Chapter 2: Materials and methods

2.1 Materials

2.1.1 Oligonucleotides

All single-stranded oligonucleotides used in this study were synthesised by Integrated DNA Technologies (IDT). Double-stranded genomic fragments (gBlocks) were also synthesised by IDT (see Table 2.3 for sequences). The sequences of single-stranded DNA oligonucleotides are given in Table 2.1. For substrates comprising three or more strands their component oligonucleotides are noted in Table 2.2, and for disintegration substrates (used in Chapter 5) the sequence around the branch point is given.

Table 2.1 Sequence of oligonucleotides used in this thesis

Oligonucleotide	Sequence 5' - 3'
Chapter 2	
sso1450 E142Af	GTTGGATAAGGATGCACCGGCTGCTGCTAG
sso1450 E142Ar	CTAGCAGCAGCCGGTGCATCCTTATCCAAC
sso1450 W150Af	CTGCTGCTAGAGTGTATGCACAGAACATATCTCAAC
sso1450 W150Ar	GTTGAGATATGTTCTGTGCATACACTCTAGCAGCAG
sso1450 N175Af	GATGGGACTGACCAATTTGCAATGGCATTGAACTACTC
sso1450 N175Ar	GAGTAGTTCAATGCCATTGCAAATTGGTCAGTCCCATC
sso1450 R166Af	GATTTGACGGTGCAGATGTGGATGG
sso1450 R166Ar	CCATCCACATCTGCACCGTCAAATC
sso1404 D10Af	CCTAATATTTTACGCCATTACTGATGATAATC
sso1404 D10Ar	GATTATCATCAGTAATGGCGTAAAAATTAGG
Chapter 3	
SSBfor	AGTTTTGGAAGCAAGCGAAG
SSBrev	GTGGTCCACGCGTTTTCTAT
1450for	ATTGGGCTAGGGTAACTGGA
1450rev	GGTCAGTCCCATCCACATCT
1443for	GTGAATCAAGTGGGGCTAAGT
1443rev	TTCTCACCACCTTGCTCACT
1986for	TAGTGGAGCTGGATGGGAAG
1986rev	GCCCCTTATCATCGATTTGA
CRISPRCfor	CGAGGAGTAGTGGTAGGAGTG
CRISPRCrev	AGAGTAGTGAACGTCAGCA
1424for	AACCGGGAAGTTGTTGTTGG
1424rev	CCCCGTTAAAGTATCGTAAACCC
sso1451 operator for	FAM- CCCTAGTAAATTCGGGAATTCTTTTACCCCCCTCCTTAAAA CGGTTTTTA
sso1451 operator rev	TAAAAACCGTTTTAAGGAGGGGGTAAAAGAATTCCTCGAA TTTACTAGGG

sso1443 operator for	FAM- TTTTTCATATTTATGAAAAGAGTTTTTCGTACACTAGAAATA GAAATGT
sso1443 operator rev	ACATTTCTATTTCTAGTGTACGAAAACCTTTTTCATAAATAT GAAAAA
sso1451promf	CACCACATCAAACGACCCCCACTT
sso1451promr	CGTCTAGGCTTTATCGGAGGC
T6r	GGTCGACGGTATCGATAAGC
Chapter 4	
B50-5'-FAM	FAM- CCTCGAGGGATCCGTCTAGCAAGCCGCTGCTACCGGAA GCTTCTGGACC
B50	CCTCGAGGGATCCGTCTAGCAAGCCGCTGCTACCGGAA GCTTCTGGACC
B50 comp	GGTCCAGAAGCTTCCGGTAGCAGCGGCTTGCTAGGACGG ATCCCTCGAGG
CRISPRAB rep (1404) DNA	FAM-GATTAATCCCAAAGGAATTGAAAG
CRISPRAB rep (1404) RNA	FAM-GAUUAAUCCCAAAGGAUUUGAAAG
CRISPRCD rep	FAM-GATAATCTCTTATAGAATTGAAAG
CRISPRCD rep rev	FAM-CTTTCAATTCTATAAGAGATTATC
CRISPRCD rep scramble	FAM-ATACTATAAGCTATAGTGAATAGT
CRISPR C LR for	ATAGTAAGAGATTAATAAACCCCTCAGATAATCTCTTATAGA ATTGAAAGCAA-FAM
CRISPR C LR rev	TTGCTTTCAATTCTATAAGAGATTATCTGAGGGTTTATTAA TCTCTTACTAT
Jbm5A	GCGTTACAATGGAAACTATTCTTGGCAGTTGCATCCAACG
Jbm5B	CGTTGGATGCAACTGCCAAGAATAGTGTGAGTTCCAGCAG
Jbm5C	CGTCTGGAAGTACACTATTCTTGGCGAATGGTCGTAAGC
Jbm5D	GCTTACGACCATTCCGAAGAATAGTTTCCATTGTAACGC
CRISPR C transcript DNA	FAM- TAGTAAGAGATTAATAAACCCCTCAGATAATCTCTTATAGAA TTGAAAG
CRISPR C transcript DNA comp	CTTTCAATTCTATAAGAGATTATCTGAGGGTTTATTAATCT CTTACTA
CRISPR C transcript RNA	FAM- UAGUAAGAGAUUAAUAAACCCUCAGAUAAUCUCUUAUAG AAUUGAAAG
3'OHprotospacer for	TCGCCATGGTGAGCACAGAGGATAATGTAACACT
3'OHprotospacer rev	TACATTATCCTCTGTGCTCACCATGGCGACGAGC
Chapter 5	
1a	TAGTAAGAGATTAATAAACCCCTCAGATAATCTCTTATAGA ATTGAAAGTTCGG
1b	TTTTTTTTTTTTTTTTTTTATTATCTGAGGGTTTATTAATCTC TTACTA
1c	CCGAACCTTTCAATTCTATAAGAG
2a	TAGTAAGAGATTAATAAACCCCTCAGATAACCTCTTATAGA ATTGAAAGTTCGG
2b	TTTTTTTTTTTTTTTTTTTGTATCTGAGGGTTTATTAATCTC TTACTA
3b	TTTTTTTTTTTTTTTTTTTAGTTATCTGAGGGTTTATTAATCTC TTACTA
4b	TTTTTTTTTTTTTTTTTTTCGTTATCTGAGGGTTTATTAATCTC TTACTA

5b	TTTTTTTTTTTTTTTTGGTTATCTGAGGGTTTATTAATCTC T TACTA
6a	TAGTAAGAGATTAATAAACCCCTCAGATAAGCTCTTATAGA ATTGAAAGTTCGG
6b	TTTTTTTTTTTTTTTTTACTTATCTGAGGGTTTATTAATCTC T TACTA
7a	TAGTAAGAGATTAATAAACCCCTCAGATAAACTCTTATAGA ATTGAAAGTTCGG
7b	TTTTTTTTTTTTTTTTTATTTATCTGAGGGTTTATTAATCTC T TACTA
8b	TTTTTTTTTTTTTTTTTAATTATCTGAGGGTTTATTAATCTC T TACTA
9a	TAGTAAGAGATTAATAAACCCCTCAGATAACATCTTATAGA ATTGAAAGTTCGG
9c	CCGAACCTTCAATTCTATAAGAT
10a	TAGTAAGAGATTAATAAACCCCTCAGATAACTTCTTATAGA ATTGAAAGTTCGG
10c	CCGAACCTTCAATTCTATAAGAA
11a	TAGTAAGAGATTAATAAACCCCTCAGATAACGTCTTATAGA ATTGAAAGTTCGG
11c	CCGAACCTTCAATTCTATAAGAC
12a	TAGTAAGAGATTAATAAACCCCTCAGATAACTCCTTATAGA ATTGAAAGTTCGG
12c	CCGAACCTTCAATTCTATAAGGA
13a	TAGTAAGAGATTAATAAACCCCTCAGATAACTGCTTATAGA ATTGAAAGTTCGG
13c	CCGAACCTTCAATTCTATAAGCA
14a	TAGTAAGAGATTAATAAACCCCTCAGATAACTACTTATAGA ATTGAAAGTTCGG
14c	CCGAACCTTCAATTCTATAAGTA
<i>SacI</i> -a	TAGTAAGAGATTAATAAACCCCTCAGATGAGCTCTTATAGA ATTGAAAGTTCGG
<i>SacI</i> -b	TTTTTTTTTTTTTCTCATCTGAGGGTTTATTAATCTCTTAC TA
1b-3'-FAM	TTTTTTTTTTTTTATTATCTGAGGGTTTATTAATCTCTTAC TA-FAM
1b-10	TTTTTTTTTTTATTATCTGAGGGTTTATTAATCTCTTACTA
1b-5	TTTTTATTATCTGAGGGTTTATTAATCTCTTACTA
1C-RNA	CCGAACUUUCAAUUCUAUAAGAG
1c-2	CCGAACCTTCAATTCTATAAG
19a	CCTCGAGGGATCCGTCCTAGCAAGCCGCTGCTACCGGA AGCTTCTGGACC
19b	GCTCGAGTCTAGACTGCAGTTGAGAGCTTGCTAGGACG GATCCCTCGAGG
19b-25	GCTTGCTAGGACGGATCCCTCGAGG
19c	GGTCCAGAAGCTTCCGGTAGCAGCG
20d-10	AGTCTAGACTCGAGC
20d-5	ACTGCAGTCTAGACTCGAGC
20d	TCTCAACTGCAGTCTAGACTCGAGC
25c-d	GGTCCAGAAGCTTCCGGTAGCAGCGTCTCAACTGCAGTC TAGACTCGAGC
1c-3'P	CCGAACCTTCAATTCTATAAGAG-PHOS
Sp3-1a	CTGGCGCGGGAACTCTCTAAAAGTATACATTTGTTCTT
Sp3-1b	TGTAATTGATAATGTTGAGAGTTCCCGCGCCAG
Sp3-1c	AAGAACAATGTATACTTTTAGA
Sp3-2a	CCAGCGGGGATAAACCGTTTGGATCGGGTCTGGAATTC

Sp3-2b	TGTTCCGACAGGGAGCCCGGTTTATCCCCGCTGG
Sp3-2c	GAAATTCCAGACCCGATCCAAAC
site1-a	CTTTCAATTCTATAAGAGATTATCTGAGGGTTTATTAATCT CTTACT
site1-b	TGCTTCATCTGGGCTAAGATAATCTCTTATAGAATTGAAA G
site1-c	AGTAAGAGATTAATAAACCCCTCA
site2-a	GATAATCTCTTATAGAATTGAAAGTCGAGGCCAGAGAAG GTGCGTTA
site2-b	AGAGGCTAGTAAGGTTGCTTTCAATTCTATAAGAGATTAT C
site2-c	TAACGCACCTTCTCTGGCCTCGA
Dumbbell1	TTTTTTTTTTTCGCGTTCAGCGGAACGCTGAACGCTCCAT ACCGGGAACCGGTATGGA
Dumbbell2	TTTTTTTTTTTCGCGAGCGGAACGCTCGCTCCATACCGG GAACCGGTATGGA
Dumbbell3	TTTTTTTTTTTCGCGTTCAGCGGAACGCTGAACGCTCCAC GGGAACCGTGA
Dumbbell4	TTTTTTTTTTTCGCGAGCGGAACGCTCGCTCCACGGGAA CCGTGGA
Chapter 6	
CRISPR D spacer dup F	CTTGAAATTACAGAAAAATAACATTCATTTACCCTGTG
CRISPR D spacer dup R	CACAGGGTAAATGAATGTTATTTTTCTGTAATTTCAAG
CRISPR D spacer 3'end F	TACAGAAAAATAACATTCATTTACCCTGTG
CRISPR D spacer 3'end R	AAATGAATGTTATTTTTCTGTAATTTCAAG
CRISPR D spacer 5'end F	CTTGAAATTACAGAAAAATAACATTCATTT
CRISPR D spacer 5'end R	CACAGGGTAAATGAATGTTATTTTTCTGTA
ssRNA spacer	CCGAACUUUCAAUUCUAUAAGAG
5 nt overhang spacer F	TTACTAGCCTCTTGTGTTGCTTCATCTGGGCTAA
5 nt overhang spacer R	CCAGATGAAGCAACACAAGAGGCTAGTAAGGTTG
4 nt overhang spacer F	TTACTAGCCTCTTGTGTTGCTTCATCTGGGCTAA
4 nt overhang spacer R	CCCAGATGAAGCAACAAGAGGCTAGTAAGGTTG
6 nt overhang spacer F	TACTAGCCTCTTGTGTTAGTTGCTTCATCTGGGCTAA
6 nt overhang spacer R	CAGATGAAGCAACTAACAAGAGGCTAGTAAGGTTG
GG PAM F	TTACTAGCCTCTTGTGTTGCTTCATCTGGGCTAAGGTT
GG PAM R	CCAGATGAAGCAACACAAGAGGCTAGTAAGGTTGGGTT
CC PAM F	TTACTAGCCTCTTGTGTTGCTTCATCTGGGCTAACCTT
CC PAM R	CCAGATGAAGCAACACAAGAGGCTAGTAAGGTTGCCTT
29 nt duplex spacer F	TTACTAGCCTCTTGTGTTGCTTCATCTGGGCTAA
29 nt duplex spacer R	CCAGATGAAGCAACACAAGAGGCTAGTAAGGTTG
24 nt duplex spacer F	TTACTAGCCTCTTGTGTTGCTTCAGCTAA
24 nt duplex spacer R	TGAAGCAACACAAGAGGCTAGTAAGGTTG
34 nt duplex spacer F	TTACTAGCCTCTTGTGTTGCTTCATCTGGAGCTAGCTAA
34 nt duplex spacer R	TAGCTCCAGATGAAGCAACACAAGAGGCTAGTAAGGTTG
PCR protospacer F	TCGCCATGGTGAGCACAGAGGATAATGTAACACT
PCR protospacer R	TACATTATCCTCTGTGCTCACCATGGCGACGAGC
Primer <i>Nco</i> I F	TCGCCATGGTGAGCACAGAGGATA
Primer <i>Xho</i> I R1	AATTCTCGAGTTGGCCGATTCATTAATGC
Primer <i>Xho</i> I R2	AATTCTCGAGGGATAACCGTATTACCGCC

Table 2.2 Complex substrates and disintegration substrate junction sequences

Substrate name	Constituent oligonucleotides	Notes			
Jbm5 Holliday	Jbm5A, Jbm5B, Jbm5C, Jbm5d	See Figure 4.4 (p.108)			
Disintegration substrates (used in Chapter 5)		Junction Sequence			
Substrate name	Constituent oligonucleotides	-2	-1	1	IC
Substrate 1	1a, 1b, 1c	A	G	A	T
Substrate 1-FAM	1a, 1b-3'-FAM, 1c	A	G	A	T
Substrate 1-gap	1a, 1b, 1c-2	A	G	A	T
Substrate 1-10flap	1a, 1b-10, 1c	A	G	A	T
Substrate 1-5flap	1a, 1b-5, 1c	A	G	A	T
Substrate 1-RNA	1a,1b, 1c-RNA	A	G	A	T
Sacl substrate	Sacl-a, Sacl-b, 1c	A	G	C	T
Substrate 2	2a, 2b, 1c	A	G	G	T
Substrate 3	2a, 3b, 1c	A	G	G	A
3'-phos substrate	2a, 3b, 1c-3'P	A	G	G	A
Substrate 4	2a, 4b, 1c	A	G	G	C
Substrate 5	2a, 5b, 1c	A	G	G	G
Substrate 6	6a, 6b, 1c	A	G	C	A
Substrate 7	7a, 7b, 1c	A	G	T	A
Substrate 8	1a, 8b, 1c	A	G	A	A
Substrate 9	9a, 3b, 9c	A	T	G	A
Substrate 10	10a, 3b, 10c	A	A	G	A
Substrate 11	11a, 3b, 11c	A	C	G	A
Substrate 12	12a, 3b, 12c	G	A	G	A
Substrate 13	13a, 3b, 13c	C	A	G	A
Substrate 14	14a, 3b, 14c	T	A	G	A
Substrate 15	2a, 4b, 1c	A	G	G	C
Substrate 16	11a, 4b, 11c	A	C	G	C
Substrate 17	10a, 4b, 10c	A	A	G	C
Substrate 18	9a, 4b, 9c	A	T	G	C
Substrate 19	19a, 19b, 19c	C	G	G	A
Nicked-19	19a, 19b-25, 19c	C	G	G	-
Gap10	19a, 19b, 19c, 20d-10	C	G	G	A
Gap5	19a, 19b, 19c, 20d-5	C	G	G	A
Nicked-Y	19a, 19b, 19c, 20d	C	G	G	A
Y-junction	19a, 19b, 20c-d	C	G	G	A
Spacer 3-1 substrate	Sp3-1a, Sp3-1b, Sp3-1c	G	A	G	A
Spacer 3-2 substrate	Sp3-2a, Sp3-2b, Sp3-2c	A	C	G	C
Site1-ss0	site1-a, site1-b, site1-c	C	A	G	A
site2-ss0	site2-a, site2-b, site2-c	G	A	C	G

2.1.2 Restriction enzymes

FastDigest restriction enzymes, purchased from ThermoFisher Scientific, were used according to manufacturer's protocols, unless otherwise stated.

2.1.3 Vectors for recombinant protein expression

cas genes were amplified by PCR from *S. solfataricus* P2 genomic DNA and cloned into expression vectors by previous members of the White lab. The *cas1_{CD}* (*ssol1450*) and *cas2_{AB}* (*ssol1404*) genes were cloned into the pEHISTEV expression vector (Liu & Naismith, 2009), using *NcoI* and *Bam*HI restriction sites. *cas1_{AB}* (*ssol1405*) and *cas2_{CD}* (*ssol1450a*) were inserted (using *NcoI* and *Bam*HI sites) into a modified vector, pV5HISTEV (produced by Reyes Sanles-Falagan, White lab), with an extended linker between polyhistidine tag and protein. The *csa3* (*ssol1445*) gene was cloned into the pDEST14 expression vector (ThermoFisher Scientific). The transcription factor β (TFB) and TATA-binding protein (TBP) constructs were amplified from *S. solfataricus* P2 DNA and inserted into pDEST14 vectors as described previously (Götz et al., 2007). DNA sequencing was provided by GATC Biotech.

2.1.4 Vectors used as substrates

The pUC19 plasmid (ThermoFisher Scientific) was modified by the insertion of gBlocks (IDT) with sequences matching the leader, repeat1 and spacer1 of CRISPR array C or A of *S. solfataricus*. The gBlocks were cloned into pUC19 using *Eco*RI and *Bam*HI restriction sites to form the pCRISPRA and pCRISPRC plasmids (see Table 2.3, for sequences and details). Versions of pCRISPRC with mutations in either the repeat (pCRISPRC rep mutant) or leader sequence (pCRISPRC pal mutant) were also made in the same way (see Table 2.3 for sequences). These plasmids were used as substrates in integration assays in Chapter 6.

Table 2.3 Vectors made in this study

Name	Insertion	Method of modification	Comments and use
pCRISPRA	GCCGGAATTCGCTTTCACGATAACGATTACA ACAGTTATTTGGTAAGAGCTGATGTATATAAT CTTTTGTATTTATGCATATATGATAAACTTA TTCTTAATTCTCAGATAAAGGATTTTCATTATA TTGGCGGTTATTAATTGGGAAAACAAACGTG CTTAAAAAGCTGTTTAAAAAGATAATGGTGC CTTAAATGAAAAATTTATAATTGAAGTCGGA ATAGTAGTAAACGATTATTTACGTGATGTAAC GGTTTTATGAAAGTAAAGAGATAAAGAGAAA ACCGGTTAAGTTCGTTTTTCATGAAGTTGTTTA AAAGTGTGAAAGTTCGAGTCTCAATGCGACC GAAACGAATCTTCTATAATAATTGAACGTTT ATAAATGATAGGGTGTATTTCAATTTAACATA AAATCCTTGCGACCAGAAATTGTTAAATTAAT	gBlock (IDT) cloned into pUC19 using <i>Eco</i> RI and <i>Bam</i> HI restriction sites (underlined)	Integration assays, Chapter 6

	TACAAC TAAAATTGGTCGCATGAAGAGTAAA GGGTAGTCATGAAGATTTATAAGTAAGAAAA GAGAAAGAAAGATAGGAAGTATAAAAACACA ACAGATTAATCCCAAAGGAATTGAAAGGAA CTAGCTTATAGTTTAGGGATCCGCCG		
pCRISPRC	GCCGGAATTCGGATTGAAAAA ACTATAAAAA AATTGAAAACGCAAACCAGAGAAAAGCTTAT AAATAACTAAGGAGAAAATAAGAAAATAGTAA GAGATTAATAAACCCCTCAGATAATCTCTTATA GAATTGAAAGCAACCTTACTAGCCTCTTGTG TTGCTTCATCTGGGCTAAGGATCCGCCG	gBlock (IDT), cloned into pUC19 using <i>EcoRI</i> and <i>BamHI</i> restriction sites (underlined)	Integration assays, Chapter 6
pCRISPRC pal mutant	GCCGGAATTCGGATTGAAAAA ACTATAAAAA AATTGAAAACGCAAACCAGAGAAAAGCTTAT AAATAACTAAGGAGAAAATAAGAAAATAGTAA GTAAAATATAAACCCCTCAGATAATCTCTTATA GAATTGAAAGCAACCTTACTAGCCTCTTGTG TTGCTTCATCTGGGCTAAGGATCCGCCG	gBlock (IDT) cloned into pUC19 using <i>EcoRI</i> and <i>BamHI</i> restriction sites (underlined)	Integration assays, Chapter 6
pCRISPRC rep mutant	GCCGGAATTCGGATTGAAAAA ACTATAAAAA AATTGAAAACGCAAACCAGAGAAAAGCTTAT AAATAACTAAGGAGAAAATAAGAAAATAGTAA GAGATTAATAAACCCCTCATATAATCTCTTATA GAATTGAAAGCAACCTTACTAGCCTCTTGTG TTGCTTCATCTGGGCTAAGGATCCGCCG	gBlock (IDT) cloned into pUC19 using <i>EcoRI</i> and <i>BamHI</i> restriction sites (underlined)	Integration assays, Chapter 6
pChi1451- T6	GCCGGGATCCTTCTCTTAACGATGAAGTAAG TTTTTCCCTAGTTTAATTATTAATCTTTATAT AGAGATGATCTTCTTAATTCTAGGTTAATCCC TAGTAAATTCGGGAATTCTTTTACCGAGTAAA GTTTAAATACTTATATAGATAGAGTATAGATA GAGGGTTCAAAAAATGGTTTCACCCCAAACC CGAAAAGAAGAAGAAGCTTATCGATACCGTC GACCTCGAGGCCG	gBlock (IDT), cloned into pBluescript SK+ using <i>BamHI</i> and <i>XhoI</i> restriction sites	<i>in vitro</i> transcription, Chapter 3
p1451prom	<u>CACCACATCAAACGACCCCCACTT</u> ACAAAAA CGGGACAAAAAATACAAAATTACTAGACTATT CTCTTAACGATGAAGTAAGTTTTTCCCTAGT TTAATTATTAATCTTTATATAGAGATGATCTTC TTAATTCTAGGTTAATCCCTAGTAAATTCGGG AATTCTTTACCCCCCTCCTTAAAACGGTTTT TAGATTTTTCAACTGCTATTATATTGTGAGGT CGCAGATAGTTAGACAGCTACGAAGACTCCA CTCATATAGGGCCTCAGACCTATTGAGGAG GAGCTTCGGGGGTGGAAGTATTATATGCCTC <u>CGATAAAGCCTAGACG</u>	PCR amplification from <i>S.</i> <i>solfatarius</i> P2 genomic DNA (primers underlined), followed by directional TOPO cloning into pET151/D- TOPO (ThermoFisher Scientific)	<i>in vitro</i> transcription, Chapter 3

The *ssol451* operator sequence was amplified (using *ssol451promf* and *ssol451promr* primers, see Table 2.1 for sequences) from *S. solfataricus* P2

genomic DNA. The amplified product was Topo cloned into the pET151/D-TOPO vector (ThermoFisher Scientific) (see Table 2.3 for details). The *S. shibitae* viral T6 promoter was amplified and cloned into the pBluescript SK+ vector (Agilent), by previous members of the White lab (Paytubi & White, 2009), to form the pT6 plasmid. A chimeric version of the *sso1451* and viral T6 promoter was ordered as a gBlock and cloned into the pBluescript SK+ vector (Agilent), using the *Bam*HI and *Xho*I restriction sites to form the pChi1451-T6 plasmid (see Table 2.3 for details).

2.1.5 Strains

Vectors used for DNA sequencing, cloning, as substrates in assays and for *in vitro* transcription were amplified in, and purified from *E. coli* DH5 α or TOP10 cells (ThermoFisher Scientific). *S. solfataricus* proteins were expressed recombinantly in C43 *E. coli* cells. Glycerol stocks for long-term storage of strains were made by mixing 1000 μ l of an overnight cell culture with 500 μ l of 60% sterile glycerol. The stocks were stored at -80°C.

2.2 Methods

2.2.1 Cloning and protein expression

2.2.1.1 Site-directed mutagenesis

Site-directed mutagenesis (SDM) was carried out to make changes in protein gene sequences so that selected residues were mutated to alanine. Primer pairs used for SDM are shown in Table 2.1 (p.47) (*sso1450* E142Af and *sso1450* E142Ar; *sso1450* W150Af and *sso1450* W150Ar; *sso1450* N175Af and *sso1450* N175Ar; *sso1450* R166Af and *sso1450* R166Ar; *sso1404* D10Af and *sso1404* D10Ar). The vector containing the wildtype gene sequence was used as a template for SDM and the reaction was carried out on a TC-512 Thermal Cycler (Barloworld Scientific). The standard QuikChange II PCR protocol (Agilent), using *Pfu* polymerase (2.5 U) (ThermoFisher Scientific), was followed. After completion of PCR, 1 μ l of *Dpn*I enzyme (10 U/ μ l) was added and the reaction mix was incubated at 37 °C for 1 hour to digest parental plasmid. Competent DH5 α cells were then transformed by addition of 1 μ l of the reaction mix. Transformants were selected by overnight growth at 37 °C on agar plates containing the appropriate antibiotic. Plasmids were

extracted by QIAprep Spin Miniprep Kit (Qiagen) and sequenced (GATC Biotech) to confirm the mutation.

SDM of *cas1* was carried out by previous University of St Andrews undergraduate students James Robson and Kotryna Temcinaite. I contributed by supervising the students and providing technical assistance.

2.2.1.2 SDS-PAGE

Proteins were resolved by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) on 4-12% NuPAGE Bis-Tris gels (ThermoFisher Scientific). Before loading, samples were mixed 4:1 with protein loading buffer (4X NuPAGE LDS sample buffer, 1 mM DTT) and heated for 2 min at 90°C. The samples were then transferred to ice, loaded on the SDS-PAGE gel and run at 200 V for 35 min in 1X MES running buffer (50 mM MES, 50 mM Tris Base, 0.1% SDS, 1 mM EDTA, (pH 7.3)) (ThermoFisher Scientific). Molecular weights were determined against PAGERuler unstained protein ladder (ThermoFisher Scientific). Gels were stained with InstantBlue coomassie dye (Expedeon) for 30 min and rinsed in water before analysis.

2.2.1.3 Restriction digests and DNA ligation

Restriction digests of plasmids and inserts were carried out according to the enzyme manufacturer's protocol. Typically, double digests were performed with FastDigest restriction enzymes (ThermoFisher Scientific). 1-2 µg of vector DNA was digested in a 50 µl reaction with 1X FastDigest buffer (ThermoFisher Scientific) and ~1 unit of desired restriction enzyme/µg of DNA for 2 hours at 37 °C. Restriction enzymes were heat inactivated where possible and the DNA product purified using the Wizard SV Gel and PCR Clean-Up System (Promega).

Ligation of insert and vector fragments following restriction digest was carried out in a 10 µl reaction with 50 ng vector and the required volume of insert to achieve a molar ratio of 1:1 – 1:3 vector:insert. 1 unit of T4 ligase and 1 µl of 10X T4 ligase buffer (ThermoFisher Scientific) were added and the reaction was made up to 10 µl with RNase-free H₂O. The ligation reactions were incubated overnight at 16 °C, before 1-5 µl of the ligation mix was transformed into competent DH5α *E. coli* cells. Positive clones were selected, prepared and sent for DNA sequencing.

2.2.1.4 Protein over-expression and purification

Starter cultures for expression were made by mixing 10 μ l of a C43 *E. coli* glycerol stock with 10 ml Luria-Bertani (LB) broth and kanamycin (35 μ g/ml) (pEHISTEV vector and derivatives), or ampicillin (100 μ g/ml) (pDEST14 vector). The cultures were grown overnight at 37 °C with 180 revolutions per minute (rpm) shaking. Expression was then scaled up to 1 l from every 10 ml starter culture (typically 4 l of culture was grown). Cultures were incubated at 37 °C, with 180 rpm shaking until the optical density at 600 nm (OD_{600}) reached 0.6 - 0.8. Expression of the gene of interest was induced by the addition of 0.4 mM isopropyl β -D-thiogalactopyranoside (IPTG). Expression was carried out at 25 °C overnight with 180 rpm shaking.

Cells were then harvested by centrifugation at 6000 rpm, at 4 °C for 20 min, with the Avanti J20-XP centrifuge and JLA 8.1000 rotor (Beckman Coulter). Cell pellets were resuspended in 100 ml lysis buffer (4.5 mM NaH_2PO_4 , 15 mM Na_2HPO_4 , 500 mM NaCl, 30 mM imidazole, 1% Triton-X (pH 7.5) and protease inhibitors (Sigma-Aldrich)) and lysed by sonication (Soniprep 150, MSE) at 12 microns for 5 x 3 min bursts. The lysate was centrifuged at 35,000 rpm for 35 min, at 4 °C, with the Optima L-90 K Ultracentrifuge and 70Ti rotor (Beckman Coulter). The supernatant was filtered through a 0.22 μ m syringe filter (Millipore) and loaded onto a 5 ml FF/HP HisTrap column (GE Healthcare), which had been pre-equilibrated with buffer A (4.5 mM NaH_2PO_4 , 15 mM Na_2HPO_4 , 500 mM NaCl, 30 mM imidazole (pH 7.5)). Unbound proteins were eluted by washing the column in buffer A on the ÄKTA protein purification system (GE Healthcare), until the absorbance at 280 nm approached baseline. The His-tagged protein of interest was eluted over a linear imidazole gradient of 30 – 500 mM, provided by buffer B (4.5 mM NaH_2PO_4 , 15 mM Na_2HPO_4 , 500 mM NaCl, 500 mM imidazole (pH 7.5)). Fractions containing the Cas protein of interest were identified by SDS-PAGE, concentrated and buffer exchanged into buffer A, using centrifugal filter units (Amicon, Millipore). The poly-histidine tag was cleaved by overnight incubation at room temperature with TEV protease, at ratio of 1:10 TEV:tagged protein. The cleaved protein was loaded on a 5 ml FF HisTrap column in buffer A and collected in the flow-through. The sample was further purified by gel filtration on a 26/60 Superdex 200 prep grade column (GE Healthcare) in gel filtration buffer (20 mM Tris-HCl (pH 7.5), 500 mM NaCl, 0.5 mM DTT, 1 mM EDTA, 10% glycerol). The purity of the Cas protein elution peak was assessed by SDS-PAGE and selected fractions were pooled and concentrated.

Concentrated protein samples were flash-frozen using liquid nitrogen and stored at -80 °C.

The *E. coli* Cas1 and Cas2 proteins used in this study were kindly provided by Dr Ed Bolt and had been cloned and purified as described previously (Rollie et al., 2015).

2.2.1.5 Protein concentration determination

Protein concentrations were calculated using the absorbance of the solution at 280 nm, measured on the NanoDrop spectrophotometer (ThermoFisher Scientific). The extinction coefficient and the protein molecular weight, calculated using the ProtParam program (Gasteiger et al., 2005), were also required to calculate concentration. Cas proteins Cas1, Cas2, and Csa3 are all present as dimers in solution, therefore their dimer molecular weights were used to calculate protein concentrations quoted in this study.

2.2.2 Substrate preparation

2.2.2.1 Gel purification

Lyophilised nucleic acids were resuspended in TE-NaCl buffer (10 mM Tris-HCl, 1 mM EDTA, 10 mM NaCl) (DNA substrates) or RNase-free H₂O (RNA substrates) to a concentration of 500 μ M and stored at -20 °C until required. 2 μ l of the oligonucleotide (500 μ M) was diluted with 8 μ l RNase-free water and 10 μ l denaturing loading buffer (100% formamide, 0.25% bromophenol blue and 0.25% xylene cyanol) and the mixture was heated at 90 °C for 5 min. The solution was then cooled on ice before being loaded on a pre-run denaturing polyacrylamide-TBE gel (20% polyacrylamide, 7 M urea). Gels were run in 1X Tris-Borate-EDTA running buffer (100 mM Tris (pH 8), 90 mM M boric acid, 1 mM EDTA) at 22 W and 45 °C for between 1.5 and 3 hours, depending on oligonucleotide length. Substrates were visualised using UV shadowing (Minerallight USV-54 UV wand) and the substrate band was excised. The gel band was soaked in 400 μ l TE-NaCl buffer/RNase-free H₂O overnight at 4 °C. The supernatant was then decanted and filtered before the nucleic acid was extracted by ethanol precipitation as described below.

2.2.2.2 Ethanol precipitation

Ethanol precipitation of DNA/RNA substrates was carried out by adding 2 volumes of cold (4 °C) 100% ethanol and 0.1 volume of 3 M (pH 5.2) sodium acetate (C₂H₃NaO₂). The solution was then centrifuged at 13,200 rpm and 4 °C (Eppendorf fixed angle F-45-24-11 rotor) for 30 min, before the supernatant was decanted. 2 volumes of cold 70% ethanol was added to the nucleic acid pellet and the solution was centrifuged for a further 30 min (Eppendorf fixed angle F-45-24-11 Rotor, at 13,200 rpm). The ethanol was carefully decanted and the pellet was air-dried and resuspended in the desired volume of RNase-free water (RNA substrates) or TE-NaCl buffer.

2.2.2.3 Nucleic acid concentration determination

Nucleic acid concentration was calculated using the extinction coefficient of the substrate (provided by IDT) and the measured absorbance of the solution at 260 nm using a Varian Cary Eclipse spectrophotometer (Agilent Technologies).

2.2.2.4 Preparation of substrates with multiple DNA strands

To produce substrates made up of multiple oligonucleotides, the purified complementary single strands were mixed at equimolar concentrations in TE-NaCl buffer and heated at 90 °C for 5 min in a heat block. The block was turned off and the reaction was left to cool overnight to room temperature, to allow annealing of the oligonucleotides. The annealed substrate was mixed 5:1 with native loading buffer (15% ficoll, 0.025% bromophenol blue, 0.025% xylene cyanol) and run on a native polyacrylamide-TBE gel. The concentration of polyacrylamide was varied from 6-12% depending on the size and complexity of the complete substrate. Native gels were run at 180 V at room temperature for 3-5 hours before visualisation of substrate by UV shadowing. The substrate was excised and extracted as described for single-stranded substrates.

2.2.2.5 Plasmid DNA preparation

Plasmid DNA for cloning or assays was extracted from cell cultures using a QIAprep Spin Miniprep Kit. Cells were harvested from 10 ml overnight cultures by centrifugation at 4000 rpm (Eppendorf A-4-62 rotor) for 10 min at 4 °C. The cell

pellet was then resuspended and the plasmid was extracted according to manufacturer's instructions (Qiagen).

Long double-stranded or supercoiled DNA substrates were purified by separation on agarose gels (0.8 – 1.5%) pre-stained with ethidium bromide. The agarose gels were typically run at 100 V for 1 to 1.5 hours in 1X TBE. Before loading, samples were mixed 6:1 with 6X DNA loading dye (ThermoFisher Scientific). A GeneRuler 1 kb DNA ladder (ThermoFisher Scientific) was also run on the agarose gel to allow approximation of substrate size. DNA visualisation was carried out on a UV transilluminator (VWR) and substrates were excised and cleaned-up with the QIAquick gel extraction kit according to the manufacturer's instructions (Qiagen).

2.2.2.6 5' end-labelling with [γ - ^{32}P] ATP

Oligonucleotides to be used in assays were ordered with a 5' or 3' fluorescein (FAM) label, or were 5' end-labelled in-house with ^{32}P . The single-stranded oligonucleotide (0.2 - 20 μM) was added to an end-labelling reaction mix containing 1X T4 PNK buffer A (ThermoFisher Scientific), 1 μl [γ - ^{32}P] ATP (10 mCi/ml) (PerkinElmer) and 1 μl T4 Polynucleotide kinase (PNK) (ThermoFisher Scientific) (10 U/ μl) in a final reaction volume of 20 μl . The assay was incubated at 37 °C for 1 hour, before the enzyme was heat inactivated at 80 °C for 10 min. The free isotope was removed by gel extraction or by purification on a MicroSpin G-25 column (GE Healthcare).

2.2.3 Substrate ladders

2.2.3.1 A + G Maxam-Gilbert DNA Ladder

5 ng of ^{32}P -radiolabelled single-stranded DNA was mixed with 1 μl (1 mg/ml) calf thymus DNA and the volume was made up to 9 μl with TE buffer. 1 μl of 4% formic acid was then added and the reaction was incubated at 37 °C for 25 min before being transferred to ice. 150 μl of 1 M piperidine was added and the solution was incubated at 90 °C for 30 min. Following this incubation the reaction was transferred to ice and the product was extracted by ethanol precipitation, resuspended in 20 μl denaturing loading dye (100% formamide, 0.025% xylene cyanol, 0.025% bromophenol blue) and stored at 4 °C.

2.2.3.2 RNA alkaline hydrolysis ladder

100 ng labelled ssRNA was heated at 90 °C in alkaline hydrolysis buffer (50 mM NaHCO₃ (pH 9.2), 1 mM EDTA) for 2, 5 or 15 min. Reactions were stopped by incubation on ice and addition of 1 volume of formamide (100%). Aliquots from each time point were mixed to achieve optimal ladder resolution, and ~2 µl of the final ladder was loaded on denaturing polyacrylamide gels.

2.2.4 Binding assays

2.2.4.1 Electrophoretic mobility shift assay (EMSA)

Increasing concentrations of the protein of interest were incubated at room temperature in 10 µl reactions containing 200 nM FAM-labelled nucleic acid substrate in binding buffer (20 mM Tris (pH 7.5), 10 mM NaCl, 1 mM DTT, 5 mM EDTA). Following a 30 min incubation, reactions were mixed 5:1 with native loading buffer (15% ficoll, 0.025% bromophenol blue, 0.025% xylene cyanol) and run on native polyacrylamide-TBE gels (6-12% polyacrylamide depending on substrate). Gels were run at room temperature and 180 V for 3-5 hours and scanned on the Typhoon FLA-5000 imaging system (GE Healthcare). Specific protein concentrations, substrates and modifications are indicated in figure legends.

2.2.4.1.1 EMSA with small molecule extract

S. solfataricus P2 cell lysate was prepared from 5.7 g of cell pellet resuspended in 20 ml of binding buffer, as described previously (Götz et al., 2007). The cleared lysate was passed through 0.45 µm and 0.22 µm filters, followed by spin concentrators of decreasing kDa cut-offs to remove large molecules and proteins. The flow-through from each step was collected and reloaded into a concentrator with a lower molecular weight cut-off (from 50 kDa to 3 kDa). A 10 min spin at 4000 rpm and 4 °C (Eppendorf A-4-62 rotor) was performed at each cut-off. From 20 ml of original homogenate, 7 ml small molecule extract remained after filtration. 1 µl aliquots of this extract were added to 10 µl EMSA assays, performed as described above, with increasing Csa3 protein concentrations from 0 - 2 µM.

2.2.4.2 Fluorescence anisotropy

Fluorescence anisotropy measurements were taken with the Varian Cary Eclipse spectrophotometer with automatic polarizers (Agilent Technologies). 5'-FAM-labelled substrates were diluted to 10-30 nM in binding buffer, and 150 μ l of the diluted substrate was loaded into a quartz cuvette. The protein of interest was titrated into the cuvette, and anisotropy (r) and total fluorescence intensity were measured at each addition. The FAM dye of the sample was excited at 490 nm and emission spectra were collected at 530 nm. The G-factor was calculated automatically before each set of measurements. Triplicate titrations were carried out for each substrate and anisotropy values were plotted against substrate concentration, using KaleidaGraph (Synergy Software). The data were fitted to a binding isotherm (Equation 3, Chapter 3) that assumes 1:1 binding of protein:nucleic acid, as described previously (Reid *et al.*, 2003).

2.2.4.3 DNaseI footprinting

500 nM FAM-labelled double-stranded DNA was incubated at room temperature for 15 min with varying concentrations of Csa3 (SSO1445) from 0.1-10 μ M in footprinting buffer (20 mM (Tris pH 8), 30 mM NaCl, 1 mM DTT, 1 mM MgCl₂, 100 μ g/ml calf thymus DNA, 100 μ g/ml BSA). After the initial incubation, 2 μ l DNaseI (1 μ g/ml) was added and the reaction was incubated at 37 °C for 1 min before the reaction was stopped by the addition of 20 μ l of 2 mM EDTA. The solution was mixed with an equal volume of formamide (100%), boiled, then run on a 20% denaturing polyacrylamide-TBE gel (20% polyacrylamide, 7 M urea). The gel was run at 90 W, 45 °C for 1.5 hours and exposed to an imaging plate (IP) overnight before visualisation. A Maxam-Gilbert A+G ladder was also run to allow mapping of the protected sites.

2.2.5 *In vitro* transcription

The *S. solfataricus* TBP, TFB and RNA polymerase (RNAP) proteins required for *in vitro* transcription were purified as described previously (Götz *et al.*, 2007). RNAP was isolated and purified from *S. solfataricus* cell extract, while TFB and TBP were expressed recombinantly in *E. coli*. All proteins were flash-frozen in aliquots at 4 μ M and stored at – 80 °C. The vectors used for *in vitro* transcription (see Table 2.3 and Paytubi & White (2009) for details) were prepared using the QIAprep Spin Miniprep

Kit and linearised by restriction digest with *SacI* (p1451prom) or *XhoI* (pT6 and pChi1451-T6) (ThermoFisher Scientific). The linearised vectors were gel-purified and extracted before being used in transcription assays. 50 ng (pT6 and pChi1451-T6) or 100 ng (p1451prom) linearized plasmid was incubated in transcription buffer (20 mM Tris (pH 8), 220 mM KCl, 10 mM MgCl₂, 2 mM DTT) with RNAP (80 nM). The putative transcriptional regulator Csa3 was added in varying concentrations and the reaction was incubated for 30 min at 55 °C. TFB and TBP (40 nM each) and BSA (14 µM) were then added and the reaction was incubated at 70 °C for 10 min before the addition of 200 µM of each rNTP (ThermoFisher Scientific) to initiate transcription, and further 20 min incubation at 70 °C.

Following the transcription reaction, 300 fmol ³²P-labelled DNA primer (T6r or sso1451promr, see Table 2.1 (p.47)), complementary to the transcription product, was incubated with 5 or 13 µl of the reaction mix at 70 °C for 5 min, then chilled on ice. Once cool, 4 µl 5X RT buffer, 0.1 µl Ribolock RNase inhibitor (40 U/µl) and 2 µl dNTPs (10 mM mix) (all from ThermoFisher Scientific) were added. The reaction was made up to 19 µl with RNase-free H₂O and incubated at 37 °C for 5 min before the addition of 1 µl (200 U) RevertAid reverse transcriptase (ThermoFisher Scientific). The reverse transcriptase reaction was incubated at 42 °C for 1 hour. The cDNA product was then phenol-extracted and run on a denaturing polyacrylamide-TBE gel (12% polyacrylamide, 7 M urea) at 90 W, 45 °C for 1.5 hours before phosphorimaging.

2.2.6 Assaying changes in gene expression

2.2.6.1 RNA extraction from *S. solfataricus* cell pellets

S. solfataricus P2 cell pellets were obtained from the Dr Susanne Erdmann, Garrett lab, University of Copenhagen. Control pellets from four time points during culture growth were provided as well as infected pellets from the same time points. The infected cultures had been grown in the presence of the *Sulfolobus* monocaudavirus 1 (SMV1) and conjugative plasmid pMGB1 as described in Erdmann et al., 2013. Approximately 30 mg samples of the cell pellets were used for RNA extraction, carried out using the Gram-positive protocol of the RNeasy Mini Kit (Qiagen). The protocol was adapted for small sample size by lowering the volume of lysis buffer to 200 µl, and for archaeal cells by digesting with proteinase K (ThermoFisher

Scientific). On-column DNaseI digests were performed using RNase-free DNase Set (Qiagen) as part of the extraction procedure. The quality of the extracted RNA was assessed by separation on a 0.8% agarose gel, which produced two well-defined bands of ribosomal RNA. The 260/280 ratio was also measured using the NanoDrop spectrophotometer (ThermoFisher Scientific) and found to be >2 for all samples.

2.2.6.2 Reverse transcription-quantitative PCR (RT-qPCR)

2.2.6.2.1 Primer efficiency determination

The efficiency of each primer pair (see Figure 3.4) to be used for RT-qPCR was evaluated in PCR reactions with 10-fold dilutions of genomic *S. solfataricus* P2 DNA. These reactions contained 10 μ l 2X iQ SYBR Green Supermix (Bio-Rad), 1 μ l forward and reverse primers (10 μ M) and 1 μ l template DNA (100, 10, 1, or 0.1 ng). The reaction was made up to 20 μ l with RNase-free H₂O. The amplification was carried out in 96 well plates (StarLab), which were sealed with X-Clear Advanced Polyolefin StarSeal film (StarLab) and the PCR reaction was run in the iCycler IQ system (Bio-Rad). A standard curve was plotted of the crossing point (Ct) values collected, against input DNA concentration. The gradient of this curve was then used to calculate primer efficiencies with Equation 1 (Chapter 3) (Pfaffl, 2001).

2.2.6.2.2 One-step RT-qPCR

One-step RT-qPCR reactions were carried out with the iScript One-Step RT-PCR Kit (Bio-Rad). Reactions contained 50 ng *S. solfataricus* RNA, 25 μ l 2X SYBR Green RT-PCR reaction mix, 1.5 μ l forward and reverse primers (10 μ M), and 1 μ l iScript reverse transcriptase and were made up to 50 μ l with RNase-free H₂O. Reactions were set up as above, with an altered PCR program beginning with a 10 min incubation at 50 °C to allow cDNA synthesis before PCR cycling began. Cycling conditions were as follows: 1 cycle at 95 °C for 3 min, followed by 35 cycles of 95 °C for 30 sec and 55 °C for 30 sec. A melt curve analysis was also performed automatically by the iCycler. Ct values calculated by the iCycler were used to calculate the fold-change in transcript levels between control and infected samples using the Pfaffl equation, as described in Chapter 3 (Equation 2) (Pfaffl, 2001). *ssb* transcripts were amplified as endogenous controls. Each transcript was assessed by triplicate assays, with each triplicate containing 3 intra-assay replicates (see Figure 3.5 and Appendix A: Triplicate Ct values from RT-qPCR (p.217)). Control

PCR reactions without template or reverse transcriptase were carried out to assess purity of the samples.

2.2.6.3 Western Blot

2.2.6.3.1 Antibody generation

Polyclonal primary antibodies were raised in sheep against the selected *S. solfataricus* proteins that had been recombinantly expressed in *E. coli*. The antibodies were supplied by the Scottish National Blood Transfusion Service, Pentlands Science Park, Midlothian.

2.2.6.3.2 Blot method and scanning

A western blot was carried out to assess differences in Cas protein levels in control and infected *S. solfataricus* cultures. 20 mg cell pellet samples were dissolved in 200 μ l lysis buffer (20 mM Hepes (pH 7.5), 250 mM NaCl, 10 mM imidazole) and sonicated at 12 microns for 3 x 12 sec. 30 μ l of the lysate from each sample was then mixed 4:1 with protein loading buffer (4X NuPAGE LDS sample buffer, 1 mM DTT), boiled for 5 min and separated by SDS-PAGE. A positive control of the recombinantly expressed protein of interest (40 nM final) was also loaded on the gel.

The SDS-PAGE gel was blotted onto a nitrocellulose membrane using the iBlot Dry Blotting System (ThermoFisher Scientific). The membrane was blocked for 10 min in blocking buffer (500 ml phosphate buffered saline (PBS) (137 mM NaCl, 2.7 mM KCl, 8.1 mM Na₂HPO₄, 1.76 mM KH₂PO₄ (pH 7.4)), 250 μ l Tween-20, 25 g marvel milk powder). Membrane was incubated with shaking in blocking buffer containing 1:1000 primary antibody for 1.5 hours, washed 3x in blocking buffer, then incubated in the dark with 1:10000 dilution of secondary antibody (IRDye 800CW Donkey Anti-Goat IgG (H+L), LI-COR) for 2 hours. The blot was washed again (3x) and imaged on the Odyssey CLx (LI-COR) scanner with excitation at 778 nm and emission at 795 nm.

2.2.7 Assessing complex formation

2.2.7.1 Isothermal titration calorimetry (ITC)

Cas1 and Cas2 (150 μ M) were dialysed into ITC buffer (200 mM KCl, 20 mM HEPES-KOH (pH 7.5), 5% glycerol and 1 mM TCEP) overnight at room temperature.

Titration was performed on the MicroCal (GE Healthcare) instrument with the reference power of 6 $\mu\text{cal}/\text{sec}$, stirring at 900 rpm and the temperature set to 25 °C. 150 μM Cas1 was loaded into the injection syringe and titrated into the cell containing 15 μM Cas2. Following a null injection of 0.4 μl , 16 injections of 2.5 μl were performed until the molar ratio of Cas1:Cas2 was $\sim 2:1$. Injections were performed every 180 sec and had a duration of 5 sec. Data were displayed using Origin software (OriginLab) and the integrated heat changes were plotted.

2.2.7.2 Gel filtration

To assess whether Cas1 and Cas2 formed a stable complex, a gel filtration elution of the proteins after incubation together, with or without DNA (duplex DNA made by annealing 3'OHprotospacer for and 3'OHprotospacer rev, see Table 2.1 for sequences), was carried out. Cas1 and Cas2 \pm DNA were mixed and incubated together at 45 °C for 15 min before being dialysed together overnight at room temperature into sample elution buffer (20 mM Tris (pH 7.5) and 150 mM NaCl). The proteins and DNA were incubated at a molar ratio of 2Cas1:4Cas2:1DNA (40:80:20 μM). 100 μl samples of single and mixed samples of proteins were run on an equilibrated Superose 12 size-exclusion column (GE Healthcare) on the ÄKTA protein purification system. Following elution, the protein content of peak fractions was assessed by SDS-PAGE. Elution profiles were plotted and peak absorbances were normalised to 1.

2.2.8 Activity assays

2.2.8.1 Nuclease assays

2.2.8.1.1 Cas1 nuclease assay

Holliday junction substrates (Jbm5 Holliday, see Table 2.2) (50 nM) were 5'-³²P-radiolabelled on one strand and incubated with 500 nM Cas1 and 5 mM manganese chloride (MnCl_2) in a 10 μl reaction with nuclease buffer (20 mM Tris (pH 7.5), 10 mM NaCl, 1 mM DTT). Following a 30 min incubation at 55 °C (*S. solfataricus* Cas1) or 37 °C (*E. coli* Cas1), 1 μl of 20 mg/ml proteinase K was added and the reaction was incubated at 37 °C for 30 min. The product was then extracted by adding 40 μl neutral phenol:chloroform:isoamyl alcohol (Sigma-Aldrich), vortexing for 10 sec then centrifuging the reaction at 13,200 rpm (Eppendorf fixed angle F-45-24-11 Rotor), 4 °C for 1 min. The upper, aqueous, phase containing the DNA was

removed and mixed 1:1 with denaturing loading buffer (100% formamide, 0.25% bromophenol blue and 0.25% xylene cyanol), heated at 90 °C for 2 min and resolved on a pre-run denaturing polyacrylamide-TBE gel (15% polyacrylamide, 7 M urea) in 1 X TBE buffer at 50 °C and 90 W for 1.5 hours, before phosphorimaging.

2.2.8.1.2 Cas2 nuclease assay

5' FAM labelled substrates (200 nM) were incubated with 5 μ M Cas2_{CD} in nuclease buffer and either 5 mM MgCl₂ or 5 mM EDTA at 55 °C for 20 min. The assay was stopped by the addition of 1 volume of formamide (100%) and heating at 90 °C for 5 min. The products were resolved on a pre-run denaturing polyacrylamide-TBE gel (20% polyacrylamide, 7 M urea) as described above.

2.2.8.2 Disintegration reactions

2.2.8.2.1 Standard disintegration reaction

Disintegration reactions were performed using branched substrates with 5' flaps (see Table 2.2). A typical reaction contained 200 nM disintegration substrate, mixed with 2 μ M Cas1 protein in nuclease buffer (20 mM Tris (pH 7.5), 10 mM NaCl, 1 mM DTT) and 5 mM MnCl₂ and was incubated at 55 °C (SsoCas1) or 37 °C (for EcoCas1). After 20 min 20 mM EDTA was added to stop the reaction, followed by the addition of 1 μ l 20 mg/ml Proteinase K (ThermoFisher Scientific) and incubation at 37 °C for 30 min. The nucleic acid was then separated from the reaction by phenol chloroform extraction with 60 μ l neutral phenol:chloroform:isoamyl alcohol (25:24:1) (Sigma-Aldrich). The upper aqueous phase, containing the DNA, was removed and mixed 1:1 with denaturing loading buffer (100% formamide, 0.25% bromophenol blue and 0.25% xylene cyanol) and heated at 95 °C for 5 min. The reaction was then chilled before being resolved on a pre-run denaturing polyacrylamide-TBE gel (20% polyacrylamide, 7 M urea). Gels were run in 1X TBE at 80 W, 45 °C for 90 min before overnight exposure to an IP and phosphorimaging.

2.2.8.2.2 Disintegration-coupled *SacI* digest

A standard disintegration reaction with the *S. solfataricus* Cas1 protein was carried out as described above with a disintegration substrate containing a *SacI* site (*SacI* substrate (see Table 2.2). 1 unit of *SacI* was then added to the disintegration products and the reaction was heated at 37 °C for 30 min in 1X FastDigest buffer. Proteinase K digest, phenol extraction, and product separation and visualization were performed as for the standard disintegration reaction described above.

2.2.8.2.3 Time course disintegration reactions

Disintegration reactions were set up with 50 nM substrate and either 50 nM (SsoCas1) or 500 nM (EcoCas1) protein. Other reaction components were as for the standard reaction described above. A mastermix was incubated at the reaction temperature (37 or 55 °C), aliquots were taken at 1, 2, 3, 5, 10, 15, 20 and 30 min, and the reaction was quenched by the addition of 20 mM EDTA, incubation on ice and phenol extraction. The time course reactions were carried out in triplicate, and resolved and visualized as described above. The average fraction cleaved at each time point was quantified using ImageGauge software (FUJIFILM Life Science) and plotted against time, with standard deviation shown as error bars. Kaleidagraph was used to fit a single exponential equation (Equation 4, Chapter 5), with either a fixed (SsoCas1) or floating end point (EcoCas1) as described previously by Niewoehner et al. (2014).

For disintegration time courses with Cas1 and Cas2, EcoCas2 (15 μ M) was pre-incubated with EcoCas1 (15 μ M) at 37 °C for 30 min, before the proteins were added to the reaction at a final concentration of 150 nM.

2.2.8.3 Integration reactions

2.2.8.3.1 Integration assay with radiolabelled protospacer

Cas1 and Cas2, both at 20 μ M, were incubated with 5³²P-radiolabelled DNA substrates for integration (20 μ M total, of which ~1% is labelled) at 55 °C for 30 min. 1 μ l of this solution was then added to a reaction containing 1 μ l (100 ng/ μ l) plasmid DNA, 1 μ l 10X integration buffer (200 mM Tris (pH 7.5), 100 mM NaCl), 1 μ l MnCl₂ (50 mM) and 5 μ l water making the total reaction volume up to 10 μ l. This reaction was then incubated at 55 °C for 30 min. Following the incubation, 1 μ l of proteinase K (20 mg/ml) (ThermoFisher Scientific) was added and the digest was incubated at 37 °C for 1 hour, before phenol extraction of the DNA. 10 μ l of the aqueous phase containing the DNA was removed, mixed with 2 μ l of 6X DNA loading dye and run on a 1% agarose gel, pre-stained with ethidium bromide, at 100 V for 1 hour in 1X TBE buffer. The wet gel was imaged before it was dried for 4 hours on a slab gel drier (Savant) and phosphorimaged.

2.2.8.3.2 Integration time course

An integration time course experiment was carried out by scaling up the standard integration reaction by 10-fold. The 10X mastermix was incubated at 55 °C, and 10

μl aliquots were taken at time points 2, 5, 10, 15, 20, 30, 60 and 120 min. The reaction was stopped at each time point by the addition of 50 mM EDTA and incubation on ice. After the 120 min time point, all samples were treated with proteinase K and phenol-extracted, before being resolved on a 1% agarose gel as described above. The time course experiment was completed in triplicate and the fraction of supercoiled substrate converted to open circle/nicked form at each time point was calculated using ImageGauge. The average fraction of substrate converted was plotted against time using Kaleidagraph, with standard deviation of the mean displayed as error bars.

2.2.8.3.3 *Bst*UI digest of integration reaction products

A 10X integration reaction was performed with radiolabelled protospacer DNA as above. The nicked integration product was visualized using UV, then excised and gel-extracted using the QIAquick gel extraction kit (Qiagen). The purified DNA was then digested with 1 U *Bst*UI restriction enzyme in 1X CutSmart Buffer (both from New England Biolabs) at 60 °C for 1 hour. The reaction products were resolved on a pre-stained ethidium bromide 1.2% agarose gel. The wet gel was scanned, before being dried and phosphorimaged.

2.2.8.4 PCR amplification of integration sites

A 9 μl reaction was prepared containing 200 ng of the pCRISPR/pCRISPRC plasmids (see Table 2.3), 5 mM MnCl_2 , 1X integration buffer and 2 μM protospacer substrate (made by annealing PCR protospacer F and PCR protospacer R, see Table 2.1 for sequences). 1 μl of a Cas1 and Cas2 mix (both at 20 μM) was added to this reaction and a 30 min incubation at 55 °C was carried out. The reaction was phenol-extracted and the aqueous phase was diluted 1:1 with RNase-free water. 1 μl of this dilution was added to a PCR reaction containing 1 μl of forward and reverse primer (Primer *Nco*I F, Primer *Xho*I R1) (10 μM), 10 μl 2X MyTaq Red Mix (Bioline) and 7 μl RNase-free water. The forward primer contained an *Nco*I restriction site and was complementary to the protospacer used in the integration assay. The reverse primer contained an *Xho*I restriction site and was complementary to a region of pUC19 downstream of the CRISPR insert. A PCR reaction was performed consisting of an initial denaturation step at 98 °C for 2 min, followed by 25 cycles of 98 °C for 30 sec, 55 °C for 30 sec and 72 °C for 1 min, with a final extension for 5 min at 72 °C and an infinite hold step at 4 °C.

The products of the PCR reaction were separated on a 1.5% agarose gel, which allowed rough localisation of the integration sites. PCR products selected for sequencing were cleaned up using the Wizard SV Gel and PCR Clean-Up System (Promega). Products were then digested with 1 μ l *Nco*I and 1 μ l *Xho*I FastDigest enzymes in a 20 μ l reaction containing 1X FastDigest buffer at 37 °C for 1 hour. 1 μ g of the pEHISTEV vector was also restricted using the same method with *Nco*I and *Xho*I to produce compatible ends for ligation of the insert. The digested inserts and plasmid were ligated (as described in section 2.2.1.3) and the ligation products were transformed into DH5 α *E. coli* cells. Transformants were selected by overnight growth at 37 °C on LB agar plates containing 35 μ g/ml kanamycin. Plasmids were extracted from positive clones by Miniprep and sent for sequencing using the T7 primer (GATC Biotech). The sequences around the insertion site (\pm 100 bp) were analysed for secondary structure using Mfold (Zuker, 2003) and the 10 bp immediately surrounding the insertion site were used to make a sequence logo on the WebLogo server (Crooks, 2004) .

The PCR amplifications described in Chapter 6 were carried out by Dr Shirley Graham. I contributed by providing the integration assay products and analysing sequence data.

2.2.8.4.1 Integration assays with *S. solfataricus* lysate

Integration assays coupled to PCR were modified by the addition of *S. solfataricus* lysate before Cas1 and Cas2 proteins. The reaction mix was set up as above without the addition of Cas proteins or RNase-free water. Different volumes of *S. solfataricus* cell lysate (1 – 5 μ l) (prepared as described previously (Götz et al., 2007)) were added to the reaction mix and the total volume was made up to 9 μ l with RNase-free H₂O before the addition of 2 μ M Cas1 and Cas2. The reaction was completed and the products resolved as described above.

Chapter 3: Regulation of the CRISPR-Cas system in response to infection

3.1 Introduction

3.1.1 Changes in *cas* gene expression during viral infection

While our understanding of the mechanisms of CRISPR-Cas immunity has advanced hugely in the last decade, studies of how this immunity is regulated remain scarce. However, there is an obvious need for the abundance of different protein elements of the system to be under tight control, in order that self-targeting and non-specific nucleic acid degradation is avoided.

To date, studies have focused on the silencing of the CRISPR system in *E. coli*, which was explored in detail in the introduction to this thesis. Briefly, the *cas* promoters of *E. coli* are cryptic due to silencing by the global repressor heat-stable nucleoid-structuring (H-NS) protein (Pul et al., 2010). This repression is lifted by the H-NS antagonist LeuO, which blocks cooperative binding of H-NS on the *cas* promoters (Westra et al., 2010). It was hypothesized that on viral infection H-NS is titrated away from *cas* promoters by AT-rich foreign DNA, which, in turn, allows LeuO to bind and de-repress transcription of the *cas* genes (Westra et al., 2010).

3.1.2 CRISPR-Cas regulation in archaea

Studies in archaea have also found that *cas* genes are kept at low levels in the absence of infection. A strong activation of these genes was reported during *Sulfolobus islandicus* rod-shaped virus 2 (SIRV2) infection, with *cas* genes and crRNA being upregulated by between 3- and 10-fold (Quax et al., 2013). In contrast, genes involved in controlling the cell cycle were strongly repressed from the onset of infection (Quax et al., 2013). Cas proteins were also found to be highly regulated during STIV (*Sulfolobus* turreted icosahedral virus) infection in *S. solfataricus* P2 (Maaty et al., 2012).

Interestingly, Erdmann & Garrett (2012) reported that infection with single-virus cultures failed to activate CRISPR adaptation in *S. solfataricus* P2. The majority of archaeal viruses studied to date can co-exist with the host without causing cell lysis.

Therefore, the authors suggest that this lack of CRISPR activation may in fact be beneficial; conserving energy and allowing favourable gene transfer (Erdmann & Garrett, 2012). They later showed that infection with an environmental sample, containing a viral cocktail and co-infecting conjugative plasmid, led to retarded culture growth and adaptation of CRISPR loci C, D and E. Intriguingly, all new spacers inserted during this co-infection experiment came from the conjugative plasmid and not from the viral DNA (Erdmann & Garrett, 2012).

Subsequent work by the Garrett lab showed that activation of adaptation at loci A and B of *S. solfataricus* was not triggered by viral infection alone, but also required environmental stress (Erdmann et al., 2013). The authors hypothesised that as environmental stress is also known to influence whether viruses adopt a lytic or lysogenic lifestyle, activation of CRISPR-Cas by the same signalling mechanism may provide an extra line of defence against host cell lysis.

More recently the *Sulfolobus* monocaudavirus 1 (SMV1) was identified as an essential component of viral co-infections in *Sulfolobus*, required for stable infection and adaptation of the CRISPR arrays (Erdmann et al., 2014). A single infection of *S. islandicus* with single-tailed fusiform *Sulfolobus* virus (STSV2), failed to trigger CRISPR-Cas adaptation. However, a co-infection with both STSV2 and SMV1 activated adaptation, with spacers originating exclusively from the STSV2 genome (Erdmann et al., 2014). The authors concluded that through an as yet unknown mechanism SMV1 'primes' spacer acquisition, whilst itself remaining immune to the host CRISPR-Cas system.

3.1.3 Transcriptional regulators in archaea

While considerable progress has been made into understanding the players involved in transcriptional regulation of CRISPR-Cas systems in bacteria, the mechanisms for control of *cas* gene expression in archaea remain enigmatic. Proteins suggested to play a part in modulating these CRISPR-Cas systems are introduced briefly below.

3.1.3.1 Cbp1

In *S. solfataricus* the CRISPR DNA repeat-binding protein (Cbp1) has been identified as a potential modulator of transcription (Deng et al., 2012). The authors

showed that overexpression of Cbp1 promotes an increase in pre-crRNA yields, while a knockout strain led to its depletion. The increase in pre-crRNA observed in the presence of Cbp1 was predicted to be due to the protein binding CRISPR repeat sequences and blocking transcriptional signals in the AT-rich spacers, which would otherwise lead to early transcriptional termination (Deng et al., 2012).

3.1.3.2 Csa3 proteins

The most concrete identification of a transcriptional regulator of the CRISPR-Cas response in archaea is that of the Csa3 proteins. The resolution of the crystal structure of a Csa3 protein from *S. solfataricus* revealed a dimeric protein with features which suggest a role in transcriptional modification (Lintner et al., 2011a). Csa3 monomer subunits have two domains: an N-terminal domain important for dimer formation and a C-terminal winged helix-turn-helix (wHTH) domain, predicted to bind DNA (Figure 3.1, A). Furthermore, a potential ligand-binding pocket was identified at the interface of the N-termini (Figure 3.1, C). It was suggested that a small, symmetric, hydrophobic molecule might bind in this pocket and regulate the activity of the protein - either to stimulate or repress transcription (Lintner et al., 2011a).

The Csa3 C-terminal fold is similar to that of the MarR transcriptional regulators, with three α -helices forming a right-handed bundle, two of which constitute the HTH motif, while the other is predicted to make sequence-specific contacts with the major groove of a regulatory DNA sequence (Lintner et al., 2011a). The two wHTH motifs of Csa3 are rich in positively charged residues (Figure 3.1, B), likely to interact with DNA. As each dimer contains two identical wHTH motifs, separated by a cleft, it was predicted that Csa3 would likely bind palindromic DNA sequences (Lintner et al., 2011a). The authors suggested that the Csa3 studied may be involved in the transcriptional regulation of the adjacent CRISPR arrays (CRISPR C and D) of *S. solfataricus*, but no binding was observed for the putative promoters associated with these arrays (Lintner et al., 2011a). As there are two Csa3 proteins in *S. solfataricus*, to avoid confusion, I will refer to the protein coded for by the *sso1445* gene, located between CRISPR C and D, as Csa3_{CD}.

The prediction that the Csa3 family of proteins act as transcriptional regulators was strengthened when a Csa3-family protein from *S. islandicus* REY15A was shown to activate transcription of a *cas* operon when overexpressed (Liu et al., 2015). The

Csa3a protein from this organism was found to bind to semi-palindromic sequences located in the promoter regions of the *csa1* and *cas1* genes, both coding for proteins involved in adaptation. Overexpression of Csa3a also led to the protein levels of Csa1 and Cas1 increasing, as measured by western blot, as well as the activation of hyperactive spacer uptake into both CRISPR arrays of *S. islandicus*. These results are supported by an earlier study that showed upregulation of both the *csa3a* gene and adaptation-related genes in response to rudivirus infection (Quax et al., 2013).

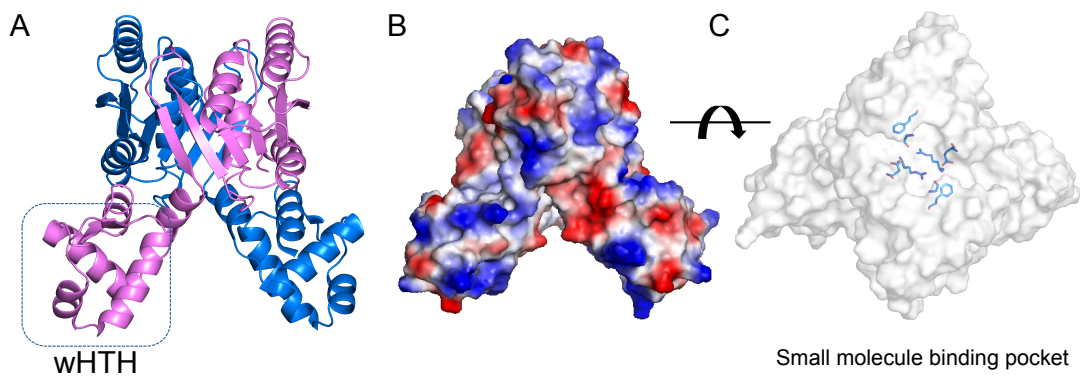


Figure 3.1 The structure of a Csa3_{CD} protein from *S. solfataricus*

A. The crystal structure of Csa3_{CD} (SSO1445) from *S. solfataricus* (PDB ID 2WTE). The protein is a dimer with a winged helix-turn-helix (wHTH) domain at the C-terminus and the dimer interface contained in the N-terminal domain. **B.** Surface electrostatic charge representation of Csa3_{CD}. The N-terminal wHTH domains are rich in positive residues and are predicted to bind DNA. **C.** Rotated view of the Csa3 surface structure showing a symmetrical putative small-molecule-binding domain at the N-terminal dimer interface. The pocket is lined with conserved residues shown in blue (Phe10, Arg98, Gly96, Glu122).

This chapter will describe work carried out to investigate the changes in expression of Cas proteins in *S. solfataricus* P2 undergoing CRISPR adaptation in response to viral infection. Previous studies have observed dramatic effects of infection on the expression of CRISPR-related proteins and I was particularly interested in examining changes in expression of the adaptation-related genes. My findings indicate that Cas proteins involved in adaptation and those which make up the Cascade complex are highly upregulated in response to infection, whereas the type III Csm and Cmr interference complexes are constitutively expressed and only show a minor increase in expression levels compared to control samples. Furthermore, I present data that indicate that the *S. solfataricus* Csa3_{CD} protein may be responsible for the regulation of adaptation genes in response to a viral infection in this system.

3.2 Results

I wish to acknowledge Dr Susanne Erdmann (University of Copenhagen) for the preparation and provision of cell pellets used in this chapter, as well as the PCR image and OD₆₀₀ values used in Figure 3.2.

3.2.1 Infection time course of *S. solfataricus* with SMV1 and pMGB

As described above, a stable *Sulfolobus* infection could be set up with the SMV1 virus and a co-infecting conjugative plasmid (pMGB) or STSV2 virus (Erdmann et al., 2013, 2014). In both of these studies SMV1 viral particles were present at the end of the infection experiment, whereas the conjugative plasmid or co-infecting virus was lost, apparently through CRISPR-Cas interference. As all *de novo* acquired spacers originated from the co-infecting elements in this study, it was concluded that SMV1 activates adaptation without being susceptible itself to the host Cas proteins (Erdmann et al., 2014).

To investigate the effect of viral infection on the expression of *cas* genes in *S. solfataricus*, infected and control *S. solfataricus* P2 cell pellets were obtained from Dr Susanne Erdmann, Garrett lab, University of Copenhagen. The infected culture had been grown in the presence of both the SMV1 virus and the pMGB conjugative plasmid. During growth the OD₆₀₀ was checked regularly, and the cultures were diluted if the absorbance exceeded ~1.5 (Figure 3.2, A). Activation of CRISPR adaptation was assessed by PCR amplification through the leader-repeat junction of two of the six CRISPR arrays of *S. solfataricus*, with any addition of new spacers leading to an increase in the length of the PCR product (Erdmann, 2013). Adaptation of CRISPR C was first observed at 8 days post-infection (dpi) (Figure 3.2, B) and was still highly active at 9 and 9.5 dpi, with bands of increasing size visible following PCR. Adaptation was thought to have slowed between 9.5 and 10 dpi, as no further increase in the length of PCR products was observed (Figure 3.2, B).

The expansion of CRISPR arrays was only apparent in the culture infected with the virus and conjugative plasmid and not in control samples (Figure 3.2, B). Adaptation of CRISPR locus A was not observed in infected cultures (Figure 3.2, B). This is not unexpected, as adaptation was previously found to be dormant at this locus during

SMV1 infection and only activated in response to environmental stress (Erdmann et al., 2013).

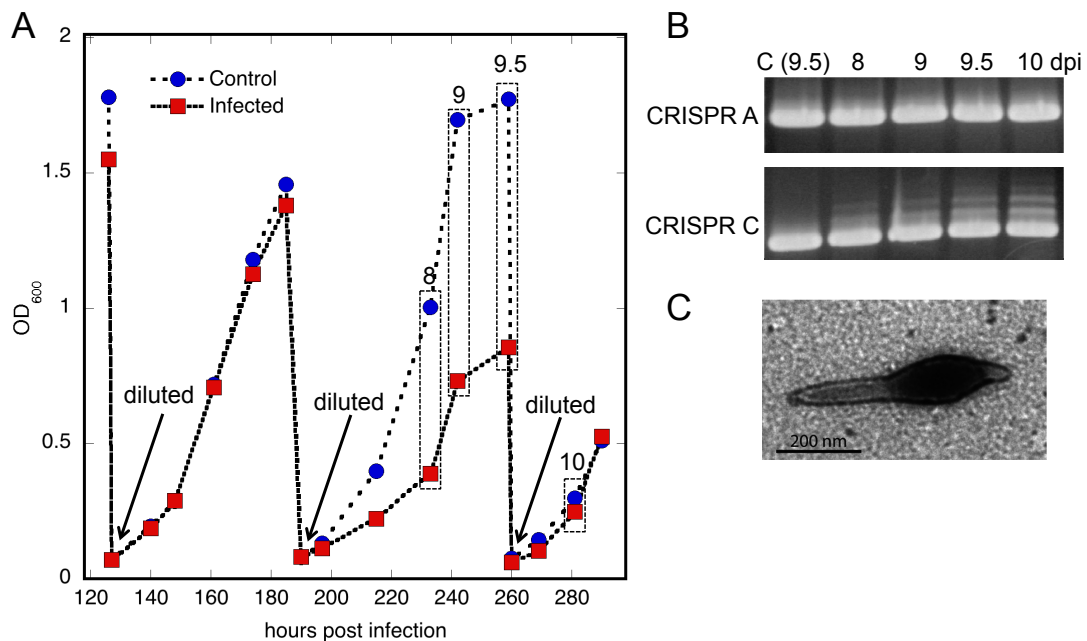


Figure 3.2 Infection time course of *S. solfataricus* with SMV1 and pMGB

A. Growth curves for control (blue circles) or infected (red squares) *S. solfataricus* cultures. OD₆₀₀ measurements were made at intervals and the cultures were diluted when OD₆₀₀ exceeded ~1.5. Dilutions are indicated at 127 and 190 and 260 hours post infection. The dashed rectangles mark time points (8, 9, 9.5 and 10 days post-infection) from which cell pellets were taken for further analysis. **B.** PCR amplification through the leader-repeat 1 junction of CRISPR loci A and C. The first sample for each array (C) is a control amplification from an uninfected sample from 9.5 dpi. The product is of the size expected for an unmodified leader-repeat junction. The subsequent lanes are PCR products from infected samples at 8, 9, 9.5 or 10 dpi. OD₆₀₀ values and PCR data were obtained and provided by Dr Susanne Erdmann, University of Copenhagen. **C.** Electron micrograph of an SMV1 viral particle, adapted from Erdmann et al., 2013.

Infected *S. solfataricus* cultures showed slower growth than the control cultures from 6 dpi (Figure 3.2, A). This growth retardation is likely to be a compound result of the viral infection and the energy expended due to the activation of adaptation. A similar growth retardation was observed during infection-induced CRISPR adaptation in *S. islandicus* (León-Sobrino et al., 2016). Following final dilution and regrowth at 10 dpi the growth rates of infected and control cultures were similar (Figure 3.2, A). This recovery could be due to the CRISPR-Cas response beginning to overcome infection and adaptation being switched off.

The cell pellet samples provided came from four time points during the infection time course, 8, 9, 9.5 and 10 days post infection (dpi) (Figure 3.2, A, indicated by the dashed boxes). The first cell pellet was harvested at the point *de novo* spacers were first found by PCR to be added to the CRISPR C array (8 dpi). The two

subsequent samples were taken during active adaptation and the final time point was taken after the adaptation was thought to have ended. For each time point a control sample of uninfected cells was also collected. The samples were pelleted, frozen and shipped for our analysis. The uninfected control culture at 10 dpi was later found to have been contaminated; therefore, samples from this time point were excluded from any further analysis.

3.2.2 Changes in Cas protein levels in response to infection

The control and SMV1 + pMGB-infected cell pellets were used to produce cell lysate in order to look for changes in abundance of the Cas proteins. Equal mass cell pellets (20 mg) were resuspended and lysed by sonication, before being cleared by centrifugation, and samples of the supernatant were separated by SDS-PAGE. Western blots were then carried out from these gels to look for differences in Cas protein levels between the control and infected samples (Figure 3.3) (see section 2.2.6.3.2 for full method). The *S. solfataricus* single-strand binding protein (SSB) was probed as a loading control for each blot. Polyclonal antibodies raised in sheep against recombinant Cas proteins were used as the primary antibodies, with a secondary donkey anti-goat fluorescently tagged antibody being used for detection.

Firstly the levels of the Cas1_{CD} (SSO1450) protein coded for by the adaptation cassette located between loci C and D in *S. solfataricus* were examined. Over the time points collected there was a very low signal obtained when probing for the Cas1_{CD} protein in control cell lysates. In contrast, for each of the time points in the infected conditions, there was a clear signal at the expected mass of the Cas1_{CD} protein (35 kDa). This indicates that Cas1_{CD} expression, and therefore potentially the expression of the other adaptation-related proteins (Cas2_{CD}, Cas4_{CD} and Csa1_{CD}), coded for by genes in the same operon, is induced strongly by infection.

Levels of the type I-A Cascade interference complex were examined by probing for subunits Cas5 and Cas7 (SSO1441 and SSO1442). The changes observed were subtler than for the Cas1_{CD} protein, with a fluorescent signal being present at each time point in the control samples, implying a degree of constitutive expression. There was, however, a clear upregulation in the Cas5-Cas7 proteins levels in the infected samples that increased between 8 and 9.5 dpi.

Antibodies against Cmr7 (SSO1986) were used to look for changes in the protein levels of the Cmr type III-B interference complex. Unlike the other Cas proteins examined, the Cmr complex seems to be expressed constitutively and not upregulated strongly in response to infection. The signal obtained when probing for the Cmr7 protein remained fairly constant at every time point and between control and infected samples. SSB protein (SSO2364) levels also remained stable in infected and control lysate samples over the time course, implying that differences seen were not due to poor sample preparation or loading.

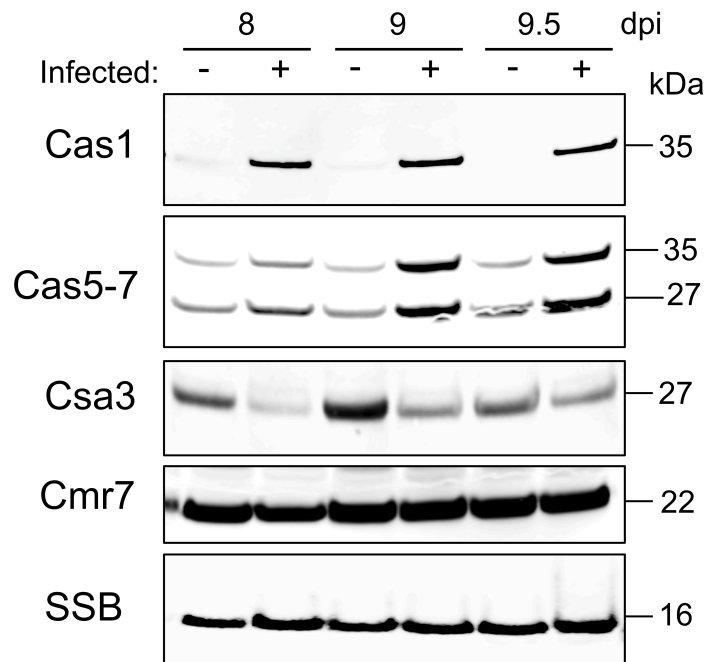


Figure 3.3 Changes in Cas protein levels in response to infection

Cell pellets (20 mg) of control or SMV1 + pMGB-infected *S. solfataricus* cultures were lysed and separated by SDS-PAGE before western blotting. The three sampled time points represent: the start of adaptation of the CRISPR C locus (8 dpi) and active adaptation of the array (9 dpi and 9.5 dpi). Primary polyclonal antibodies raised in sheep were used to probe for the *S. solfataricus* proteins: Cas1_{CD} (SSO1450), Cascade subunits Cas5 (SSO1441) (27 kDa) and Cas7 (SSO1442) (35 kDa), Csa3_{CD} (SSO1445), Cmr7 (SSO1986) and SSB (SSO2364). Secondary donkey anti-goat antibodies with a fluorescent tag were used to detect the protein signal.

Given the strong regulation of Cas proteins observed, a western blot was also carried out to examine how the levels of the putative transcriptional regulator Csa3_{CD} (SSO1445) changed during infection. Interestingly, in this case the Csa3 protein was found to be present in control samples throughout the infection time course. However, at 8 dpi the levels were much reduced in the infected samples compared to control. The level of Csa3_{CD} in infected samples increased back to close to control levels by 9.5 dpi.

These results indicated that in *S. solfataricus*, Cas proteins involved in adaptation and the Cascade complex are highly upregulated in response to infection while others, such as the type III-B Cmr complex, are constitutively expressed. Interestingly, the decreased levels of putative transcriptional regulator Csa3_{CD} during early infection coincided with a strong upregulation of Cas1_{CD} and Cascade proteins.

3.2.3 Change in *cas* gene transcript levels during infection

To investigate further, and attempt to quantify, the changes in expression levels of *cas* genes during infection, RT-qPCR was carried out. For this set of experiments control and infected samples from 9 dpi were analysed. This time point was chosen as adaptation was well established and on going. Total RNA was extracted using an RNeasy kit (Qiagen) from 30 mg of the infected and control cell pellets. Extracted RNA was then used in one-step RT-qPCR reactions (iScript One-Step RT-PCR Kit, Bio-Rad) to determine the change in transcript levels from various *cas* genes in response to infection (see section 2.2.6.2.2 (p.63) for method). The *ssb* gene (*ssb2364*) was chosen as an internal reference gene. Protein levels from this highly expressed gene were previously found to remain constant in infected and control samples, and protein and transcript levels have been reported to not to change significantly in response to DNA damage (Götz et al., 2007).

The RT-qPCR reactions performed relied on the properties of the fluorescent dye, SYBR green. This dye fluoresces when bound to a double-stranded DNA, produced during the PCR cycle. Therefore, by using sequence-specific primers, the production of a PCR product from a given transcript can be assessed in real-time by measuring the proportional increase in SYBR-green fluorescence. The iCycler (BioRad) used to perform the reactions calculates a crossing point (Ct) value for each reaction, which is the cycle number at which SYBR-green fluorescence increases past the background levels of the early cycles. Therefore, for transcripts that were in high abundance in the original sample, the threshold will be reached after few cycles, leading to a low Ct value. In contrast, for low-level transcripts, many cycles will be required to increase fluorescence above threshold, producing high Ct values.

3.2.4 Calculating primer efficiency

Sequence-specific primer pairs (SSBfor and SSBrev; 1450for and 1450rev; 1443for and 1443rev; 1986for and 1986rev; CRISPRCfor and CRISPRCrev; 1424for and 1424rev) were used to amplify a product from each transcript of interest (see Table 2.1 for sequences). Before use in the RT-qPCR reactions the efficiency of each primer pair was established to ensure that differences in product production are not due to the ability of the primers to anneal to the template and allow amplification. Each set of primers was tested in amplifications carried out with four 10-fold dilutions of genomic *S. solfataricus* DNA at 100, 10, 1 and 0.1 ng per reaction. The number of cycles taken for fluorescence to cross a threshold value (Ct) set by the instrument was then plotted against the log of initial template concentration, to produce a standard curve. From the standard curve, primer efficiency (E) was calculated using Equation 1 (Pfaffl, 2001). A gradient of -3.33 represents a 100% efficiency of amplification and efficiencies of between 90-110% are generally acceptable for use in RT-qPCR (Taylor et al., 2010). Primer efficiencies were found to be between 99.46 and 108.05% for the 6 primer sets used (Figure 3.4, A and B). Melt curves for each primer set also had one product peak (data not shown), with no off-target or large primer-dimer peaks. Given their apparent high efficiency and specificity, these primers were deemed acceptable for use in RT-qPCR experiments.

$$E = ((10^{(-1/\text{slope})} - 1) * 100)$$

Equation 1. Primer efficiency calculation from a standard curve

E = primer efficiency (%); slope = gradient of the standard curve produced from amplification from a dilution series of *S. solfataricus* genomic DNA

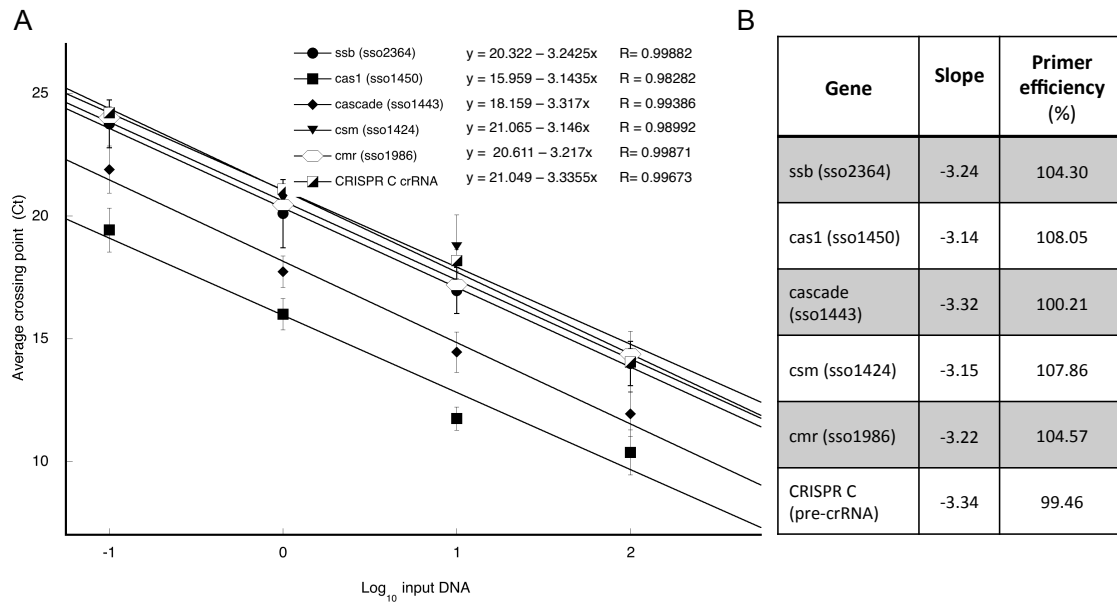


Figure 3.4 Efficiency of primer sets used in RT-qPCR

A. Standard curves of threshold values (Ct) plotted against the \log_{10} of a dilution series of genomic *S. solfataricus* DNA. Starting quantities of DNA were 100, 10, 1, 0.1 ng/reaction. Data points represent average Ct values from duplicate experiments \pm standard deviation (shown as error bars). The straight-line equation fitted for each primer pair is shown, with the R representing the correlation coefficient. The sequences of primer pairs are given in Table 2.1 (p.47). **B.** Table listing the primer efficiencies shown as percentages calculated from the slope of the standard curves using Equation 1.

3.2.5 Relative gene transcript levels from RT-qPCR

RT-qPCR reactions were carried out using RNA extracted from infected and control *S. solfataricus* cell pellets and mean Ct values for each gene of interest were calculated (see Appendix A: Triplicate Ct values from RT-qPCR (p.217)). The Pfaffl equation (Pfaffl, 2001), which takes into account the efficiency of primer pairs, was used to calculate the relative expression level of each gene between control and infected conditions (see Equation 2).

$$ratio = \frac{(E_{target})^{\Delta Ct_{target}(\text{control-sample})}}{(E_{ref})^{\Delta Ct_{ref}(\text{control-sample})}}$$

Equation 2. The Pfaffl equation for quantification of relative expression ratios from RT-qPCR. Target = the gene of interest; ref = reference gene, in this case *ssb*; E is the calculated primer efficiency; ΔCt is the difference in Ct value between control and infected samples.

Calculated fold changes in expression, normalised to the change in the reference gene (*ssb*) are shown in Figure 3.5 (A and B). A striking 12.9-fold increase in *cas1_{CD}* transcript levels was found between infected and control conditions. The

small subunit of *cascade*, *csa5* (*ssol443*), was also highly upregulated with a 12.8-fold increase in the infected sample. The Ct value obtained when amplifying from the *cas1_{CD}* transcript in control conditions was higher than for any other gene studied. This implies that this transcript was present at very low levels in the absence of infection. The results from western blotting and RT-qPCR correlate well, with Cas1_{CD} and Cascade protein and transcript levels being highly upregulated in infected samples, where adaptation is known to be on going, compared to the control.

Transcription of genes of the interference complexes Cmr and Csm was also enhanced during infection, but to a much lesser degree (2.81- and 2.5-fold, respectively) (Figure 3.5, A). In addition, the control Ct values for these genes were much lower than those for the *cas1_{CD}* gene (Figure 3.5, B). Therefore, it can be concluded that these interference complexes are expressed constitutively and only weakly upregulated during infection. The high and relatively constant Cmr protein levels observed by western blot in both infected and control conditions support this conclusion.

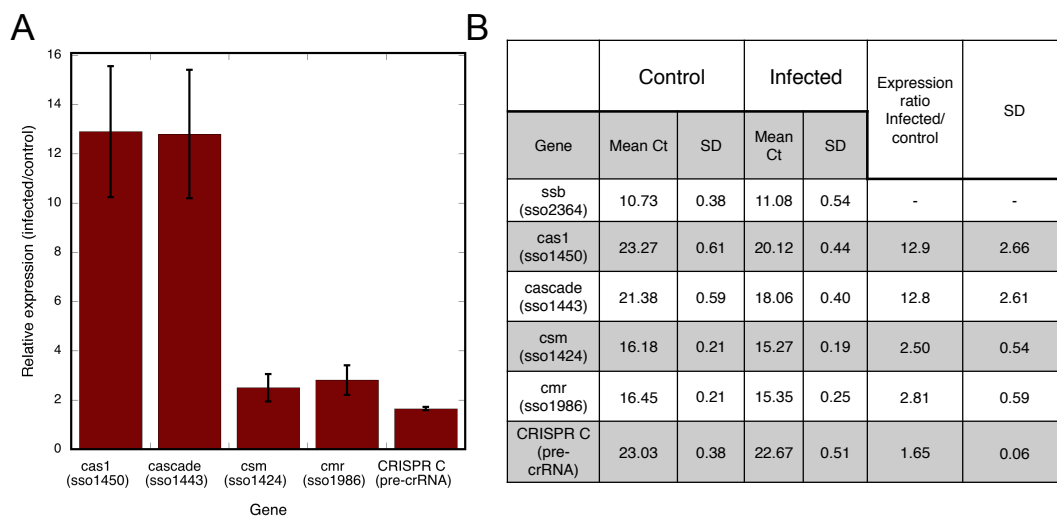


Figure 3.5 Fold changes in *cas* transcripts identified by RT-qPCR

A. Plot of the relative expression ratios of selected *cas* transcripts between infected and control *S. solfataricus* cultures. Triplicate Ct values for each sample were used to calculate the change in gene expression, taking into account primer efficiency and change in expression of a reference gene (*ssb*) using the Pfaffl equation (Equation 2) (Pfaffl, 2001). A value of 1 represents no change in expression. **B.** Table listing the mean Ct values for each transcript of interest \pm standard deviation under control and infected conditions. The ratio of expression between the two conditions \pm standard deviation as calculated by the Pfaffl method is also shown.

Finally, the levels of pre-crRNA in infected and control cultures were examined and the fold change was found to be 1.65, indicating that there was only a very slight increase in transcription of the CRISPR locus on infection. This result is not unexpected given the constitutive expression of pre-crRNA identified in the Sulfolobales (Lillestøl et al., 2009).

3.2.6 A putative transcriptional regulator

Given strong differences in protein and transcript levels observed during infection, work to try to identify transcriptional regulators responsible for this effect was a clear next step. Potential candidates were the Csa3 family proteins, as recently Csa3a from *S. islandicus* had been shown to be a transcriptional activator (Liu et al., 2015). Furthermore, the crystal structure of one of the two Csa3 proteins present in *S. solfataricus* (SSO1445) also suggested a role as a transcriptional regulator, given the presence of two wHTH DNA-binding domains and a putative small molecule regulatory site (Lintner et al., 2011a).

The *csa3_{CD}* (*sso1445*) gene is located between CRISPR array C and D and adjacent to the adaptation operon containing *cas1_{CD}*, *cas2_{CD}*, *csa1_{CD}* and *cas4_{CD}* genes (Figure 3.6, B). Previously Lintner and co-workers reported that the Csa3_{CD} protein did not bind putative promoters of the two adjacent CRISPR repeat-spacer arrays (Lintner et al., 2011a). Therefore, it was suggested that instead this Csa3 protein might be involved in the transcriptional regulation of one of the associated *cas* gene operons.

A Semi-palindromic regions upstream of *cas* operons in the Sulfolobales

Upstream of *csa5*
 SUL SOLFATARICUS P2 1443 TTTTTCATATTTATGAAAAAGAGTTTTCGTACACTAGAAATAGAAATGTTTATA TAGTGGAAAT
 SUL ISLANDICUS M 16 27 TTTTCCATATTCATGAAAAGTCTTTTCGTTTCAAGAAATAGAAAGTTTATATATTTGGGGT
 SUL ISLANDICUS M 14 25 TTTTCCATATTCATGAAAAGTCTTTTCGTTTCAAGAAATAGAAAGTTTATATATTTGGGGT
 SUL ISLANDICUS M 16 4 TTTTCCATATTCATGAAAAGTCTTTTCGTTTCAAGAAATAGAAAGTTTATATATTTGGGGT
 SUL ISLANDICUS Y G 57 14 TTTTCCATATTTATGAAAAGTCTTTTCGTTTCAAGAAATAGAAAGTTTATATATTTGGGGT
 SUL ISLANDICUS Y N 15 51 TTTTCCATATTTATGAAAAGTCTTTTCGTTTCAAGAAATAGAAAGTTTATATATTTGGGGT
 SUL ISLANDICUS L S 2 15 TTTTCCATATTTATGAAAAGTCTTTTCGTTTCAAGAAATAGAAAGTTTATATATTTGGGGT
 ***** * ***** * ***** * ***** * ***** * ***** * ***** * ***** *
Upstream of *csa1*
 SUL SOLFATARICUS P2 1451 CCCTAGTAAATTCGGGAATTCCTTTTACCCCCCTCC TAAAAACGGTTTTTA
 SUL ISLANDICUS M 14 25 CCCTAGTAAACTTGGGAATGCTTTTACCCACTCCTTAAACGGTTTTTA
 SUL ISLANDICUS M 16 27 CCCTAGTAAACTTGGGAATGCTTTTACCCACTCCTTAAACGGTTTTTA
 ***** * ***** * ***** * ***** * ***** * ***** * ***** * ***** *

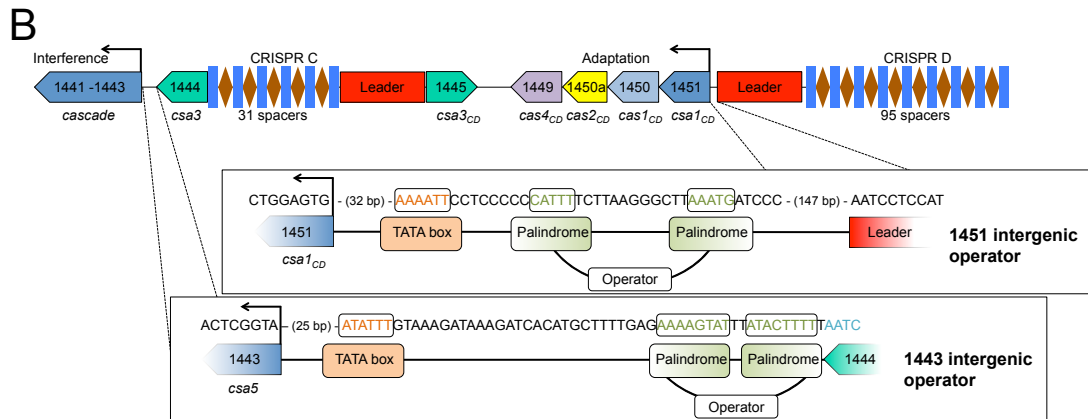


Figure 3.6 Two putative operator sequences in *S. solfataricus*

A. Alignment of palindromic operator sequences upstream of *cas* gene cassettes in the *S. solfataricus* and *S. islandicus*. The top alignment shows the sequences found upstream of *csa5* genes and the bottom those found upstream of *csa1* genes. Asterisks indicate identical residues. The TATA box is coloured orange and the palindromic residues green. **B.** Genomic location and structure of the identified putative operator regions in *S. solfataricus*. The 1451 operator is upstream of a *cas* cassette coding for proteins thought to be involved in adaptation and the 1443 operator is upstream of the *cascade* genes. The *csa3* genes shown in green are thought to code for proteins involved in transcriptional regulation. The Csa3 protein studied in this chapter is that coded for by the *csa3_{CD}* gene found between CRISPR array C and D.

Csa3a in *S. islandicus* was reported to enhance transcription by binding to palindromic operator sequences within *cas* gene promoters (Liu et al., 2015). Therefore, identifying similar sequences in *S. solfataricus* was key to understanding more about the Csa3_{CD} protein in this system. Two candidate operator sequences, containing weak palindromic regions and putative TATA boxes, were identified upstream of *cas* operons in *S. solfataricus*. The first of these sites was located upstream of the *csa5* (*sso1443*) gene of the *cascade* operon and will be referred to as the 1443 intergenic operator (shown in Figure 3.6, A). The second putative operator was identified upstream of the *csa1_{CD}* gene (*sso1451*) and will be referred to as the 1451 intergenic operator. These predicted operator regions aligned well with sequences upstream of the *csa5* or *csa1* genes in *S. islandicus* (Figure 3.6, A).

In order to assess whether the Csa3_{CD} protein interacted with any of the putative promoters identified, electrophoretic mobility shift assays (EMSAs) were carried out (see section 2.2.4.1 (p. 60) for method). Csa3_{CD} was expressed recombinantly in *E. coli* and purified to homogeneity before use in assays. The protein was found to bind strongly to a FAM-labelled double-stranded 1451 operator sequence (duplex made by annealing 1451 operator for and 1451 operator rev, see Table 2.1 (p.47) for sequences) (Figure 3.7, A), with a shifted, protein-bound substrate band visible even at the lowest concentration of protein (0.5 μM). This binding was highly sequence-specific, as Csa3_{CD} had no affinity for the 1443 operator sequence (made by annealing 1443 operator for and 1443 operator rev) at protein concentrations up to 1 μM (Figure 3.7, A). Interestingly, when the Csa3_{CD} protein was purified and the polyhistidine tag left uncleaved on the N-terminus of the protein, DNA binding was no longer observed. The N-terminal domain is known to be involved in dimer formation and contains a putative small molecule binding pocket or protein interaction site, while the C-terminal domain is thought to be directly involved in DNA binding (Lintner et al., 2011a). These results suggest that the polyhistidine tag may distort the dimer structure and prevent DNA docking. Lintner and colleagues also failed to successfully model the docking of duplex DNA *in silico* for tagged Csa3_{CD}. However, they predicted that the crystallized conformation of Csa3_{CD} was not favourable for DNA binding or that the protein interacts with bent or unwound DNA (Lintner et al., 2011a).

Fluorescence anisotropy was used to assess the binding of Csa3_{CD} to the FAM-labelled 1451 operator DNA in a more quantitative manner (see section 2.2.4.2 (p. 61) for method). Anisotropy is a measure of the change in the ratio of polarized:depolarized light detected after exciting a sample with a polarized beam. The speed of tumbling of the labelled DNA in solution determines this ratio. Free DNA tumbles fast in solution, leading to more depolarized emissions, while protein-bound DNA tumbles more slowly, resulting in a higher emission of polarized light.

$$A = A_{\min} + [(D + E + K_D) - \{(D + E + K_D)^2 - (4DE)\}^{1/2}] (A_{\max} - A_{\min}) / (2D)$$

Equation 3. Binding isotherm assuming 1:1 binding of protein:nucleic acid

A=anisotropy; E=total [protein]; D=total [DNA]; A_{min}=anisotropy of free DNA; A_{max}=maximum anisotropy of DNA-protein complex; K_D=disassociation constant (Reid et al., 2001).

Triplicate titrations with increasing concentrations of Csa3_{CD} were carried out with single- or double-stranded 1451 operator DNA and mean anisotropy values fitted to

a simple binding isotherm which assumes 1:1 protein to DNA binding (Equation 3) (Reid et al., 2001).

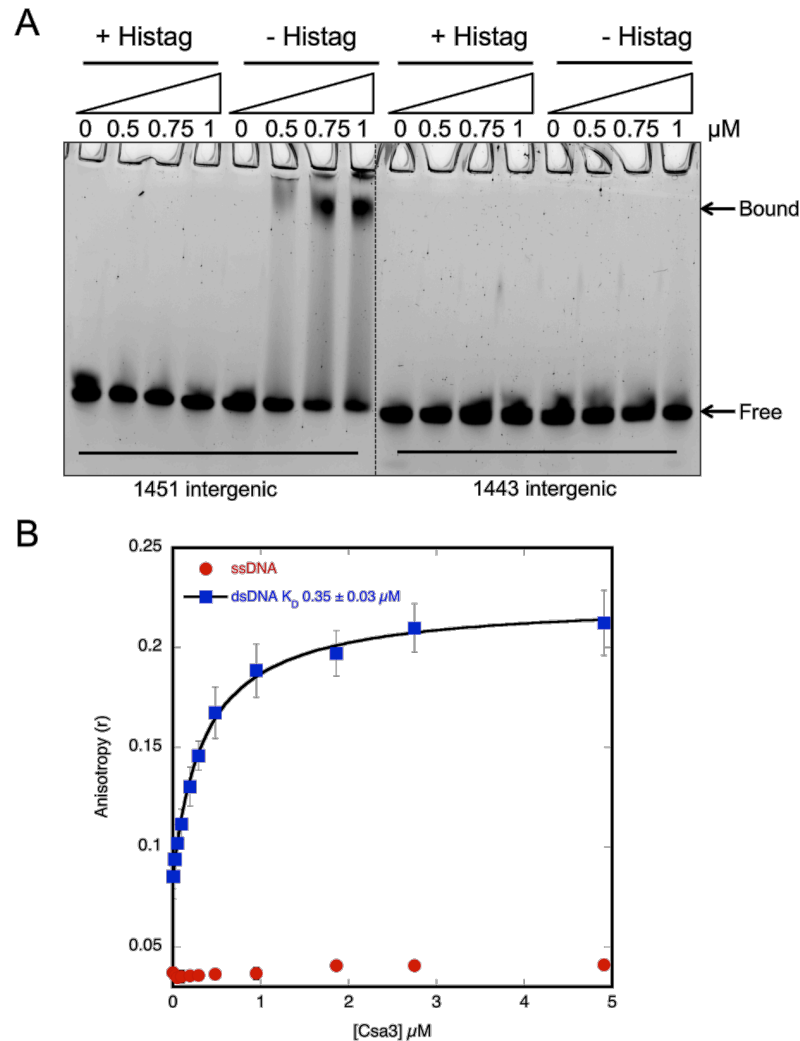


Figure 3.7 Csa3_{CD} binds the semi-palindromic 1451 operator sequence

A. EMSA assay of Csa3_{CD} binding to the 1451 (left) or 1443 (right) dsDNA FAM-labelled operators (200 nM) (duplexes made by annealing oligonucleotides: sso1451 operator for and sso1451 operator rev; sso1443 operator for and sso1443 operator rev, sequences given in Table 2.1 (p.47)). Binding was assessed over a protein gradient from 0 to 1 μM Csa3_{CD}. The first 4 lanes for each substrate show a gradient of N-terminal polyhistidine-tagged Csa3_{CD} protein and the subsequent 4 lanes are the same gradient with untagged protein. Protein and substrate were incubated together at room temperature in binding buffer for 20 min before separation on a 12% native polyacrylamide-TBE gel. **B.** Fluorescence anisotropy titration comparing Csa3_{CD} binding to double- (blue squares) or single-stranded (red dots) 1451 operator DNA (oligonucleotides used: sso1451 operator for, or a duplex made of sso1451 operator for and sso1451 operator rev, see Table 2.1 for sequences). Titrations were carried out at room temperature in binding buffer. Mean anisotropy values from triplicate titrations were fitted to a binding isotherm (Equation 3) ± SD using Kaleidagraph (Synergy Software).

Following curve fitting a K_D of ~350 nM was calculated for Csa3_{CD} binding to the double-stranded (ds) 1451 operator sequence (Figure 3.7, B). Titration of Csa3_{CD}

into single-stranded (ss) operator DNA led to a very small increase in anisotropy, indicating no, or very low-affinity, protein binding to the ss1451 operator sequence.

3.2.7 A DNaseI footprint of Csa3_{CD} binding region

In order to identify the precise binding site of the Csa3_{CD} protein on the operator sequence, DNaseI footprinting was carried out (see section 2.2.4.3 (p.61) for method). This technique relies on sequence-specific binding proteins protecting a binding site from digestion by DNaseI nuclease (Galas & Schmitz 1978). Subsequent separation of the assay components by electrophoresis allows the identification of the protein-binding site as a gap visible in the ladder of cleavage products.

Increasing concentrations of Csa3_{CD} was incubated with the FAM-labelled double-stranded 1451 operator DNA (duplex made of sso1451 operator for and sso1451 operator rev oligonucleotides, sequences given in Table 2.1 (p47)) and DNaseI nuclease before separation by denaturing gel electrophoresis. As the concentration of Csa3_{CD} was increased from 0 to 3 μ M, regions of operator DNA became protected from DNaseI digestion. The nuclease products visible between residues A7-G4 and G2-A2 in the absence of Csa3_{CD} become weaker as protein concentration is increased (Figure 3.8). The protected region is centred around the palindromic sequences making up the putative operator region, consistent with the hypothesis of that each wing of the Csa3_{CD} dimer may bind one arm of the palindrome (Lintner et al., 2011a). At the highest concentration of Csa3_{CD} (10 μ M) a wider region of protection is visible, with only the 3' end of the substrate being digested by DNaseI in this condition. This indicates that at concentrations greatly exceeding the K_D , Csa3_{CD} may bind non-specifically or co-operatively along the DNA leading to protection outwith the consensus binding site.

3.2.8 Searching for a small molecule ligand for Csa3_{CD}

The crystal structure of Csa3_{CD} revealed a potential small-molecule-binding pocket at the dimer interface of the N-termini. A component of the crystallisation buffer, polyethylene glycol (PEG), was identified in this pocket in the crystal structure. However, no small molecules that could be responsible for modulating the activity of Csa3_{CD} *in vivo* were co-crystallised (Lintner et al., 2011a). The lack of small

molecule binding by Csa3_{CD} could be due to the absence of the native ligand in the *E. coli* host used for expression. In addition, if the binding pocket is involved in modulating protein activity the small molecule signal may only be produced transiently, perhaps in response to infection.

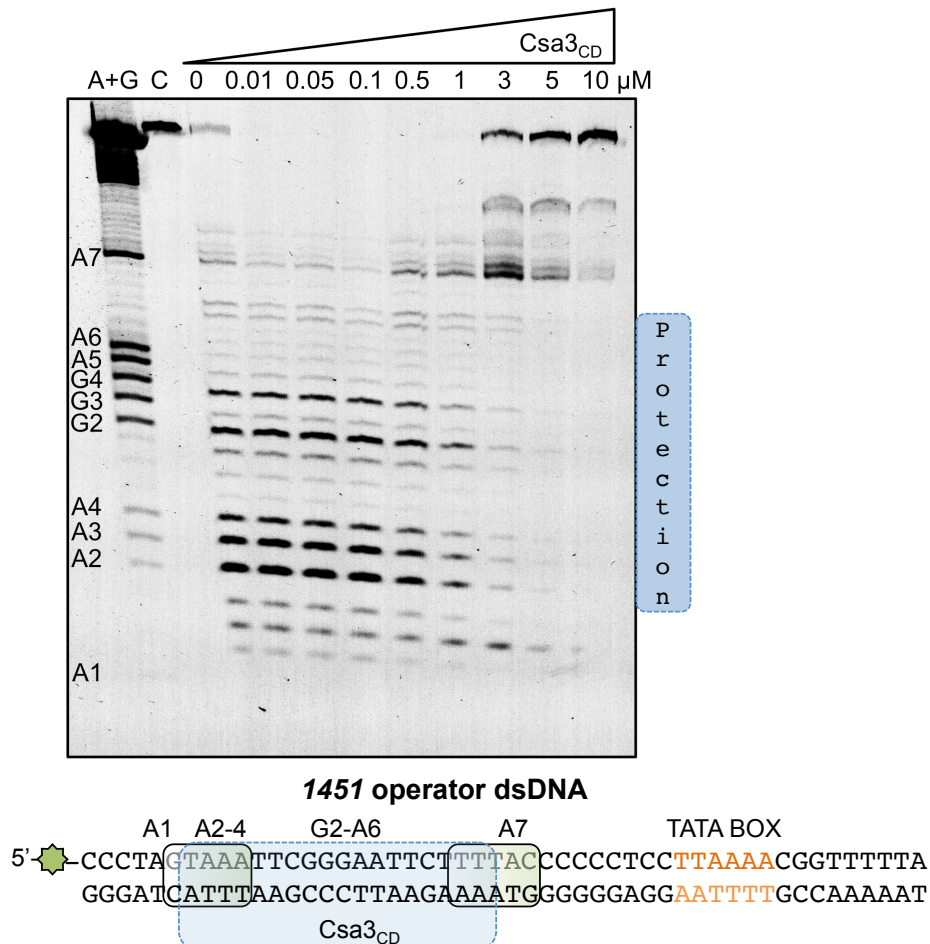


Figure 3.8 DNaseI footprint analysis of Csa3_{CD} binding to operator DNA

DNaseI digest of 5' FAM-labelled 1451 operator dsDNA (duplex made of sso1451 operator for and sso1451 operator rev oligonucleotides, see Table 2.1 (p.47) for sequences) in the presence of increasing concentrations of Csa3_{CD} (0 to 10 μ M). The first lane of the gel is a Maxam-Gilbert A+G ladder to allow mapping of the nuclease products. Lane 2 is a control without DNaseI or Csa3_{CD}, followed by a gradient of increasing Csa3_{CD} concentrations with a constant amount of DNaseI (1 μ g). Following incubation at 37 °C for 10 min products were separated on a denaturing 20% polyacrylamide-TBE gel. The operator sequence is shown at the bottom, with palindromic regions highlighted in green and TATA box in orange. The region protected from DNaseI digest is shown in a blue dashed box.

To try to identify a potential ligand for Csa3_{CD}, various small molecules were added to EMSAs to look for any changes they might induce in the binding affinity of Csa3_{CD} for the operator sequence (see section 2.2.4.1.1 (p.60) for method). The small molecules ADP, ATP and cAMP did lower the binding affinity of Csa3_{CD} for the operator marginally (Figure 3.9, A). This is interesting as cAMP has been shown to be a signalling molecule during viral infection, which in combination with the cAMP-

receptor protein (CRP) led to the de-repression of the CRISPR system in *T. thermophilus* (Shinkai et al., 2007).

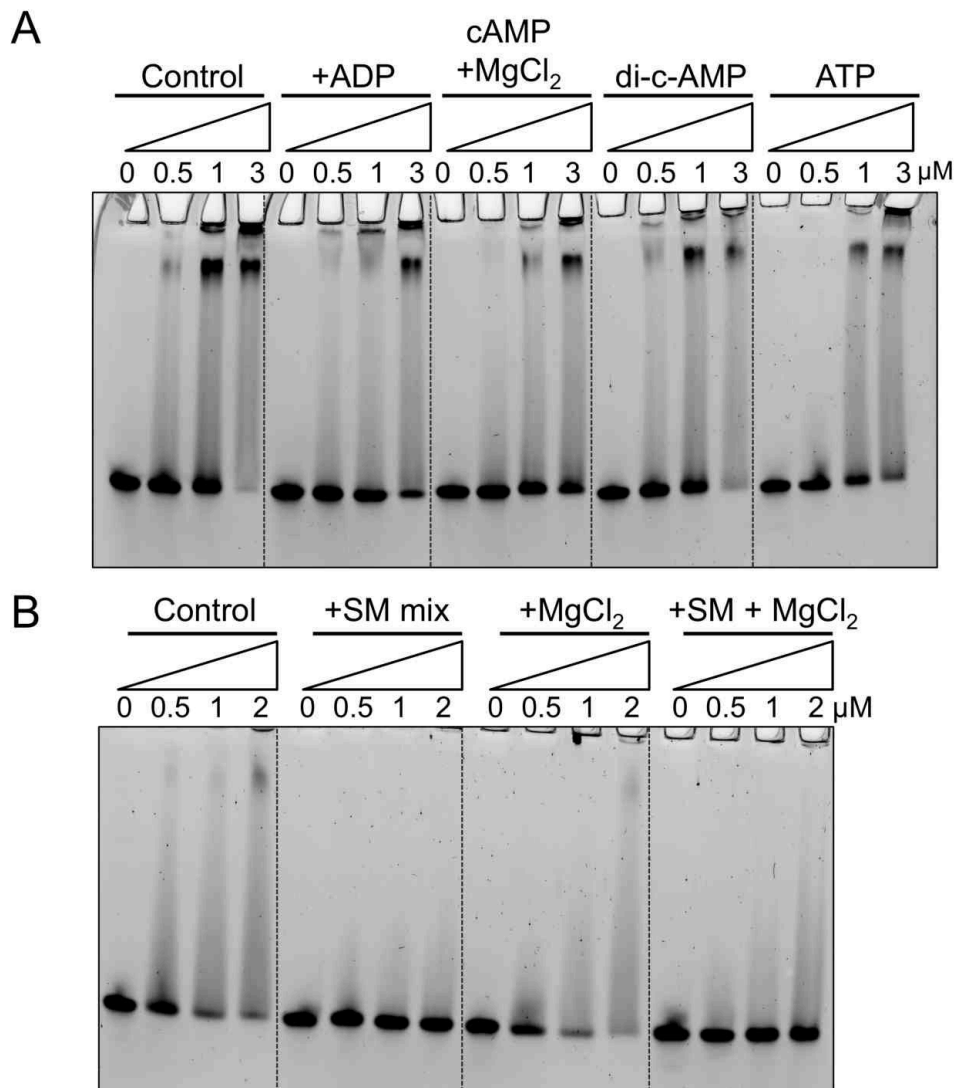


Figure 3.9 Csa3_{CD} binding in the presence of potential regulatory ligands

A. EMSA to look for changes in operator binding by increasing concentrations of Csa3_{CD} from 0 to 3 μ M, in the presence of the indicated small molecule (10 mM). Protein and small molecules were incubated at room temperature for 20 min, before the addition of 200 nM double-stranded *sso1451* operator DNA (duplex made of *sso1451* operator for and *sso1451* operator rev oligonucleotides, sequences given in Table 2.1 (p.47)) and a further 20 min incubation. Reactions were separated on a native 12% polyacrylamide gel. The first condition is a control in the absence of a small molecule ligand, followed by addition of adenosine diphosphate (ADP), cyclic adenosine monophosphate (cAMP) and magnesium chloride (MgCl₂), cyclic di-AMP (c-di-AMP) and adenosine triphosphate (ATP). **B.** EMSA showing binding of increasing concentrations of Csa3_{CD} from 0-2 μ M on *1451* operator dsDNA (same substrate as used in **A**) (200 nM) in the presence of a small molecule extract. The first condition is a control without the addition of a small molecule to the binding reaction. The second condition is Csa3_{CD} binding operator DNA following incubation with a small molecule extract (SM) obtained from size exclusion of *S. solfataricus* lysate components above 3 kDa. The following condition has MgCl₂ (10 mM) added and the final condition is after the addition of MgCl₂ (10 mM) and the small molecule extract. Other

conditions were as in **A**.

The most significant change in binding of Csa3_{CD} to operator DNA was brought about by the addition of a small molecule extract from *S. solfataricus*. This extract was prepared from *S. solfataricus* cell lysate, by repeated filtration through spin concentrators with decreasing molecular weight cut-offs (from 30 kDa to 3 kDa). Following collection of the filtrate from a 3 kDa cut-off concentrator, the small molecule extract was added to the EMSA assays with or without magnesium chloride (Figure 3.9, B). While magnesium chloride did not affect the binding of operator DNA, addition of the small molecule extract reduced binding considerably, with only a weak bandshift of the ds1451 operator observed at even the highest Csa3_{CD} concentration (2 μ M). The reduced binding on the addition of filtered lysate may be due to a small-molecule- or protein-mediated modulation of Csa3_{CD}. However, these results are very preliminary and other components of the lysate, such as salt or pH may also be responsible for the change in binding observed.

3.2.9 *In vitro* transcription

Although the binding of Csa3_{CD} to operator DNA had been established, whether this region was in fact a promoter for *cas* gene transcription, and the influence of Csa3_{CD} on this transcription still remained to be investigated. An *in vitro* system for transcription using *S. solfataricus* host proteins has been widely used to look at the influence of regulatory proteins on transcript production (Bell & Jackson, 2001). In order to set up this system to investigate the effects of Csa3_{CD}, the minimum host protein components required for active transcription were purified. TBP and TFB were both expressed recombinantly in *E. coli* and purified via nickel affinity chromatography, before tag cleavage and gel filtration. RNA polymerase (RNAP) was purified from native *S. solfataricus* cell lysate as described previously (Paytubi & White, 2009).

Primers were designed (sso1451promf and sso1451promr, see Table 2.1 (p.47) for sequences) to amplify the intergenic sequence bound by Csa3_{CD} and the start of the upstream *csa1_{CD}* (*sso1451*) gene from *S. solfataricus* P2 genomic DNA. The PCR product produced was then Topo cloned into the pET151/D-TOPO vector (ThermoFisher Scientific) to create the p1451prom plasmid (see Table 2.3 (p.52)). This construct was then digested with *SacI* to produce a suitable substrate for run-

off transcription and primer extension. A template for transcription containing the strong T6 promoter of the *Sulfolobus shibatae* virus (SSV1) (pT6), which had been used previously (Paytubi & White, 2009), was also prepared and linearised with *Xho*I, for use as a sequence non-specific control. The expected cDNA product following run-off transcription and reverse transcription primer extension from a 32 P-labelled DNA primer was 116 nt from the p1451prom substrate and 67 nt from the pT6 template (Figure 3.10, A).

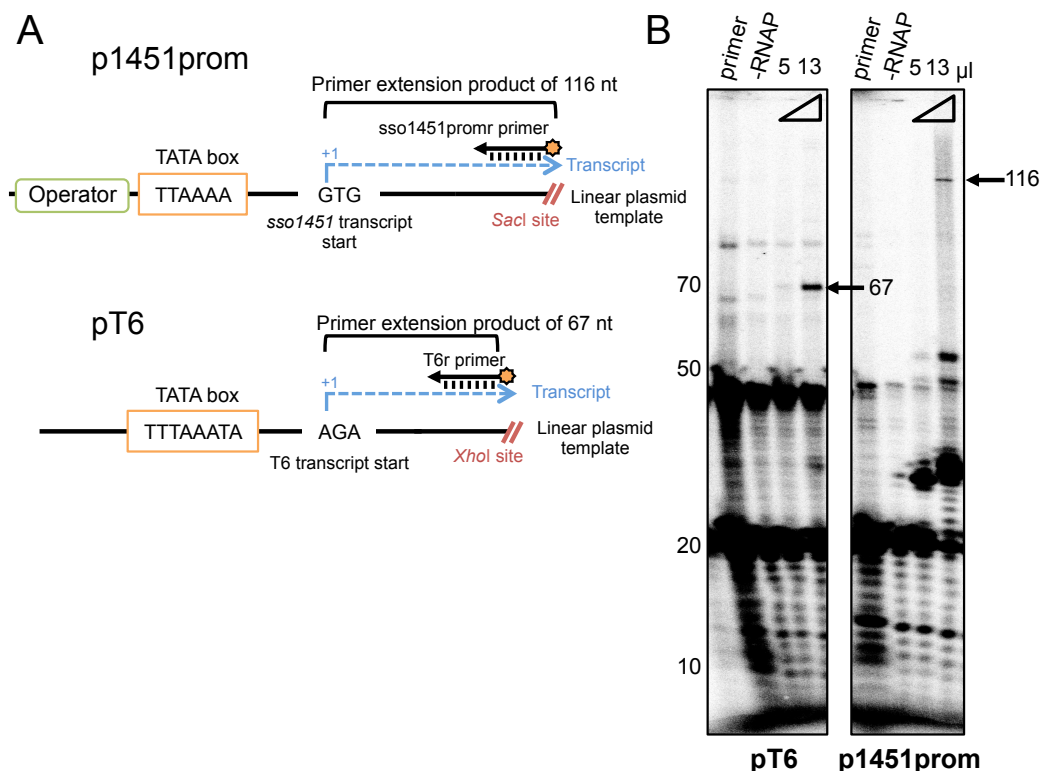


Figure 3.10 *In vitro* transcription from the 1451 promoter

A. Schematic showing the constructs used for *in vitro* transcription and primer extension. The top schematic shows the 1451 promoter region cloned into the pET151/D-TOPO vector to produce the plasmid p1451prom (see table Table 2.3 for details). The insert contains the palindromic operator sequence (green box), TATA box (orange box) and the start of the *csa1* (*sso1451*) gene. The p1451prom was linearised at a *Sac*I site (red dashes) in the *sso1451* gene to make it suitable for run-off transcription. Transcription is expected to start at a GTG sequence of the *csa1_{CD}* gene. A 5'- 32 P-labelled reverse primer (*sso1451promr*, see Table 2.1 (p.47) for sequence) complementary to 1451 transcript was used for primer extension, with an extension product of 116 nt expected. The bottom scheme shows the construct used for transcription and primer extension from the *S. shibatae* viral T6 promoter (also used in Paytubi & White (2009)). Transcription from this promoter starts at the indicated AGA sequence with a primer extension product of 67 nt expected (primer T6r used, see Table 2.1). *In vitro* transcription was carried out as described previously by Paytubi & White (2009). **B.** The products of *in vitro* transcription from linearised templates and primer extension were separated on a denaturing 12% polyacrylamide gel and phosphorimaged. The first lane of each gel is the primer used for the primer extension alone. The second lane is a further control showing the products of primer extension from 13 μ l of a transcription assay carried out with TBP and TFB, but without RNAP. The two subsequent lanes show

the result of primer extension from either 5 or 13 μ l of an *in vitro* transcription assay carried out with all essential transcription proteins (TBP (40 nM), TFB (40 nM) and RNAP (80 nM)). DNA lengths are shown on the left of the gel and full-length products of reverse transcription are indicated by arrows.

As shown in Figure 3.10 (B) a strong product of \sim 70 nt was produced following an *in vitro* transcription and primer extension from the pT6 plasmid (see section 2.2.5 (p.61) for method). This product was only formed when RNAP was added to the reaction and increased dependent on the volume of transcription reaction (5 or 13 μ l) added to the primer extension step. Following transcription from p1451prom, a band at \sim 116 nucleotides was visible, which corresponded to the expected size of a reverse transcript from this substrate (Figure 3.10, B). The 1451 product band was much fainter than the T6 product signal, which indicated that while transcription was initiated from this putative promoter region, the promoter strength was weak.

The laddering observed below the full-length product both gels of Figure 3.10 (B) is partly due to the primer not being gel-purified before use, leading to the 32 P-labelling of some of partial-synthesis or breakdown products of the ssDNA primer. In addition, in lanes where transcription was active some products of stalled or incomplete reverse-transcription are visible in addition to the primer background.

3.2.10 Effect of Csa3_{CD} on transcription *in vitro*

To assess the effect of Csa3_{CD} on transcription from p1451prom, *in vitro* transcription was carried out in the presence of increasing concentrations of the Csa3_{CD} (from 0 to 5 μ M). From Figure 3.11 (A), it is apparent that there was no clear change in the amount of the 67 nt primer-extension product produced from the pT6 or in the amount of the 116 nt product from the p1451prom substrate, even at the highest concentration of Csa3_{CD}.

Given the weak nature of the 1451 promoter, small differences in the amount of product produced were not obvious. To improve the efficiency of transcription a promoter sequence was designed with the palindromic operator sequence bound by Csa3_{CD} located immediately upstream of the minimum BRE and TATA motifs of the strong T6 promoter (Figure 3.11, B). A gBlock insert was ordered (Integrated DNA Technologies) and this chimeric promoter was cloned, using *Bam*HI and *Xho*I restriction sites, into a pBlueScript SK+ vector (Agilent), to make the pChi1451-T6 plasmid (see Table 2.3 for sequence details) for use *in vitro* transcription.

Following run-off transcription and primer extension a strong transcription product of 67 nt was produced from this hybrid promoter (Figure 3.11, C). However, the addition of Csa3_{CD} at concentrations of up to 5 μ M did not affect the amount of product produced by *in vitro* transcription (Figure 3.11, C). If the Csa3_{CD} protein is indeed a transcriptional regulator it may require additional co-factors, such as protein binding partners or a small molecule ligand, to be fully functional.

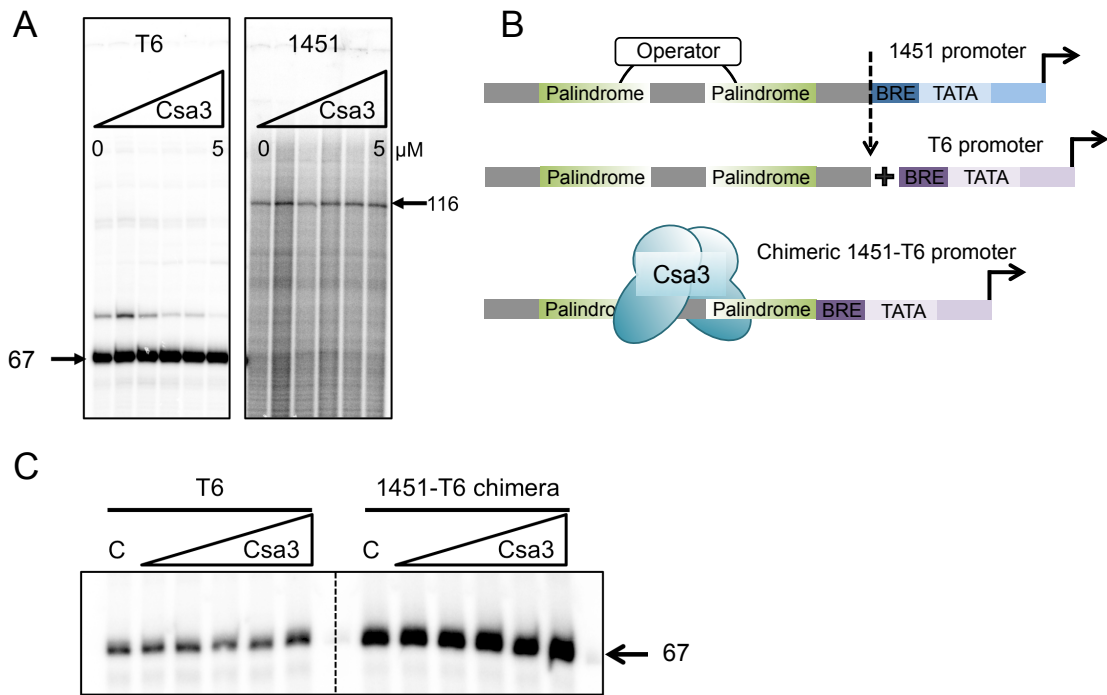


Figure 3.11 The effect of Csa3_{CD} on transcription efficiency

A. The results of *in vitro* transcription in the presence of increasing concentrations of Csa3_{CD}. The gel on the left shows the products of transcription and primer extension (from radiolabelled primers sso1451promr and T6r, see Table 2.1 (p.47) for sequences) from the linearised pT6 (50 ng) and the gel on the right shows those from p1451prom (100 ng) (see Table 2.3 (p.52) for details of these plasmids). The first lane for each substrate is a control reaction without the addition of Csa3_{CD}. The following lanes show the products of primer extension following transcription in the presence of 0.1, 0.5, 1, 3 and 5 μ M Csa3_{CD}. Full-length reverse transcripts are indicated by arrows. **B.** Schematic of the chimeric promoter construct (see Table 2.3 for details) made by replacing the BRE and TATA motif of the weak 1451 promoter with the minimal BRE and TATA and downstream initiator motifs of the strong T6 promoter. This 1451-T6 chimeric substrate was then cloned into the pBlueScript SK+ vector to form the pChi1451-T6 plasmid (see Table 2.3) and the vector linearised with *Xho*I before use in transcription. **C.** Primer extension products (T6r primer) following *in vitro* transcription from pT6 (left) or chimeric pChi1451-T6 (right) in the presence of 0 (C), 0.1, 0.5, 1, 3 and 5 μ M Csa3_{CD}.

In summary, I have shown that Csa3_{CD} binds specifically and strongly ($K_D=360$ nM) to a palindromic operator sequence upstream of a putative promoter of the adaptation cassette located between CRISPR loci C and D of *S. solfataricus*. In addition, I confirmed that transcription could be initiated from this promoter region *in*

vitro, although the promoter strength was weak. Under the conditions assayed, no obvious modulation of transcription from the 1451 promoter or the hybrid 1451-T6 promoter was observed *in vitro*.

3.3 Discussion

3.3.1 CRISPR-Cas upregulation in response to infection

While much attention has been paid to the silencing of the CRISPR-Cas system in *E. coli* by the global repressor H-NS, few studies have focused on regulation of CRISPR elements in archaea. One key report described the activation of *cas* gene transcription following infection of *S. islandicus* LAL14/1 cells with SIRV2 (Quax *et al.*, 2013). Furthermore, a proteomics study found that the abundance of a number of Cas proteins changed after STIV infection, providing the first evidence that the CRISPR-Cas system may be regulated in *S. solfataricus* (Maaty *et al.*, 2012). This chapter aimed to expand our knowledge of CRISPR-Cas regulation in *S. solfataricus* by investigating the effect of infection on the expression of Cas proteins.

3.3.2 Strong upregulation of Cas1 in response to infection

A clear upregulation of Cas protein expression in infected *S. solfataricus* cultures was observed by western blotting and RT-qPCR (Figure 3.3 and Figure 3.5), with the Cas1 protein required for adaptation being the most highly induced protein studied. Constitutively expressed Cas1 was almost undetectable by western blotting of control samples, whereas levels were increased dramatically during infection, with a strong protein signal detected at each time point investigated. A similar upregulation was apparent at the mRNA level as *cas1* transcripts were found to be increased by 12.9-fold in infected cultures compared to control samples. The type I-A Cascade interference complex was also strongly upregulated during infection, with mRNA levels increased by 12.8-fold.

These results show that, in common with the *E. coli* system, some CRISPR-Cas elements in *S. solfataricus* are or strongly repressed, or not induced, in the absence of invading genetic elements. The level of upregulation observed was similar to the 3-10-fold increase in *cas* transcripts observed during *S. islandicus* infection (Quax *et al.*, 2013). There are several potential reasons for the constitutive silencing of Cas

proteins. Firstly, several Cas proteins, such as Cas1 and Cas4, have been identified as nucleases with low sequence specificity (Wiedenheft et al., 2009; Zhang et al., 2012a); therefore, high constitutive levels of these proteins may compromise host DNA integrity. Secondly, the Cas1 and Cas2 proteins are essential for the addition of new spacers to the CRISPR array. If they are expressed in the absence of invading DNA this may increase the likelihood of erroneous insertion of self-spacers, leading to autoimmunity. Finally, CRISPR-Cas has been shown to block horizontal gene transfer (HGT) (Marraffini & Sontheimer, 2008). Therefore, another conceivable advantage of a silenced CRISPR-Cas system is the acquisition of fitness-enhancing genes by HGT.

3.3.3 Differential regulation of CRISPR-Cas components

Archaeal CRISPR systems are often more complex than those of bacteria, with several subtypes and CRISPR arrays frequently found in one host. Multiple studies have shown that the response of CRISPR subtypes can vary greatly, depending on the infecting virus and environmental conditions. León-Sobrino and colleagues showed that during *S. islandicus* infection by STSV2 the expression of adaptation genes and the type-III *cmr-β* interference cassettes were strongly induced. However, in the same study, transcripts of the *cmr-α* interference cassette remained unchanged or decreased throughout the infection time course (León-Sobrino et al., 2016). This chapter has also demonstrated a clear difference in the response of CRISPR-Cas elements to infection in *S. solfataricus*. While Cas1 and Cascade levels were found to increase markedly compared to control levels, expression of the Cmr and Csm interference complex subunits and pre-crRNA transcripts were only weakly upregulated (between 1.65- and 2.81-fold) during infection. These results echo those of a similar study in *S. islandicus*, which found that while the majority of *cas* operons were expressed very weakly under control conditions, the type I-A and III-Ba interference complexes were strongly expressed constitutively. The authors suggest that these complexes act as surveillance systems, providing the first line of defence in case of invasion (Quax et al., 2013).

In conclusion, there seems to be a level of fine-tuning of the archaeal CRISPR-Cas response to infection, which is less apparent in the simpler system of *E. coli*. The results presented here provide evidence that *S. solfataricus* has low constitutive levels of the adaptation-related proteins, which are then strongly upregulated in

response to infection, while interference complexes are expressed constitutively, perhaps fulfilling a role in the early detection of invading genetic elements.

3.3.4 A Csa3 binding site upstream of the adaptation genes

Differential expression of elements of the CRISPR-Cas system implies that a diverse interplay of regulatory elements exists in *S. solfataricus*. The crystal structure of the Csa3_{CD} protein from *S. solfataricus* revealed conserved features similar to those of the MarR transcription factors, and led to this protein family being identified as putative transcriptional regulators of archaeal CRISPR-Cas systems (Lintner et al., 2011a).

In this chapter I identified a putative promoter sequence upstream of the adaptation cassette located between CRISPR C and D of *S. solfataricus* and I confirmed that the Csa3_{CD} protein binds specifically to this motif with a K_D of ~350 nM. From DNaseI footprint analysis the binding of Csa3_{CD} was found to be centred on a weak palindromic operator sequence upstream of the TATA box of the *sso1451* promoter. Csa3_{CD} protected both halves of the palindromic sequence from DNaseI digestion, suggesting that each wHTH motif of the Csa3_{CD} dimer may interact with one half of the palindrome.

3.3.5 Reduced Csa3_{CD} during early infection

The strong and specific binding of Csa3_{CD} to the putative promoter of adaptation-related *cas* genes supports the hypothesis that this Csa3 protein may play a role in transcriptional regulation. A recent study has shown that the Csa3a protein from *S. islandicus* is a transcriptional activator (Liu et al., 2015). The authors reported that no Csa3a was present in control cell lysates and its production was activated by infection, which then enhanced the expression of other *cas* genes. In contrast, in this work the Csa3_{CD} protein was detected in control lysate and seemed to decrease during early infection, while the level of Cas1_{CD} protein was strongly upregulated. From these data it is tempting to hypothesise that in the *S. solfataricus* system Csa3_{CD} may act to repress transcription of the adaptation cassette, including *cas1_{CD}*, by binding to the identified *sso1451* operator sequence. The binding of both halves of the palindromic operator could conceivably lead to DNA-looping, known to

influence transcription in other prokaryote systems (Cournac & Plumbridge, 2013), and block the assembly of transcription machinery.

3.3.6 Csa3_{CD} does not affect transcription *in vitro*

To test this hypothesis, *in vitro* transcription reactions were carried out from the 1451 promoter region identified. The data presented here confirm that transcription can be initiated from this putative promoter region (Figure 3.10). However, the addition of purified Csa3_{CD} protein, which strongly bound this sequence, to transcription assays had no effect on the levels of transcript produced (Figure 3.11). Therefore, although transcription of the adaptation genes looks likely to be initiated from the promoter flanked by the identified Csa3_{CD}-bound operator region, the Csa3_{CD} protein could not be shown to either enhance or repress transcription under the conditions used in this study.

The N-terminal domain of Csa3_{CD} was identified as a small molecule binding pocket from the crystal structure (Lintner et al., 2011a). It is conceivable that the lack of transcriptional regulation by Csa3_{CD} observed here could be due to the absence of a regulatory ligand. The *E. coli* host used to recombinantly express the protein may lack small molecule ligands present in *S. solfataricus*, or the ligand may be a secondary messenger, produced uniquely during *S. solfataricus* infection. The importance of the N-terminal domain conformation of Csa3_{CD} was revealed in this chapter as the presence of a 5' polyhistidine tag abolished binding, although the N-terminal is not thought to directly interact with DNA. Therefore, the binding of a regulatory ligand in the N-terminal binding pocket could potentially induce a conformational change in Csa3_{CD} leading to the modulation of transcription from the nearby 1451 promoter.

Chapter 4: Characterisation of Cas1 and Cas2 from *S. solfataricus*

4.1 Introduction

Genes coding for Cas1 and Cas2 proteins are universally present in functional CRISPR-Cas systems and have therefore become known as the hallmark of these systems (Makarova et al., 2006). Early work into CRISPR-Cas structure and function showed that Cas1 and Cas2 were not required for the processing of CRISPR transcripts or for the interference response (Brouns et al., 2008). Therefore, it was hypothesized that these proteins played a key role in the integration of spacers into the CRISPR array. This hypothesis was further strengthened when Yosef and colleagues showed that in *E. coli* the minimum genetic requirement for the addition of a new spacer comprised the CRISPR leader sequence and genes coding for Cas1 and Cas2 (Yosef et al., 2012).

When the work contained in this chapter was undertaken, very little was known about the biochemical activity of the Cas1 and Cas2 proteins. However, the field has advanced enormously in recent years. In this introduction, a short summary of the initial studies of these proteins will be given, with the discussion providing the latest developments and how the work presented in this chapter fits with what we now know about the structure and function of Cas1 and Cas2.

4.1.1 Structure of the Cas1 protein

The sequences of Cas1 proteins show little conservation outside of an active site domain. Despite this, the structures of Cas1 proteins solved thus far are similar, suggesting that structural conservation does exist. Cas1 is a homodimeric protein with an overall topology resembling that of a butterfly with spread wings (Wiedenheft et al., 2009). Each Cas1 monomer has two distinct domains: an α -helical C-terminal domain, which forms the upper wing of the butterfly and a N-terminal domain rich in β -sheets making up the lower wing lobe. The two domains are joined by a flexible linker, which was suggested to allow independent movement of the domains relative to each other (Wiedenheft et al., 2009).

The C-terminal helical bundle contains the conserved metal-binding active site of the enzyme and a positively charged cleft thought to bind DNA and orient it towards the active site (Kim et al., 2013). The crystal structure of the *P. aeruginosa* Cas1 protein showed that three conserved active site residues (Glu190, His252, Asp286) co-ordinate a manganese ion (Wiedenheft et al., 2009).

4.1.2 Biochemical activity of Cas1

The characterisation of the *P. aeruginosa* Cas1 revealed a metal-dependent nuclease activity on single- and double-stranded DNA (Wiedenheft et al., 2009). The authors also showed that this Cas1 processed double-stranded DNA into ~80 bp fragments and suggested that these products may represent protospacer precursors.

The *E. coli* Cas1 protein was found to cut ssDNA and branched structures, including Holliday junctions (Babu et al., 2011). Additionally, this protein was also found to interact with components of the *E. coli* DNA repair and recombination machinery, implying that there may be a link between DNA repair and adaptation (Babu et al., 2011).

In contrast to the activities published for the *P. aeruginosa* and *E. coli* proteins, the SSO1450 (Cas1_{CD}) Cas1 protein from *S. solfataricus* had been reported to be devoid of nuclease activity against single- and double-stranded DNA and RNA (Han et al., 2009). The authors did note high-affinity nucleic acid binding by this Cas1 protein and an apparent ability to promote the reannealing of complementary DNA strands after duplex melting (Han et al., 2009). The divergent activities of these three Cas1 proteins seemed to suggest that they performed different roles in CRISPR adaptation, which given the low sequence identity (18 – 20%) was not inconceivable. However, I favoured a second hypothesis that predicted that a common activity existed for these proteins, but was yet to be reconstituted *in vitro*.

4.1.3 Structure of the Cas2 protein

The sequences of Cas2 proteins are very divergent across different CRISPR-Cas systems, with even putative active site residues being weakly conserved. Cas2 proteins are small (10 kDa) homodimeric proteins with an N-terminal ferredoxin-like fold, common to RNA-binding proteins (Beloglazova et al., 2008; Samai et al., 2010). This fold consists of four antiparallel β -strands flanked by α -helices. The β -sheets

of each monomer interact at the dimer interface to form a β -sandwich, with the C-terminal α -helical regions surrounding them to form the exterior edges of the protein.

4.1.4 Biochemical activity

The resolution of the crystal structure of Cas2 from *S. solfataricus* was accompanied by the report of a metal-dependent ssRNA cleavage activity (Beloglazova et al., 2008). The cleavage products were between 7 - 29 nt in length and the cleavage sites mapped to U-rich regions of the RNA substrates. This activity was suggested to play a role in CRISPR-Cas immunity by degrading phage transcripts (Beloglazova et al., 2008). The authors also identified a conserved aspartic acid (D10) in the putative active site of the *S. solfataricus* Cas2 to be crucial for RNA cleavage and speculated that these residues may coordinate a divalent metal ion at the dimer interface. A study of the *B. halodurans* Cas2 found no RNA cleavage, but did report a strong DNA nuclease activity. This activity was metal-dependent and the authors also hypothesised that the conserved D10 residues of Cas2 were important in coordinating a metal ion essential for catalysis (Nam et al., 2012).

However, a study of the *Desulfovibrio vulgaris* Cas2 was not able to reproduce the ssRNA or dsDNA nuclease activity reported for Cas2 proteins of other CRISPR systems. Additionally the Asp residues thought to be crucial for activity were separated by 12 Å in this structure and therefore unable to jointly bind a metal ion without massive domain reorganization (Samai et al., 2010).

Given the conflicting reports on the structure and function of the Cas1 and Cas2 proteins and the absence of any real insight into their role in CRISPR adaptation at the start of this work, there was a clear need for further investigation. This chapter aimed to study the nuclease and binding activity of a Cas1 and Cas2 from *S. solfataricus*. I observed that, unlike the activity reported for other Cas1 proteins, the Cas1 protein studied had no nuclease activity on single-stranded DNA or Holliday junctions. A striking preference for binding ssDNA was identified, which suggested that single-strand or splayed structures might be key intermediates in adaptation. In common with other Cas2 proteins, the two *S. solfataricus* Cas2 proteins studied in this work were found to degrade ssRNA. However, how this activity links to the role of Cas2 in CRISPR adaptation remains unclear. Finally, although both proteins are required for adaptation of the CRISPR locus *in vivo*, the *S. solfataricus* Cas1 and Cas2 studied in this chapter were not found to interact *in vitro*.

4.2 Results

4.2.1 Expression and purification of a Cas1 and Cas2 from *S. solfataricus*

S. solfataricus has two sets of genes coding for proteins associated with CRISPR adaptation. The first of these adaptation cassettes is located between CRISPR loci A and B. In this cassette, the *cas4* and *csa1* genes are separate from, and in an opposite orientation to, the *cas1* and *cas2* genes (Figure 4.1, A). The Cas1 and Cas2 proteins coded for by this cassette are denoted Cas1_{AB} and Cas2_{AB}. The work presented in this chapter will focus on the second set of Cas1 and Cas2 proteins, coded for by an adaptation cassette located between CRISPR loci C and D (Figure 4.1 A). In this cassette the *cas1* and *cas2* genes are adjacent and are flanked on either side by the *cas4* and *csa1* genes. The proteins coded for by this cassette will be referred to as Cas1_{CD} and Cas2_{CD}. The structure of the Cas2_{CD} protein (Proudfoot et al., 2008) is shown in Figure 4.1 (B), with a model of the Cas1_{CD} created using Phyre2 (Kelly et al., 2015) shown in Figure 4.1 (D).

The Cas1_{CD} and Cas2_{CD} proteins were both expressed recombinantly in *E. coli* (as described in Rollie et al., 2015) with N-terminal polyhistidine tags using the pEHISTEV expression vector (see section 2.2.1.4 and Liu & Naismith (2009)). Metal ion affinity chromatography was used to isolate the tagged protein from cell lysate before tag cleavage and separation. A final size exclusion chromatography step was carried out before flash-freezing the pure purified protein stocks. Both Cas1_{CD} and Cas2_{CD} proteins eluted in sharp, symmetrical peaks from the Superdex 200 size exclusion column (GE Healthcare) and were obtained in stable form, free from contaminants (Figure 4.1, C and E). Cas2_{CD} ran as a single band on an SDS-PAGE gel. However, the pure Cas1_{CD} protein migrated as a smear with a main band at the expected monomer mass (37.2 kDa) and several higher bands visible in the gel. A fraction of the sample did not enter the gel and remained trapped in the wells (Figure 4.1, E). This smearing indicated that some aggregation or oligomerisation of Cas1_{CD} was taking place. The effect was dependent on concentration, with fractions from the edges of the peak, or diluted central peak fractions, running as a single band at monomer weight. Whether these larger oligomers have any functional role remains unknown. However, this effect does seem to be a peculiarity of this particular Cas1 protein, as the Cas1_{AB} protein from *S.*

solfataricus did not exhibit the same smearing on SDS-PAGE (unpublished data, White lab).

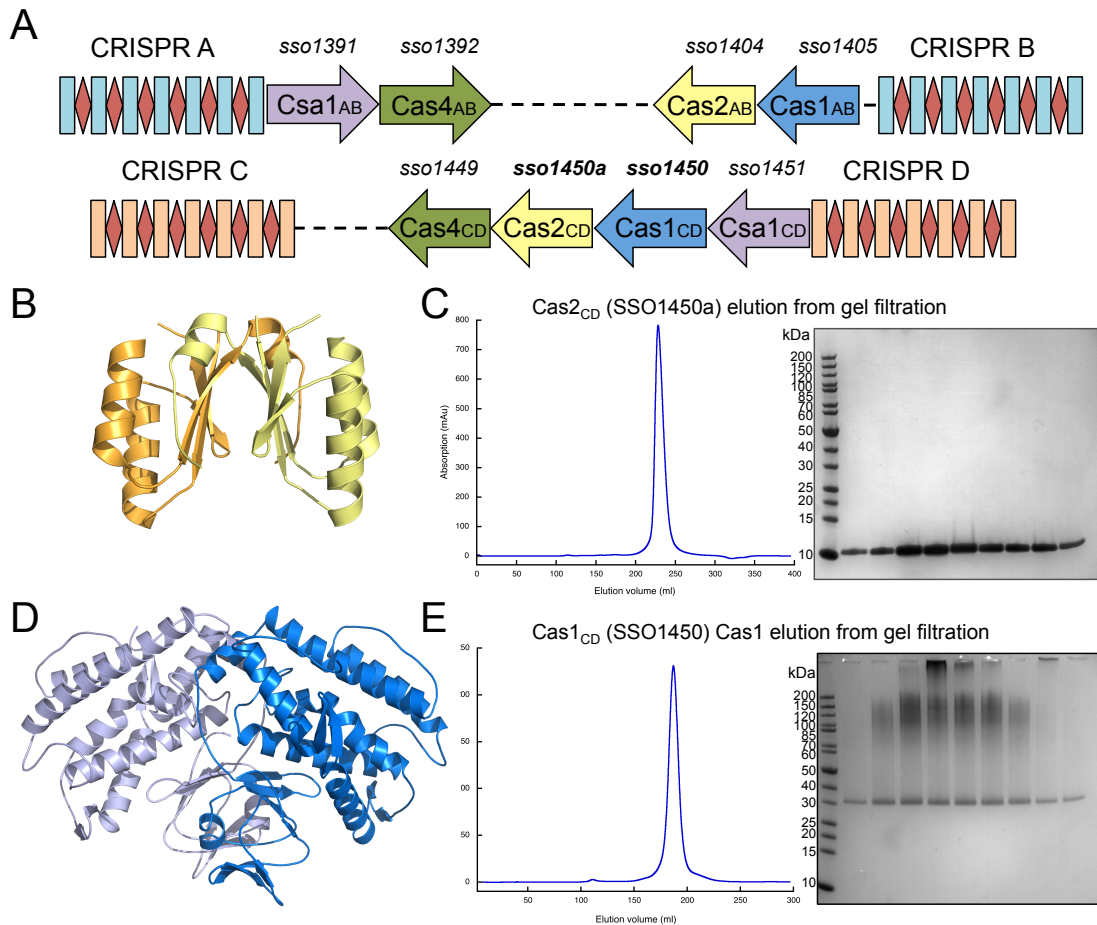


Figure 4.1 Cas1 and Cas2 proteins from *S. solfataricus*

A. A representation of the genomic locations of *cas1* and *cas2* genes of *S. solfataricus*. The *ssol1450a* and *ssol1450* genes (**bold**) code for the Cas1_{CD} and Cas2_{CD} proteins studied in this chapter. These genes are located between CRISPR loci C and D (24 bp repeats shown as rectangles and ~39 bp spacers shown as diamonds). **B.** Structure of the homodimeric Cas2_{CD} (SSO1450a) protein (PDB ID 3EXC) (Proudfoot et al., 2008). **C.** Elution profile and SDS-PAGE gel of the Cas2_{CD} protein (10.2 kDa) following the final gel filtration purification step. **D.** Model of the homodimeric Cas1_{CD} (SSO1450) protein structure created using Phyre2 and based on the structure of the *A. fulgidus* Cas1 protein (PDB ID 4N06) (Kim et al., 2013). **E.** Elution profile and SDS-PAGE gel showing the purity of the SSO1450 protein (~35 kDa) following size-exclusion chromatography.

4.2.2 Substrate preference of the Cas1_{CD} protein

To try to understand the role Cas1 plays in the adaptation process in *S. solfataricus* the substrate preference of the Cas1_{CD} protein was investigated. Electrophoretic mobility shift assay (EMSAs) were performed with a 50 nt single- or double-stranded FAM-labelled DNA of non-CRISPR origin (the B50-5'-FAM oligonucleotide and a duplex made by annealing B50-5'-FAM and B50comp were used, see Table 2.1

(p.47) for sequences), and increasing concentrations of Cas1_{CD} (Figure 4.2, A). High-affinity binding was observed for the single-stranded substrate with a shifted, protein-bound substrate band apparent at even the lowest protein concentration (0.05 μ M). In contrast, for the same concentration of protein, minimal binding of the duplex DNA was observed, with a substrate shift only at the highest concentration of Cas1_{CD} (2.5 μ M).

A more quantitative analysis was also carried out using fluorescence polarization anisotropy and the same 50 nt DNA species. Triplicate titrations with increasing concentrations of Cas1_{CD} were carried out for each substrate, and mean anisotropy values were fitted to a binding isotherm that assumes 1:1 protein to DNA binding (Reid et al., 2001) (as described in Chapter 3). The ssDNA species was found to be bound by Cas1_{CD} with a K_D of 22 ± 2 nM, whereas the dsDNA of the same sequence was bound 20-fold more weakly, with an apparent K_D of 429 ± 38 nM (Figure 4.2, B).

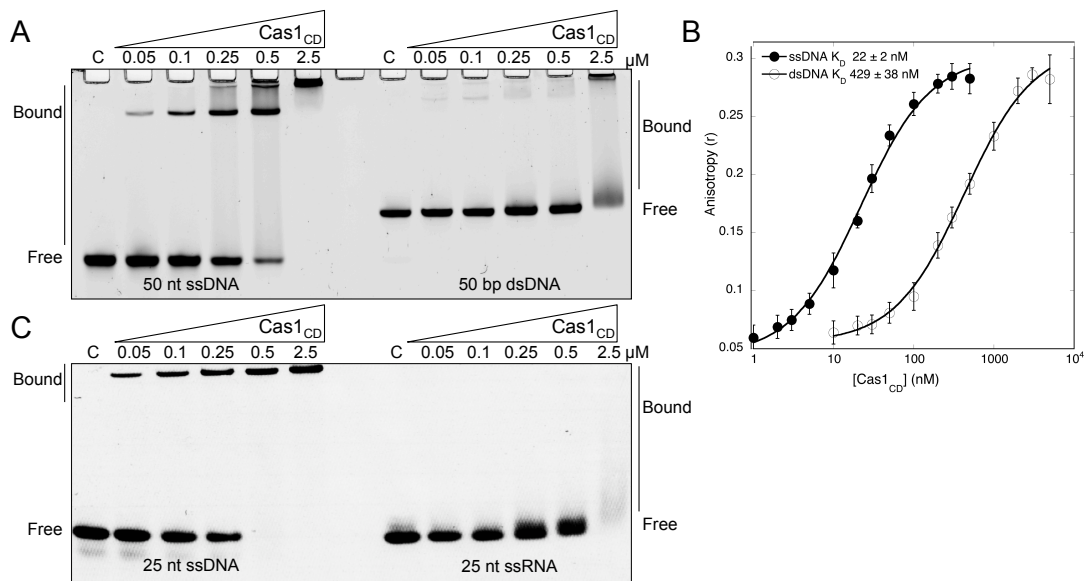


Figure 4.2 Cas1_{CD} binds preferentially to single-stranded DNA

A. EMSA assay showing the binding of Cas1_{CD} to a FAM-labelled 50-nt single- or double-stranded DNA of the same sequence (B50-5'-FAM, B50 comp, sequences shown in Table 2.1 (p.47)). Increasing concentrations of Cas1_{CD}, from 0 in the control (C) to 2.5 μ M, were incubated for 20 min at room temperature with 200 nM substrate in binding buffer (20 mM Tris (pH 7.5), 10 mM NaCl, 1 mM DTT, 5 mM EDTA). The reaction was separated on a 12% native polyacrylamide-TBE gel. **B.** Fluorescence anisotropy titration comparing Cas1_{CD} binding to a single-stranded 50 nucleotide oligonucleotide (B50-5'-FAM) and to a duplex DNA based on the same sequence. Titrations were performed at room temperature in binding buffer. All data points are the means of triplicates, with standard deviations shown. Data were fitted to binding isotherms (Equation 3 (see Chapter 3, section 3.2.6)), using Kaleidagraph (Synergy Software). **C.** EMSA to compare the binding of Cas1_{CD} to a 25 nt ssDNA or ssRNA of the same sequence (CRISPRAB rep (1404) DNA or CRISPRAB rep (1404) RNA) (see Table 2.1 (p.47)) for sequences). Conditions were as in **A.**

An EMSA assay was carried out to compare the binding affinity of Cas1_{CD} protein for a 25 nt single-strand DNA or RNA substrate of the same sequence. This assay showed that there was also strong binding of this shorter ssDNA by Cas1_{CD}, with band-shifting even at the lowest protein concentration. However, Cas1_{CD} bound very weakly to the RNA substrate, with a smear of shifted substrate present only at the highest protein concentration (Figure 4.2, C). The K_D values obtained from anisotropy indicated a higher affinity of Cas1_{CD} for the nucleic acid substrates compared to those estimated from the EMSA assays (Figure 4.2, A and B). This difference may be due to the non-equilibrium nature of EMSA.

This clear preference for single-stranded DNA over other nucleic acid substrates is very interesting and had not been shown previously for Cas1 proteins. Han and colleagues found that Cas1_{CD} bound ss and ds DNA and RNA substrates with similar affinities (Han et al., 2009). The strong preference observed in this study implied that single-stranded DNA might be a key intermediate in the adaptation response mediated by Cas1.

4.2.3 Cas1_{CD} does not bind CRISPR sequences preferentially

In order to investigate whether the strong DNA binding by Cas1_{CD} showed any sequence specificity, a 52 nt substrate was synthesised with a sequence matching the end of the CRISPR C leader and first repeat (CRISPR C LR for and CRISPR C LR rev, see Table 2.1 (p.47) for sequences). Double- and single-stranded versions of this substrate were used in EMSA assays and their binding by Cas1 compared to 50 nt non-CRISPR sequence (B50-5'-FAM, B50 comp, see Table 2.1 for sequences).

Neither the single- nor double-stranded CRISPR substrates assayed were bound with any increased affinity by Cas1_{CD} compared to the control sequence (Figure 4.3, A and B). However, once again, a clear preference was noted for the ssDNA compared to dsDNA versions of the substrates. To confirm this finding, fluorescence anisotropy was carried out with three 24 nt single-stranded fluorescein-labelled DNAs (CRISPRCD rep substrates, see Table 2.1 (p.47) for sequences). These sequences represented either the CRISPR C consensus repeat, the reverse of this sequence, or a scrambled version of the repeat. Triplicate titrations were carried out, and data points were plotted and fitted to a binding isotherm assuming 1:1 protein:DNA binding. Each of the single-stranded sequences tested was bound

with a high affinity by Cas1_{CD} (Figure 4.3, C). However, there was no obvious specificity to this binding with K_D values for the CRISPR C repeat and the scrambled sequences all being between 23 and 31 nM.

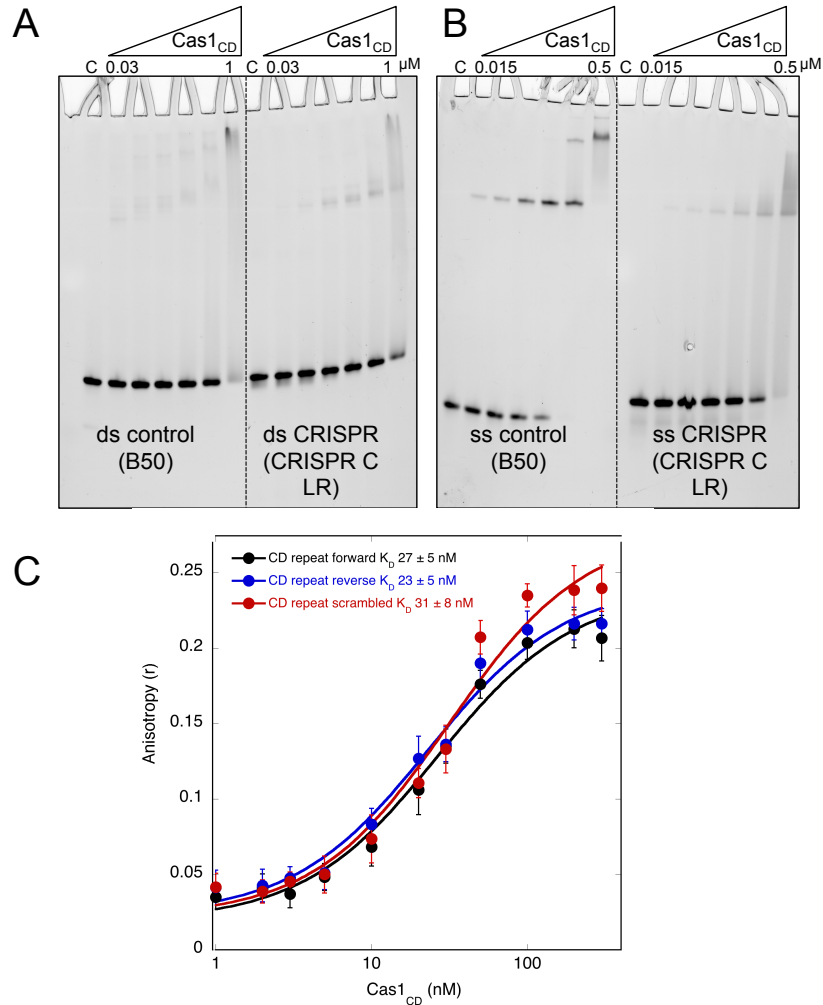


Figure 4.3 DNA binding by Cas1_{CD} is not sequence dependent

A. and **B.** EMSAs showing the binding affinity of Cas1_{CD} for a control DNA sequence (B50, 50 nt) or a DNA sequence containing the end of the leader and the first repeat from CRISPR C (CRISPR C LR, 53 nt). The first lane of each substrate is a control without protein (C). For double-strand binding (**A**) subsequent lanes contain concentrations of Cas1_{CD} of 0.03, 0.06, 0.12, 0.25, 0.5, and 1 μM . For single-strand binding (**B**) the concentrations are of 0.015, 0.03, 0.06, 0.12, 0.25 and 0.5 μM . DNA substrates are 5' (B50-5'-FAM) and 3' (CRISPR C LR for) fluorescein-labelled and are at final concentrations of 200 nM (see Table 2.1 for sequences). Substrate and protein were incubated together at room temperature for 20 min in binding buffer, then run on a 12% native polyacrylamide-TBE gel. **C.** Fluorescence anisotropy titrations comparing Cas1_{CD} binding to a 24 nt oligonucleotide corresponding to the sequence of the *S. solfataricus* CRISPR C repeat, the reverse complement of that sequence and a scrambled sequence with the same nucleotide composition (CRISPRCD rep, CRISPRCD rep rev, CRISPRCD rep scramble, see Table 2.1 for sequences). All data points are the means of triplicates, with standard deviation shown. The data were fitted to a binding isotherm (Equation 3) and K_D values are indicated for each substrate.

This lack of specificity is surprising as it is probable that integrations into the CRISPR C locus are performed by the Cas1_{CD} protein, following recognition and docking at the leader-repeat junction. Specific integrations into other CRISPR loci, in *E. coli* for example, have been shown to require a minimum of 60 bp of the leader sequence (Yosef et al., 2013). Therefore, perhaps the sequence fragments being assayed here were too short or did not contain the motif recognised by Cas1. On the other hand, the lack of specific binding by Cas1_{CD} is also perhaps an important feature of this protein. There are no known sequence motifs present inside *S. solfataricus* spacer sequences, meaning Cas1_{CD} protein must also have low-specificity binding under some circumstances to allow capture and insertion of a wide range of protospacers.

4.2.4 Cas1_{CD} is not a nuclease of ssDNA or Holliday junctions

The Cas1 protein of *E. coli* has been shown to cut Holliday junctions and branched structures, as well as single-stranded DNA (Babu et al., 2011). These nuclease reactions were strictly metal-dependent, but cleavage did not occur at sequence-specific sites. In order to assess the nuclease activity of the Cas1_{CD} protein on similar substrates, a Holliday junction (Jbm5 Holliday, see Table 2.2 for details) was made with a mobile core, which allows resolving enzymes to select favoured sequences for cleavage (Lilley & White, 2001). An active site variant of Cas1_{CD} was made by site-directed mutagenesis to mutate a highly conserved active site glutamate residue to an alanine (E142A) (work carried out by undergraduate student Kotryna Temcinaite) (see section 2.2.1.1 (p.54) for method). The equivalent mutation in the *E. coli* Cas1 protein has been shown to abolish nuclease activity (Babu et al., 2011). The variant protein acted as a control to determine whether any nuclease activity observed was due to Cas1_{CD} activity or a contaminant. The *E. coli* Cas1 protein (EcoCas1) was provided by Dr Ed Bolt (University of Nottingham), and was also assayed to compare Cas1 nuclease activity across different CRISPR-Cas systems.

The Holliday junction substrate was 5'-³²P-radiolabelled on strand A or strand B (Jbm5A or Jbm5B, see Table 2.1 (p.47) for sequences) (Figure 4.4, A) and incubated with Cas1 and divalent metal cations (see section 2.2.8.1.1 (p.65) for method). Following a 30-minute incubation at 55 °C (Cas1_{CD}) or 37 °C (EcoCas1) the assay products were separated on a denaturing polyacrylamide-TBE gel and

phosphorimaged. Each of the four ssDNA strands used to make the Holliday junction was also labelled and assayed separately with Cas1 proteins to look for ssDNA nuclease activity. Neither Holliday junction substrates, nor the single-strand components, were cut by any of the Cas1 proteins under the conditions tested here (Figure 4.4, A and B).

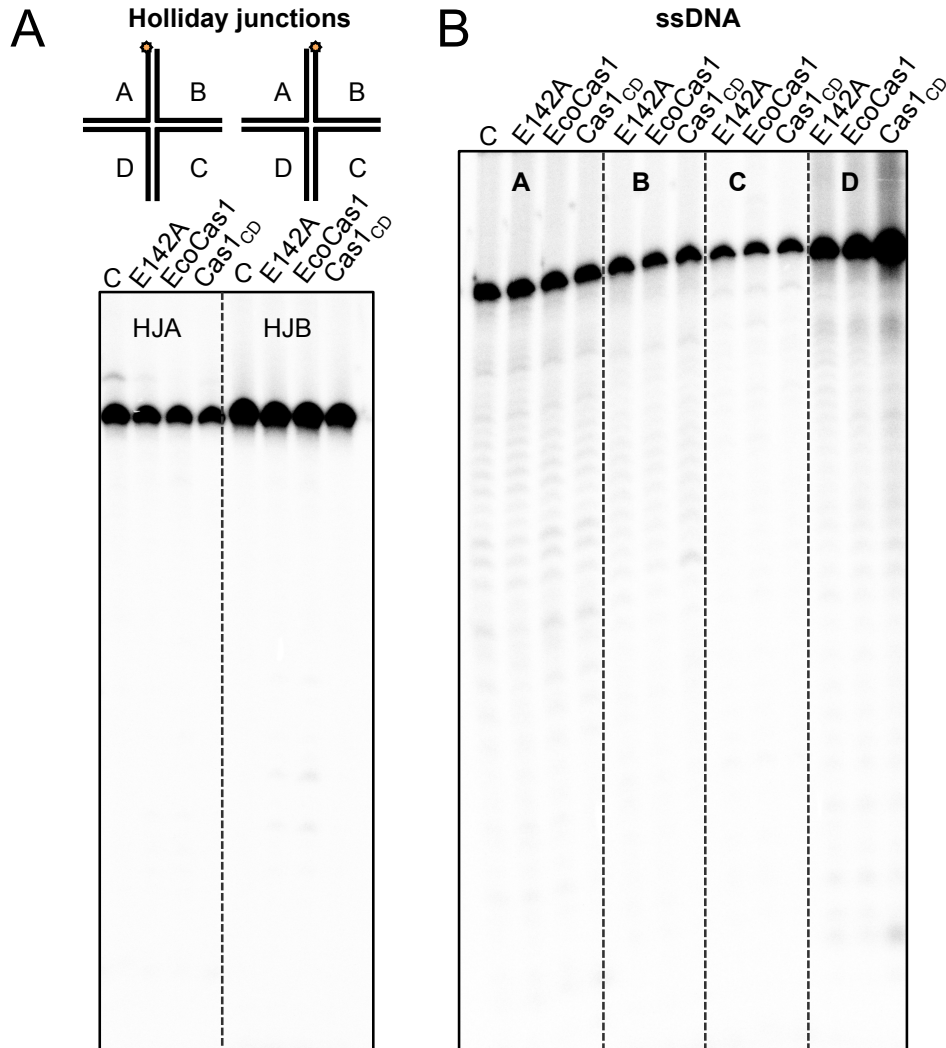


Figure 4.4 Cas1 does not cut Holliday junction or single-strand DNA sequences.

A. A Holliday junction with a mobile core (Jbm5 Holliday, see Table 2.2) was made by annealing 48 nt strands A, B, C and D (Jbm5A, B, C and D, see Table 2.1 for sequences). The junction was assayed with Cas1_{CD}, the active site variant of this protein Cas1_{CD}, E142A and the *E. coli* Cas1 (EcoCas1). The substrates (50 nM) and Cas1 proteins (500 nM) were incubated together for 30 min at 55 °C (Cas1_{CD}) or 37 °C (EcoCas1) in the presence of 5 mM MnCl₂ in nuclease buffer (20 mM Tris (pH 7.5), 10 mM NaCl, 1 mM DTT). The assay products were then separated on a 20% denaturing polyacrylamide-TBE gel before phosphorimaging. The first 4 lanes show the result of labelling strand A of the Holliday junction with ³²P and the subsequent lanes are the product of labelling strand B. The first lane in each case is a control without the addition of protein (C). B. Each of the 48 nt single strands used to make the Holliday junction were also assayed under the same conditions as in A.

The absence of non-specific DNA cleavage by the Cas1_{CD} protein found here supports the previous findings by Han and colleagues who also failed to observe nuclease activity for the *S. solfataricus* Cas1 (Han et al., 2009). However, the absence of cleavage products when assaying substrates with the EcoCas1 protein was surprising, as other studies had found robust degradation of single-stranded and branched structures by this protein (Babu et al., 2011; Wiedenheft et al., 2009).

In summary, this initial characterisation of the Cas1_{CD} found that this protein bound single-stranded DNA of varying lengths with a very high affinity. A lower specificity was observed for other nucleic acid substrates. DNA binding by Cas1_{CD} is sequence non-specific, with CRISPR sequences not bound preferentially over control sequences. In addition, this study failed to reproduce the sequence non-specific DNA nuclease activity reported for other Cas1 proteins.

4.2.5 Cas2_{CD} does not bind strongly to nucleic acids

To attempt to learn more about the Cas2_{CD} protein from *S. solfataricus*, EMSA assays were carried on 48 nt DNA, RNA or DNA:RNA hybrid substrates, made up of the end of the CRISPR C leader sequence and the first repeat (CRISPR C transcript substrates, see Table 2.1 (p.47) for sequences). No band-shift was observed for any of the nucleic acid substrates tested at concentrations up to 5 μ M Cas2_{CD} (Figure 4.5).

From these data it seems that if Cas2_{CD} is involved in DNA binding or manipulation during the capture or integration of new spacers, some kind of conformational change is required to switch the protein into a form capable of binding nucleic acids. Association with other adaptation proteins or cofactors may trigger this conformational change. Nam and co-workers showed that the binding of a metal cation induced a conformational change in the *B. halodurans* Cas2 protein. This conformational change was, in turn, shown to activate nuclease activity of this Cas2 (Nam et al., 2012a).

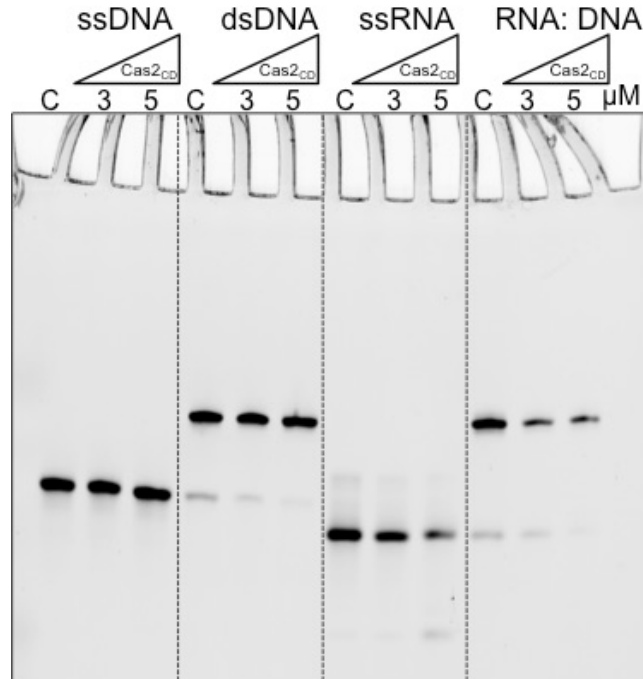


Figure 4.5 Cas2_{CD} does not show nucleic acid binding

EMSA were carried out to assess Cas2_{CD} binding activity. 48 bp sequences containing the end of the leader and the first repeat of CRISPR C (CRISPR C transcript DNA, CRISPR C transcript DNA comp, CRISPR C transcript RNA, see Table 2.1 (p.47) for sequences) were used in ss/dsDNA, RNA or RNA:DNA hybrid versions. 200 nM of 5'-FAM-labelled substrates were incubated for 20 min with 3 or 5 μ M of Cas2_{CD} protein in binding buffer at room temperature. The first lane for each substrate is a control without protein (C). The assay components were resolved on a native 12% polyacrylamide-TBE gel.

4.2.6 Cas2 proteins possess ribonuclease activity

Although no nucleic acid binding was observed for Cas2_{CD} with the CRISPR C transcript sequences assayed, some degradation of the ssRNA substrate was apparent (Figure 4.5). To further investigate this activity, nuclease assays were performed on these substrates with and without the addition of magnesium chloride (see section 2.2.8.1.2 (p.66)). Substrate and protein were incubated together at 55 °C for 30 min and the products were separated on denaturing 20% polyacrylamide-TBE gels. These experiments showed that Cas2_{CD} digested the ssRNA substrate producing one main product, even in the absence of metal ions (Figure 4.6, A). When the sequences of the CRISPR C transcript substrates were examined using the mfold RNA web server (Zuker, 2003), a palindromic region was identified which is predicted to form a hairpin with a melting temperature of 77 °C at the salt, metal ion and pH conditions used in this assay (Figure 4.6, C). If this hairpin structure exists *in vivo* in *S. solfataricus*, it could be a crucial for recognition or docking of the adaptation proteins.

In order to assess where the putative hairpin RNA structure is cut by Cas2_{CD}, an alkaline hydrolysis ladder was made and run alongside the assay products. Additionally, in order to investigate whether ssRNA cleavage activity is a common feature of Cas2 proteins from *S. solfataricus*, the Cas2_{AB} protein was expressed and purified. A variant of the Cas2_{AB} protein was also made by site-directed mutagenesis to mutate an aspartic acid residue in the putative active site to an alanine (D10A). This mutation has been reported to abolish nuclease activity in other Cas2 proteins (Nam et al., 2012; Beloglazova et al., 2008).

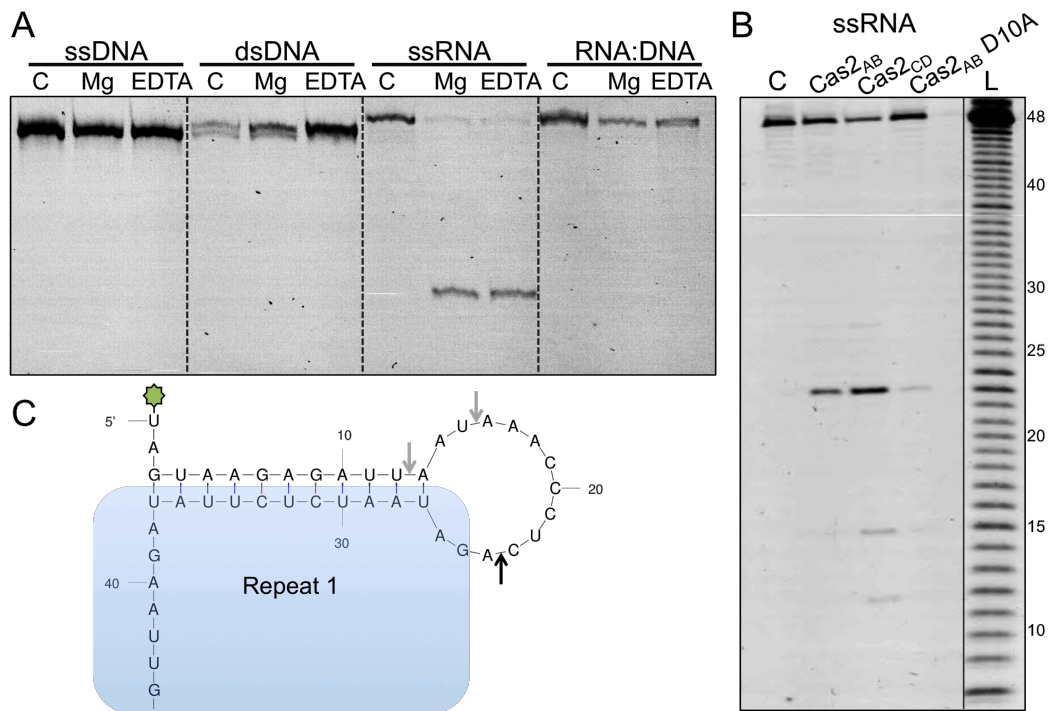


Figure 4.6 *Sulfolobus solfataricus* Cas2 proteins have ribonuclease activity

A. Nuclease assay of 48 nt ss/dsDNA, RNA and RNA:DNA hybrid duplex with Cas2_{CD}. The sequence of substrates matched the end of the CRISPR C leader and the first repeat (CRISPR C transcript DNA/RNA, see Table 2.1 (p.47) for sequences). Substrates were 5'-FAM-labelled and were used at a concentration of 200 nM in the assay. 5 μ M Cas2_{CD} was incubated with the nucleic acid substrate in nuclease buffer and either 5 mM MgCl₂ or 5 mM EDTA at 55 °C for 20 min. The first lane for each substrate is a control without protein added (C). The assay products were separated on a denaturing 20% polyacrylamide-TBE gel. **B.** Nuclease assay of *S. solfataricus* Cas2 proteins Cas2_{AB}, Cas2_{CD} and active site mutant Cas2_{AB} D10A on the 48 nt CRISPR C leader-repeat transcript (CRISPR C transcript RNA). Assay conditions were as for **A**. The final lane of the gel was an RNA ladder made by alkaline hydrolysis. The solid line indicates that the gel has been truncated to bring the ladder closer to cleavage products to allow size determination. **C.** A model of the putative hairpin structure formed by the end of the CRISPR C leader and repeat 1 in the ssRNA substrate. The partial repeat sequenced is contained within the blue box. Residues are numbered from the 5' end of the substrate and cut sites indicated by arrows, with the main site in black and minor sites in light grey.

Cleavage of the CRISPR C transcript RNA substrate occurs at the same residues for both the Cas2_{AB} and Cas2_{CD} proteins (Figure 4.6, B). The ribonuclease activity is much weaker for the Cas2_{AB} D10A variant. There is a main cleavage product of 23 nt resulting from cleavage between an A and C residue, with minor products at 15 and 12 nt resulting from cleavage between U and A residues (Figure 4.6, C). These cleavage events happen at sites that do not share an obvious sequence motif or equal nucleotide spacing between them. The significance of this putative activity to the *in vivo* role of Cas2 is not clear, as it does not directly indicate a role in integration of new spacers.

4.2.7 Cas1_{CD} and Cas2_{CD} do not interact *in vitro*

cas1 and *cas2* genes are almost always located next to each other in *cas* operons (Makarova et al., 2013a) and are both required for adaptation (Yosef et al., 2012). Before work contained in this chapter began there had been many suggestions, but no proof, that these two proteins form a complex to bring about adaptation. To try and address whether such an interaction may take place between the Cas1_{CD} and Cas2_{CD} proteins, isothermal titration calorimetry (ITC) was carried out. This technique involves titration under isothermal conditions of a protein or small molecule ligand into a sample of a putative binding partner. Any interaction of the two components leads to min heat changes, which reduce in size during the titration as the concentration of one of the binding partners becomes limiting. In a successful experiment these heat changes can then be used to plot a binding isotherm to yield useful information about the observed interaction, such as the K_D and stoichiometry.

The titration of Cas1_{CD} into Cas2_{CD} led to very small heat changes that could not be fitted to a binding isotherm (Origin software, Origin Lab) (Figure 4.7, A). These data imply that under the conditions tested the Cas1_{CD} and Cas2_{CD} proteins did not interact, or interacted extremely weakly.

I hypothesised that these two proteins may only interact in combination with a substrate DNA. To test this theory, gel filtration was carried out to look for a shift in the elution peaks of the Cas1_{CD} and Cas2_{CD} proteins, alone or pre-mixed in the presence of a DNA substrate. DNA and proteins were mixed at a ratio of 2Cas1_{CD}:4Cas2_{CD}:1DNA. A 29 bp duplex DNA substrate with single-stranded 3' overhangs (5 nt) (made by annealing 3'OHprotospacer for and 3'OHprotospacer rev

oligonucleotides (see Table 2.1 (p.47) for sequences) was used in the assay. Incubation of the proteins with or without DNA was carried out for 15 min at 45 °C before an overnight incubation at room temperature (see section 2.2.764 (p.64) for method).

The elution peaks of Cas1_{CD} and Cas2_{CD}, run separately on to a Superose12 column (GE Healthcare), overlapped slightly, with the peak Cas1_{CD} elution at 12.85 ml and peak Cas2_{CD} elution at 14.56 ml (Figure 4.7, B). Following incubation with DNA, the elution profile of Cas1_{CD} shifted to the left, with peak elution occurring at 10.46 ml. The 260/280 absorbance ratio for the Cas1_{CD} peak fraction in the +DNA condition was 1.59, compared to a ratio of 0.50 in the Cas1_{CD}-only condition, confirming that the shift observed was due to nucleic acid binding. Incubation of Cas1_{CD}, Cas2_{CD} and DNA lead to an almost identical peak elution for Cas1_{CD} to that obtained without the addition of Cas2_{CD}. The peak elution occurred at 10.49 ml in this condition, which directly overlapped the elution peak for Cas1_{CD}+DNA. A second smaller elution peak at 14.17 ml was present in the Cas1_{CD}, Cas2_{CD} and DNA condition, corresponding to the elution of Cas2_{CD}. This peak was shifted slightly to the left compared to the elution maximum in the Cas2_{CD}-only condition.

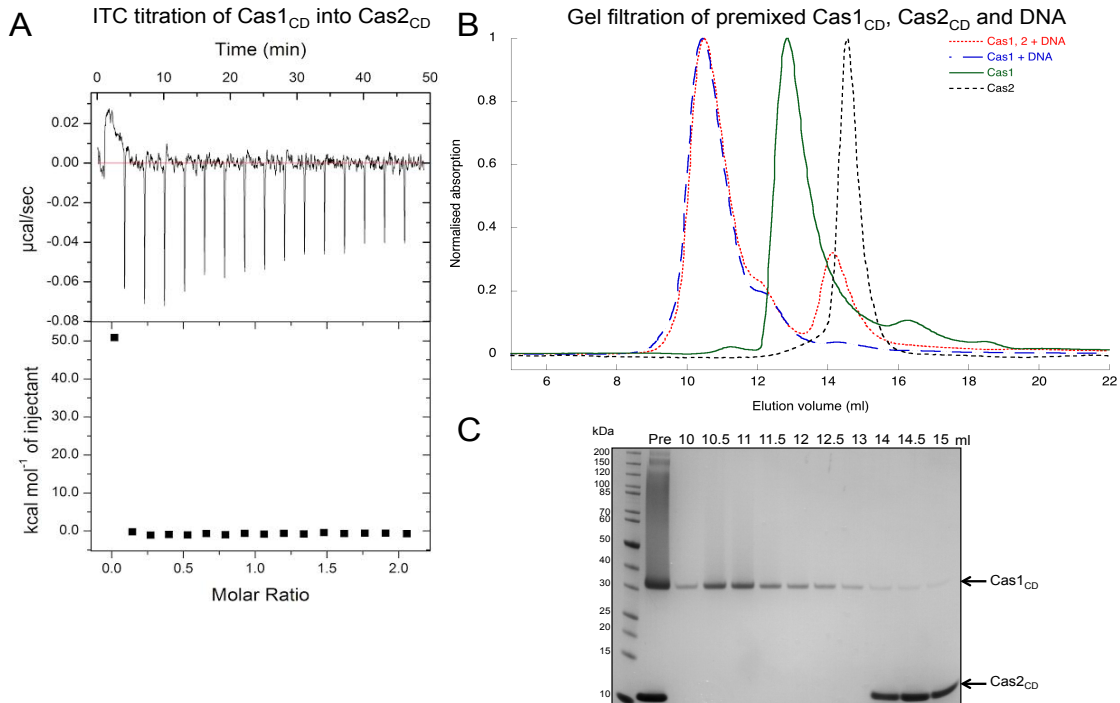


Figure 4.7 Cas1_{CD} and Cas2_{CD} do not interact *in vitro*

A. Results from an ITC titration of Cas1_{CD} into Cas2_{CD}. 16 injections of 2.5 µl Cas1_{CD} (150 µM) were made into a well containing Cas2_{CD} (15 µM). The titration was carried out at 25 °C and pH 7.5 in ITC buffer (200 mM KCl, 20 mM HEPES-KOH (pH 7.5), 5% glycerol and 1 mM

TCEP). The top panel shows a thermogram of raw heat changes following titration of Cas1 into Cas2. The bottom panel shows the integrated heat changes normalised to concentration of injectant. The first injection often carries small bubbles, which may explain the anomalous first data point, which is often discarded before plotting. **B.** Gel filtration elution profiles of Cas1_{CD} and Cas2_{CD}. The peaks correspond to the elution of Cas2_{CD} alone (black), Cas1_{CD} alone (green), Cas1_{CD} following incubation with DNA (blue) and Cas1_{CD} and Cas2_{CD} following incubation together with DNA (red). Peak maxima were normalised to 1 to allow clear visualisation of any changes in elution profiles. The DNA used was a 29 nt duplex with 5 nt single-stranded 3' overhangs (made by annealing 3'OHprotospacer for and 3'OHprotospacer rev oligonucleotides (see Table 2.1 (p.47) for sequences). The Cas1:Cas2:DNA ratio used for incubations was 2:4:1 (40:80:20 μ M). The incubations were carried out at 45 °C for 15 min before overnight incubation and dialysis at room temperature in 20 mM Tris (pH 7.5) and 150 mM NaCl. Proteins were separated on a pre-equilibrated Superose12 column (GE Healthcare) in the same buffer. **C.** SDS-PAGE gel stained with Coomassie dye showing fraction contents from the combined Cas1_{CD}, Cas2_{CD} and DNA elution shown in B (red). Bands corresponding to Cas1_{CD} (35 kDa) and Cas2_{CD} (10 kDa) are indicated by arrows.

These results implied that Cas1_{CD} and to a lesser extent Cas2_{CD} bound and potentially oligomerised on the DNA substrate, leading to an earlier peak elution from gel filtration compared to the apo-proteins. However, as no further shift was seen in the Cas1_{CD} + DNA peak on the addition of Cas2_{CD}, these proteins did not interact to form a stable, larger protein complex under the conditions tested. Furthermore, an SDS-PAGE gel with samples from across the two elution peaks obtained after mixing Cas1_{CD}, Cas2_{CD} and DNA showed that there was no Cas2_{CD} present in the peak Cas1_{CD}+DNA fractions (Figure 4.7, C). There was a small amount of Cas1_{CD} present in the fractions containing the Cas2_{CD} elution peak. However, this was likely due to a slight overlap in the elution profiles of these two proteins and not due to a true interaction.

4.3 Discussion

Following the data collection for this chapter several papers were published that provided key insights into the structure and function of the Cas1 and Cas2 proteins. These studies also shed light on some of the results presented here and provided a fuller picture of the role of Cas1 and Cas2 in adaptation of the CRISPR system.

Firstly, the Cas1 and Cas2 proteins from *E. coli* were found to interact to form a complex essential for adaptation (Nuñez et al., 2014). This complex was shown to consist of two Cas1 dimers bridged by a Cas2 dimer and while the active site residues of Cas1 were found to be necessary for adaptation, the active site of Cas2 was not required. This led to the conclusion that Cas2 is merely a structural and not a catalytic component of the complex.

A second structure of the *E. coli* Cas1-Cas2 complex bound to a protospacer substrate also greatly advanced our understanding of the mechanism of integration by Cas1 and Cas2 proteins (Nuñez et al., 2015a). This work showed that a stable complex was obtained when a protospacer length DNA (33 bp) was added to the Cas1-Cas2 complex. A central duplex DNA region of 23 bp was bound with the 5 bp at either end of the DNA being splayed by a wedge-like tyrosine residue (Figure 4.8, A). The 3' splayed ssDNA ends were tightly co-ordinated by Cas1 subunits at either end of the complex and the 3' hydroxyl residues were positioned exactly in the metal-binding active site, poised to perform nucleophilic attack (Nuñez et al., 2015a) (Figure 4.8, B). This structure was thought to represent the post-capture, but pre-integration step in adaptation of the CRISPR array. Wang and colleagues added to these findings when they reported that a Cas1 subunit of the *E. coli* Cas1-Cas2 complex makes sequence-specific contacts with, and processes, PAM sequences upstream of bound protospacers (Wang et al., 2015).

These recent discoveries have strongly influenced the interpretation of some of the results presented in this chapter and helped to explain some findings that previously had seemed to contradict what was understood about the activity of Cas1 and Cas2.

4.3.1 ssDNA binding

A key finding from the work presented here was that the Cas1_{CD} protein from *S. solfataricus* bound with 20-fold higher affinity to single-strand, compared to double-strand, DNA (Figure 4.2). This striking difference in affinities was unexpected, as a previous study had reported equal-affinity binding of Cas1 for ssDNA/RNA and dsDNA with apparent K_D values all between 20-50 nM (Han et al., 2009). This strong preference for single-stranded DNA binding implied that Cas1 encounters single-stranded intermediates while carrying out spacer capture or integration. However, the role or origin of these single strands was not clear. I hypothesized that some opening of the DNA duplex around the integration site occurs during adaptation, and Cas1 may be involved in binding and stabilizing the single strands of this structure. Furthermore, it has been suggested that single-stranded DNAs produced by the RecBCD complex may act as precursor protospacers for integration by Cas1 and Cas2 (Levy et al., 2015).

If the capture and integration of single-stranded protospacers does play a part in CRISPR adaptation, this may explain the preferences observed for Cas1 binding.

However, the publication of the crystal structure of the Cas1-Cas2 proteins from *E. coli* bound to a duplex DNA with splayed ends provided a more tangible explanation of the strong single-strand binding I observed. Cas1 monomer subunits of the Cas1-Cas2 complex were found to tightly co-ordinate 3' single-strand ends, positioning the 3'-hydroxyl residue in the Cas1 active site (Nuñez et al., 2015a) (Figure 4.8). Wang and colleagues also showed that the Cas1-Cas2 complex had a much-reduced affinity (~5-fold) for double-stranded, compared to partially single-stranded, DNA (Wang et al., 2015). Therefore, it seems that the preferential binding of Cas1_{CD} to ssDNA reported in this chapter is a true feature of Cas1 proteins and indicates that the Cas1_{CD} protein, like the Cas1 of *E. coli*, has a role in capturing and inserting at least partially single-stranded protospacers.

4.3.2 Cas1 is not a non-specific nuclease

Another finding of this work was that the Cas1_{CD} protein did not have any nuclease activity on non-CRISPR single-stranded or Holliday junction substrates (Figure 4.4). While this supported the work of Han et al. who also found no nuclease activity for the Cas1_{CD} protein (Han et al., 2009), it contradicted previous studies that reported cleavage of these substrates by the *P. aeruginosa* and *E. coli* Cas1 proteins (Wiedenheft et al., 2009; Babu et al., 2011). It is plausible that these proteins, which share little sequence similarity and come from different CRISPR subtypes, really do have different activities *in vivo*. However, the previously reported activity does not fit well with that of the purified Cas1-Cas2 complex, which has only been shown to cut single-strand DNA specifically at PAM sequences (Wang et al., 2016). It may be that the promiscuous nuclease cleavage reported by other groups was a weak secondary activity of Cas1, exposed by long incubation times and very high protein and metal ion concentrations.

Interestingly, in our hands even the *E. coli* Cas1 protein showed no activity on the substrates tested. It is perhaps worth noting that inadequately purified proteins can lead to spurious nucleic acid degradation due to contaminating nucleases, which may also explain the discrepancy in the results obtained from those published previously.

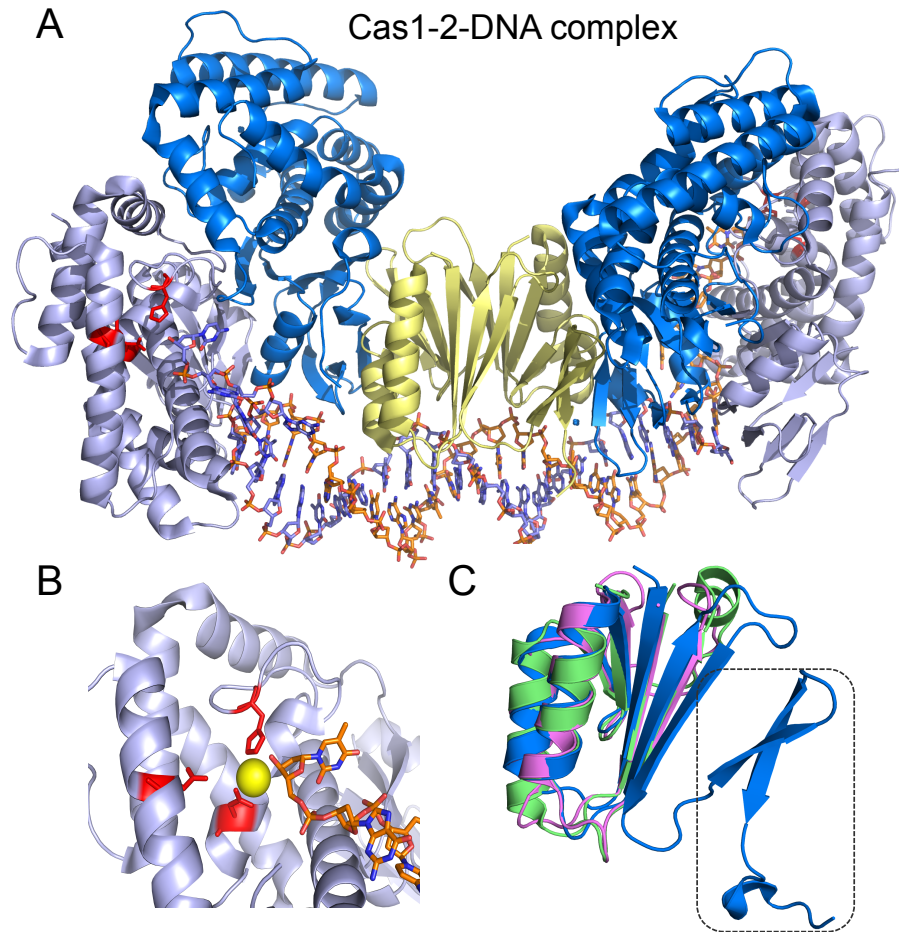


Figure 4.8 Cas1-Cas2 form a complex in *E. coli*

A. Crystal structure of the DNA-bound Cas1-Cas2 complex from *E. coli* (PDB ID 5DS6). Cas1 dimers are shown in blue and Cas2 dimer in yellow. A 33 bp protospacer is bound with 5 bp at each end splayed by the complex. **B.** The 3' hydroxyls of the single-stranded 3' ends are positioned in the Cas1 active site. A metal cation is shown in yellow and the residues required to co-ordinate this ion (E141, D221 and H208), are shown in red. **C.** Alignment of the crystal structures of Cas2 proteins from *S. solfataricus* (Cas2_{AB} (PDB ID 2I8E) (pink) and Cas2_{CD} (PDB ID 3EXC)) (green) and complex-bound Cas2 from *E. coli* (PDB ID 4P6I) (blue). The dashed box indicates the C-terminal tail of the *E. coli* Cas2, important for complex formation.

4.3.3 Cas2_{CD} does not stably bind nucleic acid

No nucleic acid binding by Cas2_{CD} was found in this work (Figure 4.5). This result is unexpected as in the protospacer-bound Cas1-Cas2 structure (Nuñez et al., 2015a) an arginine-rich 'clamp' of Cas2 was shown to interact in a sequence-independent manner with the phosphate backbone of the duplex DNA. Variant Cas2 proteins with key arginine residues replaced by alanine showed much-reduced integration activity, implying that the binding of DNA by Cas2 is key to adaptation in the *E. coli* complex (Nuñez et al., 2015a). The lack of DNA binding by Cas2_{CD} indicated that this protein might undergo a conformational change to allow DNA binding. A

conformational shift occurring on metal ion binding was reported to activate DNA degradation by Cas2 in *B. halodurans* (Nam et al., 2012a). Furthermore, a large conformational shift was also shown to happen on formation of the *E. coli* Cas1-Cas2 complex (Nuñez et al., 2014). Therefore, complex formation with Cas1 or other CRISPR proteins may be required to expose the key positively charged amino acids of Cas2 required for DNA binding.

4.3.4 The role of ribonuclease activity of Cas2

Here, both Cas2 proteins from *S. solfataricus* were shown to degrade ssRNA without the requirement for divalent metal ions (Figure 4.6). RNA substrates tested were cut at the same position by both Cas2 proteins; further work will be required to confirm any sequence or structural requirements. This non-specific ribonuclease activity has previously been reported for Cas2 proteins, in addition to dsDNA degradation (Beloglazova et al., 2008; Nam et al., 2012). These activities contradict the role Cas2 has been reported to play in the Cas1-Cas2 complex required for adaptation, where it acts only as a structural component (Nuñez et al., 2014). An explanation may lie in the suggestion that the original function of Cas1 and Cas2 is as a toxin/antitoxin system (Koonin & Makarova, 2013). The toxins of these systems often lead to cell death or dormancy by degrading RNA, with the antitoxin acting as an ‘antidote’ to the toxin. Cas2 proteins share considerable sequence and structural similarity with the VapD toxin proteins also reported to be ssRNA-specific nucleases, which cut before purine residues (Kwon et al., 2012). The nuclease activity of the Cas2 proteins may originate from their role as a toxin, and may even be important during infection in slowing viral transcription, thus giving the host time to upregulate the CRISPR-Cas response (Koonin & Makarova, 2013). The binding of the Cas1 antitoxin and complex formation may switch off Cas2’s role as a toxin by direct protein–protein interaction, as shown for other toxin/antitoxin systems (Winther & Gerdes, 2011). Finally, it also cannot be ruled out definitively that small RNases (that are often co-purified with Cas2 (unpublished data, White lab) are responsible for the varied nuclease activities attributed to Cas2 proteins.

4.3.5 Lack of Cas1-Cas2 complex formation in *S. solfataricus*

The Cas1-Cas2 complex in *E. coli* is essential for adaptation. However, as yet the Cas1 and Cas2 proteins from *S. solfataricus* have not been found to interact to form a complex. In this chapter, no interaction of the two proteins was found either by

ITC or by mixing purified Cas1 and Cas2 proteins, with or without a DNA substrate before elution from gel filtration (Figure 4.7). The lack of complex formation may indicate that other CRISPR proteins, such as Cas4 or Csa1, are needed for the *S. solfataricus* Cas1 and Cas2 proteins to interact. Furthermore, the presence of a DNA substrate with a particular sequence or structure may be important for nucleation of the complex, requirements that were potentially not met in the experiments reported here.

Interestingly, the C-terminal tail of the *E. coli* Cas2 protein, shown to be important for complex formation with Cas1, differs from that of other Cas2 proteins. Figure 4.8 (C) shows the superimposition of the structures of the two Cas2 proteins from *S. solfataricus* aligned to that of the *E. coli* Cas2 from the Cas1-Cas2 complex (Nuñez et al., 2014). The final three β -sheets of the *E. coli* Cas2 tail are at 90° to the tail of the *S. solfataricus* proteins. This difference may imply an alternative mechanism of complex formation in *S. solfataricus* involving other partners or protein contacts. However, it is hard to draw any definite conclusion from this as the final ~ 10 residues of the C-terminal tails in the apo-Cas2 structures are unresolved, which may imply this whole region is flexible and is only becomes fixed on binding Cas1.

Chapter 5: Cas1 performs a sequence-specific disintegration reaction

The *E. coli* proteins used in this chapter were provided by Dr Ed L Bolt and Anna Sophie Brinkmann (University of Nottingham).

This chapter is adapted in part from the published manuscript: **Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition** (Rollie et al., 2015).

5.1 Introduction

When the work contained in this section began, Cas1 and Cas2 were known to form a complex essential for the incorporation of foreign DNA into the host genome (Yosef et al., 2012; Nuñez et al., 2014). However, the mechanism by which this integration happened remained largely unknown. A model of how adaptation of the CRISPR locus by the Cas1-Cas2 complex was hypothesised to happen is shown in Figure 5.1.

5.1.1 Protospacer capture

The first stage in adaptation is the capture of a DNA protospacer from a foreign genetic element (Figure 5.1, 1). The resolution of the crystal structure of the Cas1-Cas2 complex bound to a protospacer substrate elucidated the means by which this occurs. Tyrosine residues from two Cas1 subunits were found to bracket a 23 bp duplex and act as wedges to splay the remaining 5 bp of duplex DNA at either end into single strands (Wang et al., 2015; Nuñez et al., 2015a). The single-stranded 3' ends are bound tightly by the protein complex and are cut five nucleotides from the end of the 23 bp duplex, at PAM complementary (5'-CTT-3') sequences in *E. coli* (Wang et al., 2015; Nuñez et al., 2015a).

However, it is still not clear how Cas1-Cas2 initially captures this segment of viral DNA prior to processing and insertion. There appears to be little influence of sequence on protospacer uptake by Cas1-Cas2, outside of a short PAM motif that is also required for interference (Díez-Villaseñor et al., 2009). A recent study reported that spacers were preferentially acquired from between replication fork-stalling

points and Chi sequences during *E. coli* adaptation (Levy et al., 2015). The RecBCD complex is known to be recruited to double-strand breaks, which frequently occur at fork-stalling points, and to degrade DNA until it reaches a Chi site (Dillingham & Kowalczykowski, 2008). The authors concluded that the products of RecBCD DNA degradation, which range in size from ten to hundreds of nucleotides, may be substrates for capture and insertion as protospacers by Cas1-Cas2 (Levy et al., 2015). This finding provided an explanation for the co-precipitation of RecB and C with Cas1 previously reported (Babu et al., 2011).

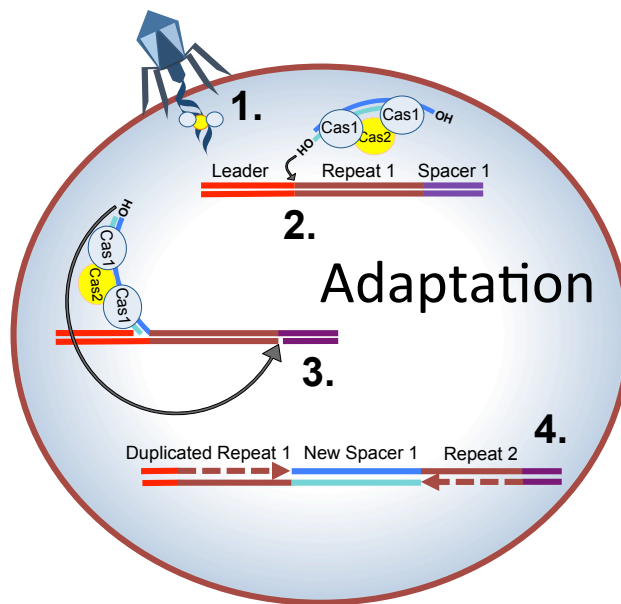


Figure 5.1 Model of the adaptation stage of CRISPR-Cas immunity

1. A protospacer is excised from incoming foreign DNA through an unknown mechanism then bound and processed by the Cas1-Cas2 adaptation complex. **2.** A specific transesterification reaction, mediated by the Cas1-Cas2 complex, at the leader-repeat junction (site 1) joins a 3' OH of the incoming spacer to the leader-proximal end of 1st repeat. **3.** A second transesterification joins the remaining 3' end of the protospacer to the end of the 1st repeat (site 2). **4.** Host replication and repair proteins are thought to fill gaps and ligate nicks.

If RecBCD is responsible for feeding Cas1-Cas2 with DNA for insertion, this raises the question of what fulfils this role in systems that lack RecBCD. It is conceivable that the exonuclease Cas4, which is essential for spacer uptake in the type I-B CRISPR-Cas system of *Haloarcula hispanica* (Li et al., 2014), may provide the DNA fragments for integration in some systems. In *S. solfataricus* the helicase/nuclease complex HerA-NurA has also been shown to degrade DNA in a 5' - 3' direction from 3' overhangs (Blackwood et al., 2012) and may substitute for the RecBCD complex to provide substrates for adaptation. Protospacers generated by these means would be single-stranded and much longer than the final spacer length. Therefore, it

seems that reannealing and processing of the protospacers is carried out before or during insertion, either by Cas1-Cas2 or further unknown host factors.

5.1.2 Spacer integration

The next step in adaptation is the specific integration of the new spacer between the leader and repeat 1 of the CRISPR array. Very little is known about how the insertion site is located by the Cas1-Cas2 proteins, although it has been shown that in *E. coli* at least one repeat sequence and the last 60 bp of the leader were essential for adaptation (Yosef et al., 2013). This implies that some key sequence or structural motifs exist in this region that are essential for the recognition and/or docking of the adaptation complex.

In adaptation, the staggered nicking of the host genome and the joining of a spacer 3' end to the 5' end of the first repeat are thought to happen simultaneously by transesterification (Figure 5.1, 2) (Arslan et al., 2014). The joining of both ends of the spacer to the host genome occurs by two 'half-site' reactions carried out by Cas1, utilising the 3' hydroxyls of the protospacer as nucleophiles. One 3' end of the incoming DNA will be joined to the 5' end of the first repeat, proximal to the leader sequence (site 1), and the second 3' end of the protospacer will be joined to the leader-distal 5' end of the first repeat on the complementary strand (site 2) (Figure 5.1, 3). Evidence to support this mechanism was provided by Arslan and colleagues who identified half-site intermediates of the adaptation reaction in *E. coli* (Arslan et al., 2014).

An outstanding question when the work in this chapter began was whether the two half-site reactions required for full spacer integration happen at once, or whether they progress in a certain order. I hypothesised that site 1 may be recognised and targeted first by the adaptation protein complex, as its sequence is unchanging, whereas the sequence and structure of site 2 is partially defined by the most recent spacer integration.

5.1.3 Repeat duplication and repair

Integration by this staggered cleavage and ligation mechanism leads to the duplication of the first repeat. The subsequent nick-ligation and gap-filling reactions are the final stages of adaptation, needed to repair the host genome, and are thought to be carried out by cellular factors (Figure 5.1, 4). Recently two essential

sequence motifs inside the first repeat have been shown to be required for the accurate duplication of the repeat (Wang et al., 2016). The authors suggested that these sequences act as docking sites for the adaptation complex, which then functions as a molecular ruler to make cuts at specific lengths from each motif.

5.1.4 Similarities with viral integrases

As adaptation involves the integration of foreign DNA into a host genome, the activity of the Cas1-Cas2 complex has been compared to that of viral integrases. One common feature is the processing of 3' ends of the protospacer DNA before insertion by Cas1, which was recently demonstrated by Wang and colleagues (Wang et al., 2015). Before integration, retroviral integrases process duplex DNA for insertion by nicking each strand at conserved 3' CA dinucleotides (Brown et al., 1989). Viral integration also requires the integrase to cut, in a staggered fashion, the host genome on complementary strands with a five nucleotide interval and join viral DNA ends by transesterification (Brown et al., 1989). In common with Cas1, this transesterification reaction does not require an external energy source and is catalysed by the same active site of the integrase responsible for 3' end processing (Engelman et al., 1991).

5.1.5 Disintegration

It is well documented that many viral integrases perform both integration, which incorporates viral DNA into the host genome, and also the reverse of this reaction, termed the disintegration reaction (Chow et al., 1992). The forward reaction consists of two strand-joining transesterification reactions, whereas the disintegration reaction is the reversal of one of these half-site reactions. Disintegration results in a nicked-duplex with a 5'-flap of viral DNA being converted back into an intact duplex with the release of the viral DNA flap. There is as yet no evidence of, or purpose for, this reaction *in vivo*. However, when provided with integration intermediates, in an isolated *in vitro* situation, viral integrases perform a robust disintegration reaction (Chow et al., 1992).

As the disintegration reaction takes place in the same active site of the integrase protein as the forward reaction, both will share similar substrate specificities. This has been shown experimentally for the HIV-1 integrase, as disintegration substrates with sequences and structures matching those of viral integration intermediates

gave rise to the most efficient disintegration reactions, although many substrates supported disintegration to some degree (Chow et al., 1992). Disintegration is often much more efficient than the integration reaction *in vitro* and, due to this attribute, disintegration is often used as a means to study the requirements and inhibitors of the forward reaction.

In this chapter it is demonstrated that, in common with viral integrases, both *S. solfataricus* and *E. coli* Cas1 proteins perform robust disintegration reactions. I show that Cas2 is not needed for, and does not influence, the disintegration activity of Cas1, suggesting a structural, rather than catalytic role for Cas2 in integration in *S. solfataricus*. Finally, I use the disintegration reaction to probe the specificity of Cas1 and reveal that preferred sequences mirror those of the *in vivo* integration sites in the corresponding CRISPR locus of the host. I conclude that investigating the disintegration reaction of Cas1 can yield important insights into the sequence specificity and mechanism of the forward reaction and through this help to delineate the process of adaptation.

5.2 Results

5.2.1 Cas1 has a high affinity for branched DNA

Cas1 from *E. coli* has previously been found to bind and cleave DNA substrates with a branched nature (Babu et al., 2011). In this study, branched structures with 5' flaps were found to lead to very efficient cleavage by Cas1 (Babu et al., 2011). In order to investigate whether the Cas1 from *S. solfataricus* shared a similar substrate preference I designed and purified a variety of forked substrates and, firstly, tested the binding affinity of Cas1 for these structures using EMSA (Figure 5.2). As only one of the two Cas1 proteins from *S. solfataricus* is used in this chapter, the Cas1_{CD} protein will be referred to as SsoCas1, and the *E. coli* Cas1 as EcoCas1. The top two panels of Figure 5.2 compare the binding affinity of SsoCas1 for three branched DNA structures and a nicked duplex DNA. The apparent affinity of SsoCas1 was highest for the branched structure with a 5' single-stranded flap, followed by the nicked Y-junction, with a complete substrate shift observed with 500 nM SsoCas1. SsoCas1 bound both the complete Y-junction and the nicked duplex DNA with a much lower affinity, with only the highest concentrations of protein (500 -1000 nM) leading to a substrate shift. The binding affinity of EcoCas1 was also highest for the

single-stranded flap substrate, whereas the three other structures were bound with a comparable, lower affinity by this Cas1.

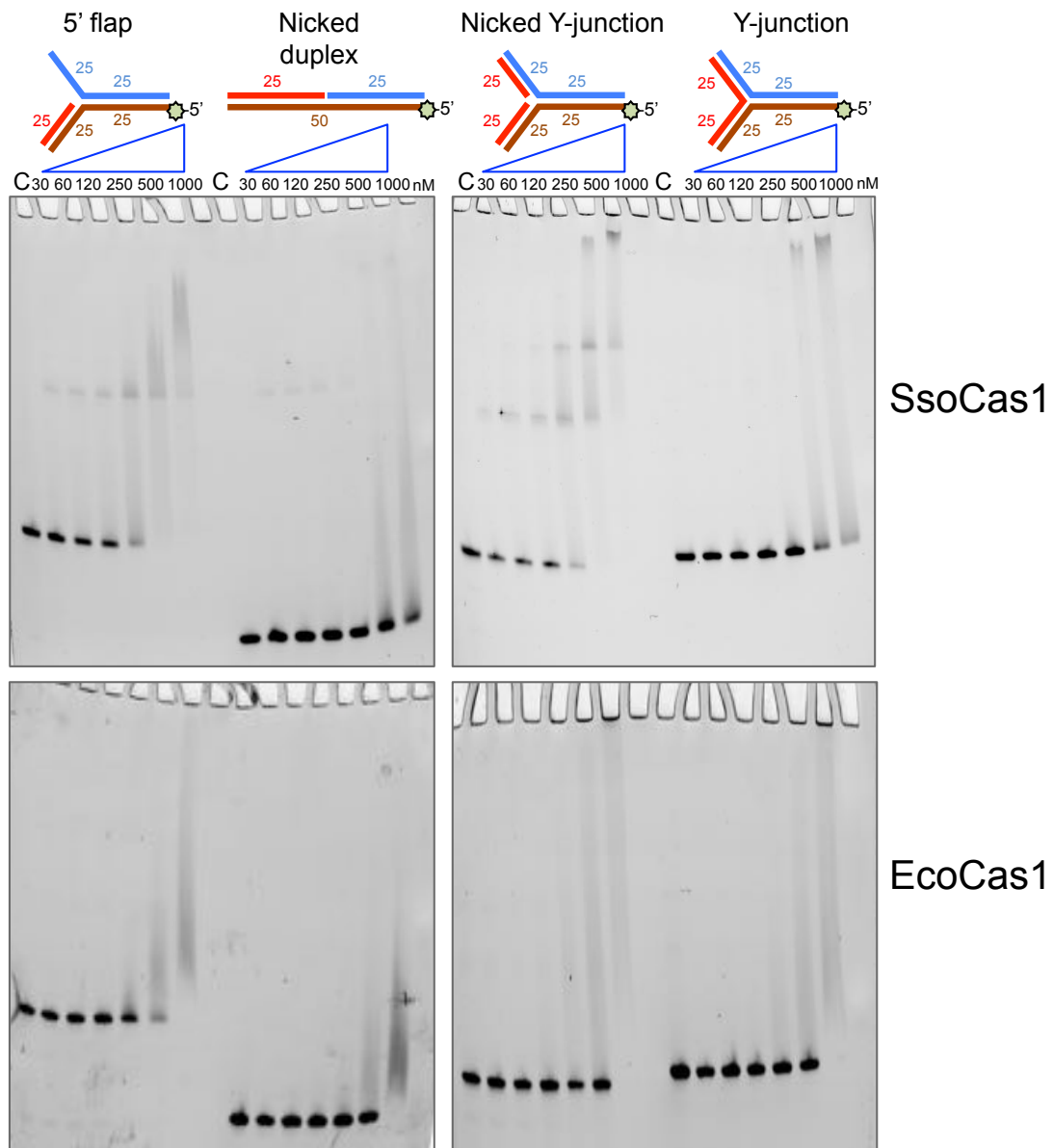


Figure 5.2 Cas1 binds branched DNA structures

The left panels show the result of an EMSA of Cas1 (SsoCas1 top, EcoCas1 bottom) binding to a branched structure containing a single-stranded 18 nucleotide flap or a nicked duplex DNA (substrate 19 or Nicked-19, see Table 2.2 for component oligonucleotides and Table 2.1 for sequences). The right-hand panels show the binding of Cas1 to a nicked or complete Y-junction (Nicked-Y and Y-junction substrates, see Table 2.2 for component oligonucleotides and Table 2.1 for sequences). The first lane of each gradient is a control without protein, followed by a concentration gradient of Cas1 from 30 nM to 1000 nM. All substrates were based on three 50 nt sequences, with strand lengths or number being altered to make the different DNA conformations. Substrates were labelled with a 5' fluorescein label on the continuous brown strand. Substrate (200 nM) and protein were incubated together at room temperature for 20 min before separation on a 12% native polyacrylamide-TBE gel.

As discussed in Chapter 4, SsoCas1_(CD) has a much higher affinity for single-stranded, compared to double-stranded DNA. The 5' single-stranded flap substrate here is also bound with the highest affinity by both Sso- and EcoCas1. However, the differences observed here did not seem to be solely accounted for by the single-stranded nature of the substrate, as SsoCas1 also bound to the nicked Y-junction with a high affinity. Branched structures with flexibility around the junction point (5' flap and nicked Y-junction substrates) led to tight binding by the SsoCas1 protein. In contrast, inflexible structures with a defined angle at the branch point and no single-stranded regions, such as the Y-junction, led to reduced binding by Cas1. It may be that flexibility at the branch point is required for DNA arms to access binding pockets of the Cas1 proteins, or that structures with a nick at the branch point allow Cas1 to open the duplex and access its favoured single-strand DNA substrate.

This preference for branched DNA over duplex DNA agreed with previous work of Babu and colleagues, which found that EcoCas1 cleaved forked structures and Holliday junctions (Babu et al., 2011). An intermediate of the integration of foreign DNA may resemble these forked structures with an attacking DNA end joined to the CRISPR integration site, which would explain the high affinity that Cas1 shows for these structures. However, this preference could also be relevant during the capture of new spacers. Recently the protospacer bound by the Cas1-Cas2 complex for integration was shown to have splayed single-stranded ends (Wang et al., 2015), which may help to explain the preference observed here.

5.2.2 Cas1 performs disintegration on branched DNA

Studies of Cas1 have found it to be a DNA nuclease that cuts single- and double-stranded as well as branched structures and Holliday junctions. In contrast to these studies, in Chapter 4 it was shown that little nuclease activity was observed for SsoCas1 on Holliday junctions or linear DNA. Given the high-affinity binding of SsoCas1 to forked structures with a 5' single-stranded flap, and the reported cleavage of these substrates by EcoCas1, I went on to test the nuclease activity of SsoCas1 on similar substrates.

The results of a nuclease assay in which Cas1 was incubated with a branched structure, containing a 5'-flap, in the presence of divalent metal ions are shown in Figure 5.3 (see section 2.2.8.2.1, (p. 66) for method). When the 5' end of the 18 nucleotide single-stranded flap was labelled, a smaller product was observed in

lanes containing Cas1 (Figure 5.3, B). The size of this product was identified using a Maxam-Gilbert A+G ladder and it was found to correspond exactly to the length of the excised 18 nucleotide single-strand flap.

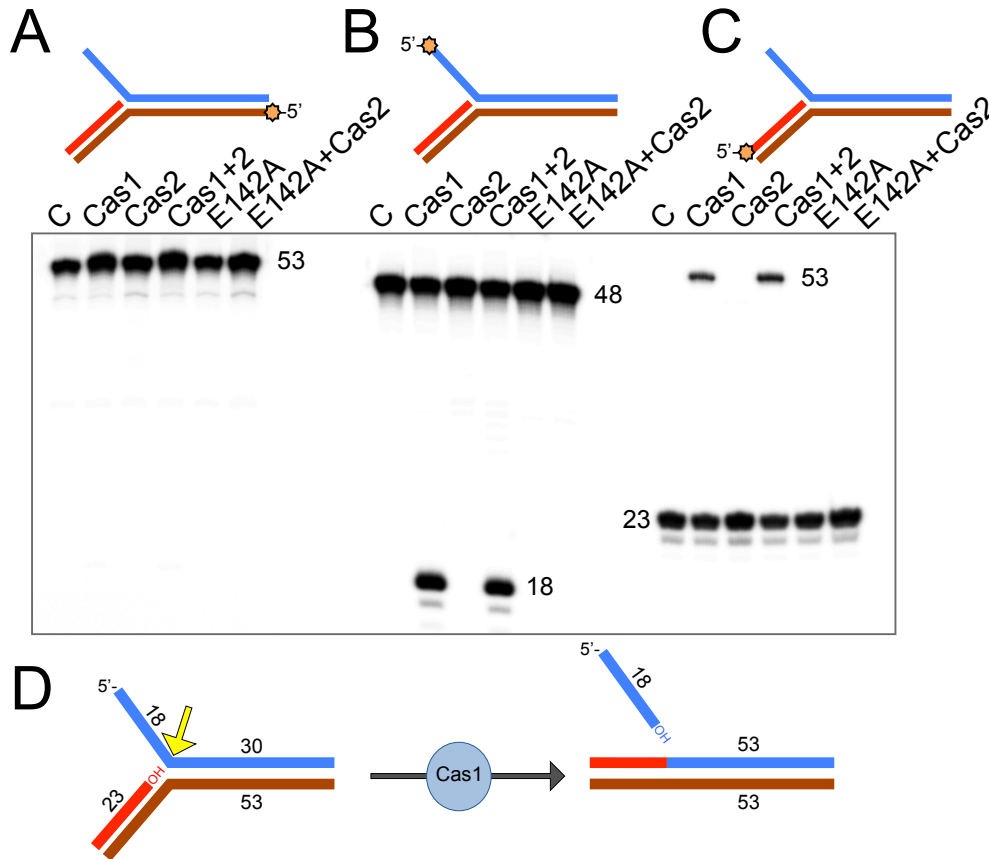


Figure 5.3 SsoCas1 activity on branched DNA

The products of an activity assay with Cas1 and a branched DNA substrate with a single-stranded flap were separated by denaturing gel electrophoresis. **A.** No nuclease products were observed when the substrate (substrate 1, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) was labelled with a 5'-³²P (orange star) on the continuous bottom strand (brown). The first lane of each assay is a control without protein (C); subsequent lanes show incubations with SsoCas1, SsoCas2, SsoCas1 and SsoCas2, active site SsoCas1 variant E142A, or E142A and SsoCas2. **B.** When the 5' end of the strand with a 5'-single-strand flap (blue) is labelled, an 18-nucleotide product is observed after incubation with SsoCas1 and SsoCas1 and Cas2. **C.** Labelling the 23-nucleotide strand (red) that presents a 3'-hydroxy at the junction point, produced a larger 53-nucleotide product on incubation with SsoCas1 and SsoCas1 and Cas2. **D.** Model scheme for the reaction showing a Cas1 mediating a transesterification reaction, in which the 3'-hydroxyl of the upstream strand attacks a phosphodiester bond at the branch point and leads to the excision of the 5' flap and the joining of the remaining DNA ends.

To follow up this finding, the remaining strands in the structure were labelled and the activity of SsoCas1 on each of these substrates was examined. Labelling the bottom strand and assaying with Cas1 did not yield any nuclease products (Figure 5.3, A). However, incubations with the short 23-bp strand labelled, produced

products that migrated more slowly through the denaturing polyacrylamide gel (Figure 5.3, C). This indicated that they were bigger in size than the substrate and I hypothesised that they were the products of a transesterification reaction performed by Cas1 in which the 3'-OH of the short 23 bp strand attacked the 5' flap at the branch point, leading to the excision of the flap and the joining of the remaining nicked DNA duplex (Figure 5.3, D).

In order to confirm that a transesterification, not a nuclease, reaction was taking place, the branched substrate was labelled on the 3' end of the strand containing the 5' flap. On assaying this substrate with SsoCas1, no nuclease products were observed (Figure 5.4, A), only transesterification products that corresponded to the joining of the 23 bp strand with the 30 bp duplex and the concomitant release of the 14 nucleotide flap.

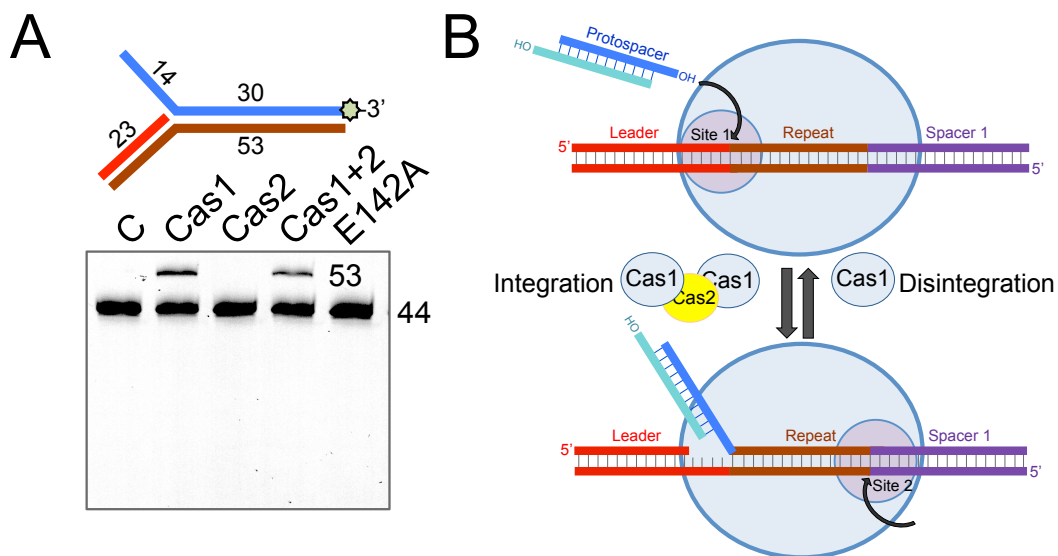


Figure 5.4 SsoCas1 performs a disintegration reaction on branched substrates

A. Labelling the disintegration substrate (substrate 1-FAM, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) with a fluorescein dye (green) on the 3' end of the strand containing the 5' flap (blue) led to the formation of one 53 nt product on incubation with SsoCas1. **B.** A model is shown for both integration and disintegration by Cas1. Adaptation involves addition of a spacer to the CRISPR locus through two half-site integration reactions. Cas1 also performs the reversal of one of these reactions, called a disintegration reaction. Disintegration is the processing of substrates with a 5' flap, to release the flap and rejoin the remaining DNA ends. In the assay shown in **A**, the 5' flap was single-stranded, however it is predicted that protospacer inserted *in vivo* are at least partially double-stranded, therefore the protospacers in the model (**B**) are depicted as partial duplexes. The reverse reaction only requires Cas1, whereas the forward integration requires a complex of Cas1-Cas2.

The transesterification reaction found to be performed by SsoCas1 on branched DNA is the reverse of integration, termed a disintegration reaction. In the forward reaction the Cas1-Cas2-protospacer complex will cut the host CRISPR at the 5' ends of the first repeat and join these ends to the 3' ends of the incoming protospacer in a coordinated transesterification reaction. The disintegration reaction shown here corresponds to the reversal of one of these half-site reactions, in which an integration intermediate is converted back into an intact duplex DNA and a free attacking protospacer DNA (Figure 5.4, B). In the disintegration reactions shown in Figure 5.3 and Figure 5.4, the 5' flap, equivalent to incoming protospacer DNA, is single-stranded in nature. However, for full integration *in vivo* it is predicted that the incoming protospacer must be at least partially double-stranded; therefore, in the model of integration/disintegration shown in Figure 5.4 (B) the 5' flap is shown as a partial duplex. This disintegration activity is abolished by the mutation of the SsoCas1 active site glutamate 142 to alanine (E142A variant), showing that the previously identified active site of Cas1, crucial for integration, is also responsible for this activity (Figure 5.4, A).

While the work in this chapter was being carried out, another group reported disintegration of branched structures by the EcoCas1 protein (Nuñez et al., 2015b), confirming that disintegration is an activity shared by Cas1 proteins from different CRISPR-Cas subtypes. Disintegration by Cas1, as for viral integrases, is not thought to happen *in vivo* or to play any part in CRISPR adaptation. However, as disintegration was reported to be more efficient *in vitro* than the forward reaction (Chow et al., 1992) and occurs in the same protein active site, it provides a simplified method to examine the requirements of integration by Cas1.

Although Cas1 and Cas2 are required for integration in *E. coli* (Yosef et al., 2012), disintegration only required the *S. solfataricus* Cas1 protein, and Cas2 did not modify or enhance the reaction. Recent structural studies have shown that *E. coli* Cas2 is required to bridge two Cas1 dimers, helping to position the active sites of Cas1 subunits at the 3' ends of the 33 nucleotide protospacers (Nuñez et al., 2015a; Wang et al., 2015). Furthermore, the active site of Cas2 was shown not to be required for integration of new spacers (Nuñez et al., 2014). Therefore, it was concluded that the *E. coli* Cas2 may play an architectural, rather than catalytic, which is potentially important for the coordination of the two half-site reactions needed for the complete integration of a spacer (Nuñez et al., 2015a). Therefore, it

makes sense that for the disintegration observed here, which mimicked a half-site integration intermediate, Cas2 had no part to play and Cas1 was able to bind and process the substrate independently.

5.2.3 Disintegration occurs precisely at the branch point

The next step was to confirm that the transesterification reaction was indeed taking place precisely at the branch point. To achieve this, a *SacI* restriction site was added so that it spanned, and was interrupted by, the branch point. After assaying this substrate, labelled on the 5' end of the single-strand flap, with the SsoCas1 enzyme, *SacI* was added to the reaction products. In Figure 5.5 (left panel), the transesterification (TES) product produced as a result of the Cas1 activity on this modified substrate is shown.

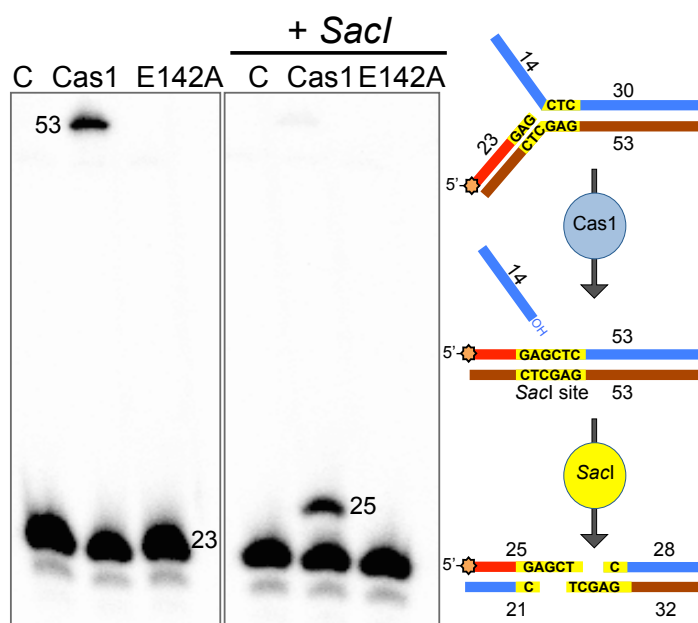


Figure 5.5 Disintegration occurs at the branch point

A disintegration assay was performed with SsoCas1 on a substrate with a nicked *SacI* site spanning the branch point (*SacI* substrate, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). Disintegration by Cas1 led to the formation of a 53 nt transesterification product (left panel). Incubation of the disintegration reaction products with *SacI* resulted in the disappearance of the 53 nt product and the formation of a 25 nt species (right panel). The first lane of each panel is a control without protein. No activity was observed with active site variant E142A. A model reaction scheme is shown on the right.

On incubation with *SacI* this reaction product is completely digested to a 25 nt species (right panel). This indicates that the transesterification performed by Cas1 is indeed occurring precisely at the branch point, as it invariably led to the formation of an intact *SacI* site, which had previously been nicked by the junction. These

experiments established that the excision of the 5' flap by Cas1 only occurred as part of a transesterification, and not nuclease, reaction, in which both the excised flap and extended duplex (TES) products were formed. Therefore, in further investigations in this chapter I often use production of the excised flap product as an indicator that disintegration is taking place without changing the label position to identify the TES product.

5.2.4 Metal dependence

I continued to investigate the mechanism and requirements of the Cas1-mediated disintegration reaction, looking firstly at the metal-dependence of the reaction for two Cas1 proteins. In common with other activities reported for Cas1 enzymes, SsoCas1 requires divalent metal ions to support disintegration (Figure 5.6, A). The strongest activity was observed with manganese; however, cobalt and, to a lesser extent, magnesium also facilitated disintegration. No disintegration was observed with calcium ions or with EDTA.

The EcoCas1 protein was also tested for disintegration activity with this range of metal ions (Figure 5.6, B). From the results it is clear that EcoCas1 also performed an efficient metal-dependent disintegration reaction, supported by magnesium, manganese and cobalt divalent metal cations.

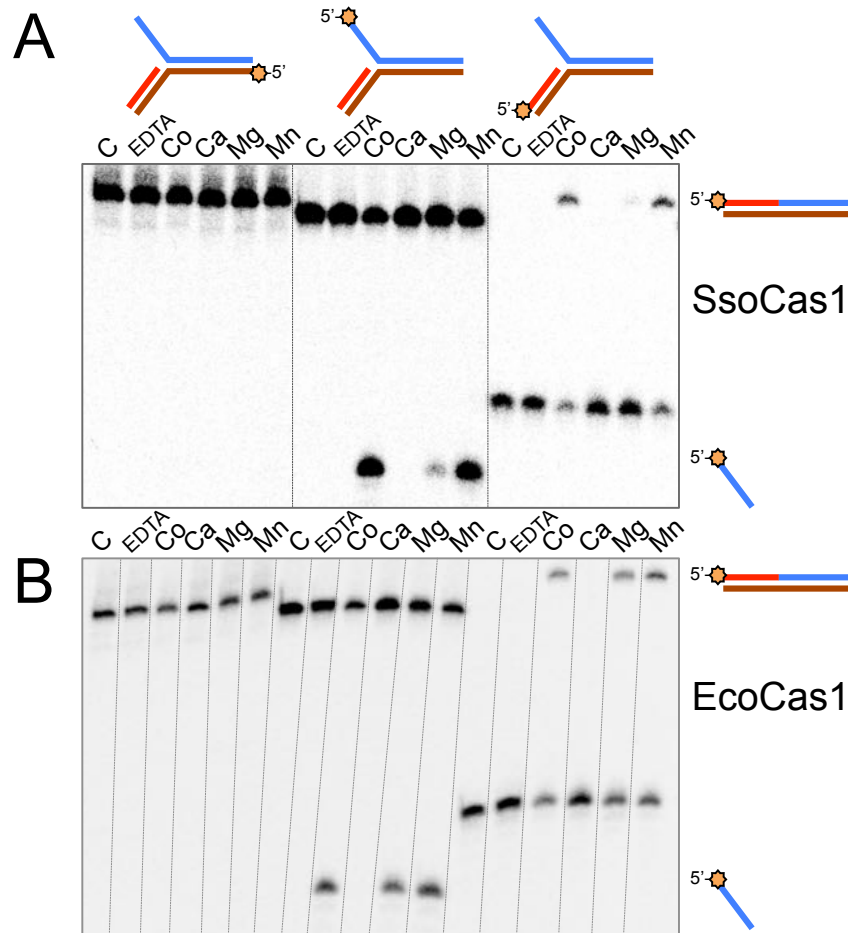


Figure 5.6 Metal-dependent disintegration by Sso- and EcoCas1

A. A range of divalent metal ions were added to an SsoCas1 assay for disintegration activity, with each of the three strands of the branched substrate (substrate 3, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) labelled with a 5'-³²P. The first lane of each assay is a control without protein, followed by incubations with 5 mM of EDTA, cobalt (Co²⁺), calcium (Ca²⁺), magnesium (Mg²⁺) or manganese (Mn²⁺). **B.** As for A except the protein used is the *E. coli* Cas1 protein.

5.2.5 Identifying the nucleophile in disintegration

It was hypothesized that the 3' hydroxyl (3' OH) at the branch point was acting as the nucleophile in this disintegration reaction. However, small polar molecules such as glycerol or water molecules are also common nucleophiles in DNA transposition events (van Gent et al., 1993). To understand the identity of the nucleophile in this transesterification reaction I exchanged the 3' OH at the junction point for a 3' phosphate. The branched substrate with this modification was no longer a disintegration substrate for Cas1 (Figure 5.7, B), indicating that the 3' OH does act as a nucleophile and is crucial for disintegration. In addition, when a gap of two nucleotides was introduced between the 3' OH and the junction point, disintegration activity was also abolished (Figure 5.7, A), confirming that Cas1 uses this 3' OH as

a nucleophile to attack the phosphodiester bond of the flap at the junction during disintegration. This suggests that during the forward reaction both the attacking hydroxyl of the incoming DNA and the phosphodiester bond to be attacked must be brought into very close proximity in the same active site of a Cas1 monomer, a theory that is now supported by the crystal structure of the protospacer-bound *E. coli* Cas1-Cas2 complex (Nuñez et al., 2015a).

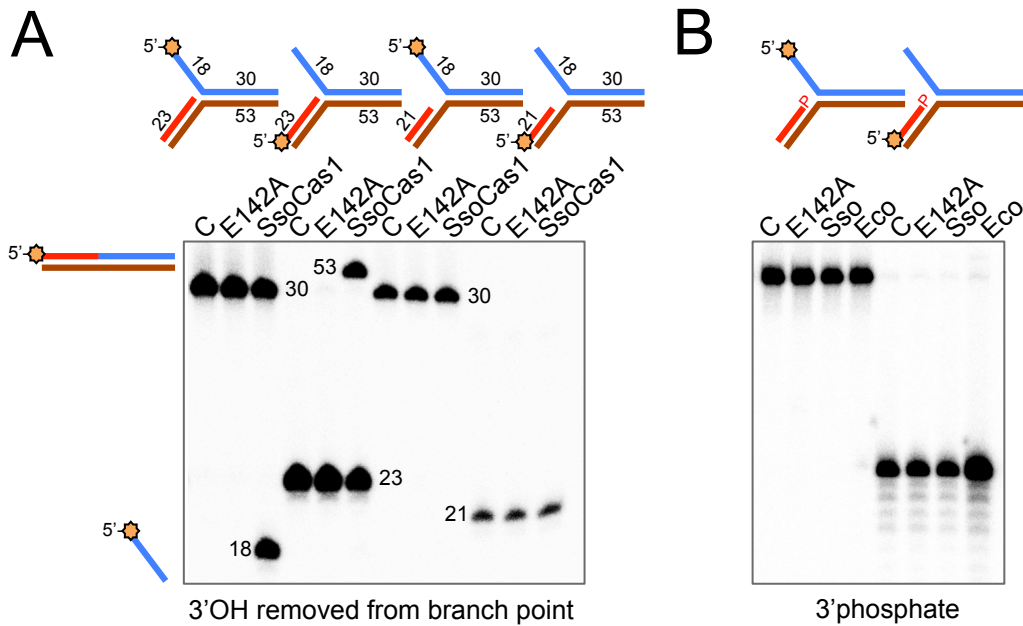


Figure 5.7 Requirements for disintegration

A. On the left, the standard disintegration substrate (substrate 1, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) is used in an assay with SsoCas1. Depending on the strand labelled, a nuclease or TES product (18 or 53 nt) is visible in lanes containing Cas1. The right side of the gel shows the effect of shortening, by removal of two nucleotides from the 3' end, the red strand ending in the 3' hydroxyl that mediates nucleophilic attack at the branch point (substrate 1-gap). The first lane of each assay is a control without protein (C), followed by a lane with products of active site variant Cas1 E142A. **B.** The effect on disintegration activity of both Sso- and EcoCas1 of exchanging the 3' hydroxyl at the branch point for a 3' phosphate (3'-phos substrate, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). The left-hand lanes show the results of labelling the strand with the 5' flap (blue) and the lanes on the right contain substrate labelled on the short strand (red) with the 3' end opposing the junction point. The first lane of each assay is a control without protein (C), followed by lanes containing active site variant Cas1 E142A, SsoCas1 and EcoCas1.

5.2.6 Concentration-dependent disintegration

As SsoCas1 (Cas1_{CD}) was found to aggregate at high concentrations during purification, a disintegration assay in which the concentration of Cas1 was increased from 5 nM to 1500 nM against a constant concentration of branched substrate (50 nM) was carried out. For SsoCas1 the optimal protein concentration was 250 nM

(5:1 ratio of protein:substrate) and above this concentration the activity was reduced by increasing protein (Figure 5.8, A). This result shows that high protein concentrations are auto-inhibitory, perhaps due to aggregation of protein and the formation of higher oligomers that are non-catalytic in these conditions. The activity of the EcoCas1 protein plateaued at 500 nM protein (10:1 ratio of protein:substrate), with only a slight decrease in product formation above this concentration (Figure 5.8, B). This difference between the two proteins is interesting; however, it may not be relevant *in vivo* as both Cas1 proteins are likely to be found in complex with Cas2, which has the potential to prevent oligomerisation.

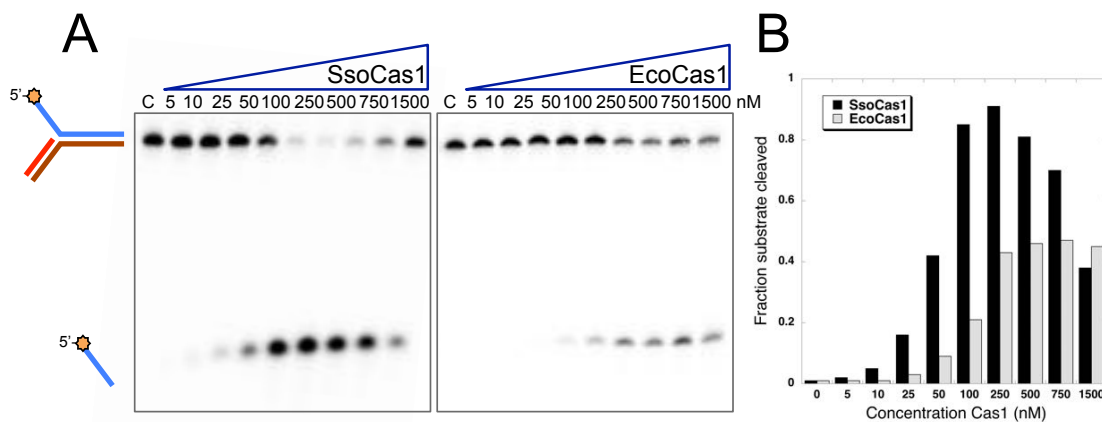


Figure 5.8 Concentration-dependent disintegration activity

A. Denaturing gels showing the effect of increasing SsoCas1 (left) or EcoCas1 (right) concentration on disintegration activity with substrate 3 (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). The first lane of each gradient is a control with no Cas1 protein (C), followed by a concentration gradient from 5 to 1500 nM protein. Results shown here are representative of duplicate experiments. **B.** Quantification of gels shown in **A.** The fraction of product cleaved at each concentration was calculated using Image Gauge software (FUJIFILM).

5.2.7 Clues to the nature of incoming DNA during integration

Considering the disintegration reaction as the reverse of the integration reaction (see proposed model in Figure 5.4, B), the 5' flap that is excised during the reaction corresponds to the incoming DNA in the forward reaction. Therefore, by modifying the 5' flap I hoped to gain insight into the structure and length of the incoming DNA during integration.

Firstly, increasing lengths of DNA were annealed to the single-strand flaps of disintegration substrates to assess whether the single-stranded nature of the flap was necessary for disintegration. All of these substrates supported disintegration

activity and no discernable difference in the efficiency of the reaction was noted for SsoCas1 (Figure 5.9). However, when a complete Y-junction was used as the substrate, disintegration activity was abolished, as there was no longer a free 3' OH available to initiate nucleophilic attack at the branch point.

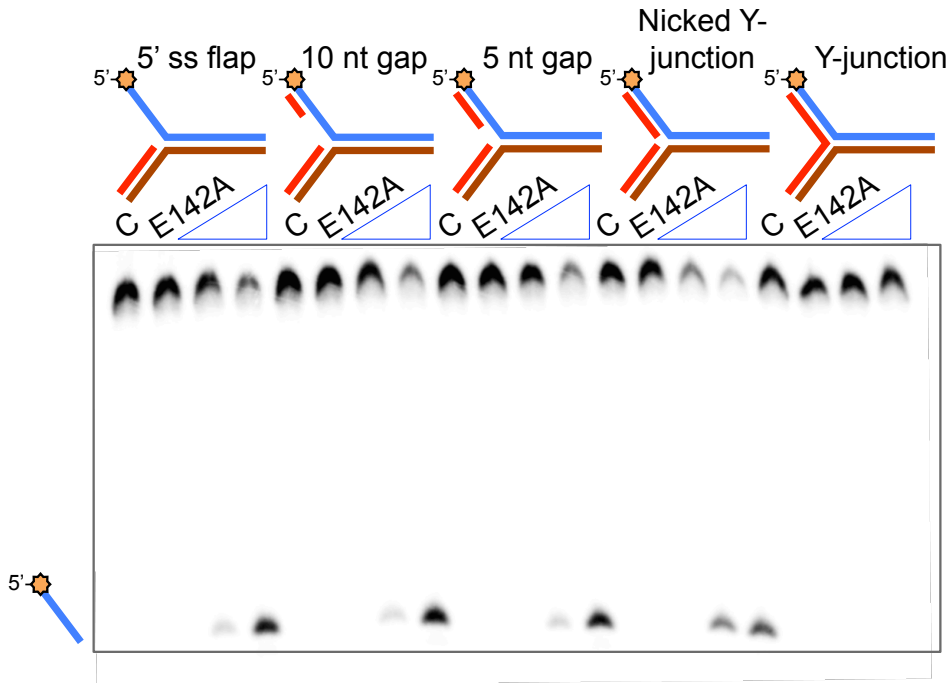


Figure 5.9 A double-stranded 5' flap supports disintegration by SsoCas1

The nature of the 5' flap of disintegration substrate 19 was altered by annealing complementary strands of varying lengths (to make gap5, gap10, nicked-Y and Y-junction substrates, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). SsoCas1 was assayed against substrates with flaps varying from fully single-stranded to completely double-stranded (left to right). A nicked and a complete Y-junction were also assayed against Cas1 for disintegration activity. The first lane of for each substrate is a control without protein (C), followed by active site variant E142A and lanes with increasing incubation times (5 and 20 min) with WT Cas1.

These data support the hypothesis that the incoming protospacer DNA during adaptation could be duplex DNA, at least in part. In the *E. coli* adaptation complex the incoming protospacer consists of a 23 bp duplex with single-stranded ends of 5 nucleotides, these data indicate that the same might be true for the *S. solfataricus* adaptation complex. While I did not observe a strong Cas1 preference for substrates with a five nucleotide single-strand region at the branch point, these substrates did undergo an efficient disintegration reaction with Cas1. The lack of discrimination between 5' flaps of varying structures may be testament to an opening of the whole junction structure on binding by Cas1, with interactions between the DNA and protein positioning the nucleophile and scissile bond in the correct position for cleavage.

The optimal protospacer substrate, that led to crystallisation of the Cas1-Cas2 complex with bound DNA, was a partial duplex of 23 bp with 5 nucleotide single-stranded 5' ends (Nuñez et al., 2015a). Another study also showed that integration progressed by the staggered nicking of the CRISPR locus and joining of protospacer ends to the 5' ends of the first repeat (Arslan et al., 2014). These studies provided evidence that protospacers are likely to be at least partially double-stranded, which complicates the theory that they are the products of the RecBCD repair complex in *E. coli* (Levy et al., 2015), as DNA fragments produced by this complex are single-stranded. If protospacers are produced by RecBCD, the single-strand by-products may re-anneal to their complementary sequence before capture and undergo processing to standardize the length of DNA inserted (Amitai & Sorek 2016).

5.2.8 Length of incoming DNA

According to our hypothesis that disintegration shares the same substrate requirements as the forward integration reaction, the length of the 5' flap required for disintegration should also correlate with protospacer length inserted during adaptation. To investigate this I modified the length of the 5' flap structure from 18 to 10 or 5 nucleotides. The transesterification reaction was reduced as flap length was shortened (Figure 5.10, A). Flaps of 10 nucleotides led to a weak disintegration reaction and the 5 nucleotide flap no longer supported disintegration. I tested flaps up to 25 nucleotides in length (those used in Figure 5.9 (substrate 19)) and these supported a robust transesterification by both SsoCas1 and EcoCas1 (Figure 5.9). These data fit well with the length of the incoming protospacer DNA during integration as protospacer average length in *S. solfataricus* is 39 nucleotides and 33 nucleotides in *E. coli*. The shorter lengths of DNA are perhaps insufficient for tight binding by Cas1, as I have shown that a branched substrate is necessary for tight binding (Figure 5.2) and transesterification activity.

Altering the disintegration substrate by replacing the strand presenting the 3' OH at the branch point with an RNA strand also abolished disintegration activity (Figure 5.10, B). This loss of activity is perhaps due to the RNA:DNA hybrid backbone interactions with Cas1 being less stable, due to interference by the 2'-hydroxyl group or a difference in curvature of the duplex, which would reduce binding and prevent transesterification. This result is not unexpected as the binding of SsoCas1 to

single-stranded RNA is much weaker than to DNA (see Chapter 4) and the *in vivo* substrate of Cas1 is not known to be RNA at any stage in the adaptation process.

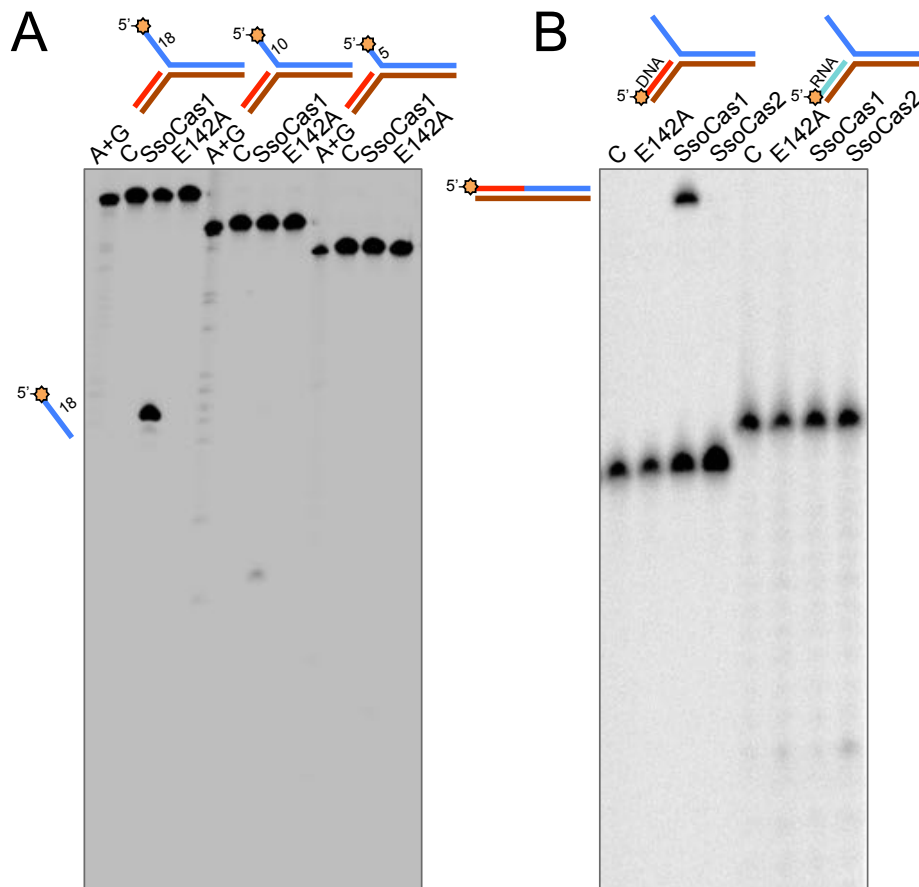


Figure 5.10 5' flap structure influences disintegration by Cas1

A. Shows the effect of shortening the 5' flap of disintegration substrate 1 from 18 to 10 or 5 nucleotides on disintegration by SsoCas1 (substrate1, substrate 1-10flap, substrate1-5flap, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). The first lane for each substrate is a Maxam-Gilbert A+G DNA ladder, the second is a control without protein, followed by lanes containing the products of assays with wildtype SsoCas1 and variant E142A. **B.** Shows the effect of exchanging the short DNA strand presenting a 3' OH at the branch-point for an RNA strand (substrate 1 and substrate 1-RNA, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). The ^{32}P label in this assay was added to the 5' end of the strand being altered (red for DNA or aqua for RNA); this means that the product of a successful disintegration will be a larger TES product. The first lane for each substrate is a Maxam-Gilbert A+G DNA ladder, C is a substrate control without protein, E142A is the active site variant of SsoCas1.

5.2.9 Disintegration by Cas1 can be used to form DNA dumbbells

Unusual disintegration DNA substrates were made by annealing a self-complementary single DNA oligonucleotide. These substrates had single-strand loops at either end of short duplex regions of varying length, and a 5' single-strand flap. Cas1 was able to perform disintegration on these substrates, evidenced by the

excision of the 5' labelled flap (Figure 5.11). The disintegration activity was reduced by shortening the duplex regions on either side of the flap.

The ability of Cas1 to perform transesterification on these substrates is an interesting finding as one of the products is a covalently-closed DNA dumbbell, which often prove very costly and time consuming to produce (Yu et al., 2015). These substrates have recently been highlighted as promising expression platforms, which could be utilised for gene therapy applications. These mini-vectors are a more attractive option for transfection than plasmid DNA because of their small size as well as posing fewer immune obstacles than viral vectors (Yu et al., 2015).

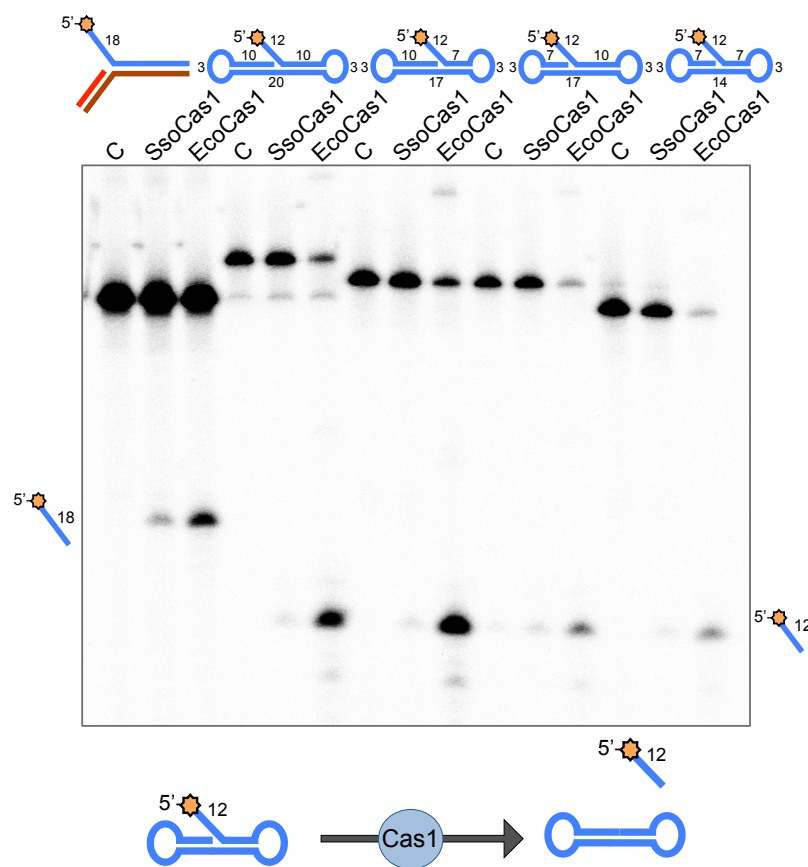


Figure 5.11 Disintegration by Cas1 produces DNA 'dumbbells' efficiently

Disintegration substrates with complementary arms, unpaired loop regions and a 5' flap were made from a single-strand of DNA and labelled on the 5' end with ^{32}P . Four of these substrates (dumbbell1, dumbbell2, dumbbell3 and dumbbell4, see Table 2.1 (p.47) for sequences) with differing length duplex arms were made and tested for disintegration activity by Cas1 alongside a standard disintegration substrate (substrate 1). The first lane for each substrate is a control without protein, followed by a lane containing SsoCas1 and a final lane containing EcoCas1. Reactions were carried out at 37 °C for EcoCas1 and 40 °C for SsoCas1. The temperature was lowered compared to the standard SsoCas1 assay conditions to prevent melting of the short duplex arms; this led to low disintegration efficiency for SsoCas1.

5.2.10 Residues required for disintegration

Site-directed mutagenesis of conserved residues outwith the active site (Figure 5.12, C) of SsoCas1 was carried out, in order to provide an insight into the DNA:protein interactions crucial for correct positioning and manipulation of the disintegration substrate. Mutagenesis was carried out by undergraduate student James Robson. Mutagenesis of residues R166, N175 and W150 to alanine greatly reduced, or completely abolished transesterification activity of Cas1 (Figure 5.12, A).

Residue R166 is situated in a flexible loop surrounded by positively charged amino acids; I hypothesized that this loop and cleft may be important for interaction with the negatively charged phosphate backbone of a DNA substrate. Mutation of the residue to alanine abolished transesterification activity, consistent with the idea that this residue is highly important for stabilising and correctly orientating the Cas1:DNA complex during the disintegration. This conclusion was recently confirmed by the solution of the crystal structure of the *E. coli* Cas1-Cas2 integration complex bound to protospacer DNA (Nuñez et al., 2015a). The corresponding residue, R163, was identified as contributing to an arginine channel, which interacts with the 3' single-stranded DNA overhang required for nucleophilic attack during integration. The fact that R166 lies in a loop region (Figure 5.12, B) may imply that there is some flexibility in position of this residue and that during integration, gross structural changes in the complex allow R166 to position the 3' nucleophile in the correct orientation to bring about the integration reaction.

Variant N175A retained some transesterification activity, suggesting that this residue may not be directly involved in stabilising the DNA:protein complex, but may act indirectly, perhaps orientating R166A into an optimal position for DNA binding. W150 lies in a positive binding cleft of Cas1 and is important for the disintegration reaction as mutagenesis to alanine greatly reduces activity. This residue may bind through base stacking to DNA during the disintegration reaction and stabilise the complex or reaction intermediates. From the *E. coli* structure this residue does not seem to be directly involved in protospacer binding (Nuñez et al., 2015a); rather, it may play a part in binding to the host genome and bringing the phosphate bond at the integration site into close proximity to the 3' hydroxyl of the protospacer.

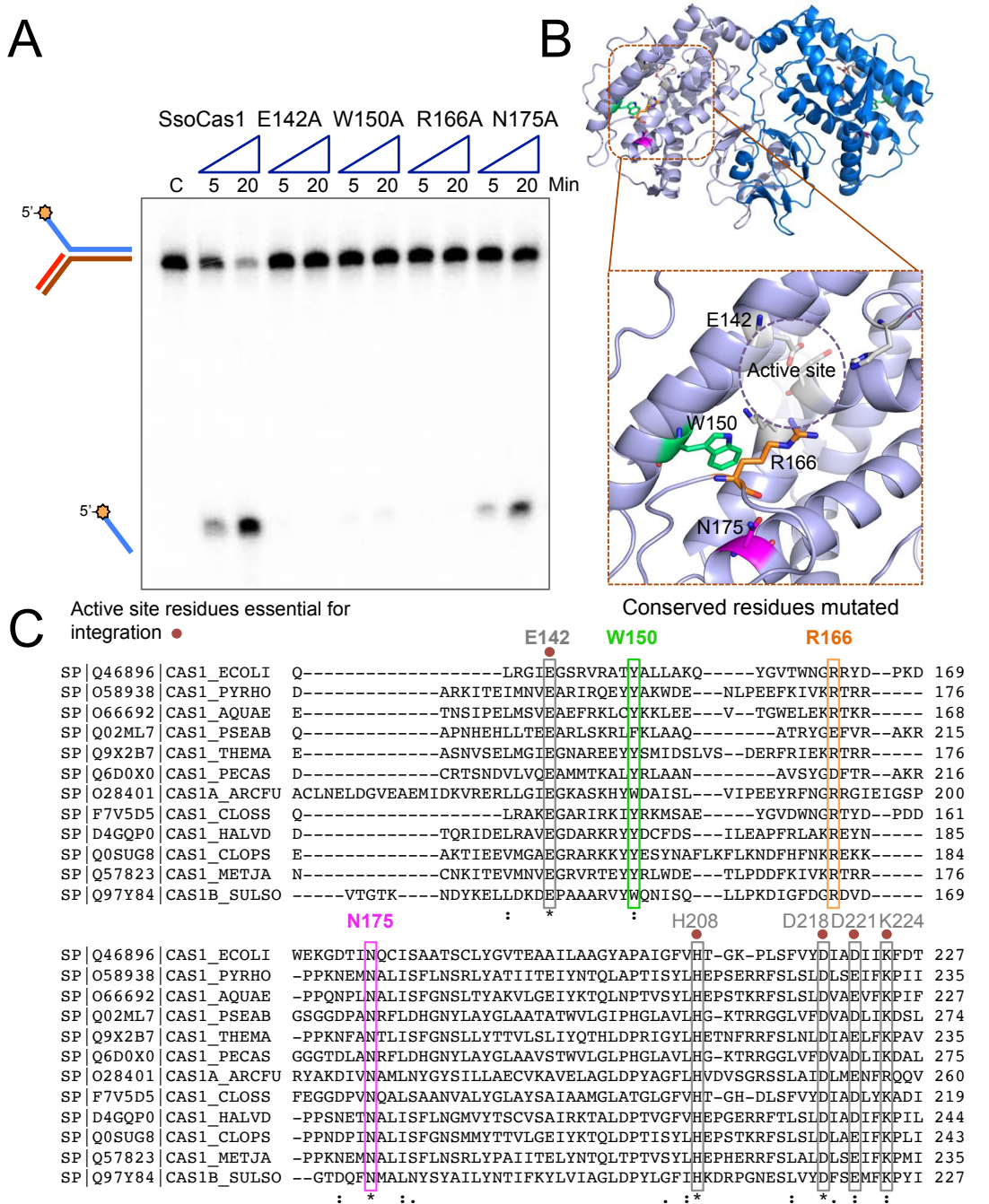


Figure 5.12 Conserved residues required for disintegration

Canonical active site residue E142 and three other residues (W150, R166, N175) conserved between Cas1 proteins, but outwith the active site, were mutated to alanine. **A.** Disintegration assay for 5 and 20 min with substrate 1 (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) to compare disintegration activity of variants compared to the wildtype protein. The first lane is a control without protein (C). **B.** Model structure of the SsoCas1 dimer (top) with a closer view of residues chosen for mutation (bottom). Conserved catalytic residues (E142, H208, D218, D221 and K224) are shown in grey and conserved residues outwith the identified catalytic site are shown in green (W150), orange (R166) or magenta (N175). **C.** Alignment of protein sequence around Cas1 active sites with residues chosen for mutation highlighted in colour and those in the already identified catalytic site shown in grey. A red dot indicates that the residue has already been shown to be essential for integration in *E. coli*.

5.2.11 Sequence specificity of Cas1 in the disintegration reaction

The integration reaction performed during adaptation of the CRISPR array *in vivo* must be highly sequence-specific as new spacers are always incorporated at the same position between the leader and first repeat. It is still not understood how the Cas1-Cas2 complex recognises this integration site. It has been suggested that two short recognition motifs in the repeat guide docking (Wang et al., 2016). Furthermore, modification of the leader and repeat 1 junction sequences leads to aberrant integration or the absence of integration (Yosef et al., 2012; Wei et al., 2015). I was interested in investigating whether Cas1 alone has a sequence specificity that may direct integration. I predicted that as the disintegration reaction is the reverse reaction to that performed during adaptation and both occur in the same active site, I would be able to detect any sequence specificity Cas1 imposes on the integration reaction by studying the reverse reaction.

In vivo integration can be divided into two half-site reactions that join each 3'-hydroxyl of the incoming spacer to the 5' ends of the first repeat. The leader-proximal 5' end of the repeat is referred to as 'site 1' and the leader-distal 5' of the first repeat as 'site 2'. A key remaining question pertaining to the integration reaction was whether these sites were targeted sequentially and, if so, in what order are they selected by the adaptation proteins. Furthermore, there are few data available to predict whether both, or just one, of these sites drives the site specificity of integration. However, as the sequence of the site 1 leader-repeat junction does not change with integration of new spacers, it seemed a more logical recognition motif for the adaptation machinery. By using disintegration substrates, which mimicked the product of one of these half-site integration reactions, and by changing nucleotides around the junction point, I hoped to be able to reveal any sequence specificity of Cas1 and use this to learn more about how spacer integration is targeted during adaptation.

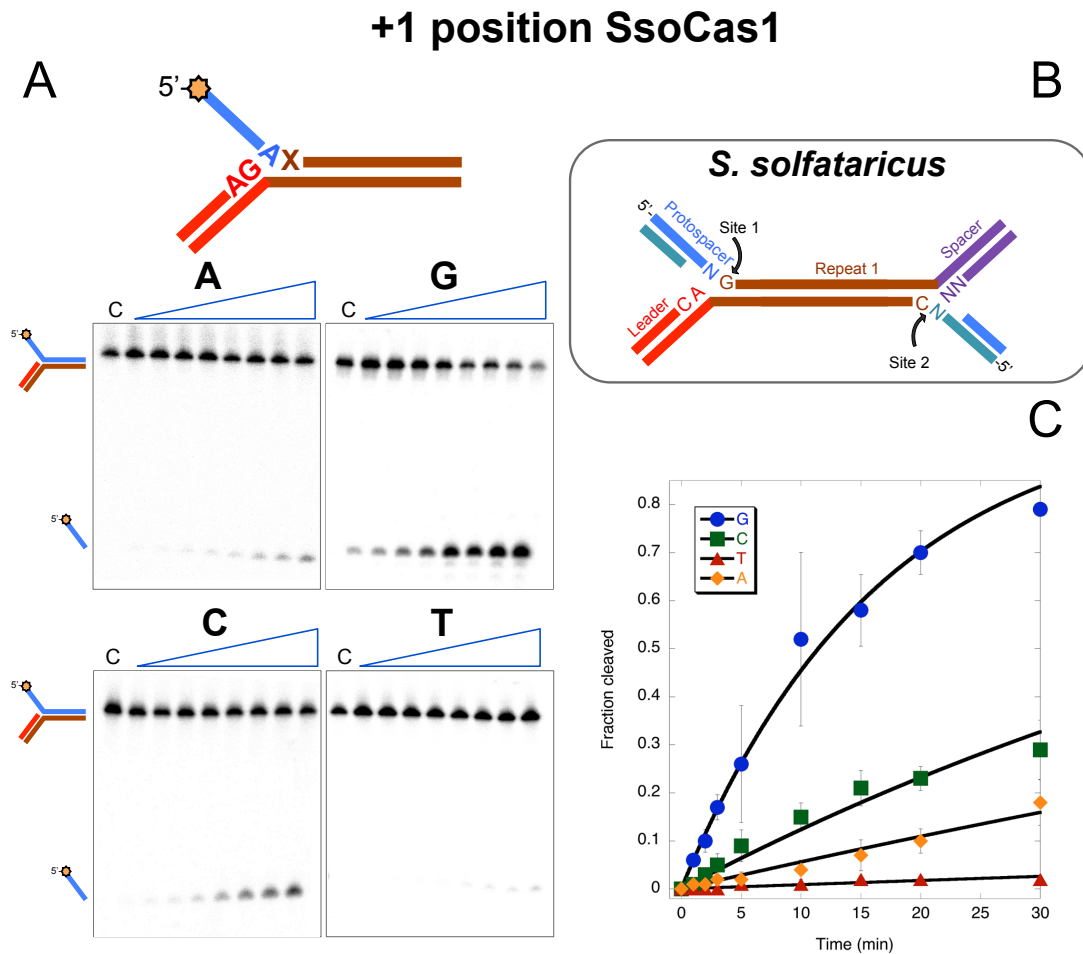


Figure 5.13 The +1 position influences disintegration activity by SsoCas1

A. The nucleotide at position +1 was varied and the substrates (substrate 8 (top left), 3 (top right), 6 (bottom left), and 7 (bottom right)) (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) assayed with SsoCas1 over a time course from 30 sec to 30 min. The first lane of every time course is a control without protein. The structure and nucleotides surrounding the junction of the substrate used in the assay are shown, X indicates the +1 nucleotide that was varied in these assays. **B.** Schematic of the two potential integration half-sites, site 1 and site 2 in the *S. solfataricus* CRISPR array. The nucleotides surrounding the sites in the *S. solfataricus* CRISPR C array are shown; the +1 nucleotide is a G at site 1 or a C at site 2. Incoming DNA is shown as being partially duplex, although the disintegration substrates used in these assays had 5' single-stranded flaps. **C.** The fraction of substrate processed by Cas1 in the disintegration assays shown in A was quantified and plotted against time. The points are representative of the mean of triplicate experiments and the standard deviation of the mean is indicated by error bars. The data were fitted to a single exponential equation (Equation 4).

5.2.11.1 +1 position

The first position I examined was the nucleotide at +1 position; this is the nucleotide to which new spacers are joined in the forward reaction (the 5' nucleotides of the repeat) or the last nucleotide before the single-strand flap in the disintegration substrate, indicated by an X in Figure 5.13 (A). The nucleotide at this position was varied to either A, G, C or T in four disintegration substrates and triplicate time

course disintegration assays with either SsoCas1 or EcoCas1 were carried out (see section 2.2.8.2.3 (p. 67) for method). The fraction of substrate which had undergone disintegration was quantified at each time point and the progress of the reaction was plotted against time, with curve fitting to the single-exponential Equation 4 with either a floating (EcoCas1) or fixed (SsoCas1) curve maximum amplitude (Niewoehner et al., 2014).

$$\text{Fraction cleaved} = A(1 - \exp(-kt))$$

Equation 4 Single exponential for reaction rate calculation

A=curve amplitude (which is fixed at 1 for SsoCas1 calculations and floating for EcoCas1 calculations); t=time; k=1st order rate constant

For SsoCas1, the processing of substrates with a G at position +1 was most efficient, with a C in this position also supporting activity (Figure 5.13, A). The reaction was much weaker with an A, and especially a T, at position +1. These observations were confirmed following the fitting of data to a single-exponential (Equation 4), which produced rates of 0.06 (G), 0.013 (C), 0.0058 (A) and 0.0009 (T) min⁻¹ (Figure 5.13, C).

For EcoCas1 there is an even more pronounced preference for a G at position +1 with all other nucleotide substitutions leading to a much-reduced activity (Figure 5.14, A and C). As the reaction does not go to completion, the fitting of the EcoCas1 data points was poor when the endpoint was fixed at 1. Therefore a floating endpoint was used in Equation 4, to allow curve fitting. A 10-fold higher reaction rate was observed with a G at position 1+, compared to any other nucleotide at this position (Figure 5.14, C).

As the disintegration reactions with SsoCas1 or EcoCas1 do not go to completion, the rates calculated are not accurate and cannot be used to draw comparisons with other proteins. However, they are sufficient for the internal comparison of the efficiency of disintegration by Cas1 and clearly demonstrate that there is an effect of altering the nucleotide at position 1, with a G being the most favoured residue at this position for both SsoCas1 and EcoCas1.

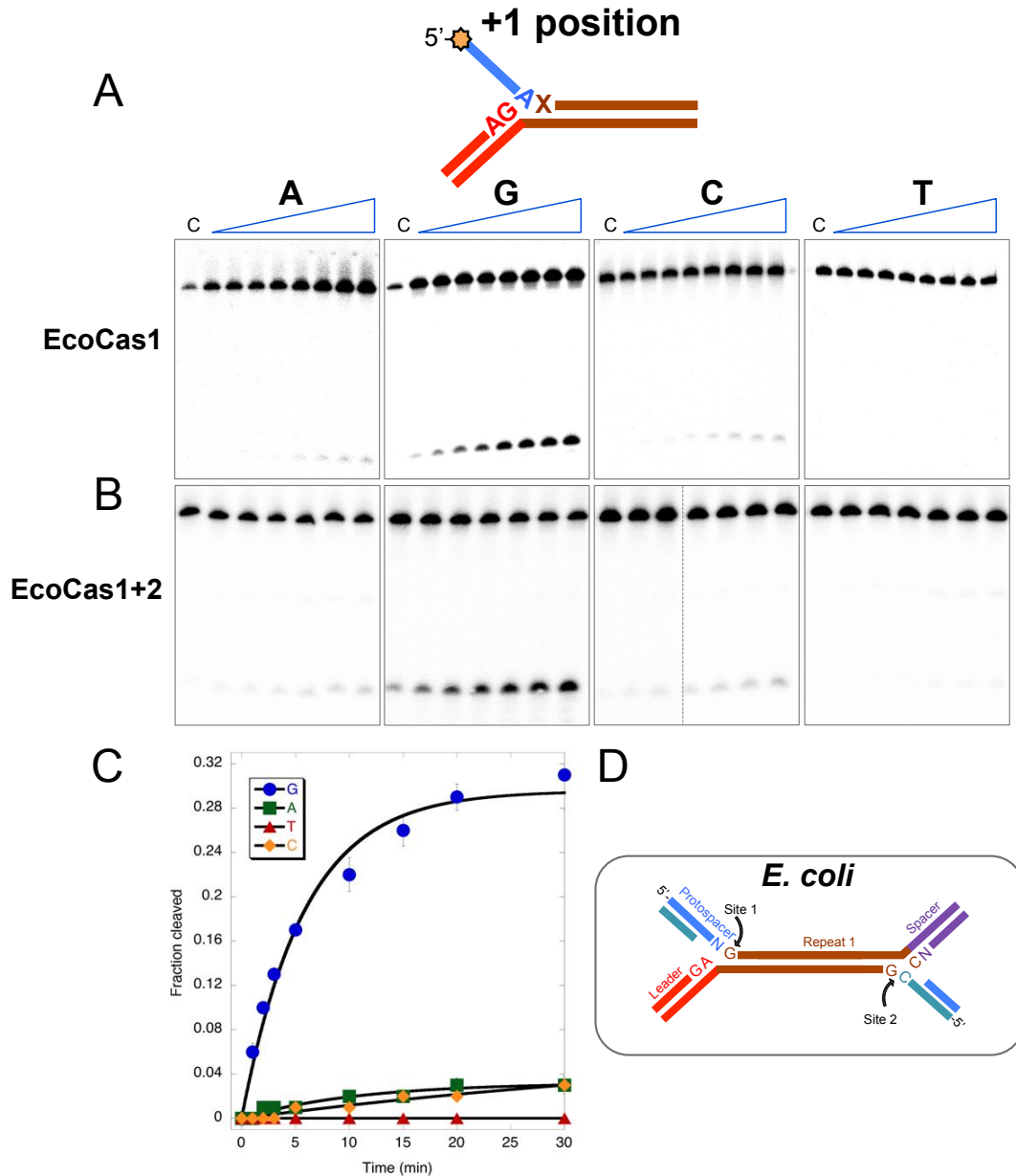


Figure 5.14 Sequence specificity for the +1 position of EcoCas1

A. The nucleotide at position +1 was varied and the substrates (from left to right - substrate 8, 3, 6, and 7) (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) assayed with EcoCas1 over a time course from 30 sec to 30 min. The first lane of each time course is a control without protein. The structure and nucleotides surrounding the junction of the substrate used are shown; X indicates the +1 nucleotide that was varied.

B. The assay in A was repeated with both EcoCas1 and EcoCas2 in the reaction (500 nM equimolar concentrations), to assess the effect of Cas2 on specificity of Cas1 for disintegration substrates. The dashed line indicates that the time course assay was split across two imaging plates and the result is a composite image.

C. The fraction of substrate processed by EcoCas1 disintegration in A was quantified and plotted against time. The points are representative of the mean of triplicate experiments and the standard deviation of the mean is indicated by error bars. The points were fitted to a single-exponential equation with a floating end point (Equation 4).

D. Schematic of the two potential integration half-sites, site 1 and site 2 in the *E. coli* K12 CRISPR array. The nucleotides surrounding the sites are shown: the +1 nucleotide is a G at site 1 and a G at site 2, as the first nucleotide of the new spacer makes up the last nucleotide of the duplicated repeat in *E. coli* adaptation.

The specificity observed for the +1 position during the disintegration reaction fits well with the sequence of the *bona fide* integration site for both *E. coli* and *S. solfataricus*. During adaptation, new spacers are always joined to the 5' nucleotide at either end of the first repeat in *S. solfataricus*, which is a G at site 1 or a C at site 2 (Figure 5.13, B). For *E. coli*, at site 1, the +1 nucleotide is a G. At site 2, the first nucleotide of the new spacer makes up the last nucleotide of the repeat, with the new spacer being joined to the penultimate nucleotide of the repeat, which is also a G (Figure 5.14, D) (Swarts et al., 2012). For SsoCas1 the rate of disintegration was highest with either a G or C at position +1 and for EcoCas1 the highest rate was obtained with a G at this position. Therefore, it seems that Cas1 harbours a sequence specificity for these disintegration substrates, varying at the +1 position, which matches that of the *in vivo* integration site during adaptation.

In order to examine if Cas2 had any effect on the specific disintegration observed, EcoCas1 and EcoCas2 were mixed for 30 min before being added to a disintegration time course assay with substrates differing at position +1 (Figure 5.14, B). The same preference was observed for the nucleotide at the +1 position with or without the addition of Cas2, with a G at this position being favoured over all other nucleotides. This finding indicated that the formation of an adaptation complex of Cas1 and Cas2 is not required for this sequence discrimination, and it may imply that Cas1 alone is responsible for identifying the conserved sequence elements of the integration sites in the forward reaction.

5.2.11.2 -1 position

The nucleotide at the -1 position in the disintegration substrate is the nucleotide that presents a 3' hydroxyl at the branch point. In the forward reaction -1 corresponds to the last nucleotide in the leader at site 1, or the first nucleotide of the previously integrated spacer at integration site 2 in *S. solfataricus*. In *E. coli* the -1 nucleotide at site 2 is the last nucleotide of the repeat due to new spacers being joined to the penultimate nucleotide of the first repeat (Swarts et al., 2012). In *S. solfataricus* the -1 nucleotide is only conserved at site 1, where it is an A. In *E. coli* it is conserved at both site 1 and 2 where it is an A or a C, respectively.

I expected to see little discrimination by SsoCas1 for substrates altered at the -1 position given that there is no conserved nucleotide in this position at site 2. However, after assaying substrates with varied -1 nucleotides over triplicate time

course reactions, I found that SsoCas1 favoured disintegration substrates with an A at this position, whereas junctions with a T at the -1 position were poor substrates (Figure 5.15, C). For EcoCas1, surprisingly, A, G and T nucleotides at the -1 position supported disintegration. However, substrates with C at -1 were not processed by Cas1 (Figure 5.15, A).

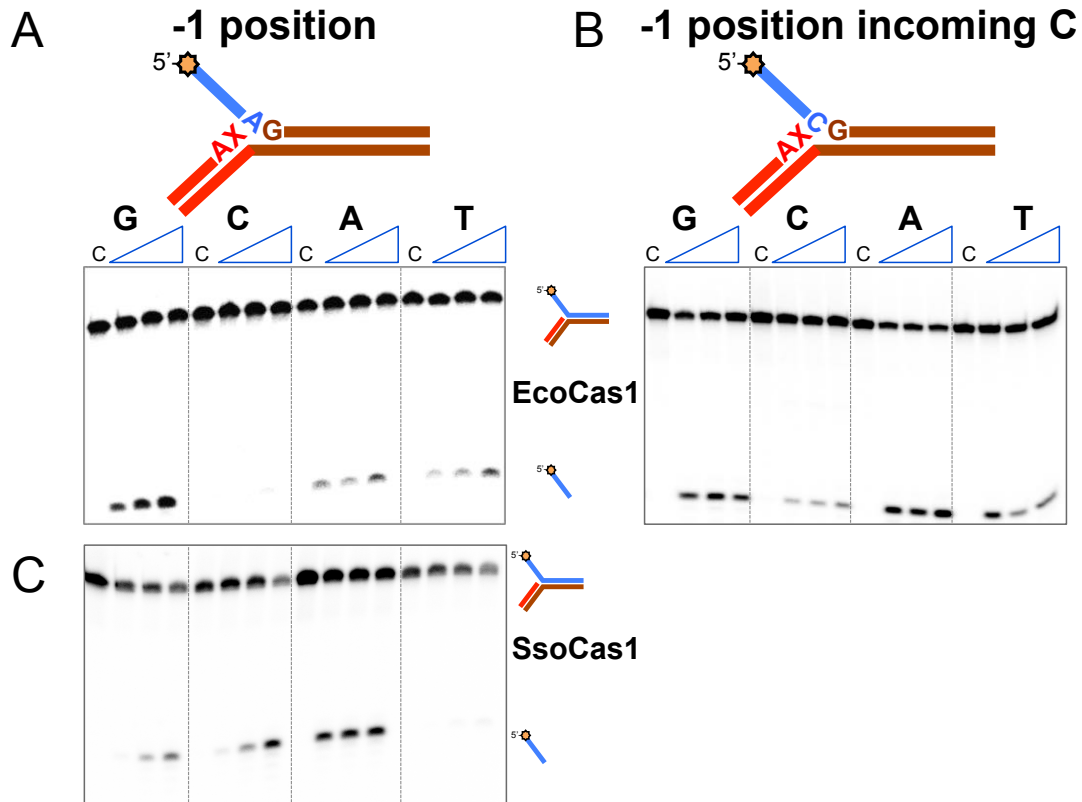


Figure 5.15 Specificity of Cas1 for the -1 position during disintegration

A. The nucleotide at position -1 of the disintegration substrate (from left to right - substrates 3, 11, 10, 9) (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) was varied and the substrates were assayed with EcoCas1 over a short time course. The first lane of each time course is a control without protein, followed by lanes containing time points at 5, 10 and 20 min. The structure of the disintegration substrate used in the assay is shown; X indicates the -1 nucleotide that was varied. The nucleotide at the -1 position is indicated above each time course. **B.** A short time course disintegration experiment was carried out with EcoCas1 and substrates varying at the -1 position (from left to right - substrates 15, 16, 17 and 18) (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences). The substrates used were as in **A**, with the exception of the first nucleotide of the 5' flap being changed to a C in these assays. **C.** A short time course assay was carried out with SsoCas1 and substrates varying at the -1 position, as in **A**. The first lane of each time course is a control without protein, followed by time points at 5, 10 and 20 min.

I hypothesised that perhaps the incoming nucleotide during adaptation, represented by the first nucleotide at the base of the 5' single-stranded flap in the disintegration substrate, might be required for specificity at position -1 for EcoCas1. During *E. coli* adaptation, the 3' end of the protospacer integrated at site 2 always ends in a C,

derived from cleavage of the PAM-complementary CTT sequence during spacer capture. To assess whether incoming nucleotide had any effect on specificity of EcoCas1 for the -1 position during disintegration I assayed substrates that varied at the -1 position and had a C as the incoming nucleotide with EcoCas1 (Figure 5.15, B). The preference observed for EcoCas1 was similar to that of the previous assays, suggesting the incoming nucleotide does not influence site specificity of Cas1. Favoured nucleotides at -1 were a G or an A, with a T leading to a weaker reaction and C strongly disfavoured at this position.

The nucleotides preferred at the -1 position during disintegration by Cas1 proteins only match those at site 1 *in vivo*. For SsoCas1 an A at the -1 position led to the most robust reaction, and an A is also present at the -1 position of site 1 *in vivo* (Figure 5.15, C). For EcoCas1 the -1 residue at site 1 is an A, and disintegration substrates with an A at -1 support disintegration. However, when I changed the -1 position to a C, the nucleotide present at the -1 position of site 2, the disintegration reaction with EcoCas1 was abolished (Figure 5.15, A).

The preferences observed for the -1 position do not match the *in vivo* integration site 2, and while nucleotides matching the -1 position at site 1 do support disintegration, there is no strong specificity observed. From these data, Cas1 seemed to prefer substrates with sequences that matched integration site 1, indicating that perhaps this position and not site 2 was key to recognition by Cas1 during integration. Furthermore, due to the weaker discrimination by Cas1 proteins at this -1 position, it seemed that this nucleotide is perhaps less important than surrounding residues for the recognition of the correct integration site by Cas1.

5.2.11.3 -2 position

I continued to investigate the effect of altering the nucleotide sequence around the junction of disintegration substrates, this time changing the nucleotide at position -2. This position corresponds to the penultimate nucleotide of the leader at site 1 and a seemingly unconserved nucleotide of the last integrated spacer at site 2. Time course assays were carried out with the EcoCas1 protein with time points taken from 30 sec to 30 min for each substrate, as shown in Figure 5.16 (A). These assays were carried out in triplicate, the fraction of substrate cleaved at each time point was calculated and the data were plotted and fitted to a single exponential

(Equation 4) with a floating endpoint. I found a clear preference for EcoCas1 for a G at the -2 position over all other nucleotides (Figure 5.16 B).

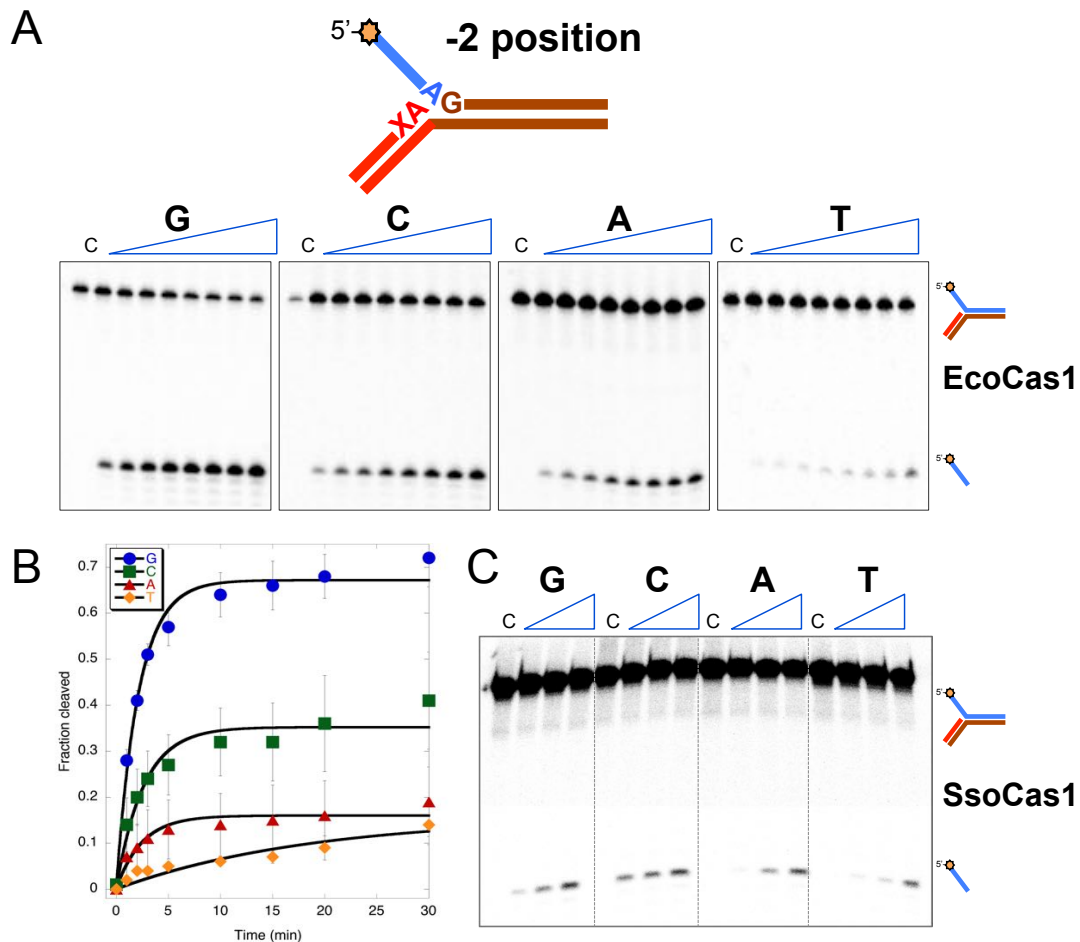


Figure 5.16 Specificity at the -2 position during disintegration by Cas1

A. The nucleotide at position -2 was varied and the substrates (left to right - substrates 12, 13, 10 and 14) (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) assayed with EcoCas1 over a time course from 30 sec to 30 min. The first lane of each time course is a control without protein. The structure and nucleotides surrounding the junction of the substrate used in the assay is shown; X indicates the -2 nucleotide that was varied. The nucleotide at the -2 position is indicated above each time course. **B.** The fraction of substrate processed by EcoCas1 disintegration in A was quantified and plotted against time. The points are representative of the mean of triplicate experiments and the standard deviation of the mean is indicated by error bars. The points were fitted to a single-exponential equation (Equation 4). **C.** A short time course assay was carried out with SsoCas1 and substrates used in A varying at the -2 position. The nucleotide at the -2 position is indicated above each time course. The first lane of each time course is a control without protein, followed by time points at 5, 10 and 20 min.

To confirm that a preference also existed for the nucleotide at this position for the SsoCas1 protein, a shorter time course reaction was carried out with time points from 5 to 20 min. These assays were carried out at least three times and a representative gel is shown in Figure 5.16 (C). The SsoCas1 was found to favour a

C at this position, with a G also leading to a strong reaction. In contrast, an A or T at this position led to a much-reduced rate of reaction.

Discrimination at the -2 position by both Cas1 proteins is clearly taking place during the disintegration reaction. The nucleotides preferred at -2 correlate with the predicted -2 nucleotides at the leader-repeat junction (site 1) of the corresponding CRISPR array, which is a G for *E. coli* and a C for *S. solfataricus*. This indicates that the sequence preferences observed during the reverse reaction match those of the integration reaction. The -2 nucleotide at integration site 2 is not conserved and depends on the sequence of the most recently integrated spacer. As Cas1 showed a sequence specificity for this position, and this preference matches the *in vivo* -2 nucleotide at site 1, it seemed that identification of the correct integration site by Cas1 might depend on the sequence of site 1, rather than that of site 2.

5.2.11.4 Incoming nucleotide

The nucleotide at the base of single-strand flap of the disintegration substrate was varied. This nucleotide represents the 3' nucleotide of the incoming spacer in the forward reaction. Again, in *S. solfataricus* there is no apparent conservation of spacer sequence, so I expected no discrimination of substrates based on this position. In keeping with this, I found that SsoCas1 performed efficient disintegration reactions no matter the nucleotide at this position (Figure 5.17, B). While each nucleotide supported disintegration, the reaction was more robust with A or T as the incoming nucleotide, compared to a G or a C.

During adaptation in *E. coli*, spacers are selected by recognition of a PAM sequence, usually a 5'-AAG-3'. Selection of a spacer leads to cleavage before the last nucleotide of the PAM, leaving one end of the newly-generated spacer with a 5' G and a complementary 3' C on the other strand (Swarts et al., 2012). The sequence at the other end of the selected spacer is thought to be unconserved and selected through a ruler mechanism. Therefore, the PAM-complementary 3' C is usually the attacking nucleotide of an incoming spacer at site 2 in the *E. coli* CRISPR and is added to the array as the last nucleotide of the repeat, whereas the incoming nucleotide of site 1 varies.

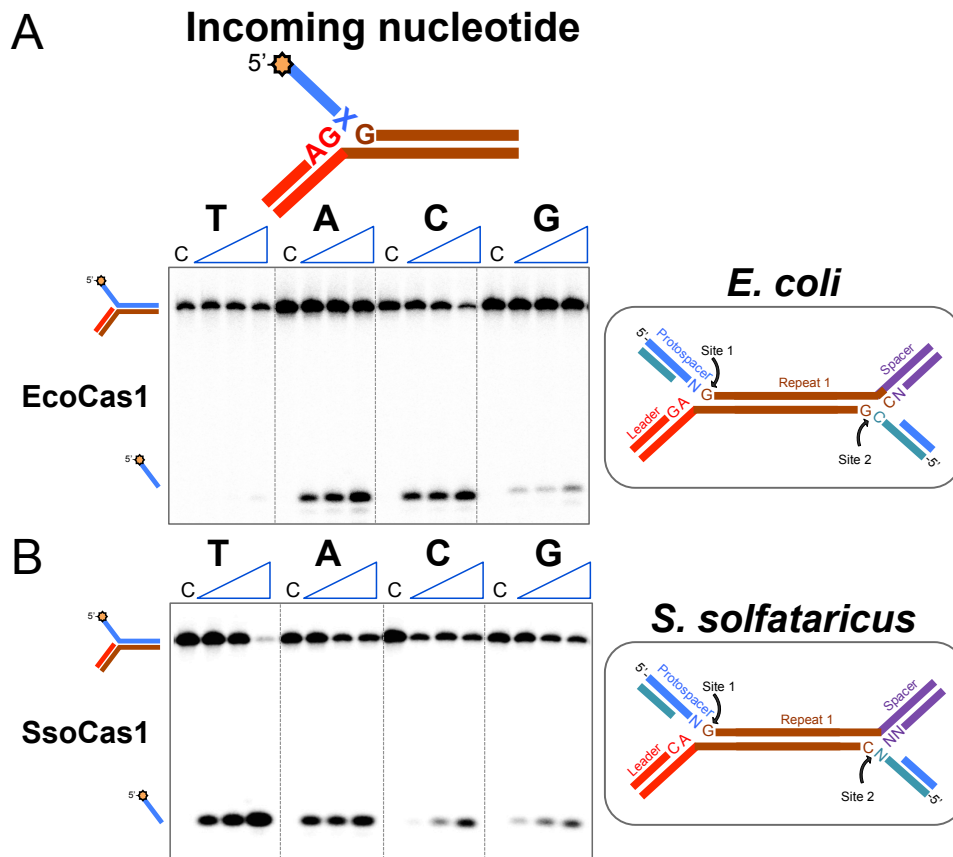


Figure 5.17 Specificity of Cas1 for the ‘incoming’ nucleotide during disintegration

A. The last nucleotide of the 5' flap (blue), corresponding to the incoming nucleotide (IC) of the protospacer in the forward reaction was varied (from left to right - substrates 2, 3, 4 and 5) (see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) and the substrates assayed with EcoCas1 over a short time course. The first lane of each time course is a control without protein, followed by time points at 5, 10 and 20 min. The structure and nucleotides surrounding the junction of the substrate used in the assay is shown, X indicates the incoming nucleotide that was varied. The nucleotide at this position is indicated above each time course. To the right is a model showing the sequence of the CRISPR array at each integration half-site. **B.** As in **A** with SsoCas1 added to the disintegration time course reactions. The sequence of the nucleotides surrounding the integration half-sites in the *S. solfataricus* CRISPR array is shown on the right.

Assaying EcoCas1 with substrates varying at the incoming nucleotide, I found that this enzyme preferentially disintegrated substrates with an incoming C, followed by A at this position (Figure 5.17, A). Incoming G and Ts were disfavoured and led to inefficient disintegration reactions. A preference for a C at this position seemed to match the *in vivo* spacer integration. However, as the *E. coli* K12 CRISPR array contains fewer than 20 spacers, it is difficult to draw strong conclusions regarding the preferential integration of substrates with a 3' A *in vivo*.

5.2.11.5 Disintegration reactions with substrates matching site 1 and 2

For both the SsoCas1 and the EcoCas1, the sequence specificity observed for the junction sequence of disintegration substrates seems to match that of integration site 1, but not site 2. There is a strong specificity for the nucleotide at the -2 position in both SsoCas1 and EcoCas1. This specificity matches the nucleotide at -2 in the leader sequence, whereas there is no conservation of this nucleotide at site 2 so it cannot be responsible for the preferences observed. Furthermore, the +1 nucleotide that led to the strongest disintegration reactions for both SsoCas1 and EcoCas1 also matched the +1 nucleotide at site 1 (the first nucleotide of the repeat). The discrimination observed at the -1 position was more loosely correlated with the sequence of the *in vivo* integration sites. However, the nucleotide present at the -1 position in site 1 led to a strong disintegration reaction. In contrast, for *E. coli* the -1 nucleotide conserved at site 2 produced the weakest reverse reaction *in vitro*.

To explore this apparent selection by Cas1 for sequences matching the leader-repeat 1 junction, substrates matching the exact sequence of site 1 and site 2 during a previous integration into the *S. solfataricus* or *E. coli* CRISPR array were designed. Time course assays were carried out and triplicate data points were plotted for EcoCas1 reactions to compare how well these substrates were processed during disintegration by Cas1.

When assayed in disintegration reactions, the substrates matching site 1 were processed at a faster rate by both SsoCas1 and EcoCas1 as shown in Figure 5.18. For the EcoCas1 reaction, triplicate data points were plotted and fitted to a single exponential (Equation 4) with a floating end point. The plot (Figure 5.18, C) confirms that the site 1 substrate is the preferred disintegration substrate, with ~80% cleaved after 30 min compared to ~45% of the site 2 substrate. The disfavoured A nucleotide at -2 and C at -1 positions may explain why the site 2 substrate leads to a weaker reaction with EcoCas1 than the site 1 sequence. The SsoCas1 also processed the site 1 sequence preferentially (Figure 5.18, D), potentially due the C at position +1 and G at position -2, which were previously shown (Figure 5.14 and Figure 5.16) to produce weaker disintegration reactions that substrates mimicking site 1, with a G at the +1 position and a C at -2. These data strengthen the argument that the disintegration reaction is a good model for integration and that Cas1's sequence preference indicates that recognition and the first half-site integration occurs at site 1, and not site 2, during adaptation.

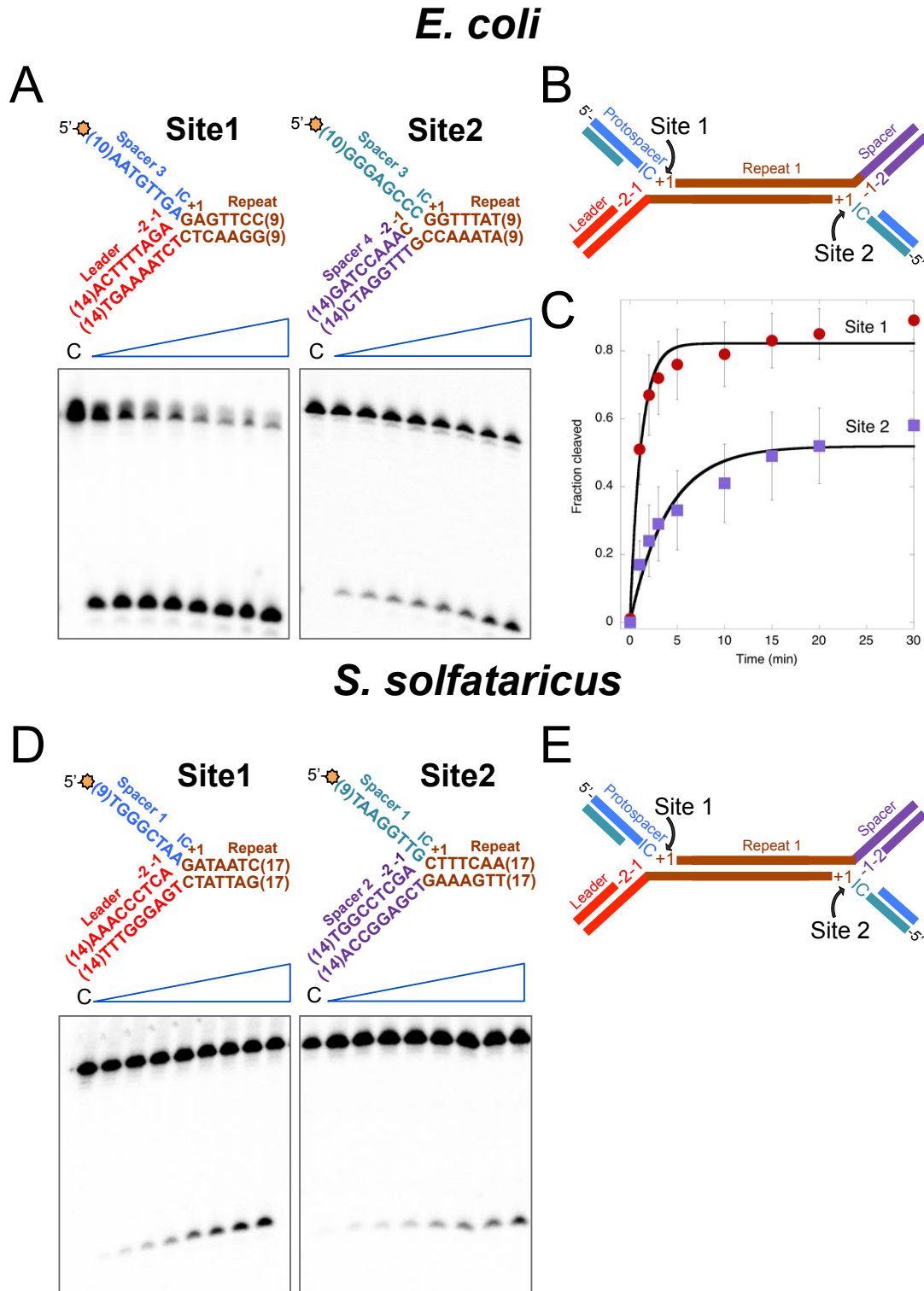


Figure 5.18 Disintegration of a site 1 versus site 2 substrate

A. Two disintegration substrates were designed with sequences matching that of integration site 1 and 2 during the integration of the current spacer 3 in the CRISPR array of *E. coli*. These substrates (spacer3-1 and spacer3-2, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) were assayed with EcoCas1 over a time course from 30 sec to 30 min. The first lane of each time course is a control without protein. The structure and nucleotide sequence of the substrates used in the assay is shown above each time course. Numbers in brackets indicate nucleotides in the substrates assayed that are not depicted in the diagram. IC stands for incoming nucleotide, which is the 3' nucleotide of the

single-stranded flap. **B.** Shows a model of the two integration sites into the CRISPR array of *E. coli*. **C.** The time course assays shown in A were carried out in triplicate and the fraction of substrate processed at each time point was calculated. The average fraction of substrate processed was plotted against time and the standard deviation from the mean is shown by error bars. The points were fitted to a single exponential (Equation 4). **D.** As for A, two disintegration substrates with sequences matching site 1 and site 2 (site1-ss0 and site2-ss0 substrates, see Table 2.2 for constituent oligonucleotides and Table 2.1 (p.47) for sequences) during the integration of current spacer 1 in the CRISPR array of *S. solfataricus* were designed and assayed with SsoCas1 over a time course from 30 sec to 30 min. The first lane of each assay is a control without protein. The sequence and structure of the site 1 and 2 substrates used in this assay are shown above the time course. **E.** Shows a model of integration site 1 and 2 in the CRISPR array of *S. solfataricus*.

5.3 Conclusions

In this chapter I have demonstrated that both SsoCas1 and EcoCas1 perform a transesterification reaction on branched substrates that results in the release of a 5' flap and the reannealing of the remaining nicked duplex (Figure 5.6). This reaction is the reverse of integration, termed disintegration, and has previously been reported for viral integrases (Chow et al., 1992). Studying the disintegration reaction of Cas1 proved an effective method to uncouple the requirements and sequence specificity of the protospacer capture and integration steps of adaptation.

5.3.1 Incoming DNA

I first examined how the structure of the disintegration substrate affected processing by Cas1. The 5' flap of the disintegration substrate, which represents the incoming DNA during the forward integration reaction, could be single-stranded, or partially or fully double-stranded to support disintegration by Cas1 (Figure 5.9). However, reducing flap length from 18 to 10 nucleotides resulted in a very weak disintegration by Cas1 (Figure 5.10). Incoming DNA protospacers in *S. solfataricus* range in size from 34 – 48 bp (Lillestøl et al., 2006) and are likely to be at least partially duplex in nature; therefore, this requirement of SsoCas1 for longer 5' flaps and the ability to disintegrate partial duplex flaps seem to correlate well with the *in vivo* activity of the protein. The ability of SsoCas1 to disintegrate fully double-stranded flaps is interesting as in the *E. coli* Cas1-Cas2 crystal structure only a single-stranded incoming DNA end is positioned in the active site of Cas1 (Nuñez et al., 2015a). The apparent lack of selection by Cas1 for single or double-stranded incoming DNA during the disintegration reaction may imply that SsoCas1 is able to open up the structure around the branch-point of the disintegration substrate. EcoCas1 splays

the ends of protospacer DNA using a wedge-like tyrosine residue (Y22) (Nuñez et al., 2015a). While this residue is not universally conserved in Cas1 proteins, the SsoCas1 does have a tyrosine residue (Y12) in the same loop region between β -sheet domains of the N-terminal. It is conceivable that this residue splays double-stranded flaps during disintegration, or incoming DNA during integration, explaining the tolerance observed here for fully duplex 5' flaps.

In this chapter, it was also demonstrated that the disintegration reaction takes place in the same active site as previously reported enzymatic activities of Cas1, and that the reaction is dependent on divalent metal cations and the canonical active site residues (Figure 5.12). During this work another group also reported that the Cas1 protein from *E. coli* performs disintegration *in vitro* (Nuñez et al., 2015b). However, the reaction observed in that study was very weak, perhaps due to an unfavourable junction sequence.

5.3.2 Cas2 is not required for disintegration

Cas2 was found not to be required for, or to enhance, disintegration *in vitro* (Figure 5.3). As disintegration represents the reversal of a half-site integration, this supports the conclusion that the identified nuclease activity of Cas2 (Beloglazova et al., 2008) plays no part in integration *in vivo*. The Cas2 protein instead acts as a scaffold between two Cas1 dimers (Nuñez et al., 2014), presumably playing a crucial role in adaptation by fixing the distance between each half-site integration reaction, carried out in Cas1 active sites. Furthermore, Cas2 has been suggested to be involved in specific selection of CRISPR sequences (Nuñez et al., 2014) and may be important in interacting with recently identified repeat motifs (Wang et al., 2016) to anchor the complex in the correct position for spacer insertion.

5.3.3 Cas1 has an intrinsic sequence specificity

Although Cas2 may contribute to the correct positioning of the complex during integration, I show here using the disintegration reaction that the Cas1 protein alone has an inherent sequence specificity for the nucleotides surrounding the leader-repeat 1 junction (site 1).

Disintegration by SsoCas1 and EcoCas1 was found to be strongly influenced by the +1 position of the disintegration substrate (Figure 5.13 and Figure 5.14), which represents either 5' nucleotide of repeat 1, to which spacers are joined during

adaptation. Both SsoCas1 and EcoCas1 had a strong preference for a guanine at this position, which matches both *in vivo* +1 nucleotides for *E. coli* and the site 1 +1 position for *S. solfataricus* (see Figure 5.19). This strong preference implied that Cas1 makes sequence contacts with the first residue of the repeat during adaptation.

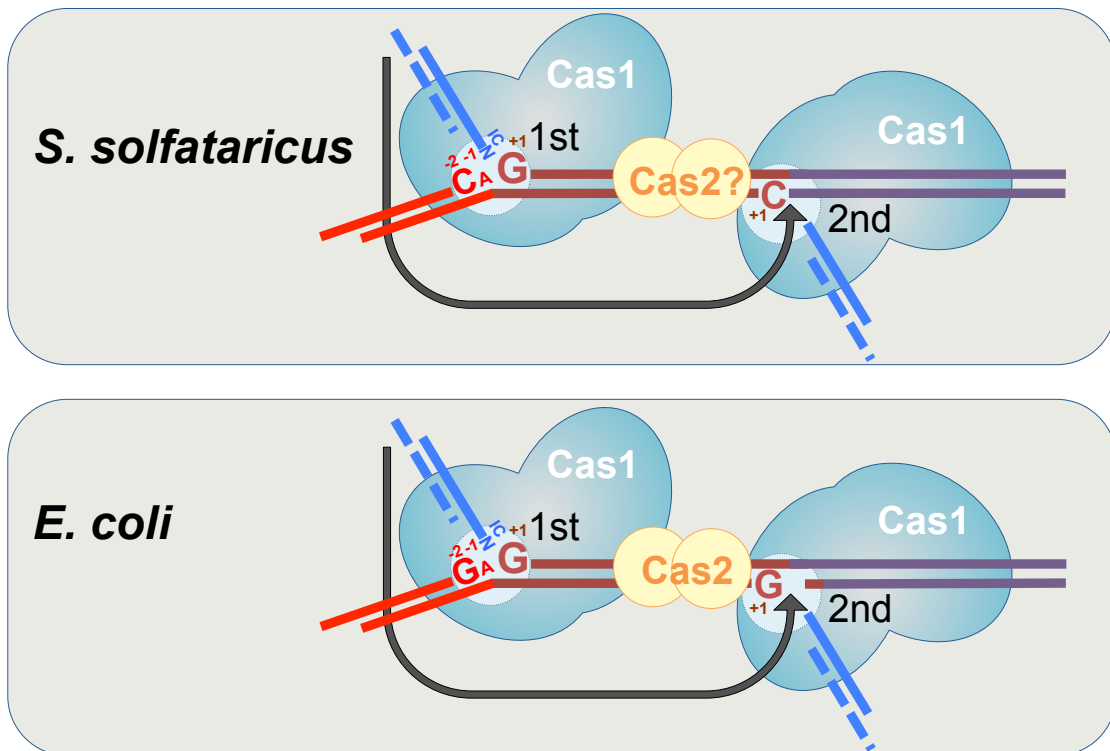


Figure 5.19 Nucleotide sequence at site 1 is key to integration and disintegration

The schematic displays a summary of the identified sequence specificity intrinsic to the Cas1 proteins of *S. solfataricus* and *E. coli*. The residues at position +1 and -2 were shown in this chapter to strongly influence disintegration by Cas1 proteins (the size of the nucleotide in the figure represents the magnitude of their influence on integration). A G at position +1 was favoured by both Cas1 proteins. A C or a G at -2 (for SsoCas1 and EcoCas1 respectively) was required for a strong disintegration reaction. In contrast, the nucleotide at position -1 and the incoming nucleotide of the protospacer (IC) were found to have little influence on the disintegration efficiency. The preferred residues matched the *bona fide* nucleotide sequences at these positions at the leader-repeat junction (site 1) during integration *in vivo*. Therefore, we conclude that: (a) these residues are important for recognition of the correct integration site by Cas1 and (b) site 1 rather than site 2 (between the first repeat and existing spacer 1) is targeted by Cas1 during the first half-site integration, with the second reaction being guided by a molecular ruler, rather than sequence specific, mechanism. As both single- and double-stranded flaps (blue) were excised during disintegration, we could not predict the nature of spacer during integration; the dashed duplex strand of the incoming DNA in the figure indicates this. Cas2 is not needed for disintegration, but is known to be required for integration in *E. coli* at least and may play a role in defining the distance between the two protospacer ends to be integrated.

The -2 nucleotide also influenced the disintegration reaction carried out by Cas1 proteins (Figure 5.16). A guanine was preferred at this position for disintegration by EcoCas1 and a cytosine for SsoCas1. This preference matches the -2 nucleotide in

the leader sequence of the associated CRISPR arrays. The influence of the -2 position, which is only conserved at the leader-repeat junction and not at site 2, implied that Cas1 specifically recognises site 1 during integration, and that the second cut is performed through a ruler mechanism and is not influenced strongly by sequence.

Surprisingly, the -1 position, which corresponds to the last nucleotide of the repeat at site 1 and the first nucleotide of the previously integrated spacer at site 2, was demonstrated not to strongly influence the disintegration reaction (Figure 5.15). A recent study of *in vitro* integration by EcoCas1-Cas2 also showed that selected integration sites in plasmid DNA were not conserved at the -1 position, whereas the +1 and -2 positions were strongly selected for (Nuñez et al., 2015b). Importantly, this study demonstrated a preference for integration at sites with +1 and -2 positions matching those that produce the strongest disintegration reactions in this work. Furthermore, the authors also found that CRISPR adaptation was weak when Cas1 and Cas2 were provided with protospacers with a 3' T residue, compared to a 3' C (Nuñez et al., 2015b). This correlated with the discrimination seen in the disintegration reaction of EcoCas1 against an incoming T residue.

The nucleotide preferences found for *in vitro* integration by Nuñez and colleagues are complementary to those described here (Nuñez et al., 2015b). However, the authors found that half-site integrations occurred most frequently at the leader-distal end of the repeat (site 2), leading them to conclude that targeting and the first half-site integration during adaptation occurs here and not at site 1 (Nuñez et al., 2015b). This conclusion is not supported by the disintegration data presented here, as the specificity observed matched well with the sequence of the leader-repeat 1 junction for both SsoCas1 and EcoCas1. This result was confirmed by performing disintegration reactions on substrates mimicking a previous half-site intermediate at site 1 or 2. Both SsoCas1 and EcoCas1 proteins disintegrated the site 1 substrate more efficiently than the site 2 substrate, confirming a preference for the nucleotides around the junction at this site. Furthermore, it is logical that the leader-repeat 1 junction is the site targeted by Cas1 in the forward reaction as it is conserved and, unlike site 2, does not depend on the sequence of the last integrated spacer.

To conclude, this chapter demonstrated that the sequence specificity of the reverse reaction carried out by Cas1 matched the *in vivo* integration sites of the corresponding CRISPR array (see Figure 5.19). These findings indicated that Cas1

alone is capable, to some degree at least, of selecting in a sequence-specific manner the site of integration during adaptation. As the preferred sequence of the disintegration substrate matched that of the leader-repeat 1 junction, I propose that this site (site 1) is selected by Cas1 as the site of the first integration. The second half-site integration, at the other 5' end of the repeat (site 2), is likely then mediated through a ruler-dependent or structure-specific mechanism.

Chapter 6: An *in vitro* reconstitution of integration

I wish to acknowledge Dr Shirley Graham who cloned and purified the Cas1_{AB} protein and carried out the PCR reactions to amplify integration sites in this chapter.

6.1 Introduction

The study of the disintegration reaction of Cas1, discussed in Chapter 5, yielded important insights into the specificity and mechanism of adaptation. However, the reconstitution of the forward reaction *in vitro* remained one of the main aims of this project as it would facilitate investigation of the contribution of Cas2 during adaptation and also allow us to learn more about the influence of sequence and structure on integration site selection. However, recreation of the integration *in vitro* has proved difficult in the viral integrase field, with the forward reaction having a much lower efficiency than the reverse reaction (Chow et al., 1992).

After trying unsuccessfully to study integration by Cas1 and Cas2 from *S. solfataricus* using shorter linear oligonucleotides, I supplied protospacer-length double-stranded DNA to Cas1-Cas2, in the presence of metal ions and a target plasmid. Cas1-Cas2 covalently attached these oligonucleotides to the target plasmid, evidenced by their migration with the nicked plasmid species during agarose gel electrophoresis. Cas2 was not essential for this integration; however, it did enhance the efficiency of the reaction.

6.1.1 Integration of spacers is not strictly sequence-specific *in vitro*

Similar findings to the above have been published for the *E. coli* Cas1-Cas2 proteins, showing that a Cas1-Cas2 complex could integrate short DNA oligonucleotides into supercoiled plasmid DNA (Nuñez et al., 2015b). Deep sequencing of the integration sites from this study revealed that while the *in vitro* reaction was not specific for the leader-repeat junction, there was a preference for integration within the CRISPR locus. The authors also identified the start of the ampicillin resistance gene as a hotspot for integration. They speculated that this site may be aberrantly selected due to its similarity to the CRISPR leader-repeat junction, as both sites were

composed of palindromic sequences flanked by an AT-rich domain (Nuñez et al., 2015b).

The high-throughput sequencing of the *E. coli in vitro* integration sites showed that 71% of all spacer insertions into the plasmid DNA (3240 bp) were at the repeat ends of the CRISPR locus (1000 bp) (Nuñez et al., 2015b). While the highest number of these integrations occurred at the leader-distal 5' end of repeat 1 (site 2), the sequence motif identified from comparing the sequence of all integration sites in plasmid DNA mirrored that of the leader-repeat junction (site 1). From these data it seemed that the Cas1-Cas2 complex alone is not sufficient *in vitro* for selection of the *bona fide* integration site between the leader and first repeat. This lack of stringent selection for the correct integration site indicated that other host factors or CRISPR elements were required to direct integration uniquely at the leader-repeat 1 junction.

6.1.2 Supercoiled plasmid DNA is important for integration

Several studies have demonstrated that the sequence of the leader, particularly the last 60 bp in *E. coli*, or 10 bp in *S. thermophilus*, and the first repeat are important for adaptation of the CRISPR (Yosef et al., 2012; Wei et al., 2015). It was speculated that the secondary structure of these elements might play a role in guiding integration by Cas1-Cas2.

Interestingly, supercoiled target DNA was shown to be essential for integration *in vitro* by *E. coli* Cas1-Cas2 (Nuñez et al., 2015b), confirming that structure in addition to sequence is crucial for integration. The authors of this study also correlated insertion hotspots at the end of each repeat and in the ampicillin resistance gene with the presence of palindromic sequences at these positions. It was hypothesized that these palindromic sequences may form hairpin structures, which are recognized by the Cas1-Cas2 complex and direct integration (Nuñez et al., 2015b). However, while the palindromic nature of the repeats in some CRISPR systems may guide integration to a degree, often the CRISPR repeats of archaea are not palindromic (Kunin et al., 2007), meaning this cannot be a generalised mode of recognition of integration sites across CRISPR subtypes. In Chapter 4, a putative palindromic region was identified at the junction of the CRISPR C leader and repeat 1 of *S. solfataricus*. It is conceivable that this palindrome forms a hairpin and guides binding and integration by the Cas1-Cas2 complex during adaptation in this system.

The results presented in this chapter demonstrate that together Cas1 and Cas2 from *S. solfataricus* integrate spacer-length DNAs into supercoiled targets. I show that protospacers with 3' single-stranded ends are the preferred substrates for integration, whereas 5' single-stranded ends abolish integration activity. Furthermore, it is demonstrated that, in the absence of other factors, the Cas1 and Cas2 proteins insert protospacers outside of the CRISPR array. However, the insertion sites selected do share a sequence motif with the leader-repeat 1 junction (site 1). Finally, I show that a factor in *S. solfataricus* lysate is required in addition to Cas1-Cas2 to direct protospacer insertion uniquely to the leader-repeat 1 junction.

As two sets of Cas1 and Cas2 proteins from *S. solfataricus* are used in this chapter, the Cas1 and Cas2 coded for by the genes located between CRISPR loci C and D will be referred to as Cas1_{CD} and Cas2_{CD}. The second set of *cas* genes associated with loci A and B will be referred to as Cas1_{AB} and Cas2_{AB}.

6.2 Results

6.2.1 Cas1-Cas2 integrates short oligonucleotides into supercoiled DNA

In contrast to the activity of other Cas1 proteins, Cas1_{CD} showed no nuclease activity on branched or linear DNA substrates (Chapter 4). As the Cas1 protein from *P. aeruginosa* was also demonstrated to digest supercoiled plasmid DNA (Wiedenheft et al., 2009), an assay of Cas1_{CD} and Cas2_{CD} on plasmid DNA was carried out in the presence of divalent metal ions. Following separation of the assay products, no degradation of the plasmid DNA was obvious and only a very weak conversion of plasmid to the nicked/open circle form was observed (Figure 6.1 A and B). However, when the same assay was carried out with Cas1_{CD} and Cas2_{CD} in the presence of short (39 bp), radiolabelled DNA duplexes, the conversion of supercoiled to nicked plasmid was greatly increased, with 50% of the substrate being nicked following a 120 min incubation (Figure 6.1 A and B) (see section 2.2.8.3.2 (p. 67) for method).

I hypothesised that the increased nicking observed may be due to Cas1_{CD} and Cas2_{CD} utilising the 3' hydroxyls of the short duplex to mediate a transesterification reaction, which would result in the nicking of the plasmid and the joining of the two DNA species. On phosphorimaging the gel containing the products of this assay, a

band of radiolabelled DNA was identified that had migrated much more slowly through the gel than the free oligonucleotides. The position of the radioactive band was localised by overlaying the agarose gel image and the phosphorimage, and was found to correspond to the position of the nicked form of the plasmid. These results demonstrated that Cas1_{CD}, potentially in complex with Cas2_{CD}, was capable of adding protospacer-length double-stranded DNA to a supercoiled plasmid, leading to the nicking of the plasmid and co-migration of radioactivity with the nicked band during separation on an agarose gel. From this assay it cannot be deduced whether the integration occurring is a half-site, where only one 3' hydroxyl mediates transesterification, or full site reaction, as both would result in a similar nicking of the plasmid being observed (Figure 6.1 C).

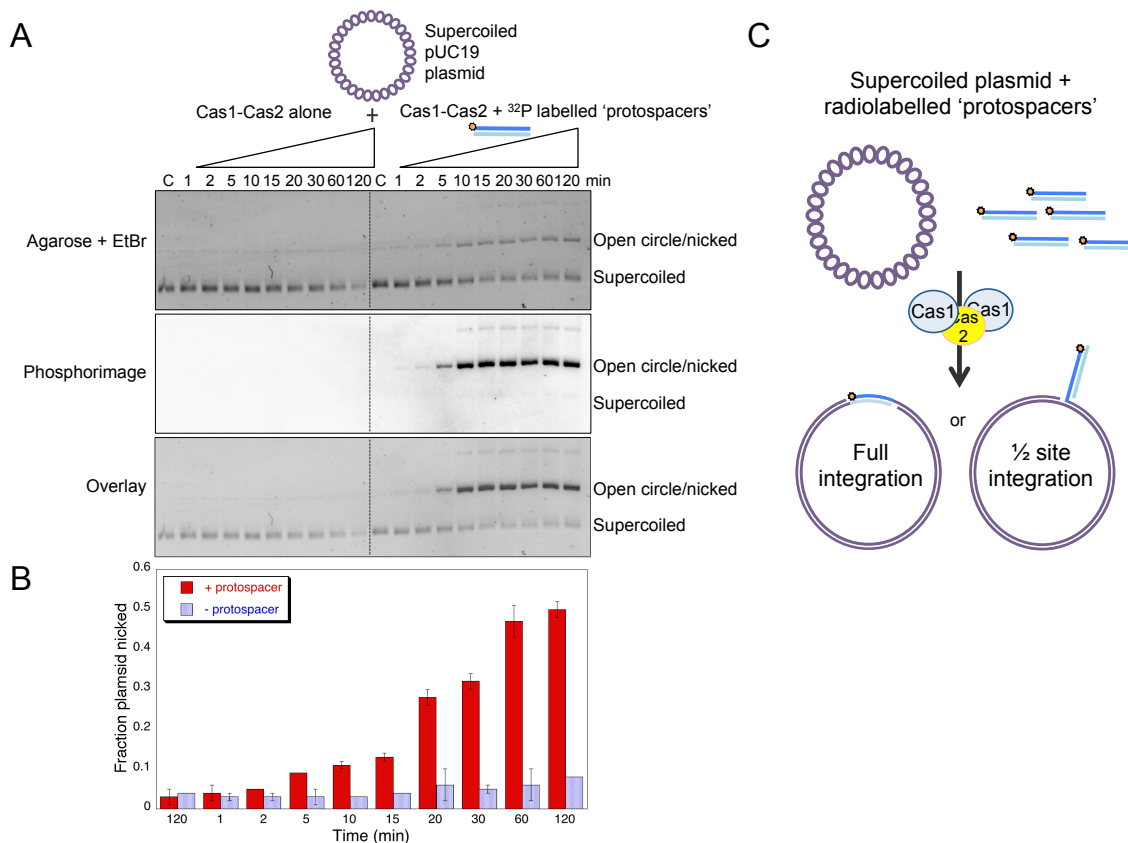


Figure 6.1 Cas1_{CD} and Cas2_{CD} joins short oligonucleotides to plasmid DNA

A. Cas1_{CD} and Cas2_{CD} (both at 2 μ M) were incubated at 55°C with supercoiled pUC19 plasmid (100 ng) and divalent metal ions (5 mM MnCl₂), with or without the inclusion of a ³²P-labelled 39 bp double-stranded oligonucleotide (2 μ M) (duplex made by annealing CRISPR D spacer dup F and CRISPR D spacer dup R oligonucleotides, see Table 2.1 (p.47) for sequences). Aliquots of the assay were removed and quenched with 50 mM EDTA and incubation on ice at the following time points: 1, 2, 5, 10, 15, 20, 30, 60, 120 min. The products of the assay were separated on a 1% agarose gel containing ethidium bromide and visualised under UV light (agarose + EtBr image, top). The gel was then dried and exposed to an imaging plate overnight before phosphorimaging (phosphorimage, middle). The

agarose gel image and the phosphorimage were overlaid to allow the radioactive signal to be aligned to the nicked/open-circle plasmid form (overlay, bottom). **B.** The fraction of plasmid converted from supercoiled to nicked/open circle was quantified using the EtBr signal and ImageGauge (FUJIFILM) software for each time point in the + or - oligonucleotide conditions in **A.** The average fraction of plasmid converted, calculated from triplicate experiments, was plotted with the standard deviation shown as error bars. **C.** A schematic of what is hypothesised to be happening in **A.** In the + oligonucleotide conditions Cas1_{CD}, potentially in complex with Cas2_{CD}, integrates labelled protospacer-like DNA molecules into supercoiled plasmid DNA. This process leads to the conversion of DNA from supercoiled to nicked form. From the experiment in **A** it is not clear whether this integration is a half-site or full integration as both would lead to nicking of the plasmid.

A recent study of the *E. coli* Cas1-Cas2 also demonstrated integration of protospacers into supercoiled acceptor plasmids *in vitro* (Nuñez et al., 2015b). The apparent requirement for supercoiled DNA might indicate that local structural distortions, such as DNA bubbles or hairpins, are recognised by Cas1-Cas2 as integration sites. Similar structural distortions have previously been shown to be important for viral integrases, with unpaired or stem loops structures being preferred integration sites (Katz et al., 1998). The absence of these structures in short duplex DNAs might explain why I was unable to reconstitute integration by Cas1_{CD}-Cas2_{CD} into these substrates *in vitro*.

6.2.2 ssDNA is a substrate for integration

The Cas1-Cas2 adaptation complex in *E. coli* has been shown to specifically integrate partial duplex DNA with single-stranded ends. Fully duplex protospacers led to weak integration activity and single-stranded DNAs abolished integration (Nuñez et al., 2015b). To establish if Cas1_{CD} and Cas2_{CD} preferred similar protospacer structures, the radiolabelled nucleic acid to be integrated was varied from double-stranded DNA to single-stranded DNA or RNA. The protospacer substrates were incubated with either Cas1_{CD}, active site variant E142A Cas1_{CD}, Cas2_{CD} or Cas1_{CD} and Cas2_{CD} (Cas1_{CD}-Cas2_{CD}) for thirty minutes before the addition of manganese ions and plasmid DNA.

An increase in the nicked form of the plasmid was apparent following this assay in agarose gel lanes containing Cas proteins (Figure 6.2). This nicking may indicate that the high temperature and metal ion conditions used led to some background degradation of the plasmid. However, integration of radiolabelled DNA into the nicked plasmid band is only observed in the presence of Cas1_{CD} or Cas1_{CD}-Cas2_{CD}. Incorporation was more robust when both Cas1_{CD} and Cas2_{CD} were present in the assay, compared to Cas1_{CD} alone (Figure 6.2). Active site Cas1_{CD} variant E142A

did not integrate spacers, which indicated that the canonical active site of the Cas1_{CD} protein was responsible for the integration activity observed.

The incorporation of radioactive signal into the nicked-plasmid band was weaker for single-stranded DNA compared to double-stranded substrates. However, it is clear that, in contrast to the *E. coli* Cas1-Cas2, the adaptation proteins from *S. solfataricus* are able to use single-stranded DNA as a substrate for integration, *in vitro* at least.

Incubation of Cas1_{CD} and Cas2_{CD} together with double-stranded protospacer led to a much stronger integration than Cas1_{CD} alone. Previous attempts to identify an interaction of the Cas1_{CD} and Cas2_{CD} proteins failed (Chapter 4). However, it seems from these results that the increase in integration observed is due to a functional interaction of the two proteins. This interaction may have been stimulated by the high-temperature incubation with protospacer-length substrates.

In Cas1_{CD}-only conditions, similar efficiencies for incorporation of single-stranded or duplex DNA were observed, whereas in Cas1_{CD}-Cas2_{CD} conditions, double-stranded oligonucleotides led to a much stronger integration reaction. This might indicate that Cas1 alone can perform only half-site integrations for which single-stranded DNA is sufficient. However, in the Cas1_{CD}-Cas2_{CD} condition a proportion of the proteins may be interacting in functional complexes, which are able to selectively bind duplex DNA and use both 3' hydroxyls to perform full-site integrations.

The ssRNA protospacer was a poor integration substrate with a very weak protospacer integration observed only in the Cas1_{CD} condition (Figure 6.2). I had previously observed that neither Cas1_{CD}, nor Cas2_{CD} bind to RNA in EMSA assays (Chapter 4). It may be this binding deficiency and the inability to position the 3' hydroxyl in the Cas1_{CD} active site that prevents integration of this substrate taking place. The preference of Cas1_{CD}-Cas2_{CD} for DNA over RNA protospacers likely reflects the fact that protospacers captured by *S. solfataricus* Cas1-Cas2 *in vivo* will originate from duplex DNA, as archaeal viruses almost all have DNA genomes (Bolduc et al., 2012). Interestingly, a bacterial type III Cas1-reverse transcriptase fusion protein was recently reported to directly incorporate ssRNA spacers into the CRISPR array, before reverse-transcription and storage of the spacer as cDNA (Silas et al., 2016). This sampling of RNA spacers was hypothesised to be involved in protection from RNA phage and perhaps even in host gene regulation.

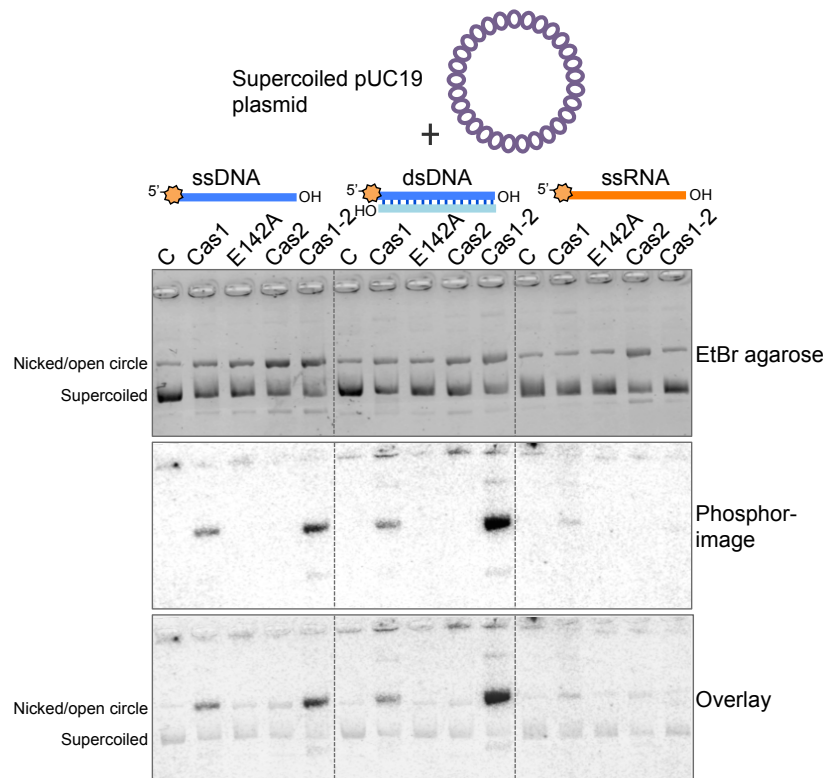


Figure 6.2 Single- and double-stranded DNA integrated by Cas1_{CD}

Integration assays were carried out with supercoiled pUC19 plasmid DNA (100 ng) and either single- or double-stranded DNA (blue) or single-stranded RNA (orange), all of which were labelled with a 5'-³²P on one strand (single-stranded CRISPR D spacer dup F, duplex of CRISPR D spacer dup F and CRISPR D spacer dup R, ssRNA spacer substrates, see Table 2.1 (p.47) for sequences). Before the addition of plasmid DNA, the nucleic acid to be integrated (2 μM) was incubated for 30 min with: Cas1_{CD}, active site variant Cas1_{CD} E142A, Cas2_{CD} or a Cas1_{CD} and Cas2_{CD} together (all at 2 μM). After a 1 hour incubation in the presence of protein and metal ions, reaction products were separated on a 1% agarose gel, pre-stained with ethidium bromide (top). The first lane for each substrate is a control without protein. The gel was dried and phosphorimaged to detect incorporation of radiolabelled oligonucleotides into the plasmid DNA (middle). The bottom image (overlay) is a composite of both the agarose gel scan and the phosphorimage.

6.2.3 Integration is not specific for a CRISPR array *in vitro*

Although integration during CRISPR adaptation is very specific and only occurs at the leader-repeat junction, in the *in vitro* reaction reconstituted here protospacers were integrated non-specifically into a pUC19 plasmid (Figure 6.2). I hypothesised that the integration observed might be more robust, or more specific, if Cas1_{CD}-Cas2_{CD} were presented with a cognate integration site. To test this I cloned a section of the CRISPR C array from *S. solfataricus* into the multiple cloning site of pUC19 (to form pCRISPRC, see Table 2.3 (p.52) for sequences and details). The insert contained the last 101 bp of the leader, the first repeat and first spacer of

CRISPR C. This plasmid was then used in an integration reaction with Cas1_{CD} and Cas2_{CD}, with the empty pUC19 vector as a control (Figure 6.3).

Both supercoiled plasmids were converted to open circle DNA following incubation with Cas proteins, with the highest degree of nicking occurring following incubation with Cas1_{CD}-Cas2_{CD}. Following phosphorimaging of the gel, double-stranded labelled oligonucleotides were found to be integrated into both the control and pCRISPRC in conditions containing Cas1_{CD} or Cas1_{CD}-Cas2_{CD} (Figure 6.3). No integration was observed following incubation with Cas2_{CD} alone, or the Cas1_{CD} active site variant E142A.

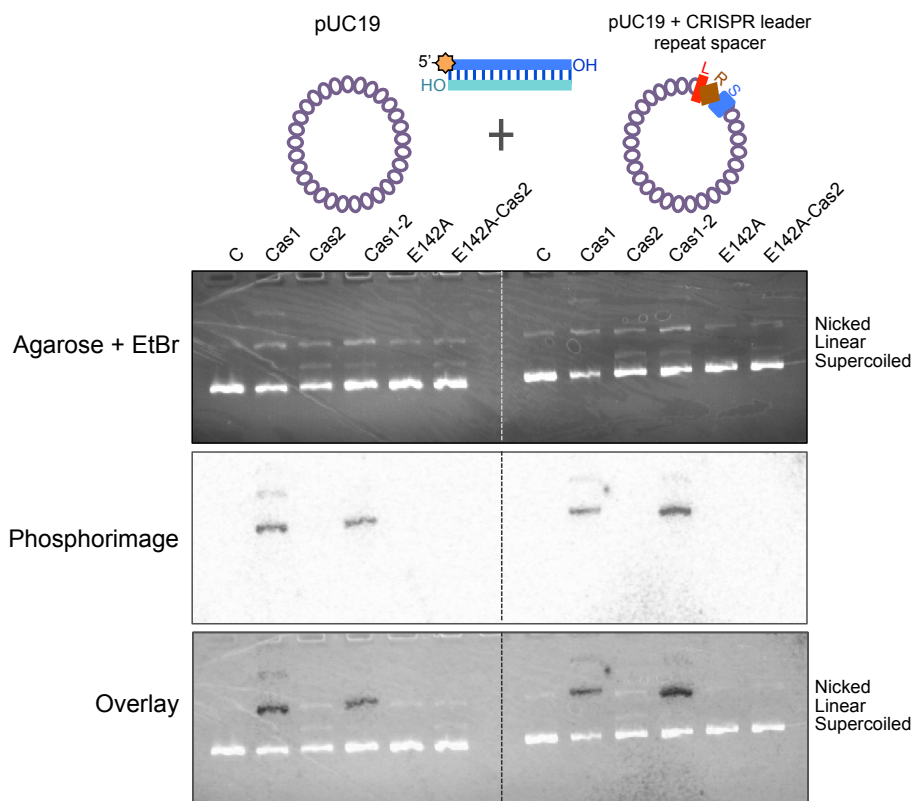


Figure 6.3 *in vitro* integration does not require a CRISPR array

An integration assay with double-stranded ³²P-labelled oligonucleotides (a duplex made by annealing CRISPR D spacer dup F and CRISPR D spacer dup R oligonucleotides, see Table 2.1 (p.47) for sequences) was carried out with either pUC19 (left), or a modified pUC19 containing a partial *S. solfataricus* CRISPR C leader, repeat and spacer insert (right) (pCRISPRC, see Table 2.3 (p.52)). The leader sequence is represented by a red rectangle, the repeat by a brown diamond and the spacer by a blue square. The protospacer (2 μM) was incubated with the indicated Cas protein (2 μM) at 55°C for 30 min before the addition of plasmid DNA (100 ng). The top gel image is a scan of the pre-stained EtBr agarose gel run to separate reaction products. The middle image is a phosphorimage of the dried agarose gel and the bottom image is a merged version of the other two scans to allow localisation of the radioactive signal to a plasmid form. The first lane for each substrate is a control without protein and the last two lanes contain Cas1_{CD} active site variant E142A, and E142A and Cas2_{CD}, respectively.

The incorporation of protospacers in the absence of CRISPR elements was surprising given the strict sequence specificity of adaptation *in vivo*. However, during these experiments another group reported similar findings, showing that integration of protospacers into plasmid DNA by the *E. coli* Cas1-Cas2 also occurred without the need for a CRISPR array (Nuñez et al., 2015b). This study found that protospacers were incorporated preferentially into a CRISPR insert in a pUC19 plasmid backbone. However, around one third of total integrations happened at other sites in the plasmid, especially within the ampicillin resistance gene (Nuñez et al., 2015b). On deep sequencing these off-target integrations were found to happen at palindromic sequences predicted to form hairpins. It was hypothesised that the palindromic nature of the repeat was crucial for recognition of the CRISPR locus by Cas1-Cas2 and that, *in vitro*, similar sequences could result in integration outwith the CRISPR array (Nuñez et al., 2015b).

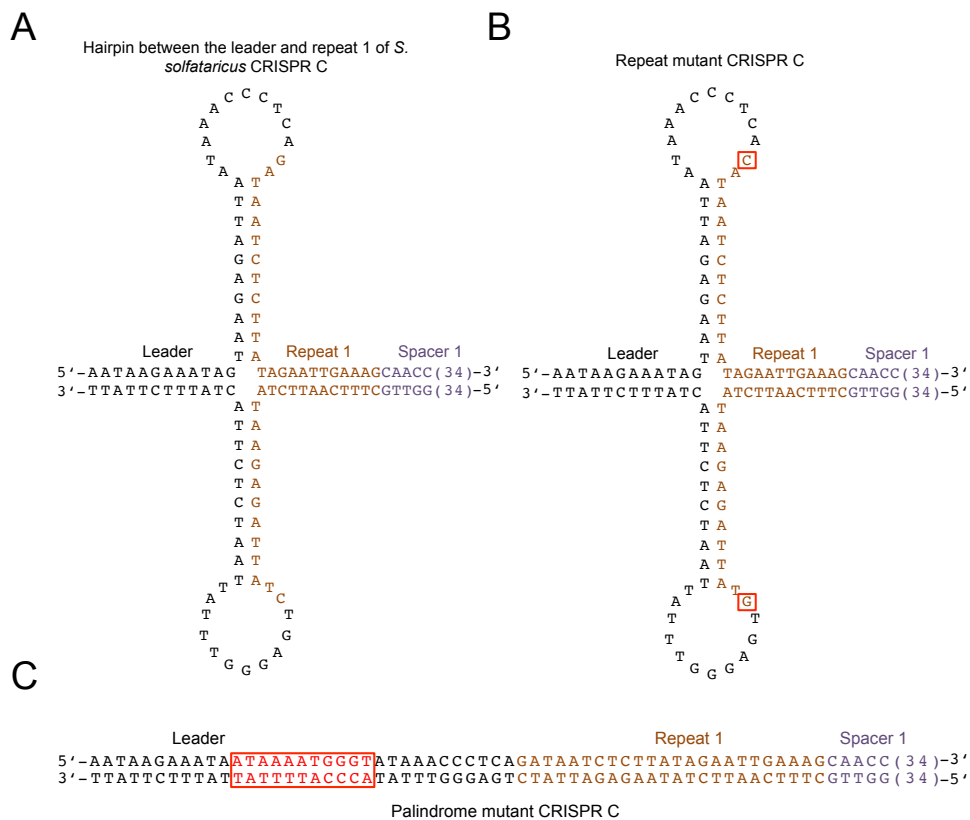


Figure 6.4 CRISPR C secondary structure and variants

A. The sequence of the CRISPR C leader, repeat 1 and spacer 1 was analysed and a palindromic region predicted to form a hairpin was identified at the leader-repeat 1 junction. The leader sequence is shown in black, the repeat in brown and the spacer in purple. Cloning was carried out to insert the last 101 bp of the leader, repeat 1 and spacer 1 of CRISPR C into the multiple cloning site of pUC19, using the *EcoRI* and *BamHI* restriction sites (to form pCRISPRC). Variants of the CRISPR C insert shown in **A** were made in which the first nucleotide of the repeat was changed from a G to a C (pCRISPRC rep

mutant) (**B**), or the palindromic sequence was altered (red boxed sequence) to disrupt the predicted secondary structure at the leader-repeat junction (pCRISPRC pal mutant) (**C**). The sequences of the CRISPR C inserts are shown Table 2.3 (p.52).

As I had previously identified a palindromic sequence between the leader and repeat 1 of CRISPR C (Chapter 4), I speculated that similar structural features might guide integration in *S. solfataricus*. To test this hypothesis in further integration experiments, variant CRISPR C inserts were designed containing either a mutation at the start of the first repeat or a disrupted palindromic sequence, which would no longer form a hairpin secondary structure. These inserts, shown in Figure 6.4, were then cloned into pUC19 (see Table 2.3 (p.52) for sequences).

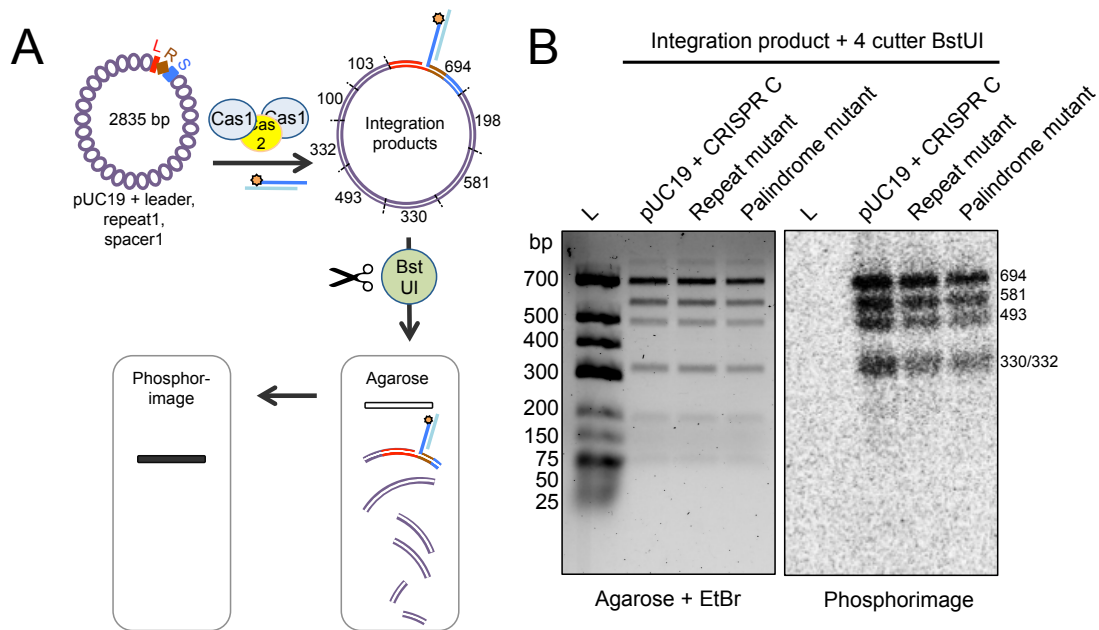


Figure 6.5 Restriction digest of Cas1_{CD}-Cas2_{CD} integration products

A. A shows a schematic of the methods used in **B**. An integration assay was carried out using supercoiled DNA, a 39 bp protospacer (a duplex made by annealing CRISPR D spacer dup F and CRISPR D spacer dup R oligonucleotide, see Table 2.1 (p.47) for sequences) and Cas1_{CD}-Cas2_{CD}. The plasmids used contained either: a wild-type CRISPR C partial leader, repeat, spacer insert (forming pCRISPRC); a version with the first nucleotide of the CRISPR repeat changed from a G to a C (pCRISPRC rep mutant); or a version with the palindromic sequence at the leader repeat junction disrupted (pCRISPRC pal mutant) (see Table 2.3 (p.52)) for details). Following the integration reaction products were digested with *Bst*UI, which produces 8 restriction products as indicated by dashed lines and sizes in bp. These products were separated on an agarose gel before phosphorimaging to identify where radioactive protospacers had been inserted. **B.** Shows the results of the assay described in **A**. The EtBr stained agarose gel of restriction products is shown on the left and the phosphorimage of this gel on the right. On digestion with *Bst*UI (10 units), the 5 longest restriction fragments (694, 581, 493, 330/332 bp) were visible by phosphorimaging for WT and mutant CRISPR C plasmids.

In order to identify if there was an obvious integration site preference for the Cas1_{CD}-Cas2_{CD} complex, integration assays were carried out with the variant CRISPR C plasmids and radiolabelled protospacers (Figure 6.5). The assay products were

separated on an agarose gel and the nicked plasmid band containing integrated protospacer was gel extracted. The extracted integration products were then subjected to restriction digest using the *Bst*UI restriction enzyme (New England Biolabs), which cut at CGCG sites to produce eight plasmid fragments (Figure 6.5, A) (see section 2.2.8.3.3 (p. 68) for method).

The restriction products were separated on an agarose gel, before phosphorimaging to reveal fragments containing radioactive protospacers (Figure 6.5, B). On phosphorimaging the five longest restriction fragments were clearly visible, indicating that they contained radioactive spacers. The longest fragment (694 bp) contained the CRISPR C or variant CRISPR C inserts. While this fragment had been the target of integration by the Cas1-Cas2 complex, the wild-type palindromic insert was not obviously enriched in radioactive protospacers compared to variant inserts. Furthermore, as the other large fragments produced by the digest all contained radioactive signal it seemed that integration by Cas1_{CD}-Cas2_{CD} was indeed taking place non-specifically all over the CRISPR C plasmids. The apparent lack of radioactivity in the shorter restriction fragments likely represented a proportional reduction in integrations relative to the length of the fragment, rather than site selection by Cas1_{CD}-Cas2_{CD}. These bands potentially contained integrated spacers, but the intensity of the signal presumably fell below the detection limit of this assay.

6.2.4 Protospacer end structure influences integration

The crystal structure of the protospacer-bound Cas1-Cas2 complex from *E. coli* revealed that the ends of protospacer DNAs are splayed and the 3' single-stranded ends are tightly bound by Cas1 subunits, with the 3' hydroxyl being positioned in the catalytic active site (Figure 6.6, A) (Nuñez et al., 2015a). In addition, *in vitro* integration by this complex was found to be most efficient with protospacer substrates with single-stranded 3' ends (Nuñez et al., 2015b).

In order to learn more about the structure of protospacers added during adaptation by the *S. solfataricus* Cas1_{CD}-Cas2_{CD} complex, the *in vitro* integration assay was carried out with protospacers with 8 nucleotide 3' or 5' overhangs or blunt ends (Figure 6.6, B). The 3' overhang protospacer led to a robust integration assay into the pUC19 plasmid. Incubation of this substrate with Cas1_{CD} or Cas1_{CD}-Cas2_{CD}

resulted in a strong nicking of the plasmid DNA and incorporation of the radioactive substrate into the plasmid.

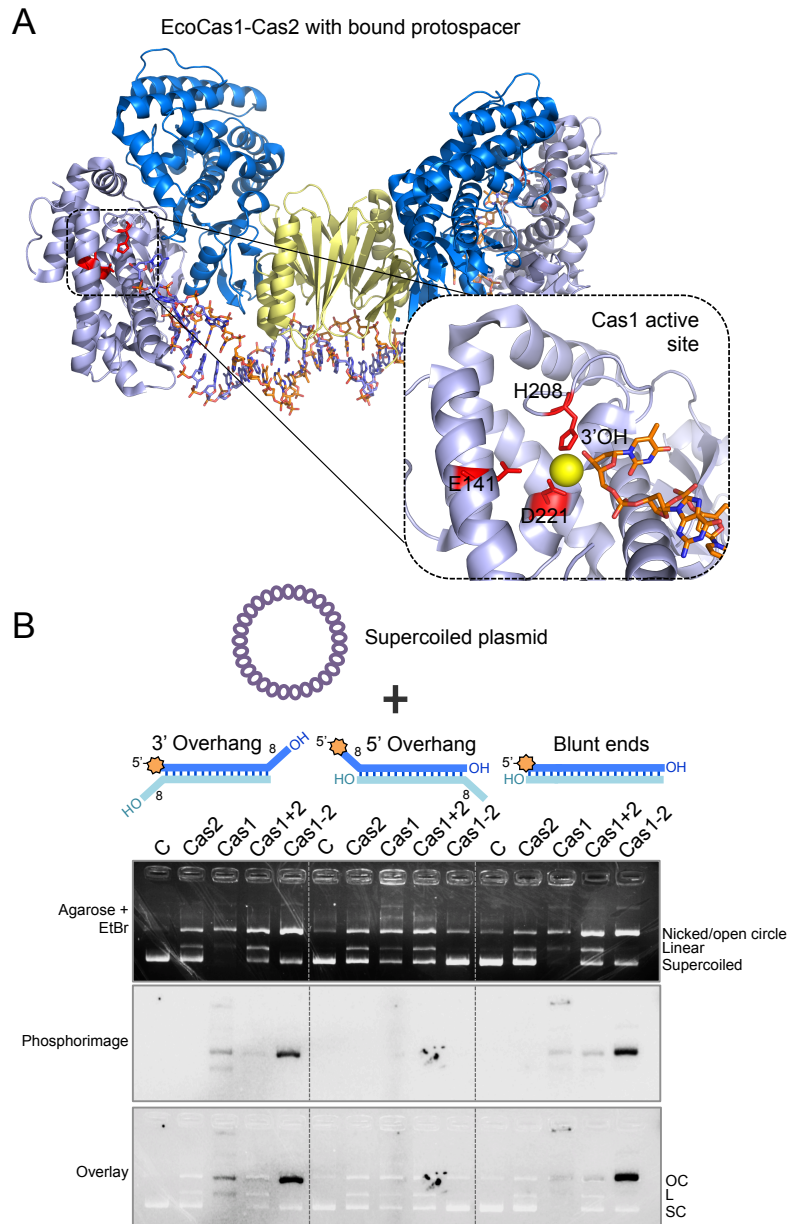


Figure 6.6 Protospacer end structure affects integration efficiency

A. The *E. coli* Cas1-Cas2 complex with bound protospacer DNA (PDB ID 5DS6) (Nuñez et al., 2015a). Two Cas1 dimers (blue and light blue) are bridged by a Cas2 dimer (yellow) with 33 bp of DNA spanning the length of the complex. Single-stranded 3' overhangs are tightly bound by the complex. The dashed box shows a zoomed view of the active site residues E141, D221 and H208 and a magnesium cation co-ordinating one of the 3' hydroxyls of the protospacer. **B.** Radiolabelled protospacers (2 μ M) with a 23 bp duplex region and either blunt ends or 3' or 5' single-stranded ends (8 nt) (made by annealing the following pairs of oligonucleotides: CRISPR D spacer dup F and R; CRISPR D spacer 3'end F and R; CRISPR D spacer 5'end F and R, see Table 2.1 (p.47) for sequences) were incubated with Cas1_{CD}-Cas2_{CD} (2 μ M) at 55 °C for 30 min. pUC19 plasmid DNA (100 ng) and MnCl₂ (5 mM) were then added before a further incubation at 55 °C. The assay products were separated by agarose gel electrophoresis (top). The first lane for each substrate is a control without

protein, followed by lanes containing Cas2_{CD}, Cas1_{CD}, Cas1_{CD} and Cas2_{CD} added to the reaction without a preincubation step (Cas1 + 2), or Cas1_{CD} and Cas2_{CD} which have been pre-incubated together with protospacer DNA and metal ions (Cas1-2). The middle image shows the result of phosphorimaging the agarose gel and the bottom image is the result of overlaying both the phosphorimage and the agarose gel scan.

A clear enhancement of the integration reaction was observed again when both Cas1_{CD} and Cas2_{CD} were added, compared to Cas1_{CD} alone. This enhancement was only observed when the two proteins were incubated together for 30 min at 55 °C in the presence of the protospacer substrate before being added to the assay (lane Cas1-Cas2 in Figure 6.6, B). When Cas1_{CD} and Cas2_{CD} were added directly into the integration assay without preincubation, a low level of integration, comparable to Cas1_{CD} alone, was observed (lane Cas1 + Cas2 in Figure 6.6, B). This supports the earlier conclusion that Cas1_{CD} and Cas2_{CD} are able to form a functional complex in the presence of protospacer DNA. However, the affinity and stability of the interaction seems to be much lower than that observed for the *E. coli* Cas1 and Cas2 proteins, which form a stable complex with a disassociation constant of ~290 nM (Nuñez et al., 2014).

Blunt-ended duplex oligonucleotides were good integration substrates for Cas1_{CD}-Cas2_{CD}, whereas protospacers with 8 nucleotide 5' overhangs led to very weak integration by Cas1_{CD}-Cas2_{CD} (Figure 6.6, B). It can be concluded from these results that the Cas1_{CD}-Cas2_{CD} complex cannot tolerate integration substrates with recessed 3' ends. If the 3' ends are recessed, the complex may not be able to position the 3' hydroxyl needed for nucleophilic attack in the active site of Cas1. In addition, it appears that under the conditions tested the complex itself cannot process the ends of the protospacer used here to access the 3' nucleophile. The strong integration observed with blunt ended protospacers may indicate that like the *E. coli* complex, Cas1_{CD}-Cas2_{CD} is able to open the ends of duplex DNA to direct single-stranded 3' ends into the active site of Cas1_{CD}.

6.2.5 Modifying 3' overhang length

The modification of protospacer ends by changing the 3' overhangs to be longer or shorter than 5 nucleotides led to a reduction in integration activity by *E. coli* Cas1-Cas2 *in vitro* (Nuñez et al., 2015a). This effect is likely due to the 3' hydroxyl no longer being exactly positioned in the active site of a Cas1 monomer, leading to an impaired ability to act as a nucleophile and to bring about a transesterification reaction.

As protospacer 3' single-stranded ends also seemed to be important for bringing about integration by the Cas1_{CD}-Cas2_{CD} proteins, I was interested in examining the effect of changing the length of these ends. Substrates were designed with a 29 nucleotide duplex region and variable length 3' ends, from 4 to 6 nucleotides. These substrates were added to integration assays with a pUC19 plasmid containing a CRISPR C insert, or a CRISPR C insert with a mutated repeat or palindromic sequence. Integration into each of the three plasmids was strongest when the 3' overhang length was five or six nucleotides in length (Figure 6.7). Duplex ends or four nucleotide overhangs led to a much-reduced level of integration. There was no obvious reduction of integration into the plasmids containing the mutated CRISPR C inserts.

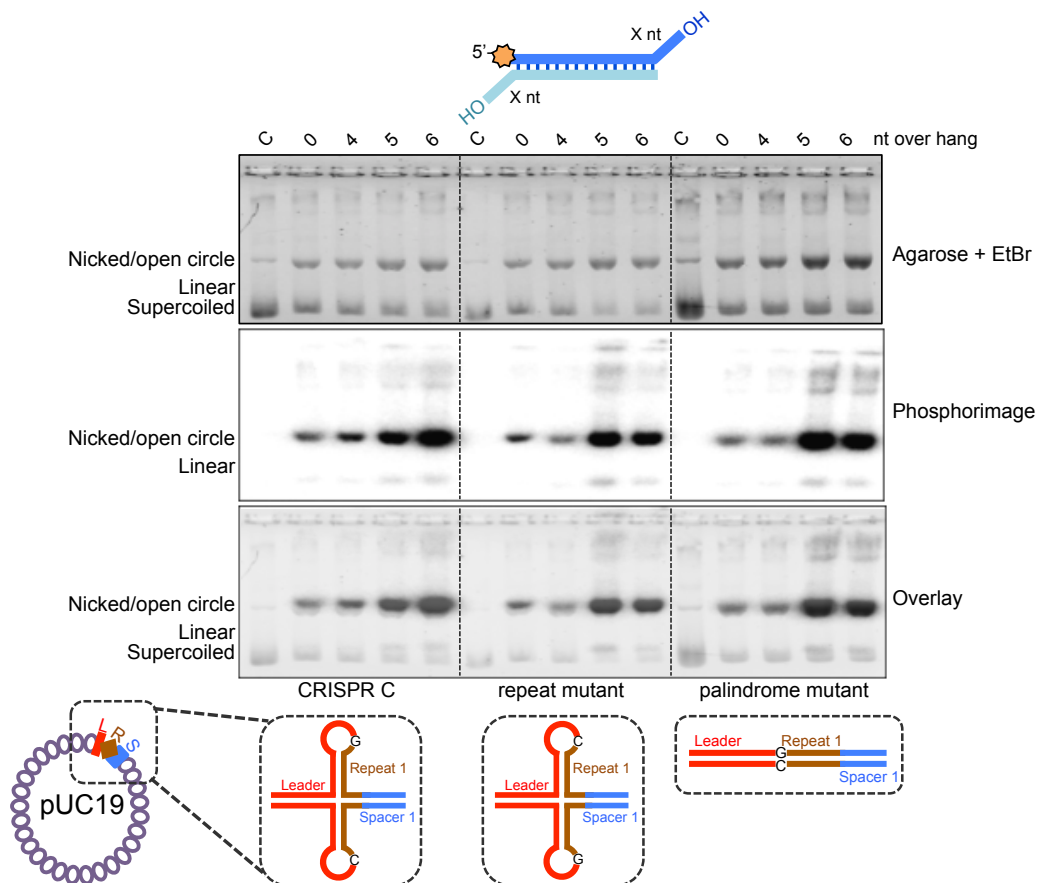


Figure 6.7 Modifying 3' overhang length affects integration

³²P-labelled protospacer substrates were made either with blunt ends or with 3' overhangs from 4 to 6 nucleotides in length (made by annealing the following pairs of oligonucleotides: CRISPR D spacer dup F and R; 4 nt overhang spacer F and R; 5 nt overhang spacer F and R; 6 nt overhang spacer F and R, see Table 2.1 (p.47) for sequences). These substrates (2 μ M) were added to Cas1_{CD}-Cas2_{CD} integration reactions containing pUC19 plasmid (100 ng) with either a wildtype CRISPR C insert (pCRISPRC) or a mutated version of the CRISPR C insert (pCRISPRC rep mutant and pCRISPRC pal mutant, see Table 2.3 (p.52) for details). The schematic below the gel images indicates how the mutant CRISPR C inserts differ from the wildtype. A pre-incubation of Cas1_{CD}, Cas2_{CD} (both at 2 μ M) and protospacer was carried

out at 55 °C for 30 min before the addition of MnCl₂ and plasmid DNA and a further 30 min incubation. The top image is a scan of the agarose gel used to separate the reaction products, the middle image is a phosphorimage of this gel and the bottom image is a composite of the two. The first lane of each assay is a control with protospacer with 4 nt ends added, but without protein.

Therefore, from this assay it seems that although protospacer 3' ends of 5-6 nt are favoured by the Cas1_{CD}-Cas2_{CD}, other end structures are also tolerated. This is similar to the preference observed for the *E. coli* integration complex, where 4-5 nt ends led to between 70 - 75% integration, whereas changing end length to 3 nt or 6 nt led to a 10 or 40% drop in integration, respectively (Nuñez et al., 2015a). This implies that the binding and co-ordination of protospacer 3' ends by the Cas1_{CD} subunits during integration occurs in a similar manner to that in the integration complex of *E. coli*.

6.2.6 No PAM processing during *in vitro* integration

In addition to the recent publication of the *E. coli* Cas1-Cas2 structure in complex with DNA, the Wang laboratory also reported that the complex processed protospacer 3' ends within PAM motifs (Wang et al., 2015). PAM complementary 5'-CTT-3' motifs at position +6 to +8 in the 3' protospacer ends were bound in a sequence-specific manner by residues close to the active site of Cas1 and a cut was made between the C and T residues to trim the overhang to the optimal five nucleotide length for integration (Wang et al., 2015).

This finding prompted the design of two further protospacer substrates to be tested with the Cas1_{CD}-Cas2_{CD} complex. These substrates had longer, nine nucleotide, 3' ends with a *S. solfataricus* PAM sequence (a GG or CC dinucleotide) at position +6 and +7 of the 3' single-strand flap. Integration reactions were carried out with the new substrates to investigate whether PAM processing also occurred in the Cas1_{CD}-Cas2_{CD}-protospacer complex before integration. An integration assay demonstrated that both the GG and CC PAM substrates were poor integration substrates compared to protospacers with five nucleotide flap length (Figure 6.8, A). This result confirms that end lengths that differ from the preferred five nucleotides cannot be used efficiently by Cas1_{CD}-Cas2_{CD} to mediate a nucleophilic attack during integration.

A nuclease assay was performed with Cas1_{CD}-Cas2_{CD} on substrates with different 3' end structures, including the longer CC and GG 3' PAM overhang substrates, to determine whether they were processed for integration. As shown in Figure 6.8 (B),

none of the substrates with or without PAM motifs were processed by the Cas1_{CD}-Cas2_{CD}. This result might indicate that cleavage requires the PAM sequence to be at a different position in the 3' single-strand, to allow precise binding and processing in the Cas1_{CD} active site. Furthermore, in *S. solfataricus* recognition and processing of PAM motifs might be carried out by other factors before the protospacer is inserted by Cas1_{CD}-Cas2_{CD}. For example, the exonuclease DNA processing activity of Cas4 (Zhang et al., 2012a) may play a part in processing foreign DNA to spacer length before insertion.

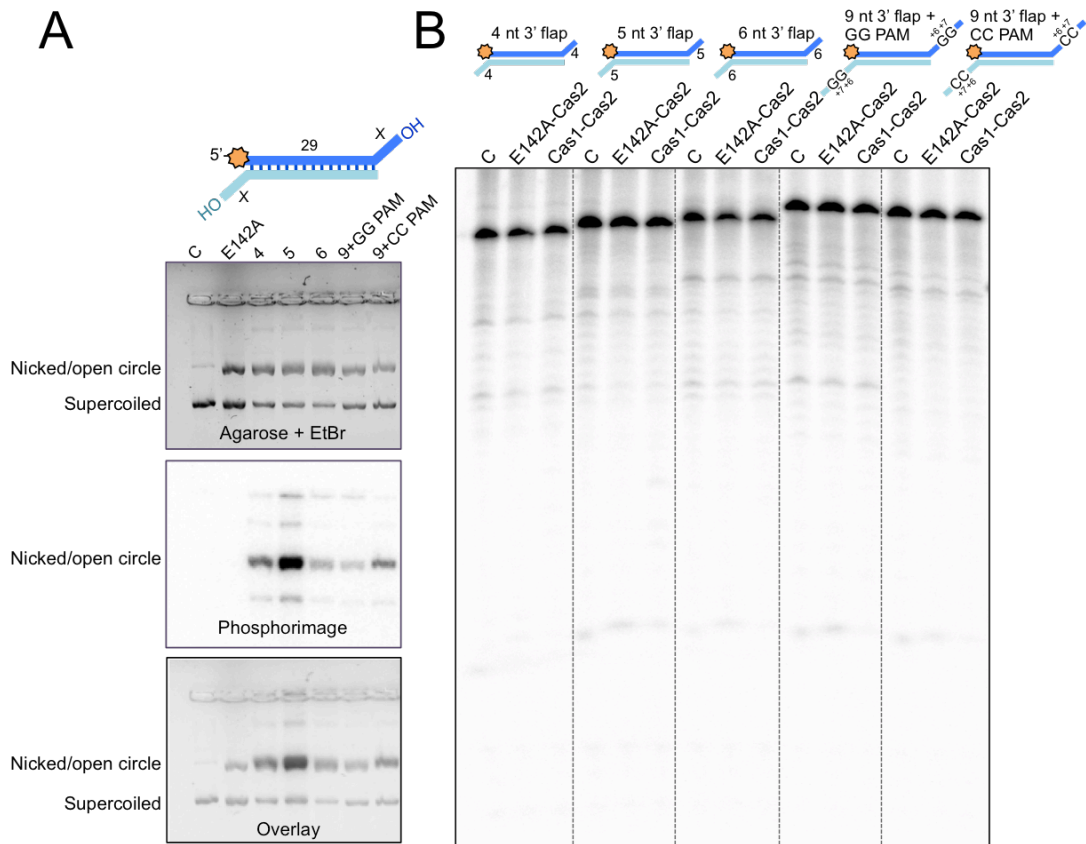


Figure 6.8 PAM sequences in protospacer ends are not processed

A. 5'-³²P labelled substrates with differing 3' single-stranded overhangs were added to integration assays with Cas1_{CD}-Cas2_{CD} and a pUC19 plasmid with a CRISPR C insert (pCRISPRC, see Table 2.3). 3' protospacer ends were 4, 5, 6 or 9 nt in length and the duplex region was 29 nt (made by annealing the following pairs of oligonucleotides: 4 nt overhang spacer F and R; 5 nt overhang spacer F and R; 6 nt overhang spacer F and R, GG PAM F and R; CC PAM F and R, see Table 2.1 (p.47) for sequences). The 9 nucleotide ends had a CC or GG dinucleotide PAM at position +6 and +7 of the overhang. The first lane is a control without Cas proteins and the second lane is a control with Cas1_{CD} active site variant E142A and Cas2, both with 4 nt overhangs added. Cas proteins (2 μM) and protospacer substrates (2 μM) were incubated together at 55 °C for 30 min before the addition of 5 mM MnCl₂ and 100 ng plasmid DNA and a further 30 min incubation. The top image is a scan of the agarose gel used to separate assay products. The middle image is a phosphorimage of the first gel, and the bottom image is the result of combining the other two

scans. **B.** A nuclease assay was performed on the substrates used in **A** to determine whether 3' overhangs were processed by Cas1_{CD}-Cas2_{CD} before integration. Protospacers (50 nM) were incubated at 55 °C for 30 min with Cas1_{CD}-Cas2_{CD} (200 nM each) in the presence of divalent metal ions (50 mM MnCl₂). The products of the reaction were separated on a 20% denaturing polyacrylamide-TBE gel before phosphorimaging. The first lane is a control without proteins added and the second lane a control with Cas1_{CD} substituted for active site mutant E142A.

6.2.7 The effect of protospacer duplex length on integration

To assay the effect of protospacer length on the *in vitro* integration reaction, protospacer substrates were designed with five nucleotide 3' overhangs and duplex regions of 24, 29 or 34 bp. Each of these substrates was integrated into supercoiled DNA with a low efficiency by the Cas1_{CD} protein alone, with the 24 bp duplex resulting in the weakest integration (Figure 6.9, A). When both Cas1_{CD} and Cas2_{CD} were present in the integration assay, a much-enhanced level of integration was observed. The 29 and 34 bp duplex substrates were strongly incorporated into plasmid DNA, whereas the 24 bp duplex protospacer led to weaker integration by the complex (Figure 6.9, A).

Protospacer size in the *E. coli* K12 CRISPR arrays is very defined, with all spacers being either 32 or 33 bp in length (Lintner et al., 2011b). However, it was recently demonstrated that the *E. coli* Cas1-Cas2 complex can integrate a wide range of spacer lengths (33 ± 15 bp) *in vitro*, leading the authors to suggest that another factor is involved in determining the strict spacer length before integration in this system (Nuñez et al., 2015a). In *S. solfataricus* the spacers in CRISPR arrays A to F range in size from 34 to 48 bp, with 39 bp being the most common spacer length (Figure 6.9, B) (Lintner et al., 2011b). The three substrates tested here cover the range of lengths of the majority of *S. solfataricus* spacers, with total lengths (including the two 5 nt 3' overhangs) of 34, 39, 44 nt. The Cas1_{CD}-Cas2_{CD} complex integrated each of these substrates, although there was a clear preference for the 39 nt and 44 nt spacers.

The range of spacer sizes integrated *in vitro* by Cas1-Cas2 from *S. solfataricus* and *E. coli* implies a flexibility in the structure of the complex and DNA binding that was not apparent from the crystal structure of the *E. coli* proteins. However, the ability of the adaptation complex to perform half-site integrations *in vitro* no matter the spacer length may also have led to the lack of specificity observed here. The wide protospacer range in the CRISPR arrays of *S. solfataricus* also suggests that

protospacer selection happens through a less defined mechanism in this system compared to *E. coli*. However, whether Cas1_{CD}-Cas2_{CD} is involved in protospacer size determination remains to be investigated.

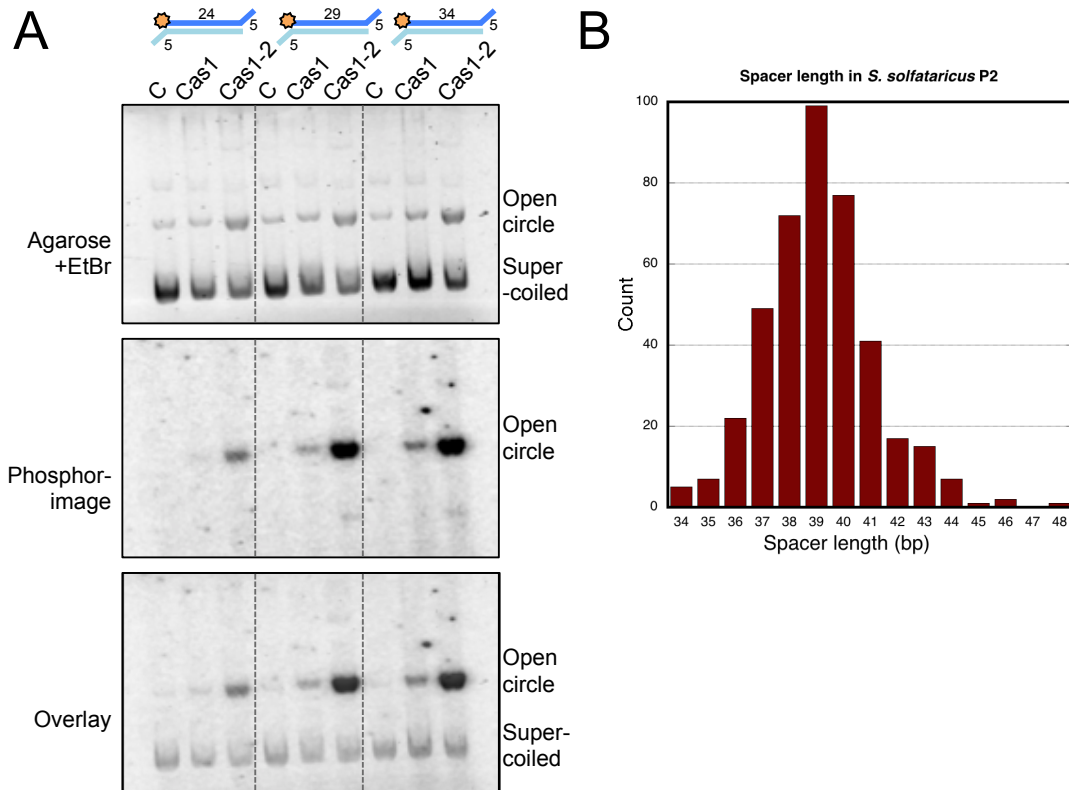


Figure 6.9 Effect of protospacer duplex length on integration

A. Substrates were designed with 5 nt 3' single-strand overhangs and varying length duplex regions of 24, 29 or 34 bp. These ³²P-labelled substrates (made by annealing the following pairs of oligonucleotides: 24 nt duplex spacer F and R; 29 nt duplex spacer F and R; 34 nt duplex spacer F and R, see Table 2.1 (p.47) for sequences) were added to integration reactions with Cas1_{CD}-Cas2_{CD} and the pCRISPRC plasmid. Cas proteins (both at 2 μM) were incubated with protospacer (2 μM) for 30 min at 55 °C before the addition of plasmid DNA (100 ng) and MnCl₂ (5 mM) and a further 30 min incubation. The first lane for each substrate is a control without protein, followed by lanes containing the Cas1_{CD} protein alone or the Cas1_{CD}-Cas2_{CD} complex. The top image is the agarose gel used to separate assay products, the middle image is a phosphorimage of this gel and the bottom image is a composite of the two. **B.** Histogram of spacer lengths present in the CRISPR arrays (A - F) of *S. solfataricus* P2, based on data from Lintner et al. (2011b).

6.2.8 Mutation of conserved residues outwith the Cas1_{CD} active site

In Chapter 5, conserved residues in proximity to the Cas1_{CD} active site were shown to be required for the disintegration of branched structures. These residues, W150, R166 and N175, are well conserved among Cas1 proteins and surround a positively

charged cleft, which is involved in binding the 3' end of protospacers (Nuñez et al., 2015a) and may also be involved in binding and orientating host DNA.

Variant Cas1_{CD} proteins with these conserved residues mutated to alanine were assayed in integration reactions to investigate their role in integration. In this assay none of the mutant proteins gave rise to an integration product (Figure 6.10, A). Mutation of residue R166 was also found to abolish disintegration activity by Cas1_{CD}, and mutation of W150 led to a strong reduction in product obtained. In contrast, the N175A variant protein was still able to disintegrate branched structures with an efficiency approaching that of the wildtype Cas1_{CD}.

The protospacer-bound *E. coli* Cas1-Cas2 structure (PDB ID 5DS6) demonstrated that the R163 residue, equivalent to arginine R166, is located on a flexible loop region in the arginine-rich cleft leading to the Cas1 active site (Nuñez et al., 2015a). In the DNA-bound structure R163 is in position to hydrogen bond with protospacer 3' single-strands, and contributes to situating them in the Cas1 active site for integration (Figure 6.10, B). The flexible loop may indicate that R166 is repositioned during DNA binding or integration to orientate DNA ends. The predicted key interaction of R166 with incoming DNA ends explains the complete absence of both integration and disintegration activity of the R166A variant.

The N175A mutation had only a minor effect on disintegration, but abolished integration. The equivalent residue in *E. coli* (N173) does not seem to interact directly with incoming DNA, but rather to hydrogen bond with and position the R163 residue of the flexible loop.

Tryptophan W150 has been shown to be essential for integration, and important for disintegration, by the Cas1_{CD} protein. As planar aromatic amino acids are conserved at this position in Cas1 proteins it is likely that this residue may be involved in π -stacking interactions with DNA bases. From the *E. coli* crystal structure, the equivalent Y149 residue is not involved in binding the protospacer DNA (Figure 6.10, B). Therefore, perhaps this residue is important in binding and positioning the host CRISPR array close to the incoming protospacer during adaptation.

To summarise this section, *in vitro* reconstitution of the integration reaction performed by Cas1_{CD}-Cas2_{CD} yielded insights into substrate requirements during adaptation. I demonstrated that blunt or 3' single-stranded protospacer ends are

required for integration, with recessed 3' ends abolishing the reaction. Furthermore, the Cas1_{CD}-Cas2_{CD} complex integrated a range of protospacer sizes, which matches the *in vivo* diversity in spacer length in *S. solfataricus*. I have also shown that Cas1_{CD} alone is essential for integration, while Cas2_{CD} enhances the strength of the reaction, implying that a functional complex is formed under the conditions tested. Finally, a surprising finding was that the *in vitro* integration reaction did not share the specificity of *in vivo* adaptation, which only occurs between the leader and first repeat, and instead, seemed to occur non-specifically, independently of CRISPR elements.

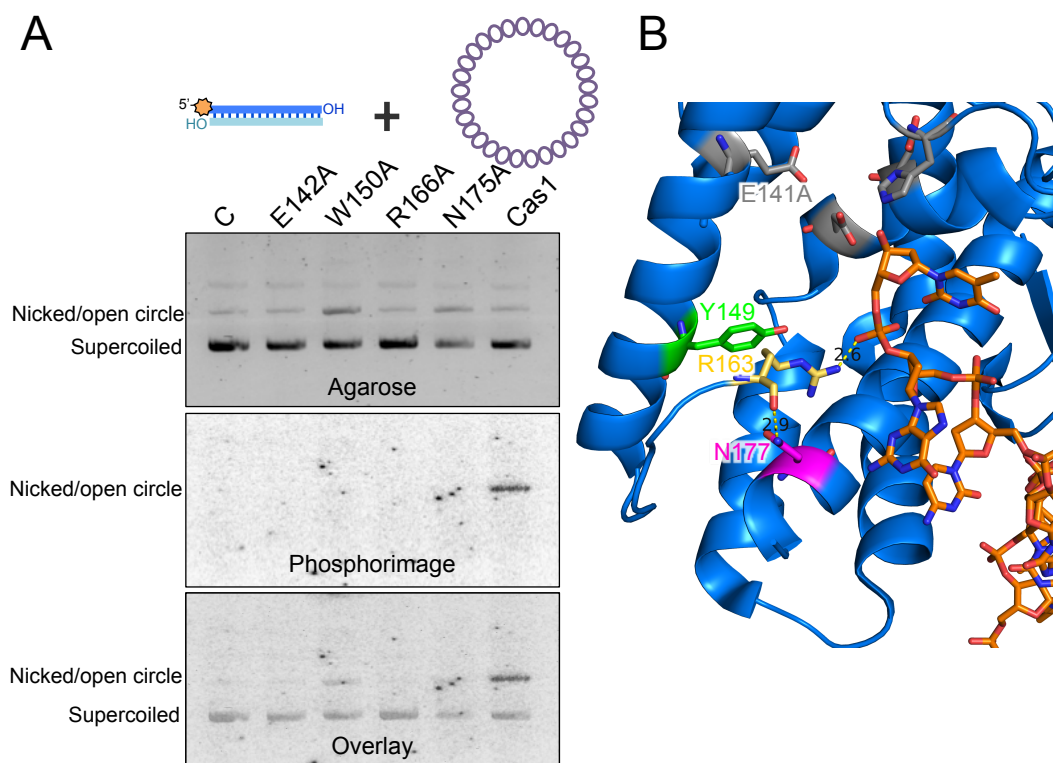


Figure 6.10 Cas1_{CD} residues important for integration *in vitro*

A. Conserved residues outwith the canonical active of Cas1_{CD} were identified and single residue variants were created by mutation of W150, R166 and N175 to alanine. The variant proteins (2 μ M) were assayed for integration activity. A duplex ³²P-radiolabelled protospacer (made by annealing CRISPR D spacer dup F and CRISPR D spacer dup R oligonucleotides, see Table 2.1 (p.47) for sequences) and pUC19 DNA (100 ng) were used as substrates for the integration. The first lane for each substrate is a control without protein, followed by lanes containing the active site variant E142A and each of the new variants to be tested. The last lane is a positive control containing active Cas1_{CD}. The products of the assay were separated by agarose gel (top), before gel drying and phosphorimaging (middle). The bottom image is a composite of the agarose and phosphorimage. **B.** The *E. coli* Cas1 (PDB ID 5DS6) with bound protospacer DNA. Conserved residues equivalent to those mutated in **A** are highlighted. Metal binding active site residues H208, D221 and E141 are coloured in grey, with E141A labelled. N177 is shown in pink, R163, present on a partially solved flexible loop region, is shown in yellow and Y149 in green (equivalent to N175, R166 and W150 in *S. solfataricus*).

6.2.9 Sequencing of the integration sites of *S. solfataricus* Cas1-Cas2

High-throughput sequencing of *in vitro* integration sites of the Cas1-Cas2 complex had suggested that structural features may guide adaptation, as palindromic regions were found to be hotspots for spacer insertion (Nuñez et al., 2015b). To look for similar features that might explain the apparent lack of specificity of the Cas1_{CD}-Cas2_{CD} integration reaction, an assay was developed that combined *in vitro* integration with PCR amplification and sequencing of the insertion site (see section 2.2.8.4 (p. 68) for method).

The first step of this assay was a standard Cas1_{CD}-Cas2_{CD} integration reaction with an optimal protospacer substrate comprising a 39 bp duplex region and 5 nt 3' single-stranded ends. Next, a forward primer complementary to one strand of the inserted protospacers with an internal *Nco*I site, and a reverse primer complementary to the pCRISPR C plasmid with an internal *Xho*I site were used to amplify through the protospacer insertion sites in plasmid DNA (Figure 6.11. A). PCR products from this assay were only produced from reactions containing Cas1_{CD} and protospacer DNA (Figure 6.11, B). Reactions with only Cas2_{CD} or variant Cas1_{CD} E142A did not yield any PCR products, confirming that amplification only occurs when the protospacer has been covalently joined to the plasmid in the active site of Cas1_{CD}. In accordance with the seemingly nonspecific nature of the integration reaction *in vitro*, a smear of PCR products was obtained. This smear resulted from integrations taking place at hundreds of sites at different distances from the reverse primer, leading to the amplification of a range of different-length products.

Following PCR amplification of integration sites, PCR products were digested at the primer restriction sites and ligated into the pEHISTEV plasmid (Liu & Naismith, 2009). Transformants were checked for inserts by colony PCR (Figure 6.11, C). Positive clones were found to contain a range of insert sizes, confirming that integration had taken place at many sites in the pCRISPRC plasmid. There was no apparent specificity for insertion at the leader-repeat 1 junction, which would have resulted in an insert size of 350 bp.

To look for conserved sequences or structural features at the integration sites, 45 positive clones were sent for sequencing (GATC Biotech). Integrations were found

to have happened all around the plasmid DNA with no apparent selection for plasmid features, such as the ampicillin resistance gene, which was found to be a hotspot for spacer insertion by the *E. coli* Cas1-Cas2 (Nuñez et al., 2015b). Sequences upstream and downstream of the integration sites were analysed for structural motifs; however, no secondary structure was identified, indicating that, *in vitro* at least, structure does not guide integration by Cas1_{CD}-Cas2_{CD}.

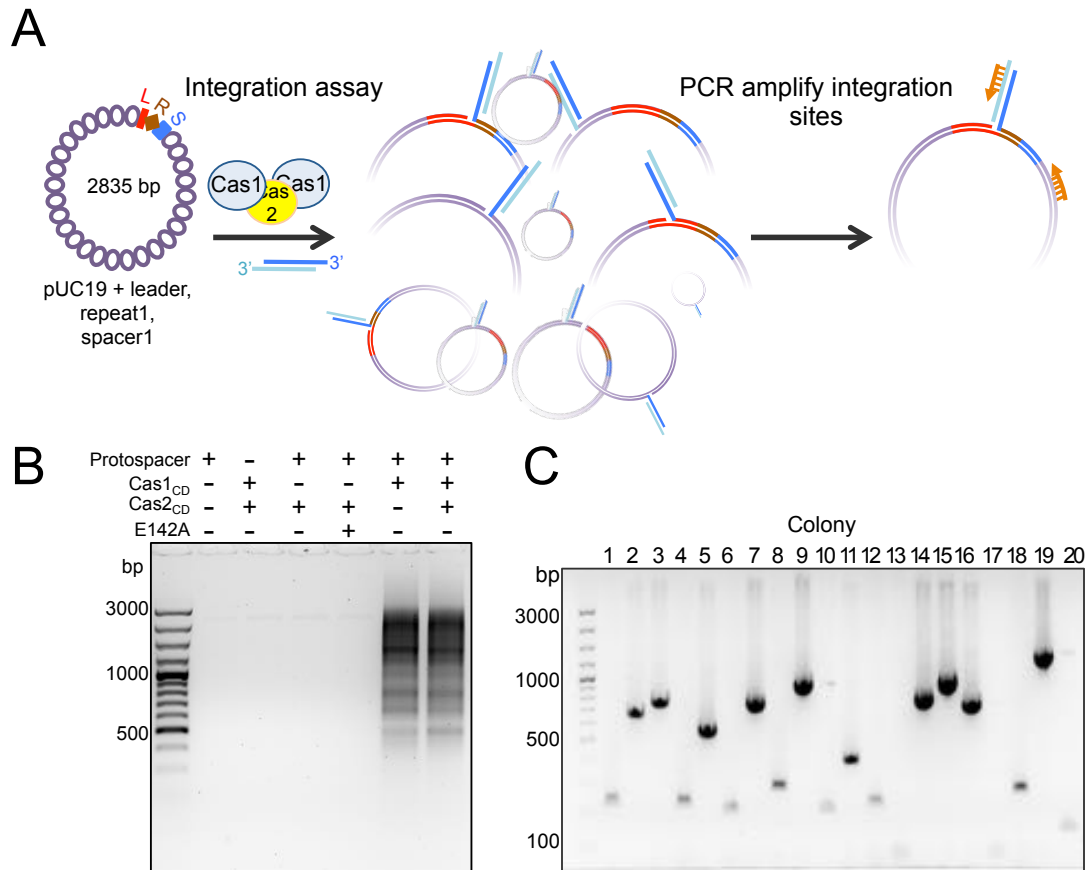


Figure 6.11 A PCR assay to amplify integration sites

A. Schematic of the strategy used to amplify integration sites. An integration reaction into the pCRISPRC plasmid (see Table 2.3 (p.52)) was performed with Cas1_{CD}-Cas2_{CD} and protospacers with 5 nucleotide 3' single-stranded ends (made by annealing PCR protospacer F and PCR protospacer R oligonucleotides, see Table 2.1 (p.47) for sequences), as described previously. The products of this assay were then amplified by PCR using a forward primer complementary to one strand of the inserted protospacer (Primer *NcoI* F) and a reverse primer complementary to the pCRISPRC plasmid (Primer *XhoI* R1) (see Table 2.1 for primer sequences). **B.** An agarose gel pre-stained with ethidium bromide showing the results of PCR amplification of integration sites. **C.** An example screen from colony PCR following ligation and cloning of amplified integration sites. A variety of insert sizes are visible, which correspond to protospacer integration by at sites varying in distance from the reverse primer used to amplify the integration product.

The 10 nucleotides around the integration sites were also compared for sequence similarities and a sequence logo was generated using the WebLogo server (Crooks,

2004) (Figure 6.12, B). This logo revealed a clear motif present at the integration sites chosen by Cas1_{CD}-Cas2_{CD}. A preference for a C or G residue at the +1 position, the nucleotide to which new spacers are joined, was observed, as well as a strong selection for a C at the -2 position (Figure 6.12, A and B). The motif preferred by the Cas1_{CD}-Cas2_{CD} for integration closely matches that of the *bona fide* integration site 1 between the leader and repeat 1 of the CRISPR C array. Furthermore, the sequence motif identified here is the same as that selected by Cas1_{CD} during the disintegration reaction (Chapter 5), implying that the specificity observed here is due to selection by Cas1_{CD}, with no influence from Cas2_{CD}.

There are six CRISPR arrays and two sets of *cas1* and *cas2* genes in the *S. solfataricus* genome. The second set of adaptation proteins, Cas1_{AB}-Cas2_{AB} (SSO1405-SSO1404) were expressed and purified in order to examine their integration site specificity. These proteins were then assayed in the same coupled integration, PCR and sequencing reaction that had been carried out for the Cas1_{CD}-Cas2_{CD} proteins. A pUC19 plasmid was used as the acceptor plasmid with a CRISPR A leader, repeat 1, spacer 1 insert (pCRISPR A see Table 2.3 (p.52)) for details). Again, the Cas1_{AB}-Cas2_{AB} proteins integrated spacers outside of the CRISPR insert at short sequence motifs similar to site 1. For these proteins an even stronger preference for a G was observed at the +1 position, and a C residue at -2 was also frequently selected (Figure 6.12, A and C).

At site 1 of CRISPR A, a G residue is found at position +1 and a C at position -2, while at integration site 2, the +1 nucleotide is a C and the -2 nucleotide is variable. This confirms again that Cas1-Cas2 complexes in *S. solfataricus* have an inherent sequence specificity that leads to the selection of sequence motifs matching site 1. Therefore, it is likely that the first half-site integration during adaptation takes place at this sequence rather than at the leader-distal site 2. Interestingly, as for the disintegration reaction, no selection by Cas1 was observed for the -1 position, indicating that Cas1 does not interact specifically with this nucleotide during the joining of new spacers.

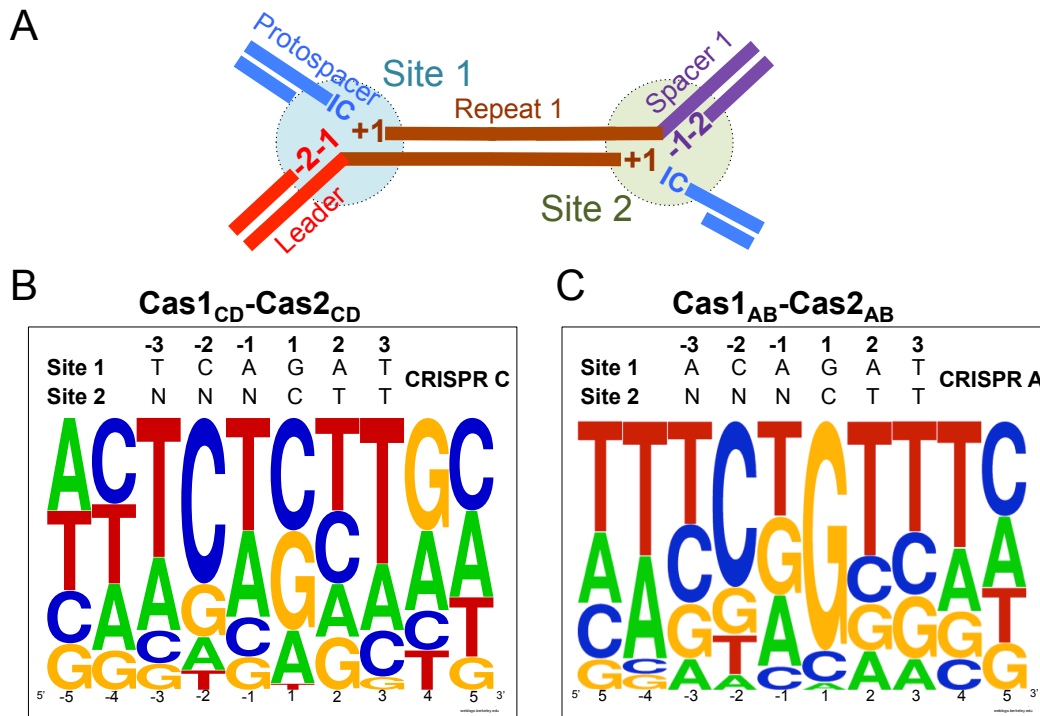


Figure 6.12 Sequence motifs at Cas1-Cas2 integration sites

A. Schematic showing the structure of the two half-site integrations carried out by Cas1-Cas2 during adaptation. The first half-site reaction takes place at the leader-proximal 5' end of repeat 1 (site 1) and the second at the leader-distal 5' end of repeat 1 (site 2). Incoming protospacer ends are shown in blue. The nucleotide to which new spacers are joined is denoted the +1 nucleotide. **B.** A sequence logo was generated on the WebLogo server (Crooks, 2004) following PCR amplification, cloning and sequencing ($n=45$) of the integration sites selected by Cas1_{CD}-Cas2_{CD} proteins. The residues are numbered from the +1 nucleotide to which new spacers were joined. The sequence of the *in vivo* integration site 1 and 2 of CRISPR C is shown. The relative height of the residues indicates the frequency at which this nucleotide is found at this position in inserts sequenced. **C.** As for **B**, except sequences analysed were generated integration sites selected by the Cas1_{AB}-Cas2_{AB} proteins ($n=45$). The sequence of CRISPR A site 1 and 2 is shown above the logo.

These results demonstrated that the *S. solfataricus* Cas1-Cas2 complexes did impose a level of selection on the adaptation process. However, alone these complexes integrated spacers at short motifs, present hundreds of times in the pUC19-CRISPR plasmids, similar to the *in vivo* leader-repeat 1 junction. This suggested that, *in vivo*, additional Cas proteins or other host factors are required to guide integration uniquely to the leader-repeat 1 junction.

6.2.10 Searching for a host factor

While the experimental work in this chapter was being carried out a small non-Cas protein in *E. coli*, called the integration host factor (IHF), was reported to greatly enhance the specificity of the Cas1-Cas2 integration reaction *in vitro* (Nuñez et al., 2016). I hypothesised that a similar host protein factor may be involved in directing

specific integration by the *S. solfataricus* Cas1-Cas2 proteins. To test this hypothesis *in vitro* integration reactions were carried out, followed by PCR amplification of the insertion sites as described previously. However, this time increasing volumes of cleared *S. solfataricus* cell lysate (prepared as described by Götz et al. (2007)) were added to the initial integration reactions (see section 2.2.8.4.1 (p.69) for method). Cas1_{AB}-Cas2_{AB} were assayed against a pUC19 plasmid containing the associated CRISPR A array. As increasing volumes of lysate were added to the Cas1_{AB}-Cas2_{AB} reactions, the smear of non-specific products obtained following PCR amplification was reduced and a specific band appeared at the size expected (450 bp) for integrations uniquely at the CRISPR A leader-repeat 1 junction (Figure 6.13, A). A faint band at this size was also observed in the assay with cell lysate only and no added Cas proteins, which might indicate that Cas1_{AB}-Cas2_{AB} are present at low levels in the lysate and led to some integration of the added protospacers.

To confirm the location of the specific integrations taking place in the presence of lysate, the PCR products obtained from these reactions were cloned and sequenced. To date only nine insertion sites from these assays have been sequenced; however, for each of these sequences the protospacer insertion site was at the leader-repeat 1 junction of the CRISPR A array. These results demonstrated that there was a factor in cell lysate essential for guiding specific integration during adaptation by Cas1_{AB}-Cas2_{AB} proteins in *S. solfataricus*.

The sequencing of insertion sites also revealed that added protospacers had undergone processing during the integration reaction. Each of the nine protospacers sequenced had had between 1 and 5 nucleotides removed from the inserted single-stranded 3' end before or during the integration reaction (Figure 6.13, B). The nature of the PCR amplification used to sequence integration sites only covers one of the 3' ends of the inserted protospacer. Therefore, further work will be required to investigate whether the integrations observed are full or half-site reactions and whether both 3' ends are processed.

The addition of *S. solfataricus* lysate did not confer the same specificity to the integration reaction performed by Cas1_{CD}-Cas2_{CD} (unpublished data, White lab). The absence of specificity of the Cas1_{CD}-Cas2_{CD} integration in the presence of cell lysate may suggest that other factors, perhaps induced by viral infection, are required for specific integration at the leader-repeat junction. I previously found that

the Cas1_{CD} protein was present at very low levels in absence of infection (chapter 3). Silencing in the absence of infection might also affect associated Cas proteins and host co-factors required for integration. In addition, the CRISPR C leader insert in the pUC19 plasmid was much shorter than the CRISPR A leader insert (107 bp compared to 537 bp). It could be that there is an essential motif missing in the truncated CRISPR C leader sequence that acts as a recognition motif for host factors required to guide specific integration.

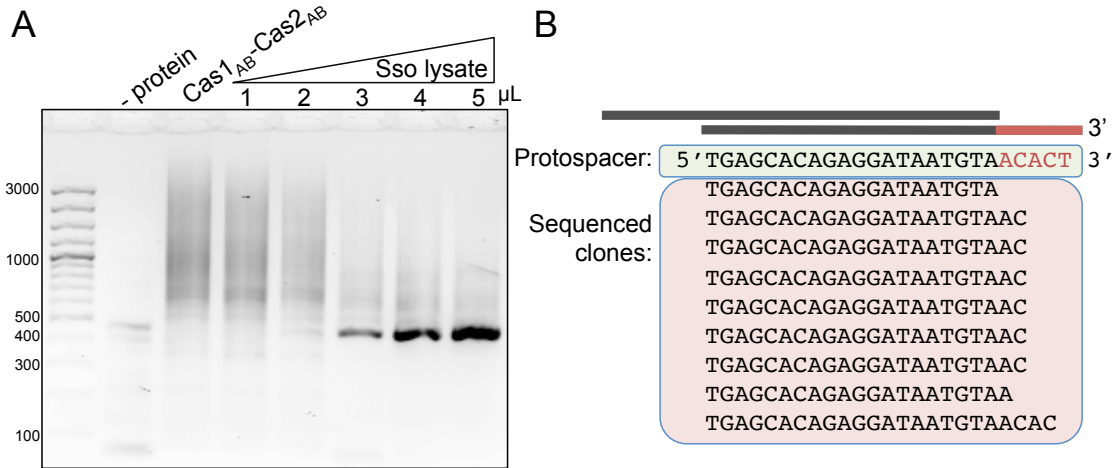


Figure 6.13 A component of *S. solfataricus* lysate needed for specific integration

A. A PCR amplification of insertion sites following integration of protospacer DNA (made by annealing PCR protospacer F and PCR protospacer R oligonucleotides, see Table 2.1 (p.47) for sequences) into the pCRISPR A plasmid (see Table 2.3) by Cas1_{AB}-Cas2_{AB} with or without the addition of *S. solfataricus* cell lysate. The first lane is a DNA ladder, lane two is a lysate-only control without added Cas proteins and the third lane is an amplification carried out following an integration reaction with only Cas1_{AB}-Cas2_{AB}. The subsequent lanes are PCR amplifications from Cas1_{AB}-Cas2_{AB} integration reactions (10 μl) carried out in the presence of increasing volumes of cell lysates, from 1 to 5 μl. Following a 25-cycle PCR amplification, products were separated on a 1.5% agarose gel. **B.** The PCR products obtained in the 5 μl lysate condition were cloned and sequenced. At the nine integration sites sequenced, the inserted 3' single-stranded end of the protospacer had been processed. The structure and sequence of the protospacer used in the assay is shown (top) with the inserted 5 nt 3' overhang shown in red. The protospacer sequences following integration are shown below in the red box. Between one and five nucleotides from the 3' end had been removed during integration.

6.3 Discussion

Integration of new spacers between the leader and repeat 1 of the CRISPR array is essential to CRISPR immunity. In this chapter I studied the requirements of protospacer and host DNA for successful integration *in vitro* by Cas1_{CD}-Cas2_{CD} from *S. solfataricus*. I also revealed for the first time that the Cas1-Cas2 proteins from *S.*

sofataricus are not sufficient for site-specific integration *in vitro* and that a factor in the *S. sofataricus* cell lysate is required to guide integration to the leader-repeat 1 junction.

6.3.1 Substrate structure is important for integration

I firstly demonstrated that Cas1_{CD}-Cas2_{CD} integrated protospacers into plasmid DNA *in vitro* (Figure 6.1). Integration of protospacers into supercoiled DNA by the *E. coli* Cas1-Cas2 proteins has also recently been shown *in vitro* (Nuñez et al., 2015b). Structural deformation, such as DNA bending or opening, is known to be important for integration of viral DNA (Surette & Chaconas, 1989; Maertens et al., 2010) as it positions integration sites with the correct spacing required for binding and processing by integrase active sites. In adaptation, it is conceivable that DNA supercoiling might be crucial in opening the host genome and allowing the Cas1-Cas2 complexes to access and interact sequence specifically with bases around both integration sites.

A previous study had implicated palindromic hairpin structures in integration *in vitro* (Nuñez et al., 2015b). However, here I found no such structural features at integration sites selected by the Cas1_{CD}-Cas2_{CD} or Cas1_{AB}-Cas2_{AB} proteins from *S. sofataricus*. This difference may stem from the structure of the CRISPR array *in vivo*, as the *E. coli* CRISPR repeats contain a palindrome and predicted hairpin, whereas the *S. sofataricus* repeats are unstructured (Kunin et al., 2007). Therefore, it is conceivable that hairpins play a role in directing integration to repeat ends in *E. coli*, but not in *S. sofataricus*.

In this chapter I also demonstrated that the Cas1_{CD}-Cas2_{CD} complex preferentially integrated protospacers with 3' overhangs (Figure 6.6). Duplex protospacers supported the reaction, whereas 5' overhangs abolished protospacer integration. From these results I concluded that protospacer DNA is bound by the Cas1_{CD}-Cas2_{CD} complex in a similar way to in the *E. coli* Cas1-Cas2 complex. The preference for 3' ends was likely due to the Cas1_{CD} subunits binding tightly to single-stranded protospacer ends and utilizing the 3' hydroxyl positioned in the enzyme active-site to mediate nucleophilic attack during integration. Another clue to the similarity in pre-integration complex structures in *E. coli* and *S. sofataricus* was that the mutation of a conserved arginine residue (R166) of Cas1_{CD}, which was shown to form part of the positively charged cleft which binds protospacer ends in *E. coli*

(Nuñez et al., 2015b), abolished integration by Cas1_{CD}-Cas2_{CD}. This supports the theory that a similar 3' overhang co-ordination by Cas1 proteins from *E. coli* and *S. solfataricus* is required for integration.

The 5' overhang substrates have recessed 3' ends, which potentially do not extend into the active site of Cas1_{CD}, and thus impede integration. The ability of duplex protospacers to bring about integration suggested that the Cas1_{CD} protein could splay the ends of the duplex DNA to allow a single-stranded 3' end to enter the active site. While there is no absolute conservation of the tyrosine residue shown to be important in splaying protospacer ends in *E. coli* (Y22) (Wang et al., 2015; Nuñez et al., 2015b), a tyrosine residue (Y12) in Cas1_{CD} is located in a homologous loop region of the protein and may play a similar role during integration in *S. solfataricus*.

6.3.2 Cas2 enhances integration *in vitro*

Cas2_{CD} was found to play no part in the disintegration reaction, which represents the reversal of a half-site integration. I found that while Cas1_{CD} alone was able to integrate protospacers, this reaction was greatly enhanced by the addition of Cas2_{CD}. The fact that Cas1_{CD} alone supported integration confirms that the predicted active site of Cas2_{CD} is not required for this reaction, and instead it seems to play a structural role in adaptation. The enhancement of integration in the presence of Cas2 was likely due to the formation of a functional complex in the presence of protospacer DNA, which was able to co-ordinate both half-site reactions and resulted in complete integration of protospacer DNA.

6.3.3 Intrinsic specificity of Cas1 guides integration

Experiments in this chapter also revealed that while Cas1-Cas2 proteins from *S. solfataricus* are not sufficient for specific integration at the leader-repeat 1 junction, the Cas1 proteins do possess an intrinsic specificity for short sequence motifs similar to this site (Figure 6.12). This is consistent with the sequence specificity of Cas1_{CD} identified during the disintegration reaction (Rollie et al., 2015). The specificity relies on sequence-specific recognition of the nucleotides at position +1 and -2 of the integration site, with minor contributions from other residues. The sites selected by Cas1_{CD} with homology to the leader-repeat 1 junction occur hundreds of times in the pUC19-CRISPR plasmids used here and obviously many more times in the *S. solfataricus* genome. Therefore, while the specificity of Cas1 was important

in designating the exact integration site and favouring initial integration at site 1 over site 2, other factors are required to guide the integration complex uniquely to the leader-repeat 1 junction during adaptation.

6.3.4 Integration host factor

The Cas1-Cas2 complex of *E. coli* was also shown to lack specificity for the leader-repeat 1 junction *in vitro*. While ~70% of all integrations were found to be specific for the CRISPR array, only ~35% took place at the ends of repeat 1 (Nuñez et al., 2015b). A recent study demonstrated that addition of the integration host factor protein (IHF) to these assays was sufficient to increase specificity for the ends of repeat 1 so that ~80% of all integrations took place there (Nuñez et al., 2016).

The integration host factor is a small heterodimeric protein which binds ~35 bp of DNA causing a 160° kink in the strands (Rice et al., 1996). IHF was first identified as a factor that promotes integration of phage λ and has since been implicated in DNA binding that triggers the assembly of large protein complexes involved in transcriptional modulation and replication (Rice et al., 1996). The IHF protein interacts with a consensus sequence through minor groove recognition (Rice et al., 1996). Nuñez and colleagues identified an IHF consensus binding site in the -9 to -35 leader region, which had previously been shown to be important for spacer integration (Nuñez et al., 2016; Yosef et al., 2012). The authors confirmed the role of IHF in CRISPR adaptation by demonstrating that deletion of the α or β IHF subunits led to *in vivo* adaptation into the CRISPR locus being abolished. The structural deformation caused by IHF has previously been shown to facilitate viral integration (Surette & Chaconas, 1989) and it is predicted to perform a similar role in the CRISPR-Cas system of *E. coli* (Nuñez et al., 2016). The protein is thought to promote recognition and targeting by the Cas1-Cas2 complex to the leader-repeat junction by bending and potentially opening the leader consensus site. The finding that IHF also stimulates integration into linear DNA substrates containing the CRISPR insertion site supports this hypothesis (Nuñez et al., 2016). As only 80% of integrations *in vitro* happen at repeat 1 ends, even in the presence of IHF, it seems that other elements must also contribute to recognition of the correct insertion site by Cas1-Cas2 *in vivo* (Nuñez et al., 2016).

6.3.5 A host factor in *S. solfataricus*

As the IHF protein is only present in Gram-negative bacteria, its role in guiding specific Cas1-Cas2 integration outside of these organisms must be fulfilled by other factors. In this chapter, the first step in identifying this factor in *S. solfataricus* was taken. I showed that integration by the Cas1_{AB}-Cas2_{AB} proteins is directed uniquely to the leader-repeat 1 junction by the addition of a factor contained in *S. solfataricus* cell lysate (Figure 6.13). Therefore, specific integration in *S. solfataricus* seems to rely, firstly, on the intrinsic specificity of the Cas1 proteins, which have specificity for nucleotides around the site 1 junction and direct specific integration at this 'micro' level. The second level of specificity is orchestrated by the unknown factor in *S. solfataricus* lysate that seems to be responsible for recognizing the leader-repeat 1 junction and targeting Cas1-Cas2 uniquely to this site.

Whether this role in enhancing specificity is carried out either by a Cas protein, or a small DNA remodeling protein, as in *E. coli*, or some other host factor remains to be identified. However, it is conceivable that there is more than one factor required for specificity in *S. solfataricus*, perhaps with one component blocking off-target integrations and another guiding integration at the leader-proximal repeat. Interestingly, the Cbp1 protein in *S. solfataricus* was previously shown to bind CRISPR repeats and open the DNA duplex around these sites (Peng et al., 2003). The authors hypothesized that this small DNA bending protein may trigger the binding of other specific factors. Therefore, this protein could potentially be involved in guiding integration by Cas1-Cas2 during adaptation in *S. solfataricus*. Further work on this project will focus on identifying this/these factor/s in *S. solfataricus* cell lysate and also on determining the features of the CRISPR array required for this increased specificity.

Finally, sequencing integration products produced from specific integrations at the leader-repeat 1 junction in the presence of cell lysate revealed that the 3' ends of inserted protospacers had been processed in this reaction. In *E. coli* the Cas1-Cas2 complex was shown to carry out 3' end processing uniquely at PAM residues (Wang et al., 2015). However, in this study the Cas1_{CD}-Cas2_{CD} complex in isolation showed no activity on 3' PAM-containing protospacer ends. Instead, processing was only observed in the presence of Cas1_{AB}-Cas2_{AB} and cell lysate, and the presence of a GG or CC PAM motif was not required. Therefore, it seems that selection of PAM residues happens at an earlier stage in the adaptation process,

before the protospacer is bound in the pre-integration Cas1-Cas2 complex. Whether this 3' processing is stimulated by cell lysate, but carried out by the Cas1_{AB}-Cas2_{AB} complex, or whether a component of the lysate cleaves 3' ends independent of the adaptation proteins remains to be investigated.

Chapter 7: Conclusions and future directions

7.1 Summary

The work described in this thesis aimed to advance our understanding of how new spacers are acquired during the adaptation stage of CRISPR-Cas immunity. While considerable progress had been made in elucidating the mechanisms involved in the expression and interference stages of the response, this initial step in CRISPR-Cas-mediated immunity remained poorly understood at the start of this project. The initial experimental plan for this project involved characterising the activity of the Cas1 and Cas2 proteins of *S. solfataricus*, in order to delineate their role in the addition of new spacers. Furthermore, I aimed to investigate regulation of the adaptation response and the modulation of the wider CRISPR-Cas system in this organism.

In Chapter 3 I showed that there was a robust upregulation of expression of Cas1_{CD} and type I-A Cascade subunits during viral infection. In contrast, levels of the type III interference complexes were shown to only increase marginally between control and infected samples. The identification of specific binding of Csa3_{CD} to a promoter sequence upstream of a *cas* adaptation cassette implicated this protein in transcriptional regulation of the CRISPR-Cas system in *S. solfataricus*.

Data presented in Chapter 4 revealed that the activity of the Cas1_{CD} and Cas2_{CD} proteins differed considerably from what had been reported previously. The Cas1_{CD} protein did not exhibit nuclease activity and unusually, had a strong preference for single-stranded DNA. Furthermore, unlike the equivalent proteins in *E. coli*, the Cas1_{CD} and Cas2_{CD} proteins did not interact under the conditions tested.

Chapters 5 and 6 described the *in vitro* reconstitution of both the forward and reverse integration reaction performed by Cas1 proteins from *S. solfataricus* and *E. coli*. These reactions were used to probe Cas1 sequence specificity and led to the discovery that while Cas1 alone is capable of integrating spacers at sequences with homology to the leader-repeat 1 junction, an additional host factor is required to direct spacer integration uniquely to the *bona fide* integration site at the CRISPR leader-repeat junction.

7.2 CRISPR-Cas regulation

An important contribution of this work was the discovery that elements of the CRISPR-Cas system in *S. solfataricus* were tightly regulated, while others were constitutively expressed. These results implied that different regulatory mechanisms exist in archaea compared to bacteria, as the CRISPR-Cas system of *E. coli* had previously been found to be globally silenced by the H-NS repressor in the absence of infection (Pul et al., 2010).

An earlier study in *S. solfataricus* had reported that the levels of several Cas proteins were regulated in response to infection (Maaty et al., 2012). However, the proteins identified did not include those involved in adaptation. Here I showed that, in fact, these proteins are among the most strongly upregulated. Under control conditions, levels of the Cas1_{CD} protein and mRNA transcripts were found to be very low, implying transcriptional silencing in the absence of infection. In contrast, following infection, Cas1_{CD} transcripts increased by around 12-fold and protein levels increased markedly. The upregulation of Cas1_{CD} was accompanied by the integration of new spacers into the CRISPR C locus. Subunits of the type I-A Cascade interference complex were also strongly upregulated, whereas the type III complexes and pre-CRISPR RNA levels were found to be expressed constitutively and only weakly upregulated during infection.

These findings demonstrated that there is a level of fine-tuning to the regulation of CRISPR-Cas in archaea, which does not seem to exist in bacteria. The more complex nature of archaeal CRISPR systems, with multiple systems and arrays, seems to allow independent regulation of different *cas* gene modules. The silencing of the *cas1_{CD}* gene in the absence of infection may be a mechanism employed to reduce incorporation of self-matching spacers and avoid the associated autoimmune penalties. In contrast, the constitutively expressed type III complexes have been suggested to act as surveillance systems (Quax et al., 2013), which alert the cell to new invaders and provide constant protection against previously encountered threats.

The second key finding in Chapter 3 was that the Csa3_{CD} protein, which had previously been predicted to be a transcriptional regulator (Lintner et al., 2011a), bound specifically to a *cas* promoter thought to control expression of Cas1_{CD} and

Cas2_{CD} proteins. Although Csa3_{CD} did not affect *in vitro* transcript levels from this promoter, a reduced level of Csa3_{CD} was obvious during the early stages of infection, which coincided with the upregulation of Cas1_{CD} expression. These results suggested that Csa3 has a role as a transcriptional repressor of this *cas* operon, but that a binding partner or small-molecule ligand, required for this activity, may be missing *in vitro*.

Taken together, I conclude from these results that levels of Cas1_{CD} (and potentially Cas2_{CD}, Cas4_{CD} and Csa1_{CD}), required for adaptation, are upregulated in response to infection and that Csa3_{CD} is likely to be involved in this modulation. A previous study demonstrated that while adaptation of the CRISPR C array is triggered by infection, activation of adaptation in CRISPR A also required environmental stress (Erdmann et al., 2013). In the future it would be intriguing to examine whether the CRISPR A-associated *cas1*_{AB} and *cas2*_{AB} genes are also transcriptionally silenced in control conditions and how protein levels of Cas1_{AB} change following infection and under environmental stress. Future work could also involve characterising the activity of a second Csa3 protein, coded for by a *cas* gene located close to the *cascade* operon. This second Csa3 protein may bind the *cascade* promoter and have a role in the strong upregulation of transcription identified during infection in this study. Finally, RNA sequencing of samples from both infected and control *S. solfataricus* samples would also expand on the data collected here and allow a more global analysis of gene regulation during the CRISPR-Cas response to infection.

7.3 Characterisation of Cas1_{CD} and Cas2_{CD}

The characterisation of Cas1_{CD} and Cas2_{CD} proteins was reported in Chapter 4. Key differences were revealed between the activity of Cas1_{CD} and the activity reported for Cas1 proteins from other organisms. Cas1_{CD} did not exhibit non-specific nuclease activity, which had been associated with the *E. coli* and *P. aeruginosa* Cas1 proteins and implicated in the processing of protospacer DNA before insertion (Babu et al., 2011; Wiedenheft et al., 2009). The lack of Cas1_{CD} nuclease activity suggested that the previously reported nucleic acid degradation was not conserved across Cas1 proteins from different CRISPR-Cas systems, or that the nuclease activity of the Cas1_{CD} protein was sequence-specific and the required motif was not contained in the substrates tested. The data presented in this chapter also revealed the preferential binding of single-stranded DNA by Cas1_{CD}. Single-stranded

substrates were bound with ~20-fold higher affinity than duplex substrates, which suggested that single-stranded DNA played a key role in adaptation. This hypothesis was validated when the structure of the *E. coli* Cas1-Cas2 complex with bound DNA revealed that the protospacer ends were splayed by the complex and Cas1 subunits tightly co-ordinated the 3' single-stranded ends (Wang et al., 2015; Nuñez et al., 2015a). The high-affinity single-strand binding exhibited by the Cas1_{CD} protein suggests that a similar protospacer binding may be required for adaptation in the *S. solfataricus* system and co-crystallisation studies will be crucial to further investigate this theory.

The Cas1_{CD} and Cas2_{CD} proteins were shown not to interact under the conditions tested. This was a surprising result as the *E. coli* Cas1 and Cas2 proteins interact to form a stable complex required for integration and the Cas1 and Cas2 of the *T. tenax* CRISPR-Cas system exist as a single fused protein, implying that close association of these proteins is key to their function (Nuñez et al., 2014; Plagens et al., 2012). I concluded from these data that perhaps a specific DNA substrate, structure or protein component is required to initiate complex formation in *S. solfataricus*. Genes coding for Cas4 proteins are present in the same operon as *cas1* and *cas2* genes in many CRISPR-Cas systems and the Cas4 protein has been shown to interact to form a complex with the Cas1-Cas2 fusion protein in *T. tenax* (Plagens et al., 2012). Therefore, the Cas4_{CD} protein of *S. solfataricus* may also play a crucial role in facilitating the interaction of Cas1_{CD} and Cas2_{CD} *in vivo*. A clear step required to advance this work is the co-expression of the Cas1_{CD}, Cas2_{CD}, Cas4_{CD} and Csa1_{CD} proteins, coded for by the same *cas* operon, to look for physical and functional interactions important for CRISPR-Cas adaptation.

The characterisation of Cas2_{CD} revealed a metal-independent endoribonuclease activity on single-stranded RNA substrates. Although similar activities have been reported previously (Beloglazova et al., 2008), integration by the Cas1-Cas2 complex of *E. coli* was shown not to require the Cas2 active site and this protein was suggested to play a structural, rather than catalytic role in adaptation (Nuñez et al., 2014). Cas2 is thought to originate from a toxin/antitoxin system and shares homology with the VapD toxins (Koonin & Makarova, 2013), which are known to cleave ssRNA (Kwon et al., 2012). Therefore, it is conceivable that the RNase activity observed in this study relates to Cas2's role as a toxin and potentially contributes to native prokaryote immunity by degrading foreign transcripts and

buying time for the activation of the adaptive CRISPR-Cas response. In this model, the Cas1 protein may act as an antitoxin, which interacts with Cas2 to induce a conformational change - neutralising RNase activity and promoting protospacer binding.

7.4 Disintegration

The remaining chapters of this thesis described research carried out to delineate the mechanism of integration of new spacers to the CRISPR array. Chapter 5 revealed that Cas1_{CD} performed a transesterification reaction on branched substrates. This reaction represented the reversal of a half-site integration reaction, termed disintegration, and had previously been described for viral integrase (Chow et al., 1992). A significant breakthrough came when Cas1 proteins from *S. solfataricus* and *E. coli* were found to impose a strict sequence specificity on the disintegration reaction, with preferred sequences matching the *in vivo* leader-repeat 1 junction of the associated CRISPR array. This finding led to the conclusion that it is this site (site 1) and not the leader-distal end of repeat 1 (site 2) that is recognised and targeted by Cas1 during adaptation. The model suggested by these data involves an initial sequence-specific half-site integration of a new spacer at the leader-repeat junction, followed by a second half-site joining of the remaining end of the new spacer to the leader-distal end of the repeat. The second half-site reaction is suggested to be guided by a distance-, rather than sequence-dependent mechanism. A very recent study showed that aberrant half-site integrations are the substrates for disintegration by the *S. pyogenes* Cas1, and suggested that disintegration could have a role in reversing off-target integrations *in vivo* (Wright & Doudna, 2016); this is an intriguing possibility that requires further validation.

7.5 Integration

The first reconstitution of protospacer integration by *S. solfataricus* Cas1 and Cas2 proteins was reported in Chapter 6. This reaction required supercoiled DNA and while Cas1 alone was capable of integrating protospacers, the reaction was greatly enhanced by the presence of Cas2. This increase in integration efficiency suggested that under these conditions, Cas1 and Cas2 interact to mediate integration. Given the broad range of spacer lengths present in the *S. solfataricus*

CRISPR arrays (34 – 48 bp), if these proteins do form a complex during adaptation, the structure adopted must differ considerably from the strict molecular ruler present in *E. coli*. Sequencing the sites of integration selected by Cas1_{CD}-Cas2_{CD} and Cas1_{AB}-Cas2_{AB} revealed that, as observed for the disintegration reaction, these proteins had a specificity for sites with sequence similarity to the leader-repeat 1 junction. For both sets of adaptation proteins the identity of the residue at position +1, corresponding to the first nucleotide of the repeat, and -2, corresponding to the penultimate residue of the leader, had the strongest influence on integration efficiency. In contrast, the nucleotide at the -1 position, which is the last nucleotide of the leader at the *bona fide* integration site, had little influence on integration or disintegration efficiency.

This chapter also revealed that specific integration by Cas1_{AB}-Cas2_{AB} exclusively at the leader-repeat 1 junction required a host factor present in *S. solfataricus* lysate. A similar requirement had recently been reported in *E. coli*, where the IHF protein was found to be crucial in guiding Cas1-Cas2 to the correct integration site (Nuñez et al., 2016). Host factors have been implicated in the final repair of the genome after spacer integration; however, the discovery that they are also key to specific integration in two different CRISPR-Cas systems suggests that there is a more extensive interaction between host and CRISPR-Cas elements than previously believed. The IHF protein influences integration site-selection by binding a consensus sequence in the CRISPR leader, causing a kink in the duplex, which is then thought to trigger recruitment of Cas1-Cas2 (Nuñez et al., 2016). The identity of the host factor in *S. solfataricus* remains under investigation; however, if structural deformation of the leader is also required in this system, small DNA-binding and -bending proteins are likely to be involved. The DNA-binding proteins Alba and Sso7 are potential candidates for this activity, as well as the Cbp1 protein, which was previously shown to specifically bind repeat sequences in *S. solfataricus* (Deng et al., 2012; Peng et al., 2003). It is also possible that the host factor in *S. solfataricus* brings about specificity through a different mechanism from IHF, perhaps through direct protein-protein interaction with the adaptation machinery. Further research will concentrate on investigating the effect of the potential host factor candidates, mentioned above, and on fractionating *S. solfataricus* lysate to isolate and identify the active component/s.

An unresolved question with regard to the *in vitro* integration reaction is whether the observed integrations are half- or full-site reactions. The joining of either one or both 3' hydroxyls of the protospacer would lead to indistinguishable nicked plasmid products following separation on agarose gels. Planned work to rectify this problem involves modifying the PCR amplification step by selecting primers that span the leader-repeat junction. These primers will result in a product increased in length by ~60 bp (spacer + duplicated repeat) compared to the product amplified from an unmodified plasmid, only if a new protospacer has been fully integrated.

The addition of *S. solfataricus* lysate to integration assays was also shown to lead to the removal of between 1 and 5 nucleotides from the 3' ends of integrated protospacers. 3' processing at PAM residues had been demonstrated previously for the *E. coli* Cas1-Cas2 complex. However, unusually, the processing described here did not occur sequence-specifically, and Cas1_{AB} in the absence of lysate was not capable of carrying out the trimming. Whether 3' processing is a requirement for integration of the protospacer in *S. solfataricus*, and the identity of the nucleases required, remains unknown. However, several potential scenarios exist in which either the Cas1_{AB} protein is somehow 'activated' for 3' nuclease activity by a factor in cell lysate, or a nuclease in the lysate is directly responsible for the cleavage. Additionally, Cas1_{AB} may bind and protect variable regions of the protospacer for integration leading to the imprecise trimming observed at 3' protospacer ends. This trimming role might be fulfilled by the exonuclease activity of Cas4 (Zhang et al., 2012a).

Protospacers with single-stranded 3' overhangs were found to be optimal substrates for *in vitro* integration. This suggested that *S. solfataricus* Cas1 co-ordinates protospacer DNA for integration in a similar way to the *E. coli* complex, with 3' single-stranded protospacer ends being tightly bound in the active site of the enzyme and the 3' hydroxyl residues being used to mediate nucleophilic attack at the integration site. In support of this theory, 5' overhangs were found to abolish integration, probably due to Cas1 not being able to access the recessed 3' hydroxyl required to catalyse transesterification. The finding that duplex protospacers also support integration suggested that Cas1 proteins from *S. solfataricus* are able to splay DNA ends in order to access the catalytic 3' hydroxyls. An interesting next step might be to investigate the effect on integration of mutating conserved aromatic residues of *S. solfataricus* Cas proteins at equivalent positions to the tyrosine

residue (Y22) of the *E. coli* Cas1 protein, required to splay protospacer ends in that system.

The development of the *in vitro* integration assay, described in Chapter 6, will allow several outstanding questions to be addressed. Firstly, the importance of sequences in the first repeat will be probed. A repeat-anchored ruler mechanism has been shown to control integration site selection in *H. hispanica* (Wang et al., 2016) and it will be interesting to assess whether tethering of Cas1-Cas2 by the first repeat is also required for specific integration in *S. solfataricus*. Secondly, the *in vitro* integration assay will allow identification of crucial regions of the *S. solfataricus* leader sequences, required for Cas-protein and host-factor docking. The length of leader required for integration in *S. solfataricus* is poorly understood. Alignment of leaders in *S. solfataricus* identified a natural deletion in the CRISPR E leader, between positions -47 and -70, that was implicated in the absence of specific integration at this locus *in vivo* (Garrett et al., 2015). Leader sequences in archaea are often much longer than in bacteria, with the CRISPR A leader comprising over 500 bp. The purpose of maintaining these extensive leaders, and how distant sequence motifs affect integration, are questions that we can now address using the *in vitro* integration reaction reconstituted in Chapter 6. Future work will involve truncating the CRISPR A leader to identify the minimal length required for integration, and to locate sequences crucial for the binding of Cas proteins and host-factors.

7.6 Capture

Adaptation can be separated into two steps – the capture of new protospacers from foreign DNA and their insertion into the host genome. While this work focused on elucidating the mechanism of insertion, the process of capture remains poorly understood and its investigation would be a priority of further work. In *E. coli*, spacer precursors are thought to be generated by the RecBCD complex during naïve adaptation, and a recent report demonstrated that the products of helicase/nuclease Cas3 may fuel primed adaptation (Levy et al., 2015; Künne et al., 2016). However, in archaea the proteins and processes involved in the generation of substrates for adaptation is a key outstanding question. The HerA-NurA helicase-nuclease complex, which performs a homologous function to RecBCD in archaea (Blackwood et al., 2012), is a potential candidate involved in protospacer generation in *S. solfataricus*. Additionally, the 5' - 3' exonuclease activity of Cas4 proteins in this

system may also be involved in processing DNA for integration (Zhang et al., 2012a). These theories will be tested using the *in vitro* integration assay, with the eventual aim being the reconstitution of both the protospacer generation and insertion steps required for CRISPR-Cas adaptation.

7.7 Conclusion

Adaptation is the process by which new immunological memories are generated by the CRISPR-Cas system to provide resistance to future infections. Over the past five years, our understanding of this stage of CRISPR-Cas immunity has advanced considerably. The work described in this thesis has contributed by examining the activation, requirements and specificity of adaptation. A key novel finding was that specific integration of new spacers in *S. solfataricus* required both the intrinsic sequence specificity of Cas1 and a host factor present in cell lysate. Many new questions have arisen based on the recent developments in the field of CRISPR-Cas adaptation. However, we can be confident that if the current rate of progress continues, answers to these questions will be forthcoming in the not too distant future.

References

- Agari, Y., Sakamoto, K., Tamakoshi, M., Oshima, T., Kuramitsu, S. & Shinkai, A. (2010). Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. *Journal of Molecular Biology*. 395 (2). p.pp. 270–281.
- Ali, S.S., Beckett, E., Bae, S.J. & Navarre, W.W. (2011). The 5.5 protein of phage T7 inhibits H-NS through interactions with the central oligomerization domain. *Journal of Bacteriology*. 193 (18). p.pp. 4881–4892.
- Amitai, G. & Sorek, R. (2016). CRISPR–Cas adaptation: insights into the mechanism of action. *Nature Reviews Microbiology*. 14 (2). p.pp. 67–76.
- Arslan, Z., Hermanns, V., Wurm, R., Wagner, R. & Pul, Ü. (2014). Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Research*. 42 (12). p.pp. 7884–7893.
- Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., Phanse, S., Joachimiak, A., Koonin, E. V., Savchenko, A., Emili, A., Greenblatt, J., Edwards, A.M. & Yakunin, A.F. (2011). A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Molecular Microbiology*. 79 (2). p.pp. 484–502.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. a & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 315 (5819). p.pp. 1709–1712.
- Bell, S.D. & Jackson, S.P. (2001). Mechanism and regulation of transcription in archaea. *Current Opinion in Microbiology*. 4 (2). p.pp. 208–213.
- Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E. V., Edwards, A.M., Savchenko, A. & Yakunin, A.F. (2008). A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *Journal of Biological Chemistry*. 283 (29). p.pp. 20361–20371.
- Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A. & Yakunin, A.F. (2011). Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *The EMBO Journal*. 30 (22). p.pp. 4616–4627.
- Blackwood, J.K., Rzechorzek, N.J., Abrams, A.S., Maman, J.D., Pellegrini, L. & Robinson, N.P. (2012). Structural and functional insights into DNA-end processing by the archaeal HerA helicase-NurA nuclease complex. *Nucleic Acids Research*. 40 (7). p.pp. 3183–3196.
- Bolduc, B., Shaughnessy, D.P., Wolf, Y.I., Koonin, E. V., Roberto, F.F. & Young, M. (2012). Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *Journal of Virology*. 86. p.pp. 5562–5573.
- Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*. 151 (8). p.pp. 2551–2561.
- Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y.,

- Wiedenheft, B., Maxwell, K.L. & Davidson, A.R. (2015). Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature*. 526 (7571). p.pp. 136–139.
- Bondy-Denomy, J., Pawluk, A., Maxwell, K.L. & Davidson, A.R. (2012). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*. 493 (7432). p.pp. 429–432.
- 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 (5891). p.pp. 960–964.
- Brown, P.O., Bowerman, B., Varmus, H.E. & Bishop, J.M. (1989). Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proceedings of the National Academy of Sciences*. 86 (8). p.pp. 2525–2529.
- Carte, J., Wang, R., Li, H., Terns, R.M. & Terns, M.P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & Development*. 22 (24). p.pp. 3489–3496.
- Charpentier, E., Richter, H., van der Oost, J. & White, M.F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiology Reviews*. 39 (3). p.pp. 428–441.
- Chen, F.-M. (1988). Effects of A:T base pairs on the B–Z conformational transitions of DNA. *Nucleic Acids Research*. 16 (5). p.pp. 2269–2281.
- Chopin, M.C., Chopin, A. & Bidnenko, E. (2005). Phage abortive infection in *Lactococci*: Variations on a theme. *Current Opinion in Microbiology*. 8 (4). p.pp. 473–479.
- Chow, S.A., Vincent, K.A., Ellison, V. & Brown, P.O. (1992). Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science*. 255 (5045). p.pp. 723–726.
- Cournac, A. & Plumbridge, J. (2013). DNA looping in prokaryotes: Experimental and theoretical approaches. *Journal of Bacteriology*. 195 (6). p.pp. 1109–1119.
- Crooks, G.E. (2004). WebLogo: A sequence logo generator. *Genome Research*. 14 (6). p.pp. 1188–1190.
- 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. *Nature Communications*. 3 (945). p.pp. 1–7.
- Deng, L., Garrett, R.A., Shah, S.A., Peng, X. & She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in *Sulfolobus*. *Molecular Microbiology*. 87 (5). p.pp. 1088–1099.
- Deng, L., Kenchappa, C.S., Peng, X., She, Q. & Garrett, R.A. (2012). Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in *Sulfolobus*. *Nucleic Acids Research*. 40 (6). p.pp. 2470–2480.
- Deveau, H., Barrangou, R., Garneau, J.E., Labonte, 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 (4). p.pp. 1390–1400.
- Díez-Villaseñor, C., Almendros, C., Mojica, F.J.M. & García-Martínez, J. (2009).

- Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*. 155 (3). p.pp. 733–740.
- Díez-Villaseñor, C., Guzmán, N.M., Almendros, C., García-Martínez, J. & Mojica, F.J.M. (2013). CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biology*. 10 (5). p.pp. 792–802.
- Dillingham, M.S. & Kowalczykowski, S.C. (2008). RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiology and Molecular Biology Reviews*. 72 (4). p.pp. 642–671.
- Dillon, S.C., Cameron, A.D.S., Hokamp, K., Lucchini, S., Hinton, J.C.D. & Dorman, C.J. (2010). Genome-wide analysis of the H-NS and Sfh regulatory networks in *Salmonella Typhimurium* identifies a plasmid-encoded transcription silencing mechanism. *Molecular Microbiology*. 76 (5). p.pp. 1250–1265.
- Engelman, A., Mizuuchi, K. & Craigie, R. (1991). HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell*. 67 (6). p.pp. 1211–1221.
- Erdmann, S. & Garrett, R. a (2012). Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Molecular Microbiology*. 85 (6). p.pp. 1044–1056.
- Erdmann, S., Le Moine Bauer, S. & Garrett, R.A. (2014). Inter-viral conflicts that exploit host CRISPR immune systems of *Sulfolobus*. *Molecular Microbiology*. 91 (5). p.pp. 900–917.
- Erdmann, S., Shah, S.A. & Garrett, R.A. (2013). SMV1 virus-induced CRISPR spacer acquisition from the conjugative plasmid pMGB1 in *Sulfolobus solfataricus* P2. *Biochemical Society Transactions*. 41 (6). p.pp. 1449–1458.
- Estrella, M.A., Kuo, F.-T. & Bailey, S. (2016). RNA-activated DNA cleavage by the Type III-B CRISPR–Cas effector complex. *Genes & Development*. 30 (4). p.pp. 460–470.
- Fineran, P.C., Blower, T.R., Foulds, I.J., Humphreys, D.P., Lilley, K.S. & Salmond, G.P.C. (2009). The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proceedings of the National Academy of Sciences*. 106 (3). p.pp. 894–899.
- Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature*. 532 (7600). p.pp. 517–521.
- Forde, A. & Fitzgerald, G.F. (1999). Analysis of exopolysaccharide (EPS) production mediated by the bacteriophage adsorption blocking plasmid, pCI658, isolated from *Lactococcus lactis* ssp. *cremoris* HO2. *International Dairy Journal*. 9 (7). p.pp. 465–472.
- Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H. & Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 468 (7320). p.pp. 67–71.
- Garrett, R., Shah, S., Erdmann, S., Liu, G., Mousaei, M., León-Sobrino, C., Peng, W., Gudbergsdottir, S., Deng, L., Vestergaard, G., Peng, X. & She, Q. (2015). CRISPR-Cas adaptive immune systems of the *Sulfolobales*: Unravelling their complexity and diversity. *Life*. 5 (1). p.pp. 783–817.
- Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., B.A..

- (2005). *Protein identification and analysis tools on the ExPASy server*. J. M. Walker (ed.). Totowa, NJ: Humana Press.
- van Gent, D.C., Groeneger, A.A. & Plasterk, R.H. (1993). Identification of amino acids in HIV-2 integrase involved in site-specific hydrolysis and alcoholysis of viral DNA termini. *Nucleic Acids Research*. 21 (15). p.pp. 3373–3377.
- Georgiou, T., Yu, Y.-T.N., Ekunwe, S., Buttner, M.J., Zuurmond, A.-M., Kraal, B., Kleanthous, C. & Snyder, L. (1998). Specific peptide-activated proteolytic cleavage of *Escherichia coli* elongation factor Tu. *Proceedings of the National Academy of Sciences*. 95 (6). p.pp. 2891–2895.
- Godde, J.S. & Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. *Journal of Molecular Evolution*. 62 (6). p.pp. 718–729.
- Goldberg, G.W., Jiang, W., Bikard, D. & Marraffini, L. a (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature*. 514 (7524). p.pp. 633–637.
- Goldberg, G.W. & Marraffini, L.A. (2015). Resistance and tolerance to foreign elements by prokaryotic immune systems — curating the genome. *Nature Publishing Group*. 15 (11). p.pp. 717–724.
- Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C. & Koonin, E. V (2015). No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. *The ISME Journal*. 9 (9). p.pp. 2021–2027.
- Goren, M.G., Yosef, I., Auster, O. & Qimron, U. (2012). Experimental definition of a clustered regularly interspaced short palindromic duplicon in *Escherichia coli*. *Journal of Molecular Biology*. 423 (1). p.pp. 14–16.
- Götz, D., Paytubi, S., Munro, S., Lundgren, M., Bernander, R. & White, M.F. (2007). Responses of hyperthermophilic crenarchaea to UV irradiation. *Genome Biology*. 8 (10). p.p. R220.
- Grissa, I., Vergnaud, G. & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*. 35 (Web Server). p.pp. W52–W57.
- Hale, C., Kleppe, K., Terns, R.M. & Terns, M.P. (2008). Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA*. 14 (12). p.pp. 2572–2579.
- Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. & Terns, M.P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*. 139 (5). p.pp. 945–956.
- Han, D., Lehmann, K. & Krauss, G. (2009). SSO1450 - A CAS1 protein from *Sulfolobus solfataricus* P2 with high affinity for RNA and DNA. *FEBS Letters*. 583 (12). p.pp. 1928–1932.
- Hatoum-Aslan, A., Maniv, I., Samai, P. & Marraffini, L.A. (2014). Genetic characterization of antiplasmid immunity through a type III-A CRISPR-cas system. *Journal of Bacteriology*. 196 (2). p.pp. 310–317.
- Hatoum-Aslan, A., Samai, P., Maniv, I., Jiang, W. & Marraffini, L.A. (2013). A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *Journal of Biological Chemistry*. 288 (39). p.pp. 27888–27897.
- Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J.A. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease.

- Science*. 329 (5997). p.pp. 1355–1358.
- Hermans, P.W.M., Van Soolingen, D., Bik, E.M., De Haas, P.E.W., Dale, J.W. & Van Embden, J.D.A. (1991). Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infection and Immunity*. 59 (8). p.pp. 2695–2705.
- Hickman, A.B. & Dyda, F. (2015). The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. *Nucleic Acids Research*. 43 (22). p.pp. 10576–10587.
- Hochstrasser, M.L. & Doudna, J.A. (2014). Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends in Biochemical Sciences*. 40 (1). p.pp. 58–66.
- Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternberg, S.H., Nogales, E. & Doudna, J. a (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proceedings of the National Academy of Sciences*. 111 (18). p.pp. 6618–6623.
- Horvath, P., Romero, D.A., Coute-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. & Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriology*. 190 (4). p.pp. 1401–1412.
- Houte, S. van, Buckling, A. & Westra, E.R. (2016). Evolutionary ecology of prokaryotic immune mechanisms. *Microbiology and Molecular Biology Reviews*. 80 (3). p.pp. 745–763.
- Hoyland-Kroghsbo, N.M., Maerkedahl, R.B. & Svenningsen, S. Lo (2013). A quorum-sensing-induced bacteriophage defense mechanism. *mBio*. 4 (1). p.pp. e00362-12-e00362-12.
- Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. (1987). Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of Bacteriology*. 169 (12). p.pp. 5429–5433.
- Ivančić-Baće, I., Cass, S.D., Wearne, S.J. & Bolt, E.L. (2015). Different genome stability proteins underpin primed and naïve adaptation in *E. Coli* CRISPR-Cas immunity. *Nucleic Acids Research*. 43 (22). p.pp. 10821–10830.
- Jansen, R., Embden, J.D., Gaastra, W. & Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*. 43 (6). p.pp. 1565–1575.
- 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*. 337 (6096). p.pp. 816–821.
- Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., Beijer, M.R., Barendregt, A., Zhou, K., Snijders, A.P.L., Dickman, M.J., Doudna, J.A., Boekema, E.J., Heck, A.J.R., van der Oost, J. & Brouns, S.J.J. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular Biology*. 18 (5). p.pp. 529–536.
- Kapitonov, V. V & Jurka, J. (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences*. 103 (12). p.pp.

- 4540–4545.
- Katz, R.A., Gravuer, K. & Skalka, A.M. (1998). A preferred target DNA structure for retroviral integrase in vitro. *Journal of Biological Chemistry*. 273 (37). p.pp. 24190–24195.
- Kelly, L.A., Mezulis, S., Yates, C., Wass, M. & Sternberg, M. (2015). The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*. 10 (6). p.pp. 845–858.
- Kim, T.-Y., Shin, M., Huynh Thi Yen, L. & Kim, J.-S. (2013). Crystal structure of Cas1 from *Archaeoglobus fulgidus* and characterization of its nucleolytic activity. *Biochemical and Biophysical Research Communications*. 441 (4). p.pp. 720–725.
- Klauck, E., Böhringer, J. & Hengge-Aronis, R. (1997). The LysR-like regulator LeuO in *Escherichia coli* is involved in the translational regulation of rpoS by affecting the expression of the small regulatory DsrA-RNA. *Molecular Microbiology*. 25 (3). p.pp. 559–569.
- Koonin, E. V. & Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nature Reviews Genetics*. 16 (3). p.pp. 184–192.
- Koonin, E. V & Makarova, K.S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biology*. 10 (5). p.pp. 679–86.
- Krüger, D.H. & Bickle, T.A. (1983). Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiological Reviews*. 47 (3). p.pp. 345–360.
- 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 (1). p.p. 36.
- Kunin, V., Sorek, R. & Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology*. 8 (4). p.p. 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 (5). p.pp. 852–864.
- Kwon, A.R., Kim, J.H., Park, S.J., Lee, K.Y., Min, Y.H., Im, H., Lee, I., Lee, K.Y. & Lee, B.J. (2012). Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. *Nucleic Acids Research*. 40 (9). p.pp. 4216–4228.
- Lawrence, C.M. & White, M.F. (2011). Recognition of archaeal CRISPR RNA: No P in the alindromic repeat? *Structure*. 19 (2). p.pp. 142–144.
- Lemak, S., Beloglazova, N., Nocek, B., Skarina, T., Flick, R., Brown, G., Popovic, A., Joachimiak, A., Savchenko, A. & Yakunin, A.F. (2013). Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from *Sulfolobus solfataricus*. *Journal of the American Chemical Society*. 135 (46). p.pp. 17476–17487.
- León-Sobrino, C., Kot, W.P. & Garrett, R. a (2016). Transcriptome changes in STSV2-infected *Sulfolobus islandicus* REY15A undergoing continuous CRISPR spacer acquisition. *Molecular Microbiology*. 99 (4). p.pp. 719–728.

- Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U. & Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature*. 520 (7548). p.pp. 505–510.
- 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 (4). p.pp. 2483–2492.
- Lillestøl, R., Redder, P., Garrett, R. a & Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. *Archaea*. 2 (1). p.pp. 59–72.
- Lillestøl, R.K., Shah, S.A., Brügger, K., Redder, P., Phan, H., Christiansen, J. & Garrett, R.A. (2009). CRISPR families of the crenarchaeal genus *Sulfolobus*: Bidirectional transcription and dynamic properties. *Molecular Microbiology*. 72 (1). p.pp. 259–272.
- Lilley, D.M. & White, M.F. (2001). The junction-resolving enzymes. *Nature Reviews Molecular Cell Biology*. 2 (June). p.pp. 433–443.
- Lintner, N.G., Frankel, K. a, Tsutakawa, S.E., Alsbury, D.L., Copié, V., Young, M.J., Tainer, J. a & Lawrence, C.M. (2011a). The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. *Journal of Molecular Biology*. 405 (4). p.pp. 939–55.
- Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copie, V., Young, M.J., White, M.F. & Lawrence, C.M. (2011b). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *Journal of Biological Chemistry*. 286 (24). p.pp. 21643–21656.
- Liu, H. & Naismith, J.H. (2009). A simple and efficient expression and purification system using two newly constructed vectors. *Protein Expression and Purification*. 63 (2). p.pp. 102–111.
- Liu, T., Li, Y., Wang, X., Ye, Q., Li, H., Liang, Y., She, Q. & Peng, N. (2015). Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. *Nucleic Acids Research*. 43 (2). p.pp. 1044–1055.
- Maaty, W.S., Steffens, J.D., Heinemann, J., Ortmann, A.C., Reeves, B.D., Biswas, S.K., Dratz, E. a, Grieco, P. a, Young, M.J. & Bothner, B. (2012). Global analysis of viral infection in an archaeal model system. *Frontiers in Microbiology*. 3 (411). p.pp. 1–15.
- Maertens, G.N., Hare, S. & Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature*. 468 (7321). p.pp. 326–329.
- Makarova, K.S., Anantharaman, V., Aravind, L. & Koonin, E. V (2012). Live virus-free or die: coupling of antiviral immunity and programmed suicide or dormancy in prokaryotes. *Biology Direct*. 7 (1). p.p. 40.
- Makarova, K.S., Aravind, L., Grishin, N. V, Rogozin, I.B. & Koonin, E. V (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Research*. 30 (2). p.pp. 482–496.
- Makarova, K.S., Grishin, N. V, Shabalina, S.A., Wolf, Y.I. & Koonin, E. V (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with

- eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct*. 1 (7).
- 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., van der Oost, J. & Koonin, E. V (2011a). Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*. 9 (6). p.pp. 467–477.
- Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Terns, R.M., Terns, M.P., White, M.F., Yakunin, A.F., Garrett, R.A., van der Oost, J., Backofen, R. & Koonin, E. V. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*. 13 (11). p.pp. 722–736.
- Makarova, K.S., Wolf, Y.I. & Koonin, E. V (2013a). Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research*. 41 (8). p.pp. 4360–4377.
- Makarova, K.S., Wolf, Y.I. & Koonin, E. V. (2013b). The basic building blocks and evolution of CRISPR–Cas systems. *Biochemical Society Transactions*. 41 (6). p.pp. 1392–1400.
- Makarova, K.S., Wolf, Y.I., van der Oost, J. & Koonin, E. V (2009). Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology Direct*. 4 (1). p.p. 29.
- Makarova, K.S., Wolf, Y.I., Snir, S. & Koonin, E. V. (2011b). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of Bacteriology*. 193 (21). p.pp. 6039–6056.
- Manica, A., Zebec, Z., Steinkellner, J. & Schleper, C. (2013). Unexpectedly broad target recognition of the CRISPR-mediated virus defence system in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Research*. 41 (22). p.pp. 10509–10517.
- Manica, A., Zebec, Z., Teichmann, D. & Schleper, C. (2011). In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. *Molecular Microbiology*. 80 (2). p.pp. 481–491.
- Marraffini, L.A. & Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in *Staphylococci* by targeting DNA. *Science*. 322 (5909). p.pp. 1843–1845.
- Marraffini, L.A. & Sontheimer, E.J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature*. 463 (7280). p.pp. 568–571.
- Meaden, S., Paszkiewicz, K. & Koskella, B. (2015). The cost of phage resistance in a plant pathogenic bacterium is context-dependent. *Evolution*. 69 (5). p.pp. 1321–1328.
- Medina-Aparicio, L., Rebollar-Flores, J.E., Gallego-Hernández, A.L., Vázquez, A., Olvera, L., Gutiérrez-Ríos, R.M., Calva, E., Hernández-Lucas, I., Va, A., Olvera, L., Calva, E. & Herna, I. (2011). The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* serovar Typhi. *Journal of Bacteriology*. 193 (10). p.pp. 2396–2407.
- Meyer, J.R., Dobias, D.T., Weitz, J.S., Barrick, J.E., Quick, R.T. & Lenski, R.E. (2012). Repeatability and contingency in the evolution of a key innovation in

- phage Lambda. *Science*. 335 (6067). p.pp. 428–432.
- Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E. V. & van der Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science*. 353 (6299).
- Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution*. 60 (2). p.pp. 174–182.
- Mojica, F.J.M., Díez-Villasenor, 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 (1). p.pp. 244–246.
- Mojica, F.J.M., Ferrer, C., Juez, G. & Rodríguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Molecular Microbiology*. 17 (1). p.pp. 85–93.
- Mojica, F.J.M., Juez, G. & Rodríguez-Valera, F. (1993). Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. *Molecular Microbiology*. 9 (3). p.pp. 613–621.
- Mulepati, S., Heroux, A. & Bailey, S. (2014). Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science*. 345 (6203). p.pp. 1479–1484.
- Nam, K.H., Ding, F., Haitjema, C., Huang, Q., DeLisa, M.P. & Ke, A. (2012a). Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *Journal of Biological Chemistry*. 287 (43). p.pp. 35943–35952.
- Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., Delisa, M.P. & Ke, A. (2012b). Cas5d protein processes Pre-crRNA and assembles into a Cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure*. 20 (9). p.pp. 1574–1584.
- Niewoehner, O., Jinek, M. & Doudna, J.A. (2014). Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Research*. 42 (2). p.pp. 1341–1353.
- 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 (6). p.pp. 824–833.
- 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 (7579). p.pp. 535–538.
- Nuñez, J.K., Kranzusch, P.J., Noeske, J., Wright, A. V, Davies, C.W. & Doudna, J.A. (2014). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Structural & Molecular Biology*. 21 (6). p.pp. 528–534.
- Nuñez, J.K., Lee, A.S.Y., Engelman, A. & Doudna, J.A. (2015b). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature*. 519 (7542). p.pp. 193–198.
- Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. & Aravin, A. (2013). Bacterial Argonaute samples the transcriptome to identify foreign DNA. *Molecular Cell*. 51 (5). p.pp. 594–605.

- van der Oost, J., Westra, E.R., Jackson, R.N. & Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nature Reviews Microbiology*. 12 (7). p.pp. 479–492.
- Osawa, T., Inanaga, H., Sato, C. & Numata, T. (2015). Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. *Molecular Cell*. 58 (3). p.pp. 418–430.
- Patterson, A.G., Chang, J.T., Taylor, C. & Fineran, P.C. (2015). Regulation of the type I-F CRISPR-Cas system by CRP-cAMP and GalM controls spacer acquisition and interference. *Nucleic Acids Research*. 43 (12). p.pp. 6038–6048.
- Pawluk, A., Bondy-Denomy, J., Cheung, V.H.W., Maxwell, K.L. & Davidson, A.R. (2014). A new group of phage Anti-CRISPR genes inhibits the Type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*. *mBio*. 5 (2). p.pp. e00896-14-e00896-14.
- Pawluk, A., Staals, R.H.J., Taylor, C., Watson, B.N.J., Saha, S., Fineran, P.C., Maxwell, K.L. & Davidson, A.R. (2016). Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nature Microbiology*. 1 (8). p.p. 16085.
- Paytubi, S. & White, M.F. (2009). The crenarchaeal DNA damage-inducible transcription factor B paralogue TFB3 is a general activator of transcription. *Molecular Microbiology*. 72 (6). p.pp. 1487–1499.
- Peng, X., Brügger, K., Shen, B., Chen, L., She, Q. & Garrett, R.A. (2003). Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. *Journal of Bacteriology*. 185 (8). p.pp. 2410–2417.
- Perez-Rodriguez, R., Haitjema, C., Huang, Q., Nam, K.H., Bernardis, S., Ke, A. & DeLisa, M.P. (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Molecular Microbiology*. 79 (3). p.pp. 584–599.
- Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*. 29 (9). p.pp. 2002–2007.
- Plagens, A., Richter, H., Charpentier, E. & Randau, L. (2015). DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiology Reviews*. 39 (3). p.pp. 442–463.
- Plagens, A., Tjaden, B., Hagemann, A., Randau, L. & Hensel, R. (2012). Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon *Thermoproteus tenax*. *Journal of Bacteriology*. 194 (10). p.pp. 2491–500.
- Pourcel, C., Salvignol, G. & Vergnaud, G. (2005). CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*. 151 (3). p.pp. 653–663.
- Proudfoot, M., Brown, M., Singer, A., Skarina, T., Tan, K., Kagan, O., Edwards, A., Joachimiak, A., Savchenko, A. & Yakunin, A.F. (n.d.). Structure of the RNA'se SSO8090 from *Sulfolobus solfataricus*. *To be Published*.
- Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N. & Wagner, R. (2010). Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Molecular Microbiology*. 75 (6). p.pp. 1495–512.

- Quax, T.E.F., Voet, M., Sismeiro, O., Dillies, M.-A., Jagla, B., Coppée, J.-Y., Sezonov, G., Forterre, P., van der Oost, J., Lavigne, R. & Prangishvili, D. (2013). Massive activation of archaeal defense genes during viral infection. *Journal of Virology*. 87 (15). p.pp. 8419–28.
- Reeks, J., Sokolowski, R.D., Graham, S., Liu, H., Naismith, J.H. & White, M.F. (2013). Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. *Biochemical Journal*. 452 (2). p.pp. 223–230.
- Reid, S.L., Parry, D., Liu, H.H. & Connolly, B.A. (2001). Binding and recognition of GATATC target sequences by the *EcoRV* restriction endonuclease: A study using fluorescent oligonucleotides and fluorescence polarization. *Biochemistry*. 40 (8). p.pp. 2484–2494.
- Rice, P.A., Yang, S.W., Mizuuchi, K. & Nash, H.A. (1996). Crystal structure of an IHF-DNA complex: A protein-induced DNA U-turn. *Cell*. 87 (7). p.pp. 1295–1306.
- Richter, C., Chang, J.T. & Fineran, P.C. (2012). Function and regulation of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR associated (Cas) systems. *Viruses*. 4 (12). p.pp. 2291–2311.
- 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 (13). p.pp. 8516–8526.
- Riede, I. & Eschbach, M.L. (1986). Evidence that TraT interacts with OmpA of *Escherichia coli*. *FEBS Letters*. 205 (2). p.pp. 241–245.
- Roberts, R.J., Vincze, T., Posfai, J. & Macelis, D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*. 38 (Database). p.pp. D234–D236.
- 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 (AUGUST 2015). p.pp. 1–19.
- Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C. V., Spagnolo, L. & White, M.F. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Molecular Cell*. 52 (1). p.pp. 124–134.
- Rutkauskas, M., Sinkunas, T., Songailiene, I., Tikhomirova, M., Siksnys, V. & Seidel, R. (2015). Directional R-loop formation by the CRISPR-cas surveillance complex Cascade provides efficient off-target site rejection. *Cell Reports*. 10 (9). p.pp. 1534–1543.
- Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A. & Marraffini, L.A. (2015). Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. *Cell*. 161 (5). p.pp. 1164–1174.
- Samai, P., Smith, P. & Shuman, S. (2010). Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. *Acta Crystallographica Section F Structural Biology and Crystallization Communications*. 66 (12). p.pp. 1552–1556.
- Sashital, D.G., Wiedenheft, B. & Doudna, J.A. (2012). Mechanism of foreign DNA

- selection in a bacterial adaptive immune system. *Molecular Cell*. 46 (5). p.pp. 606–615.
- Semenova, E., Jore, M.M., Datsenko, K. a, Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J. & Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences*. 108 (25). p.pp. 10098–10103.
- Shah, S.A. & Garrett, R.A. (2011). CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Research in Microbiology*. 162 (1). p.pp. 27–38.
- Shah, S.A., Hansen, N.R. & Garrett, R.A. (2009). Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. *Biochemical Society Transactions*. 37 (1). p.pp. 23–28.
- She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C.-Y., Clausen, I.G., Curtis, B. a, De Moors, A., Erauso, G., Fletcher, C., Gordon, P.M.K., Heikamp-de Jong, I., Jeffries, a C., Kozera, C.J., Medina, N., Peng, X., Thi-Ngoc, H.P., Redder, P., Schenk, M.E., Theriault, C., Tolstrup, N., Charlebois, R.L., Doolittle, W.F., Duguet, M., Gaasterland, T., Garrett, R. a, Ragan, M. a, Sensen, C.W. & Van der Oost, J. (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proceedings of the National Academy of Sciences*. 98 (14). p.pp. 7835–7840.
- Shinkai, A., Kira, S., Nakagawa, N., Kashihara, A., Kuramitsu, S. & Yokoyama, S. (2007). Transcription activation mediated by a cyclic AMP receptor protein from *Thermus thermophilus* HB8. *Journal of Bacteriology*. 189 (10). p.pp. 3891–3901.
- Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., Zhang, F. & Koonin, E. V. (2015). Discovery and functional characterization of diverse Class 2 CRISPR-Cas systems. *Molecular Cell*. 60 (3). p.pp. 385–397.
- Silas, S., Mohr, G., Sidote, D.J., Markham, L.M., Sanchez-Amat, A., Bhaya, D., Lambowitz, A.M. & Fire, A.Z. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science*. 351 (6276). p.p. aad4234-aad4234.
- Sokolowski, R.D., Graham, S. & White, M.F. (2014). Cas6 specificity and CRISPR RNA loading in a complex CRISPR-Cas system. *Nucleic Acids Research*. 42 (10). p.pp. 6532–6541.
- Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Terns, M., Li, H. & Stagg, S. (2013). Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Molecular Cell*. 52 (1). p.pp. 146–152.
- Staals, R.H.J., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D.W., VanDuijn, E., Barendregt, A., Vlot, M., Koehorst, J.J., Sakamoto, K., Masuda, A., Dohmae, N., Schaap, P., Doudna, J.A., Heck, A.J.R., Yonekura, K., Van der Oost, J. & Shinkai, A. (2013). Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Molecular Cell*. 52 (1). p.pp. 135–145.
- Staals, R.H.J., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., Sakamoto, K., Suzuki, T.,

- Dohmae, N., Yokoyama, S., Schaap, P.J., Urlaub, H., Heck, A.J.R., Nogales, E., Doudna, J.A., Shinkai, A. & VanderOost, J. (2014). RNA targeting by the Type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Molecular Cell*. 56 (4). p.pp. 518–530.
- Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. (2010). Self-targeting by CRISPR: Gene regulation or autoimmunity? *Trends in Genetics*. 26 (8). p.pp. 335–340.
- Stern, A. & Sorek, R. (2012). The phage-host arms-race: Shaping the evolution of microbes. *Bioessays*. 33 (1). p.pp. 43–51.
- Sternberg, S.H., Haurwitz, R.E. & Doudna, J. a. (2012). Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*. 18 (4). p.pp. 661–672.
- Surette, M.G. & Chaconas, G. (1989). A protein factor which reduces the negative supercoiling requirement in the Mu DNA strand transfer reaction is *Escherichia coli* integration host factor. *Journal of Biological Chemistry*. 264 (5). p.pp. 3028–3034.
- Suttle, C. a (2007). Marine viruses--major players in the global ecosystem. *Nature Reviews Microbiology*. 5 (10). p.pp. 801–812.
- Swarts, D.C., Jore, M.M., Westra, E.R., Zhu, Y., Janssen, J.H., Snijders, A.P., Wang, Y., Patel, D.J., Berenguer, J., Brouns, S.J.J. & van der Oost, J. (2014a). DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*. 507 (7491). p.pp. 258–61.
- Swarts, D.C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R.F., Koonin, E. V, Patel, D.J. & van der Oost, J. (2014b). The evolutionary journey of Argonaute proteins. *Nature Structural & Molecular Biology*. 21 (9). p.pp. 743–753.
- Swarts, D.C., Mosterd, C., van Passel, M.W.J. & Brouns, S.J.J. (2012). CRISPR interference directs strand specific spacer acquisition I. Mokrousov (ed.). *PLoS ONE*. 7 (4). p.p. e35888.
- Tamulaitis, G., Kazlauskienė, M., Manakova, E., Venclovas, Č., Nwokeoji, A.O., Dickman, M.J., Horvath, P. & Siksnys, V. (2014). Programmable RNA shredding by the Type III-A CRISPR-Cas system of *Streptococcus thermophilus*. *Molecular Cell*. 56 (4). p.pp. 506–517.
- Tang, T.-H., Bachellerie, J.-P., Rozhdestvensky, T., Bortolin, M.-L., Huber, H., Drungowski, M., Elge, T., Brosius, J. & Huttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proceedings of the National Academy of Sciences*. 99 (11). p.pp. 7536–7541.
- Taylor, S., Wakem, M., Dijkman, G., Alsarraj, M. & Nguyen, M. (2010). A practical approach to RT-qPCR—Publishing data that conform to the MIQE guidelines. *Methods*. 50 (4). p.pp. S1–S5.
- Took, M.R. & Dryden, D.T.F. (2005). The biology of restriction and anti-restriction. *Current Opinion in Microbiology*. 8 (4). p.pp. 466–472.
- Veening, J.-W., Smits, W.K. & Kuipers, O.P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. *Annual Review of Microbiology*. 62 (1). p.pp. 193–210.
- Wang, J., Li, J., Zhao, H., Wang, M., Yin, M., Wang, Y., 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

- (4). p.pp. 1–14.
- 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 (9). p.pp. 4266–4277.
- Wang, R., Preamplume, G., Terns, M.P., Terns, R.M. & Li, H. (2011). Interaction of the Cas6 ribonuclease with CRISPR RNAs: Recognition and cleavage. *Structure*. 19 (2). p.pp. 257–264.
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. & Barton, G.J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 25 (9). p.pp. 1189–1191.
- 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 (3). p.pp. 1749–58.
- Westra, E.R., van Erp, P.B.G., Künne, T., Wong, S.P., Staals, R.H.J., Seegers, C.L.C., Bollen, S., Jore, M.M., Semenova, E., Severinov, K., de Vos, W.M., Dame, R.T., de Vries, R., Brouns, S.J.J. & van der Oost, J. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Molecular cell*. 46 (5). p.pp. 595–605.
- Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K. & Brouns, S.J.J. (2013). Type I-E CRISPR-Cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition P. H. Viollier (ed.). *PLoS Genetics*. 9 (9). p.p. e1003742.
- Westra, E.R., Ümit, P., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., Mastop, M., Wagner, E.G.H., Schnetz, K., Van Der Oost, J., Wagner, R. & Brouns, S.J.J. (2010). H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. *Molecular Microbiology*. 77 (6). p.pp. 1380–1393.
- Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., van der Oost, J., Doudna, J.A. & Nogales, E. (2011). Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*. 477 (7365). p.pp. 486–489.
- 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 (6). p.pp. 904–912.
- Winther, K.S. & Gerdes, K. (2011). Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proceedings of the National Academy of Sciences*. 108 (18). p.pp. 7403–7407.
- Wright, A. V & Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. *Nature Structural & Molecular Biology*. 23 (10). p.pp. 876–883.
- 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 (12). p.pp. 5569–5576.
- Yosef, I., Shitrit, D., Goren, M.G., Burstein, D., Pupko, T. & Qimron, U. (2013). DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proceedings of the National Academy of Sciences*. 110 (35). p.pp.

14396–14401.

- Yu, H., Jiang, X., Tan, K.T., Hang, L. & Patzel, V. (2015). Efficient production of superior dumbbell-shaped DNA minimal vectors for small hairpin RNA expression. *Nucleic Acids Research*. 43 (18). p.pp. e120–e120.
- Yuan, Y.R., Pei, Y., Ma, J.B., Kuryavyi, V., Zhadina, M., Meister, G., Chen, H.Y., Dauter, Z., Tuschl, T. & Patel, D.J. (2005). Crystal structure of *A. aeolicus* argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Molecular Cell*. 19 (3). p.pp. 405–419.
- Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V. & Zhang, F. (2015). Cpf1 is a single RNA-guided endonuclease of a Class 2 CRISPR-Cas system. *Cell*. 163 (3). p.pp. 759–771.
- Zhang, J., Graham, S., Tello, A., lu, H. & White, M.F. (2016). Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. *Nucleic Acids Research*. 44 (4). p.pp. 1789–1799.
- Zhang, J., Kasciukovic, T. & White, M.F. (2012a). The CRISPR associated protein Cas4 is a 5' to 3' DNA exonuclease with an iron-sulfur cluster M. G. Marinus (ed.). *PLoS ONE*. 7 (10). p.pp. 1–8.
- Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.-V., Naismith, J., Spagnolo, L. & White, M. (2012b). Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Molecular Cell*. 45 (3). p.pp. 303–313.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*. 31 (13). p.pp. 3406–3415.

Appendices

Appendix A: Triplicate Ct values from RT-qPCR

Gene	Control					Infected				
	Ct 1	Ct 2	Ct 3	Mean	SD	Ct 1	Ct 2	Ct 3	Mean	SD
<i>ssb</i> (<i>sso2364</i>)	10.75	10.85	10.60	10.73	0.38	10.8	11.4	11.05	11.08	0.54
<i>cas1</i> (<i>sso1450</i>)	23.95	23.5	22.35	23.27	0.61	19.85	21.15	19.35	20.12	0.44
<i>cascade</i> (<i>sso1443</i>)	20.45	21.7	22	21.38	0.59	17.35	18.05	18.8	18.06	0.4
<i>csm</i> (<i>sso1424</i>)	16.15	16.85	15.55	16.18	0.21	15.15	16.1	14.55	15.27	0.19
<i>cmr</i> (<i>sso1986</i>)	16.5	16.55	16.3	16.45	0.21	15.7	14.95	15.4	15.35	0.25
<i>CRISPR C</i> (<i>pre-crRNA</i>)	22.5	16.55	16.3	23.03	0.38	22.15	23.15	22.7	22.67	0.51

Appendix B: Published work from this thesis

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 (AUGUST 2015). p.pp. 1–19.

Please see overleaf for full manuscript.

Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition

Clare Rollie¹, Stefanie Schneider², Anna Sophie Brinkmann³, Edward L Bolt³, Malcolm F White^{1*}

¹Biomedical Sciences Research Complex, University of St Andrews, St Andrews, United Kingdom; ²Faculty of Medicine, Institute of Cell Biology, University of Duisburg-Essen, Essen, Germany; ³School of Life Sciences, Queen's Medical Centre, University of Nottingham, Nottingham, United Kingdom

Abstract The adaptive prokaryotic immune system CRISPR-Cas provides RNA-mediated protection from invading genetic elements. The fundamental basis of the system is the ability to capture small pieces of foreign DNA for incorporation into the genome at the CRISPR locus, a process known as Adaptation, which is dependent on the Cas1 and Cas2 proteins. We demonstrate that Cas1 catalyses an efficient trans-esterification reaction on branched DNA substrates, which represents the reverse- or disintegration reaction. Cas1 from both *Escherichia coli* and *Sulfolobus solfataricus* display sequence specific activity, with a clear preference for the nucleotides flanking the integration site at the leader-repeat 1 boundary of the CRISPR locus. Cas2 is not required for this activity and does not influence the specificity. This suggests that the inherent sequence specificity of Cas1 is a major determinant of the adaptation process.

DOI: [10.7554/eLife.08716.001](https://doi.org/10.7554/eLife.08716.001)

Introduction

The CRISPR-Cas system is an adaptive immune system present in many archaeal and bacterial species. It provides immunity against viruses and other mobile genetic elements mediated through sequence homology-directed detection and destruction of foreign nucleic acid species (reviewed in [Sorek et al., 2013](#); [Barrangou and Marraffini, 2014](#); [van der Oost et al., 2014](#)). Organisms with an active CRISPR-Cas system encode one or more CRISPR loci in their genomes. These comprise a leader sequence followed by an array of short, direct, often palindromic repeats interspersed by short 'spacer' sequences, together with a variable number of CRISPR associated (*cas*) genes. Spacers are derived from mobile genetic elements and constitute a 'memory' of past infections. The CRISPR locus is transcribed from the leader, generating pre-CRISPR RNA (pre-crRNA) that is subsequently processed into unit length crRNA species and loaded into large ribonucleoprotein complexes. These complexes, known as 'Interference', 'Effector' or 'Surveillance' complexes, utilize crRNA to detect and degrade cognate invading genetic entities, providing immunity against previously encountered viruses and plasmids.

The process of foreign DNA capture and integration into the CRISPR locus is known as 'Acquisition' or 'Adaptation' (reviewed in [Fineran and Charpentier, 2012](#); [Heler et al., 2014](#)). This process underpins the whole CRISPR-Cas system, but is the least well understood aspect. Fundamentally, acquisition can be separated into two steps: the capture of new DNA sequences from mobile genetic elements, followed by the integration of that DNA into the host genome. New spacers are incorporated proximal to the leader sequence and result in the duplication of the first repeat ([Goren et al., 2012](#); [Yosef et al., 2012](#)). The leader sequence proximal to the repeat is important for integration, but transcription of the locus is not required ([Yosef et al., 2012](#)). New spacers are frequently incorporated in both possible orientations ([Shmakov et al., 2014](#)). The integration process in *Escherichia coli* results in staggered cleavage of the first CRISPR repeat, where the 3'-end of one

*For correspondence:
mfw2@st-and.ac.uk

Competing interests: The authors declare that no competing interests exist.


Funding: See page 18

Received: 14 May 2015

Accepted: 17 August 2015

Published: 18 August 2015

Reviewing editor: Timothy W Nilsen, Case Western Reserve University, United States

 Copyright Rollie et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

eLife digest In most animals, the adaptive immune system creates specialized cells that adapt to efficiently fight off any viruses or other pathogens that have invaded. Bacteria (and another group of single-celled organisms called archaea) also have an adaptive immune system, known as CRISPR-Cas, that combats viral invaders. This system is based on sections of the microbes' DNA called CRISPRs, which contain repetitive DNA sequences that are separated by short segments of 'spacer' DNA. When a virus invades the cell, some viral DNA is incorporated into the CRISPR as a spacer. This process is known as adaptation. CRISPR-associated proteins (or 'Cas' proteins) then use this spacer to recognize and mount an attack on any matching invader DNA that is later encountered.

Exactly how a spacer is inserted into the correct position in the CRISPR array during adaptation remains poorly understood. However, it is known that two CRISPR proteins called Cas1 and Cas2 play essential roles in this process.

Rollie et al. took Cas1 proteins from a bacterial cell (*Escherichia coli*) and an archaeal species (*Sulfolobus solfataricus*) and added them to branched DNA structures in the laboratory. These experiments revealed that Cas1 from both organisms can break the DNA down into smaller pieces. Cas2, on the other hand, is not required for this process. This 'disintegration' reaction is the reverse process of the 'integration' step of adaptation where the CRISPR proteins insert the invader DNA into the CRISPR array.

Rollie et al. also found that the disintegration reaction performed by Cas1 takes place on specific DNA sequences, which are also the sites where Cas1 inserts the spacer DNA during adaptation. Therefore, by examining the disintegration reaction, many of the details of the integration step can be deduced.

Overall, Rollie et al. show that selection by Cas1 plays an important role in restricting the adaptation process to particular DNA sites. The next step will be to use the disintegration reaction to examine the DNA binding and manipulation steps performed by Cas1 as part of its role in the adaptation of the CRISPR system.

DOI: [10.7554/eLife.08716.002](https://doi.org/10.7554/eLife.08716.002)

strand of the incoming DNA is joined to the end of the CRISPR repeat in a trans-esterification (TES) reaction, with another TES reaction occurring on the other strand (*Diez-Villasenor et al., 2013*) (**Figure 1**, numbered yellow arrows). Intermediates in this reaction have been observed in *E. coli*, and the sequence of the leader-repeat1 junction is important for the activity (*Arslan et al., 2014*). The order of the two integration steps shown in **Figure 1** is not yet known, although it has been suggested that the reaction on the minus strand (site 2) occurs first in *E. coli* (*Nuñez et al., 2015*). The sequence at site 2 is flanked by the end of the first repeat and the first spacer, and is therefore inherently less well conserved. In *E. coli*, the last nucleotide of the newly copied repeat is derived from the first nucleotide of the incoming spacer, which imposes further sequence requirements on that system (*Swarts et al., 2012*).

Although shown in **Figure 1** as fully double-stranded, the incoming spacer DNA could have a partially single-stranded character. Recent work by Sorek and colleagues has shown that the *E. coli* RecBCD helicase–nuclease complex, which processes DNA double-strand breaks for recombination and degrades foreign DNA, is implicated in the generation of viral DNA fragments captured by Cas1 and incorporated into the CRISPR locus as new spacers (*Levy et al., 2015*). This confirms previous observations linking Cas1 with RecBCD (*Babu et al., 2011*) and raises some intriguing questions as RecBCD generates ssDNA fragments asymmetrically, generating fragments tens to hundreds of nucleotides long from the 3' terminated strand and much longer fragments from the 5' terminated strand (reviewed in *Dillingham and Kowalczykowski, 2008*). The Cas4 enzyme, which is a 5' to 3' ssDNA exonuclease (*Zhang et al., 2012; Lemak et al., 2013*), may provide ssDNA fragments for Cas1 in systems lacking RecBCD. However, it is difficult to see how two integration reactions could occur without two 3' hydroxyl termini (**Figure 1**) and half-site integration is not observed with a ssDNA substrate (*Nuñez et al., 2015*). Potentially, the ssDNA fragments generated by these nucleases may re-anneal and experience further processing to generate partially duplex molecules of defined size prior to integration by Cas1.

Adaptation requires the products of the *cas1* and *cas2* genes in vivo and these are the most universally conserved of all the *cas* genes (*Makarova et al., 2006*). Cas1 is a homodimeric enzyme

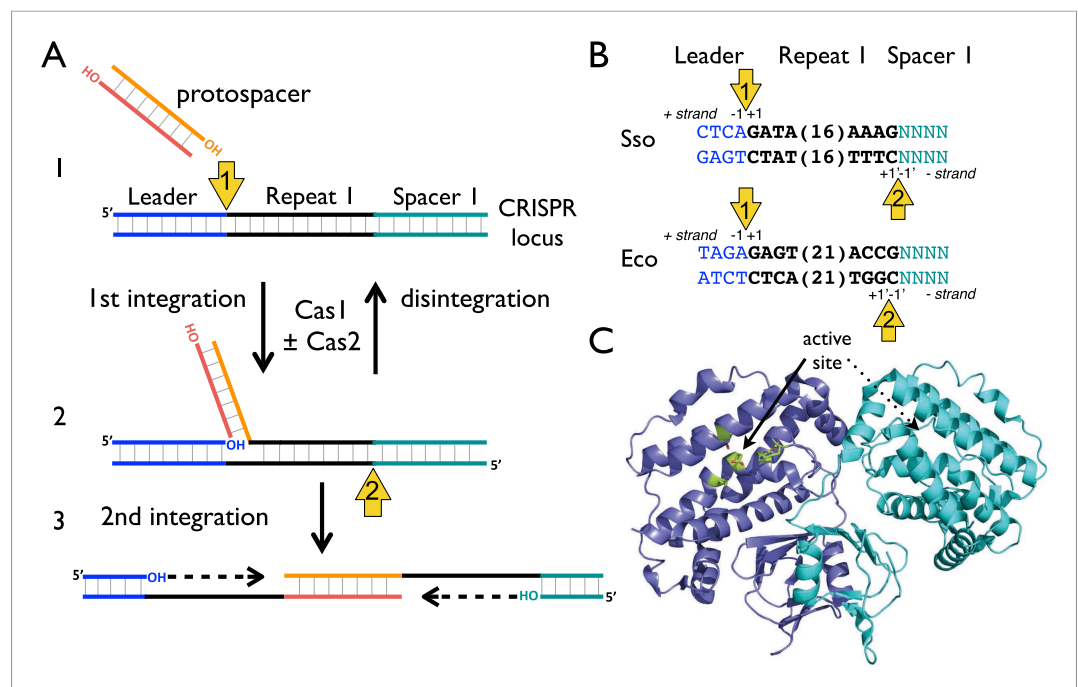


Figure 1. CRISPR spacer acquisition and Cas1. **(A, 1)** The 3'-end of an incoming protospacer attacks the chromosomal CRISPR locus at the boundary between the leader sequence and repeat 1. A trans-esterification (TES) reaction (yellow arrow 1) catalyzed by Cas1 joins the protospacer to the 5' end of repeat 1. For many integrases a (reverse) disintegration reaction can be observed in vitro. **(2)** Another TES reaction (yellow arrow 2) joins the other strand of the protospacer to the 5' end of repeat 1 on the bottom (minus) strand, resulting in the formation of a gapped DNA duplex. **(3)** The gapped duplex is repaired by the host cell DNA replication machinery, resulting in the addition of a new spacer at position 1 and replication of CRISPR repeat 1. **(B)** Sequences flanking the two TES reaction sites at repeat 1 in *Sulfolobus solfataricus* and *Escherichia coli* are shown. The leader is in blue, repeat in black and spacer 1 in teal. The number of central nucleotides of the repeat omitted from the sequence is shown in parentheses. **(C)** Structure of Cas1 from *Pyrococcus horikoshii* (PDB 4WJ0) with subunits coloured blue and cyan, showing the dimeric 'butterfly' conformation with the active site residues highlighted in green.

DOI: [10.7554/eLife.08716.003](https://doi.org/10.7554/eLife.08716.003)

with a two-domain structure and a canonical metal dependent nuclease active site in the large domain formed by a cluster of highly conserved residues (Wiedenheft *et al.*, 2009) (Figure 1C). *E. coli* Cas1 has nuclease activity in vitro, with activity observed against double- and single-stranded DNA and RNA substrates (Babu *et al.*, 2011). Some specificity was observed for branched DNA substrates, in particular for DNA constructs resembling replication forks (Babu *et al.*, 2011). Initial biochemical analyses of a panel of archaeal Cas2 enzymes revealed an endonucleolytic activity against ssRNA substrates that could be abrogated by mutation of conserved residues (Beloglazova *et al.*, 2008). In contrast, Cas2 from *Bacillus halodurans* has been shown to be specific for cleavage of dsDNA substrates (Nam *et al.*, 2012). Recently however, Doudna and colleagues demonstrated that *E. coli* Cas1 and Cas2 form a stable complex in vitro and that the 'active site' of Cas2 was not required for spacer acquisition, suggesting that Cas2 may not have a catalytic role in spacer acquisition in vivo (Nuñez *et al.*, 2014). It is probable that Cas2 acts as an adaptor protein, either bringing two Cas1 dimers together or mediating interactions with other components necessary for spacer acquisition. Recently, Nuñez *et al.* demonstrated that *E. coli* Cas1 can integrate a protospacer into a supercoiled plasmid in vitro, in a reaction stimulated by Cas2. Integration events were observed at the boundaries of most CRISPR repeats and in other areas of the DNA close to palindromic regions, implicating a role for palindromic DNA structure in the adaptation process (Nuñez *et al.*, 2015).

In order to further mechanistic understanding of the spacer acquisition process, we have undertaken a systematic analysis of the underlying biochemistry. We demonstrate that Cas1 catalyses TES of branched DNA substrates efficiently in vitro in a reaction that represents the reverse- or disintegration of an incoming spacer from the CRISPR locus. The disintegration reactions catalysed by

diverse integrases have proven a powerful model system for the understanding of the mechanism of integration. For Cas1, the reaction is strongly sequence dependent with the specificity matching the predicted integration site 1 for both *E. coli* and *Sulfolobus solfataricus* Cas1, and does not require Cas2.

Results

Cas1 catalyzes a TES reaction on branched DNA substrates

The Cas1 and Cas2 proteins from *S. solfataricus* (CRISPR-Cas subtype IA) and *E. coli* (CRISPR-Cas subtype IE) were expressed in *E. coli* with N-terminal polyhistidine tags and purified to homogeneity by metal affinity and gel filtration chromatography. Previously, it was demonstrated that *E. coli* Cas1 (EcoCas1) cleaved branched DNA substrates preferentially (Babu et al., 2011). We investigated the activity of *S. solfataricus* Cas1 (SsoCas1) against a DNA substrate with a 5'-flap structure (Figure 2). By labelling the single-stranded 5'-flap of the downstream duplex at the 5'-end with radioactive phosphorus, we observed cleavage of the flap by SsoCas1, releasing an 18 nt product (Figure 2B). A variant of SsoCas1 where the active site metal ligand glutamic acid 142 was changed to an alanine (E142A) was inactive, implicating the canonical nuclease active site of Cas1 in the reaction. This result appeared consistent with the earlier studies for EcoCas1 (Babu et al., 2011) and suggested that the ssDNA flap was removed at or close to the branch point. The activity was independent of the presence or absence of SsoCas2, suggesting that Cas2 is not involved in this nuclease activity in vitro.

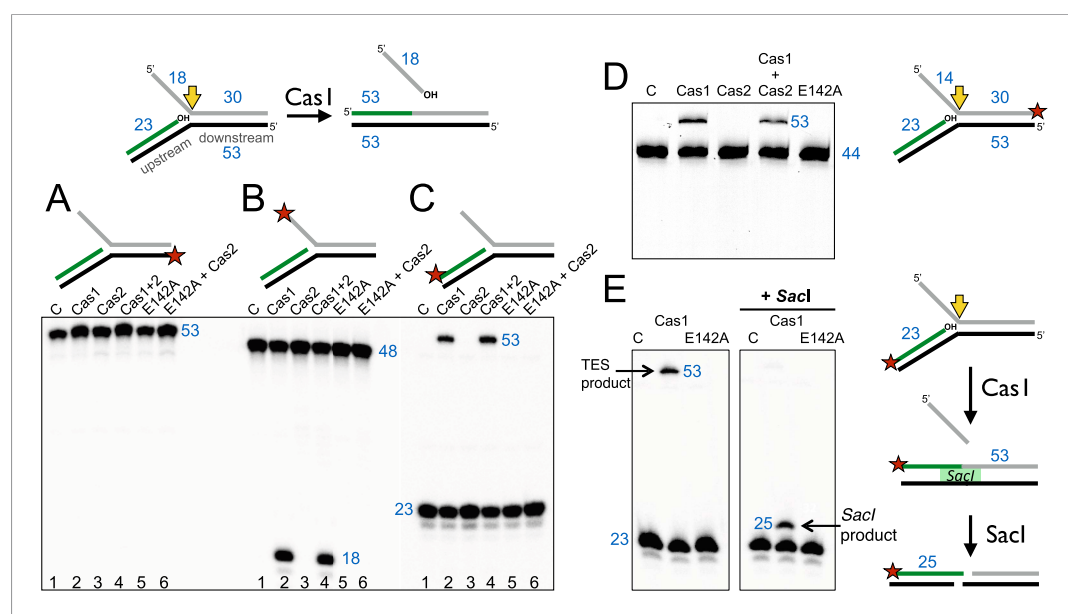


Figure 2. Disintegration of a branched DNA substrate by SsoCas1. Denaturing gel electrophoresis was used to analyse the products generated by SsoCas1 with a branched DNA substrate (Substrate 1). The 5' flap (18 nt) was released when the phosphodiester backbone was attacked by the 3'-hydroxyl group at the branch point. The reaction required active Cas1 and was independent of Cas2. DNA lengths are shown in blue (nt). The TES site is indicated with a yellow arrow and the labelled strand with a red star. (A) Shows reactions with the continuous strand (black) labelled; (B) with the flap (grey) strand labelled and (C) with the upstream (green) strand labelled, each on the 5' end. Lanes: 1, control with no added protein; 2, WT Cas1; 3, Cas2; 4, Cas1 + Cas2; 5, Cas1 E142A variant; 6, E142A Cas1 + Cas2. (D) The 5'-flap strand was labelled on the 3' end with a fluorescein moiety, and the flap reduced to 14 nt (Substrate 1-FAM). Cas1 catalyses the TES reaction generating a 53 nt labelled strand. Lane C: control incubation in absence of Cas1. (E) TES reactions were carried out with SsoCas1, or the E142A active site mutant, on a fork substrate containing a nicked *SacI* restriction site spanning the branch point (*SacI* substrate). A TES product of 53 nucleotides is visible in lane 2 containing Cas1. The right-hand panel shows the effect of adding *SacI* restriction enzyme after the TES reaction. The TES product is no longer visible, but a shorter product of 25 nucleotides is present indicating regeneration of the *SacI* recognition sequence by the TES reaction.

DOI: 10.7554/eLife.08716.004

We proceeded to label the other strands of the substrate to follow the reaction products in more detail. The continuous (black) strand was not cleaved by SsoCas1 (**Figure 2A**). However, when the 23 nt (green) strand of the upstream duplex presenting a 3'-hydroxyl end at the junction point was labelled we observed the formation of a new, larger DNA species (**Figure 2C**). This observation was consistent with a joining or TES of the upstream DNA to the downstream DNA strand by Cas1. By switching to a label at the 3' end of the downstream duplex we confirmed that the reaction catalyzed by Cas1 involved the joining of the upstream and downstream DNA duplexes with concomitant loss of the 5'-flap (**Figure 2D**). The lack of evidence for any shorter DNA species in **Figure 2D** was consistent with the conclusion that we were observing a TES rather than a nuclease reaction. Again, the TES reaction was unaffected by the presence or absence of Cas2 and was dependent on the wild-type active site of Cas1, as the E142A variant was inactive.

In order to define the product of the TES reaction in more detail, we modified the sequence of the branch point to introduce an interrupted site for the restriction enzyme *SacI* across the junction. A precise TES reaction at the branch point to generate duplex DNA would result in the 'repair' of the restriction site, generating a substrate for *SacI*. SsoCas1 processed this substrate generating the 53 bp TES reaction product. On addition of the *SacI* restriction enzyme, the 53 bp species was converted to a 25 bp product (**Figure 2E**). This confirmed that the Cas1-mediated reaction resulted in the formation of a functional *SacI* site in vitro, consistent with a precise TES reaction at the branch point.

It is likely that the TES reaction catalyzed by Cas1 with branched DNA substrates in vitro represents the reverse or disintegration of an integration intermediate, as observed recently for EcoCas1 (**Nuñez et al., 2015**). We therefore tested EcoCas1 with the same set of branched substrates, showing that manganese, magnesium and cobalt all supported the same robust disintegration activity in the absence of Cas2, with no nuclease activity observed (**Figure 3A**). Given that Eco and SsoCas1 are divergent members of the Cas1 family, this suggests that the disintegration activity is likely to be a general property of Cas1 enzymes. Experiments where the concentration of Cas1 was titrated against a fixed concentration of substrate (50 nM), showed that maximal activity plateaued above 250 nM for EcoCas1 and was maximal from 100 to 500 nM for SsoCas1 (**Figure 3B,C**).

To characterise the specificity of the disintegration reaction in more detail, we examined SsoCas1 activity for a variety of substrates differing in the nature of the 5'-flap or duplex arm released by the reaction (**Figure 4**). SsoCas1 was indifferent to the presence of duplex or single-stranded DNA in the 5'-flap, processing a nicked 3-way junction with a similar efficiency to the 5'-flap substrate. The disintegration reaction required the presence of the 3'-hydroxyl moiety at the branch point as a three-way (or Y) junction was not a substrate for Cas1. To confirm this observation we studied a 5'-flap substrate with a phosphate moiety at the 3' end of the upstream strand in place of a hydroxyl group. As expected, this substrate did not support disintegration activity for either Sso or EcoCas1, with no larger TES product detected (right hand lanes). Tellingly, neither enzyme cleaved the 5'-flap of this substrate to generate shorter products (left hand lanes), confirming that Cas1 catalyses a concerted TES reaction rather than a sequential 'cut and join' activity.

Sequence specificity of the disintegration reaction

Disintegration reactions are commonly seen for integrases and transposases, and represent a very useful means to study the underlying integration mechanism (**Chow et al., 1992; Delelis et al., 2008**) as the local arrangement of the DNA in the integrase active site is typically the same for the two reactions (**Gerton et al., 1999**). One prediction of this hypothesis is that the disintegration reaction could demonstrate some sequence preference if integration, which must happen at a specific, defined site in the CRISPR locus, is partly driven by the sequence specificity of Cas1. We therefore designed a family of related disintegration substrates by systematically varying the nucleotides flanking the TES site and tested how efficiently Cas1 could disintegrate these substrates. Having demonstrated conclusively that we could follow the progress of the disintegration reaction by monitoring the liberation of a displaced DNA strand from a 5'-flap substrate, we used this assay for all subsequent investigations.

The +1 position

We first tested the importance of the nucleotide acting as an acceptor for the attacking 3'-hydroxyl of the incoming DNA strand (the +1 position) by varying the nucleotide at this position in the model

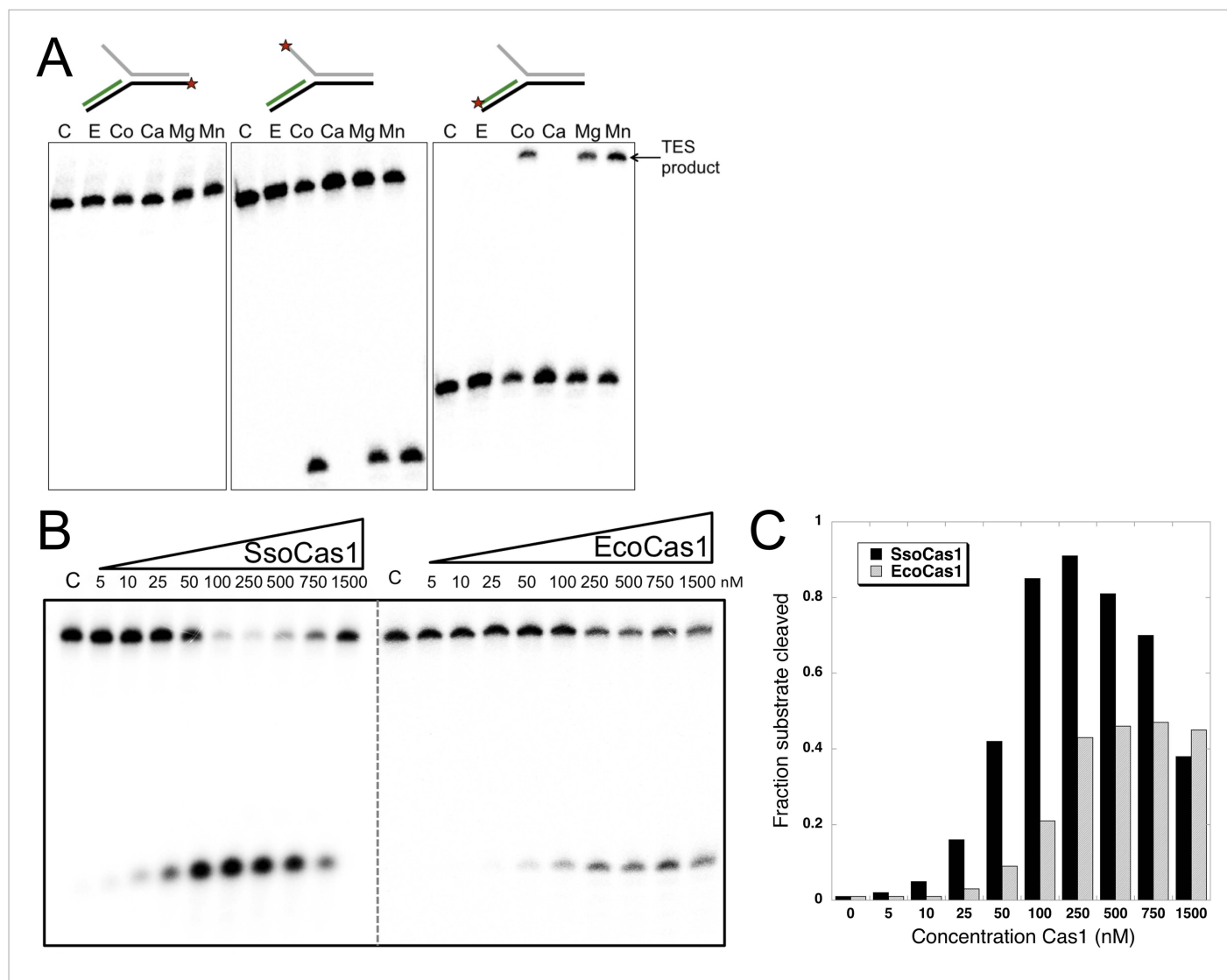


Figure 3. TES activity of *E. coli* and *S. solfataricus* Cas1. **(A)** *E. coli* Cas1 also catalyses an efficient metal dependent disintegration reaction. TES reactions were carried out under standard conditions, using Substrate 3 and varying the divalent metal ion as indicated. EcoCas1 showed robust TES activity in the presence of cobalt, magnesium and manganese. Each of the three strands of the substrate was labelled individually as for **Figure 2** (5' label indicated by a star). Lanes were: c, substrate alone; substrate incubated with Cas1 and 5 mM of E, EDTA; Co, cobalt chloride; Ca, calcium chloride; Mg, magnesium chloride; Mn, manganese chloride for 30 min at 37°C. **(B)** Concentration dependence of Cas1 TES activity. Substrate 3 (50 nM) was incubated with the indicated concentration of Sso or EcoCas1 for 30 min under standard assay conditions and the reactants were analysed by denaturing gel electrophoresis and phosphorimaging. SsoCas1 showed maximal activity at 250 nM, representing a fivefold molar excess of enzyme over substrate, with a decline in activity above 500 nM enzyme. EcoCas1 had maximal activity that plateaued above 250 nM enzyme. **(C)** Quantification of the data (raw data provided in **Figure 3—source data 1**). These data are representative of duplicate experiments.

DOI: [10.7554/eLife.08716.005](https://doi.org/10.7554/eLife.08716.005)

The following source data is available for figure 3:

Source data 1. Concentration Cas1.

DOI: [10.7554/eLife.08716.006](https://doi.org/10.7554/eLife.08716.006)

substrates (**Figure 5**). In vivo, this acceptor nucleotide should represent the position in the CRISPR locus where new spacers are joined to the end of the repeat. The *S. solfataricus* CRISPR repeat has a guanine at one end and a cytosine at the other, each of which are predicted to act as acceptors for a new phosphodiester bond formed with the incoming spacer (**Figure 1B**). Consistent with this, we observed the most efficient disintegration activity with SsoCas1 when substrates had a guanine at the +1 position (**Figure 5A**). Assays were carried out in triplicate and the reaction products quantified,

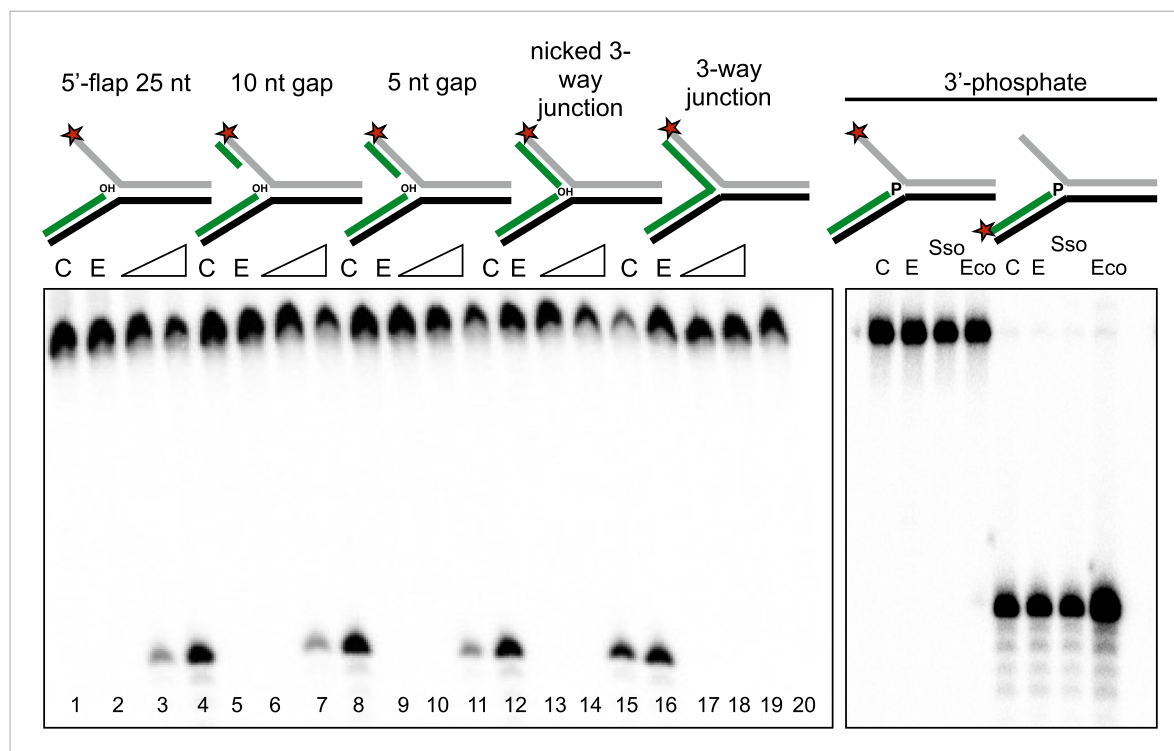


Figure 4. Importance of flap and 3' terminus structure. The importance of the released 25 nt 5'-flap structure was investigated by varying the length of duplex DNA in that arm from 0 (canonical 5'-flap) to a full 25 bp (nicked 3-way junction) (left hand panel, all based on substrate 19). All supported robust disintegration activity by SsoCas1. An intact Y-junction did not support TES activity. Lanes: C, substrate alone (1, 5, 9, 13, 17); E, SsoCas1 E142A variant 30 min incubation (2, 6, 10, 14, 18); incubation with wild-type SsoCas1 for 10 and 30 min (other lanes). The right hand panel shows the effect of replacing the attacking 3'-hydroxyl moiety at the branch point with a phosphate group (3' phos substrate) no TES or nuclease activity was observed for either Sso or EcoCas1. C, substrate alone; E, SsoCas1 E142A variant.

DOI: [10.7554/eLife.08716.007](https://doi.org/10.7554/eLife.08716.007)

confirming the qualitative observation of a preference for guanine, followed by cytosine, adenine and thymine, with rates of 0.06, 0.013, 0.0058 and 0.0009 min^{-1} , respectively (**Figure 5B**).

For *E. coli*, the first nucleotide of the repeat is a guanine. Although the corresponding position at the other end of the repeat on the minus strand is a cytosine, it has been demonstrated that the new spacer joins at the penultimate residue, which is also a guanine (**Swarts et al., 2012**). EcoCas1 displayed a striking preference for a guanine at the +1 position for the disintegration reaction, with all three alternative nucleotides strongly disfavoured at this position (**Figure 5C**), in good agreement with the prediction based on the repeat sequence. For EcoCas1 the reaction did not go to completion and accordingly we fitted the data with a variable end-point as described in the 'Materials and methods' (**Figure 5D**). Although the reaction rates could not be determined accurately, guanine at position +1 supported rates at least 10-fold higher than any other nucleotide. We also tested the effect of inclusion of Cas2 on the sequence specificity of EcoCas1, and observed that Cas2 had no effect, with strong preference for a guanine at +1 still observed (**Figure 5E**).

The -1 position

We proceeded to investigate the sequence dependence of the nucleotide at the 3'-end of the attacking DNA strand (the -1 position) in the disintegration reactions. For the first integration site, this should correspond to the last nucleotide of the leader sequence, which is an adenine in both *S. solfataricus* and *E. coli*. Although both Cas1 enzymes catalyzed the disintegration of substrates with an adenine at this position, clear sequence preference was not apparent (**Figure 6**). For *S. solfataricus*, the -1' position on the minus strand is variable in sequence. However, in *E. coli*, site 2 occurs at the penultimate nucleotide of the repeat and therefore has a cytosine at the -1' position (**Swarts et al., 2012**). In this situation, the incoming 3' terminal nucleotide of the spacer has to be a cytosine in order

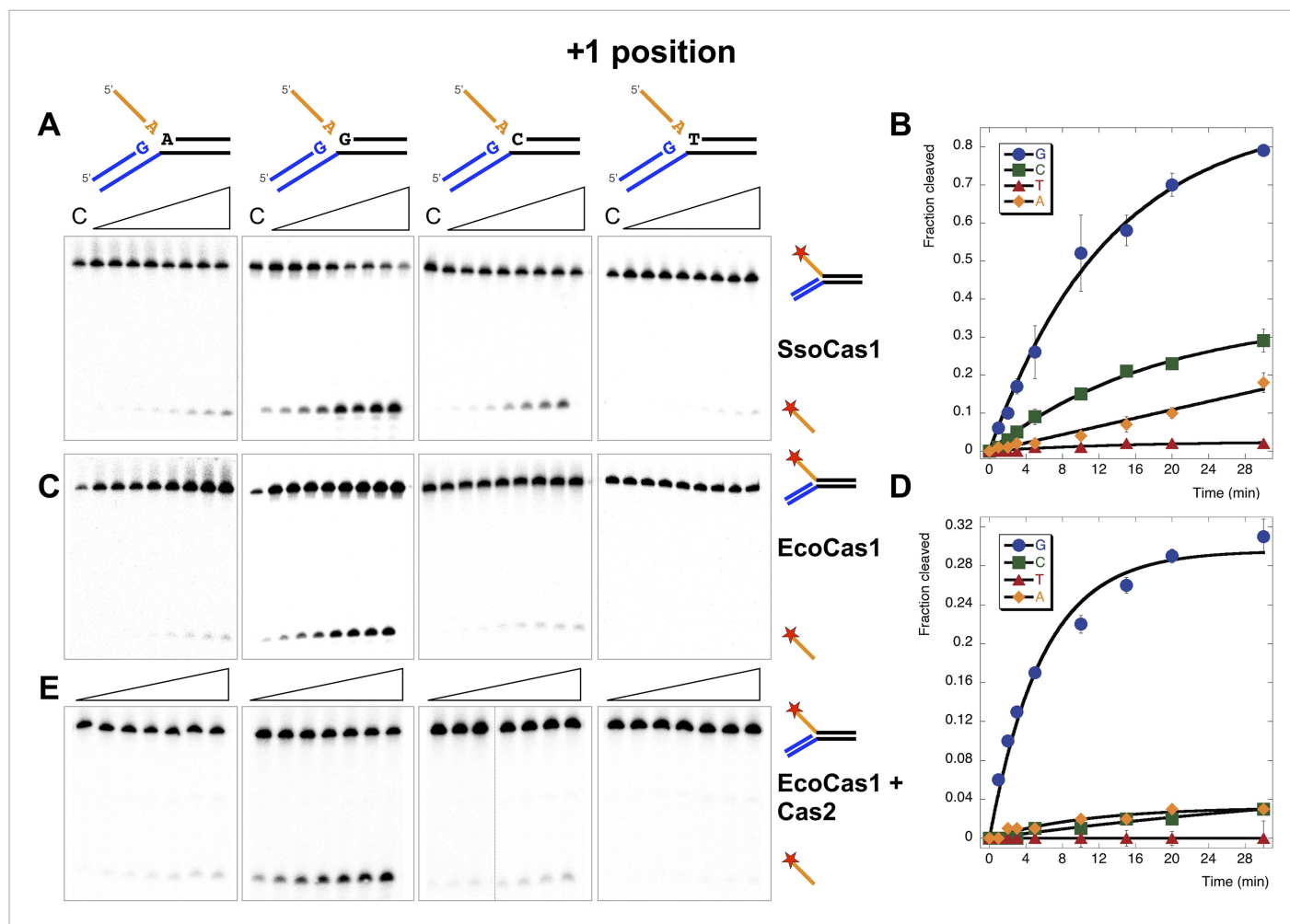


Figure 5. Sequence specificity of the disintegration reaction at the +1 position. The nucleotide at the acceptor (+1) position was varied systematically to assess the sequence dependence of the disintegration reaction carried out by Cas1 from *S. solfataricus* (A, B) and *E. coli* (C, D) (Substrates 3, 6, 7, 8). In the gels on the left (A, C) each substrate was incubated with Cas1 for 1, 2, 3, 5, 10, 15, 20 and 30 min in reaction buffer prior to electrophoresis to separate the cleaved 5'-flap from the intact substrate. C—control with no Cas1 added. The plots on the right (B, D) show quantification of these assays. Data points represent the means of triplicate experiments with standard errors shown (raw data provided in [Figure 5—source data 1](#) and [Figure 5—source data 2](#)). The data were fitted to an exponential equation, as described in the 'Materials and methods', and for EcoCas1 a variable floating end point was included to allow fitting as the reaction did not go to completion. The effect of Cas2 (150 nM) on EcoCas1 (150 nM) sequence specificity for substrates (50 nM) varying at position +1 (substrates 3, 6, 7, 8) was also tested (E). The second panel from the right is a composite image from two phosphorimages of the same time course as indicated by a black line.

DOI: [10.7554/eLife.08716.008](https://doi.org/10.7554/eLife.08716.008)

The following source data are available for figure 5:

Source data 1. Nucleotide at +1 position.

DOI: [10.7554/eLife.08716.009](https://doi.org/10.7554/eLife.08716.009)

Source data 2. Nucleotide at +1 position.

DOI: [10.7554/eLife.08716.010](https://doi.org/10.7554/eLife.08716.010)

to complete the original repeat sequence. To mimic the TES substrate at this site more closely, we tested substrates where the first nucleotide of the 5' flap, equivalent to the incoming nucleotide in the forward reaction, was a cytosine, but a cytosine at position -1 was still not favoured by EcoCas1 ([Figure 6C](#)). This may suggest that the disintegration reaction is not a good model for integration at site 2, which is further discussed later.

The -2 position

The nucleotide at the -2 position, which is part of the conserved leader sequence for integration site 1, is also a potential determinant of integration specificity. Accordingly, we varied this residue

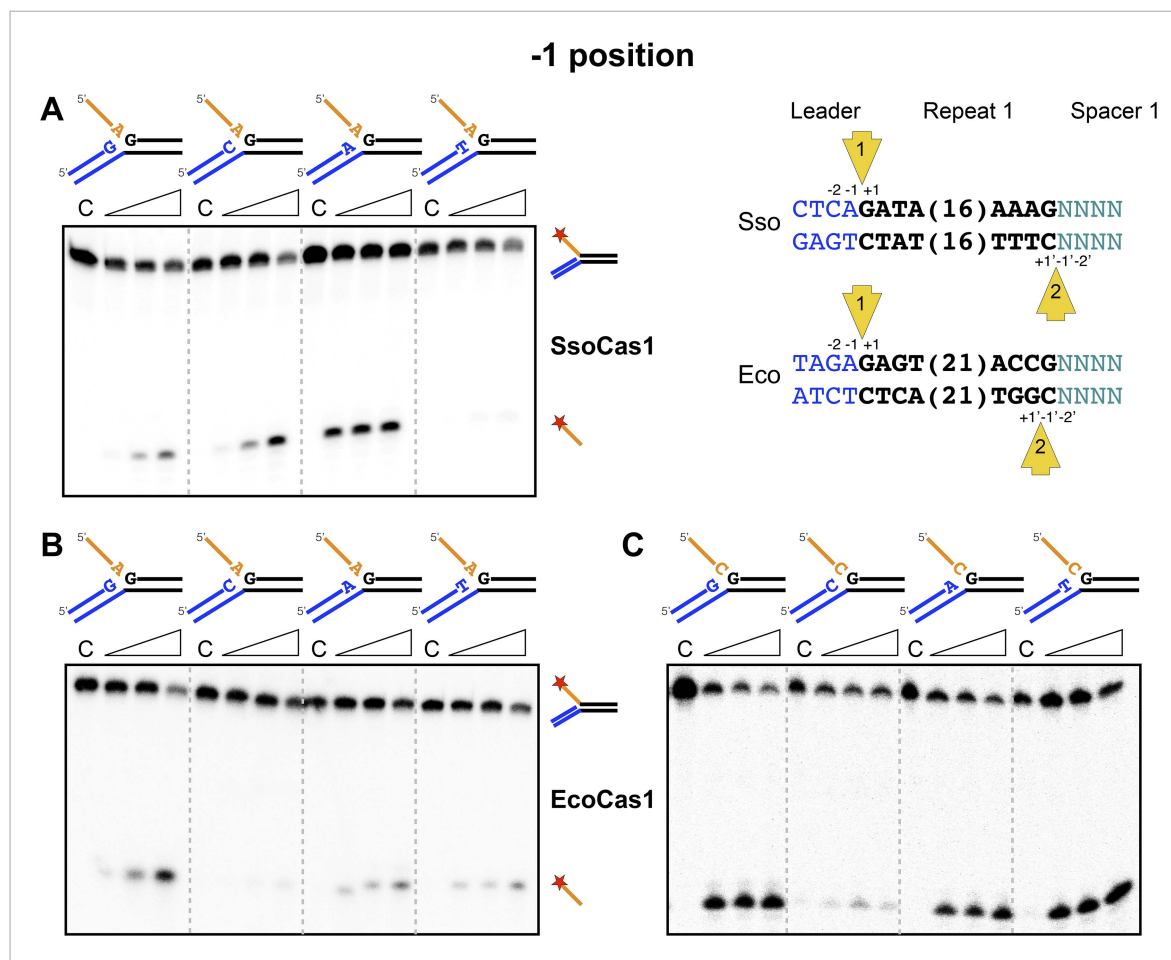


Figure 6. Sequence specificity of the disintegration reaction at the -1 position. The nucleotides participating in the disintegration reaction were varied systematically at the -1 position (substrates 3, 9, 10, 11). For SsoCas1 (**A**) there was some preference for adenine at this position, consistent with integration site 1. For EcoCas1 (**B**, **C**), a cytosine at position -1 was disfavoured over all other possibilities, even when the residue equivalent to the 'incoming' nucleotide was also a cytosine (substrates 15, 16, 17, 18). Each substrate was incubated with Cas1 for 5, 10 and 30 min in reaction buffer prior to electrophoresis. C—control with no Cas1 added.

DOI: 10.7554/eLife.08716.011

systematically and investigated the efficiency of the disintegration reaction for both Cas1 enzymes (**Figure 7**). In *S. solfataricus*, the -2 position in the leader is a cytosine, which supported the strongest disintegration activity (**Figure 7A**). In *E. coli*, the -2 position in the leader is a guanine. A clear preference for guanine over all other nucleotides was observed for EcoCas1 (**Figure 7B**), confirmed by a more detailed kinetic analysis (**Figure 7C**) which was fitted as for **Figure 5D**. These data are consistent with a role for sequence discrimination at the -2 position by both enzymes, which is relevant for integration site 1 but not site 2, where this position varies depending on the sequence of the last spacer inserted.

The incoming nucleotide

We next checked for specificity at the 3' end of the 5' flap in the disintegration product, which corresponds to the 3' end of the incoming spacer during integration. No sequence preference was detected for SsoCas1 (data not shown), which is consistent with the essentially random nature of the incoming DNA. During adaptation in *E. coli*, the incoming nucleotide at integration site 1 is expected to be random, but at site two is always a cytosine, where it completes the repeat sequence (**Swarts et al., 2012**). For EcoCas1, an adenine or cytosine was strongly favoured over guanine and particularly thymine (**Figure 8**), suggesting discrimination by EcoCas1 at this position.

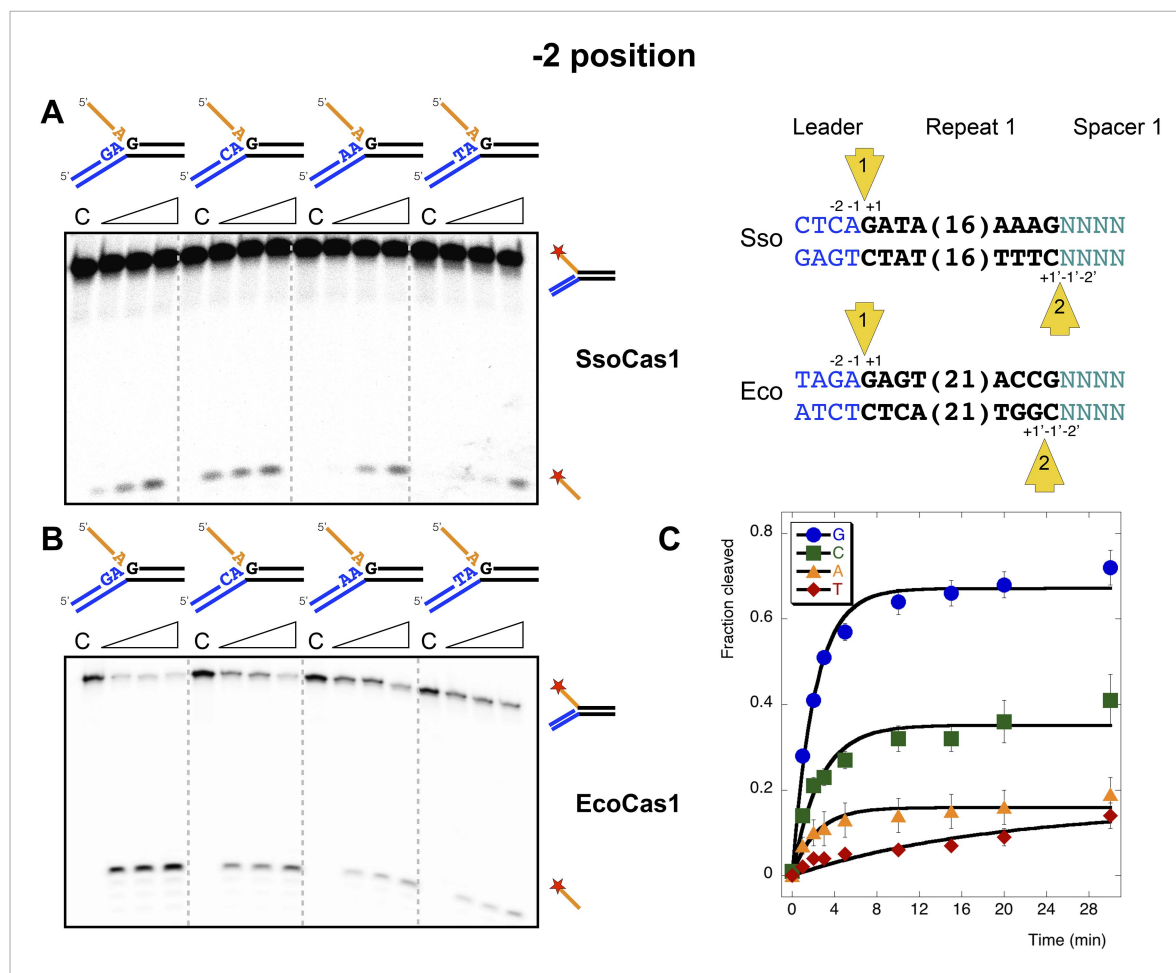


Figure 7. Sequence specificity of the disintegration reaction at the -2 position. The nucleotides participating in the disintegration reaction were varied systematically at the -2 position, which is a cytosine (Sso) or guanine (Eco) at integration site 1, and variable at integration site 2 (substrates 10, 12, 13, 14). (A) SsoCas1; (B) EcoCas1. Each substrate was incubated with Cas1 for 5, 10 and 30 min in reaction buffer prior to electrophoresis. C—control with no Cas1 added. (C) For EcoCas1, the clear preference for guanine at position -2 was confirmed by more detailed kinetic analysis (raw data provided in [Figure 7—source data 1](#)) as described for [Figure 5](#).

DOI: [10.7554/eLife.08716.012](https://doi.org/10.7554/eLife.08716.012)

The following source data is available for figure 7:

Source data 1. Nucleotide at -2 position.

DOI: [10.7554/eLife.08716.013](https://doi.org/10.7554/eLife.08716.013)

Disintegration of authentic *E. coli* integration intermediates

The substrates examined so far in this study do not correspond to the actual sequences encountered by EcoCas1 during integration. Accordingly, we constructed a pair of substrates that correspond to the products of integration when spacer 3 in the *E. coli* CRISPR array is integrated at site 1 (top strand) or site 2 (bottom strand) ([Figure 9](#)). These were constructed from oligonucleotides as before to generate disintegration substrates with a 5' flap. Disintegration was analysed by denaturing gel electrophoresis, phosphorimaging and quantification of triplicate experiments. EcoCas1 disintegrated the site 1 substrate quickly, with the reaction reaching approximately 75% conversion in 3 min, which compares favorably with the best model sequence studied. Site 2 was also a good disintegration substrate, though it was converted significantly more slowly than site 1, perhaps due to the presence of a disfavored cytosine at position -1 .

Discussion

Studies of the CRISPR spacer acquisition process in vivo have yielded many key insights, but they are complicated by the fact that it is very difficult to separate the two distinct steps of spacer capture and

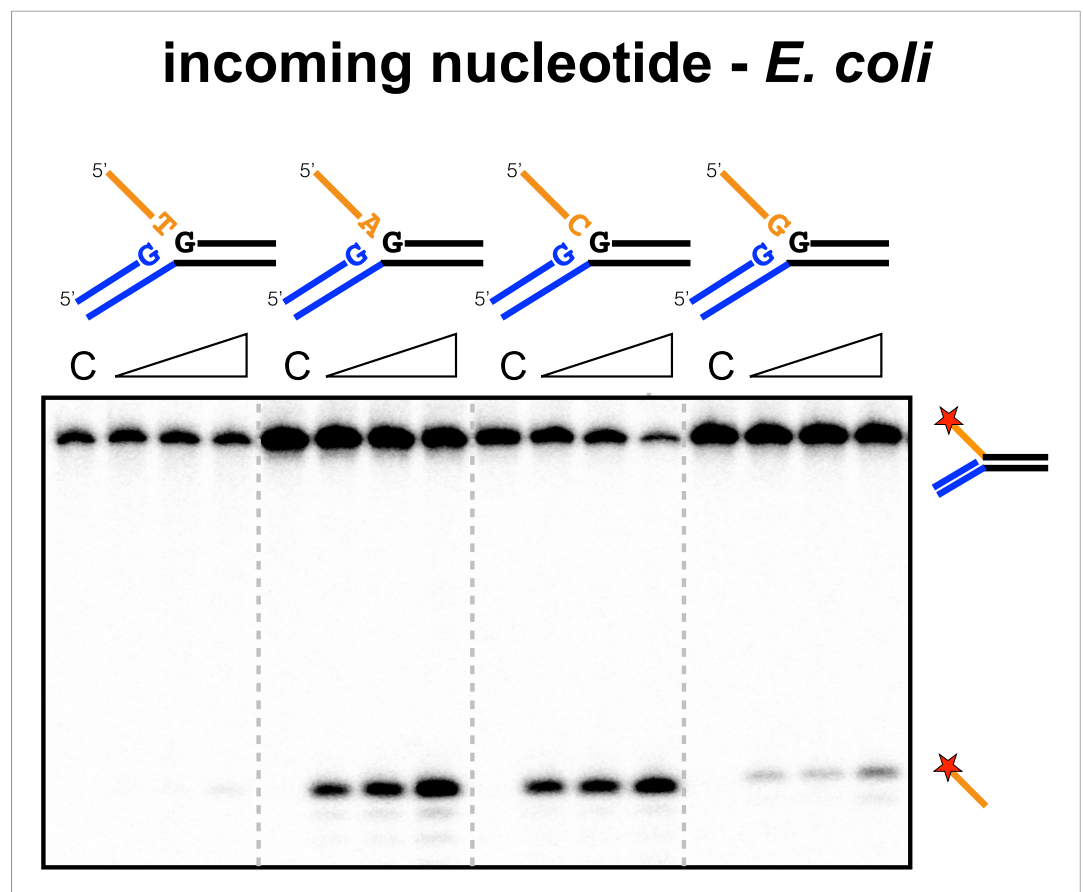


Figure 8. Sequence specificity of the EcoCas1 disintegration reaction for the incoming nucleotide. The nucleotide corresponding to the incoming 3' end of the new spacer, which is the nucleotide at the 3' end of the 5'-flap in the disintegration substrate, was varied systematically to determine its effect on the disintegration reaction catalysed by EcoCas1 (substrates 2, 3, 4, 5). C—control with no Cas1 added. Time points were 5, 10 and 30 min.
DOI: [10.7554/eLife.08716.014](https://doi.org/10.7554/eLife.08716.014)

spacer integration. Consequently we still do not have a clear understanding of the roles of Cas1, Cas2 and host proteins in the acquisition mechanism. In this study we investigated the integration reaction by focusing on the biochemistry of the Cas1 protein from *S. solfataricus* and *E. coli*. Efficient TES of branched DNA substrates with a 5'-flap or duplex arm is clearly possible for both *S. solfataricus* and *E. coli* Cas1 in vitro. This is a very precise reaction requiring attack by a 3'-hydroxyl at the branch point, generating a perfect DNA duplex. The reaction almost certainly represents the disintegration reaction that is the reverse of the spacer integration step, as observed for many integrases and transposases where it represents a very useful means to study the underlying integration mechanism (Gerton et al., 1999). Evidence for a disintegration activity was recently described by Doudna and colleagues for EcoCas1, but the activity observed was relatively weak, most likely because the branched substrate studied had a non-optimal DNA sequence around the branch point (Nuñez et al., 2015).

For the CRISPR adaptation process in vivo, integration occurs at the junction between the first repeat and the leader sequence, which immediately suggests a role for sequence specificity in the reaction. It has also been suggested that CRISPR repeat sequences, which are often palindromic, form four-way DNA junctions in supercoiled DNA, acting as a recognition signal for Cas1, a possibility that is supported by the observation that EcoCas1 can cut four-way DNA junctions in vitro (Babu et al., 2011), and the finding that spacer integration in a plasmid lacking a CRISPR locus occurs preferentially next to a palindromic site (Nuñez et al., 2015). However, a palindrome alone is not sufficient to support spacer insertion in *E. coli* in vivo (Arslan et al., 2014), and this also holds for the type II CRISPR system of *Streptococcus thermophilus* (Wei et al., 2015), suggesting that local sequence helps determine the integration site. Furthermore, some CRISPR repeat families, including many in

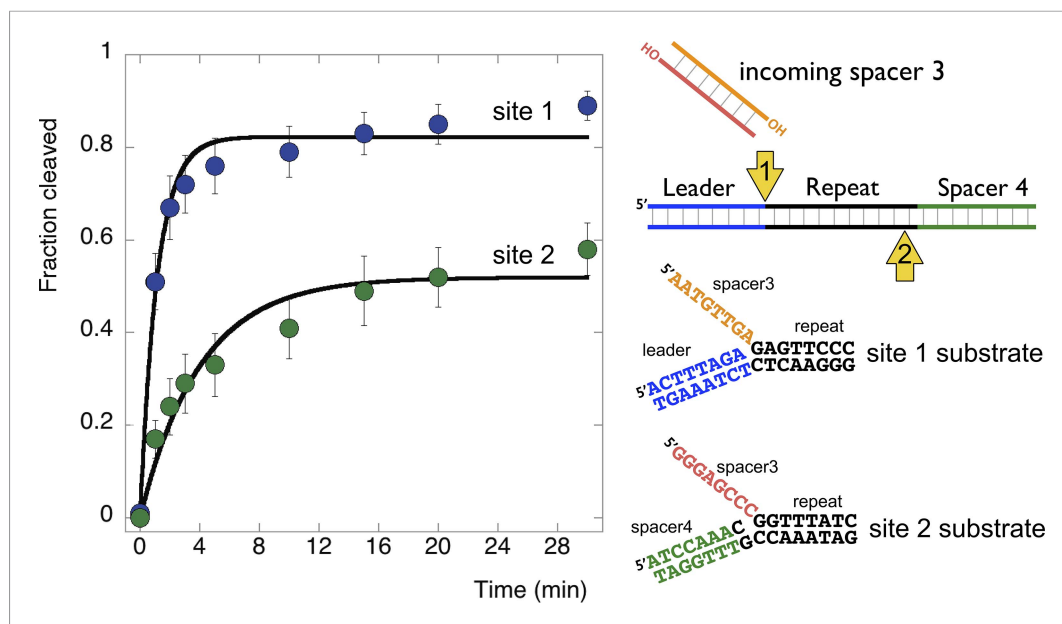


Figure 9. Disintegration of authentic *E. coli* integration intermediates. Disintegration substrates corresponding to the expected site 1 and site 2 integration products arising from the integration of spacer 3 into the CRISPR array were constructed and tested (spacer 3-1 and spacer 3-2 substrates). EcoCas1 processed both, with the rate of reaction significantly higher for the substrate corresponding to site 1 (the top strand) at the leader-repeat junction. Data points represent the means of triplicate experiments with standard errors shown (raw data provided in [Figure 9—source data 1](#)).

DOI: [10.7554/eLife.08716.015](https://doi.org/10.7554/eLife.08716.015)

The following source data is available for figure 9:

Source data 1. *E. coli* Site 1 vs Site 2 time course.

DOI: [10.7554/eLife.08716.016](https://doi.org/10.7554/eLife.08716.016)

archaea, have little or no palindromic nature and thus cannot form stable hairpin structures (Kunin *et al.*, 2007). Cas1 therefore might be expected to act as a sequence specific integrase, although local DNA structure could also play a part.

In support of this hypothesis, both *S. solfataricus* and *E. coli* Cas1 catalyse a disintegration reaction with distinct, sequence specific properties. In particular, there is a clear preference for a guanine at position +1, corresponding to the first nucleotide of the repeat, suggesting that this residue is recognised specifically in the active site of Cas1. The specificity is particularly strong for EcoCas1, consistent with the presence of a guanine in the repeat sequence at the +1 site in both plus and minus strands. Preference for a guanine at the +1 nucleotide for EcoCas1 catalysed integration events has also been observed (Nuñez *et al.*, 2015). For SsoCas1, a guanine at this position was preferred over a cytosine, which is the nucleotide present at the +1' position on the minus strand, by a factor of five. Although the -1 position might also be expected to play a role in the selection of integration sites, deep sequencing data for integration catalysed by EcoCas1 revealed no sequence preference at this position (Nuñez *et al.*, 2015). In agreement with this finding, we observed little evidence for sequence discrimination at the -1 position for the disintegration reactions catalysed by either enzyme, with the exception that EcoCas1 disfavors cytosine at this position (Figure 6B). A cytosine at position -1 is the expected residue on the minus strand, suggesting that the disintegration reaction may better reflect the reversal of integration at site 1 in the leader-repeat junction. Deep sequencing data for integration reactions catalysed by EcoCas1 in vitro did reveal a marked preference for a guanine at position -2 in the integration site (Nuñez *et al.*, 2015). This corresponds well with the -2 position in the plus strand, which is part of the leader sequence and is a guanine in *E. coli*, but cannot hold for the minus strand where the -2' position is inherently variable in nature. Disintegration of substrates mimicking the integration intermediates relevant for the integration of spacer 3 into the *E. coli* CRISPR array reinforce these conclusions, with site 1 on the plus strand processed significantly more quickly

than site 2 on the bottom strand (**Figure 9**). Taken together, both the disintegration specificity and the deep sequencing data for integration support the hypothesis that integration is targeted to the leader-repeat 1 junction on the plus strand at least in part by the inherent sequence specificity of Cas1, which presumably involves specific recognition of these bases within the active site of the enzyme (**Figure 10**).

For *E. coli* integration reactions in vitro, a marked preference for cytosine over thymine at the 3' end of the protospacer was observed (**Nuñez et al., 2015**). Furthermore, protospacers with a cytosine at the 3' end were preferentially incorporated into the minus strand at the junction between repeat 1 and spacer 1. These data are consistent with the requirement for protospacers to supply the final cytosine of the repeat on the minus strand during integration (**Swarts et al., 2012**). The deep sequencing data also revealed a marked preference for adenine over thymine at the 3' end of the protospacer (**Nuñez et al., 2015**). For disintegration by EcoCas1, we observed a clear preference for adenine or cytosine at the equivalent position, whilst thymine did not support the disintegration reaction (**Figure 8**). Thus EcoCas1 appears to recognise the nucleotide at the 3' end of the incoming DNA, although no such discrimination was observed for SsoCas1. A dinucleotide sequence 'AA' motif over-represented at the 3' end of protospacers incorporated in *E. coli* strain BL21 has been described previously (**Yosef et al., 2013**).

Although the CRISPR spacer integration system has been compared to the integration and transposition reactions carried out by mobile genetic elements, there is one key difference in the two processes—the length of DNA integrated. The persistence length of DNA, the distance over which it behaves as a fairly rigid rod, is estimated as 35–50 nm (100–150 bp) under conditions found in cells (**Hagerman, 1988; Brinkers et al., 2009**). This means that the two ends of a viral genome of several kilobases can be looped around and brought close together relatively easily, but a new spacer of 30–40 base pairs of dsDNA cannot be manipulated in the same way. Considering the scheme in **Figure 1**, the molecular origami required to achieve the second TES reaction looks challenging. Several related enzymes, including Mu transposase (**Savilahti et al., 1995**), V(D)J recombinase (**Ramsden et al., 1996**) and HIV integrase (**Gerton et al., 1999**) are known to disrupt base pairing of DNA substrates and make sequence-specific contacts during the integration reaction. It is likely that Cas1 also manipulates the local DNA duplex structure, which may help in positioning the DNA strands

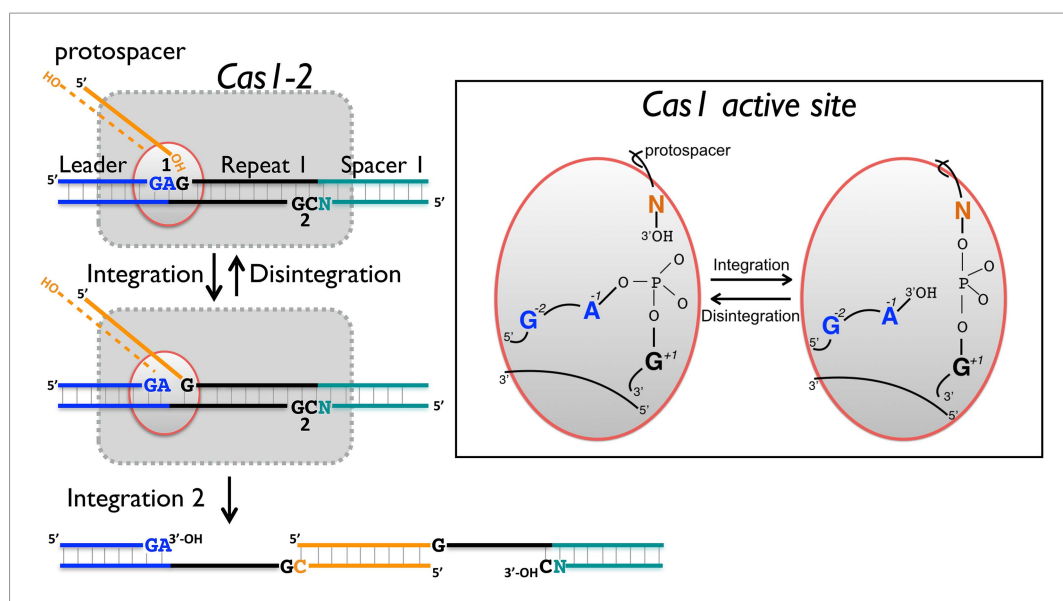


Figure 10. Reaction scheme for spacer integration and disintegration by *E. coli* Cas1. The Cas 1-2 complex integrates new spacers via two joining reactions (1 and 2) at either end of the first CRISPR repeat, which differ in their sequence context. Disintegration activity by *E. coli* Cas1 shows clear preference for the sequence at site 1, with the guanines at position +1 and -2 particularly important. At site 2, the sequence context is not optimal for disintegration in vitro, leading to slower reaction rates. In the active site of Cas1, these nucleobases likely make specific interactions with catalytic residues, and the DNA duplex structure may be distorted.

DOI: [10.7554/eLife.08716.017](https://doi.org/10.7554/eLife.08716.017)

correctly for the TES reaction. The observation that the incoming DNA (the 5'-flap in the disintegration reaction) can be single stranded, partially or fully duplex in nature suggests that there is some flexibility in the recognition of the incoming spacer. This is also consistent with the recent observation of a link between RecBCD, which generates ssDNA products, and Cas1 in *E. coli* (Levy *et al.*, 2015). There is no formal requirement that the protospacer should be fully duplex in nature, although current understanding of the integration reaction requires that new spacers have two intact 3'-ends for the two integration reactions so must be at least partially duplex in nature. Many integrases process the ends of the integrated DNA using a nuclease activity, which occurs at the same active site as the integrase activity (Gerton *et al.*, 1999). There is no reason to expect that Cas1 will differ in that regard, and indeed the reaction products of the RecBCD nuclease are on average much longer than the DNA molecules integrated by Cas1, suggesting the requirement for further processing.

In conclusion, we have shown that Cas1 from both *S. solfataricus* and *E. coli* have robust TES activities *in vitro* which reflect the reversal, or disintegration, of the integration reaction. Disintegration is strongly sequence specific, and the specificity fits with the expected sequence for the plus strand at the leader-repeat junction (Figure 9). This site is the logical place for the initiation of integration, as it has a unique, defined sequence, in contrast to the repetitive and more variable sequence found at the second integration site on the minus strand. Doudna and colleagues recently proposed a model based on an initial attack at site 2 on the minus strand (Nuñez *et al.*, 2015). However, this preference was only significant for spacers with a cytosine at the 3' terminus, and does not explain the marked preference observed by the authors for a guanine at position +2, which is a feature of the positive strand. Future studies of both the forward and reverse integration reactions catalyzed by Cas1 will help to address these issues and delineate the mechanism of spacer integration in the CRISPR system.

Materials and methods

Cloning, gene expression and protein purification

The *ssol1450* (*cas1*) gene and *ssol1450a* (*cas2*) genes were amplified from *S. solfataricus* P2 genomic DNA by PCR using primer pairs (5'-ATATAACCATGGCAAGCGTGAGGACTT; 5'-TATTGGATCCTCA CATCAACCAACTTGAACCC) and (5'-GCGCCATGGTTACACTAACCATTCCCTCTAATC; 5'-GGCCGG ATCCTTGAAATTATTGGTAGTATATGAC), respectively. The amplified genes were cloned into the pEHisTEV vector (Liu and Naismith, 2009) downstream of a cleavable His₆-tag using *Nco*I and *Bam*HI restriction sites. Site-directed mutagenesis was carried out on the vector containing *ssol1450* to mutate active site residue glutamic acid 142 to an alanine using the oligonucleotide sequence 5'-GTTGGATAAGGATGCACCGGCTGCTGCTAG. Standard site-directed mutagenesis protocols (QuikChange, Stratagene, La Jolla, CA, United States) were followed. Mutations were confirmed by sequencing (GATC Biotech, Constance, Germany). The constructs were expressed in C43 (DE3) *E. coli* cells grown in LB (Luria-Bertani) medium supplemented with 35 µg/ml kanamycin to an OD₆₀₀ of 0.6–0.8 at 37°C. Expression was then induced by the addition of 0.4 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and overnight incubation with shaking at 25°C. Cells were harvested (8000×g, 20 min) and resuspended in lysis buffer (4.5 mM NaH₂PO₄, 15 mM Na₂HPO₄ [pH 7.5], 500 mM NaCl, 30 mM imidazole, 1% Triton-X and protease inhibitors [Roche Applied Science, Basel, Switzerland]). Cells were lysed by sonication, the lysate cleared by ultracentrifugation (90,000×g, 35 min) and the supernatant filtered through a 0.22 µm syringe filter and loaded on to a HP HisTrap 5 ml column (GE Healthcare, Little Chalfont, United Kingdom) equilibrated in buffer A (4.5 mM NaH₂PO₄, 15 mM Na₂HPO₄ [pH 7.5], 500 mM NaCl, 30 mM imidazole). The His-tagged protein of interest was eluted with a linear gradient from 30 to 500 mM imidazole. Fractions containing Cas1 (or Cas2) were concentrated and buffer exchanged into buffer A, using centrifugal concentrators (30 kDa cutoff, Vivaspin, Sartorius Stedim Biotech GmbH, Goettingen, Germany). His-tags were cleaved by the addition of TEV protease (1:10 ratio of TEV:protein) and overnight incubation at room temperature. The cleaved protein was passed through a HisTrap column in buffer A, and the cleaved protein collected in the flow through. The final purification step was gel filtration on a 26/60 Superdex 200 prep grade column (GE Healthcare) in buffer C (20 mM Tris-HCL (pH 7.5), 500 mM NaCl, 0.5 mM DTT, 1 mM EDTA, 10% glycerol). Purified and concentrated protein samples were flash frozen and stored at –80°C.

Genes encoding *E. coli* K-12 MG1655 Cas1 (*ygbT*) and Cas2 (*ygbF*) were amplified by PCR from genomic DNA using the PCR primer pairs (5'-GGAATTCATATGACCTGGCTTCCCCTTAATC;

5'-GGAATTCTCAGCTACTCCGATGGCCTGC) and (5'-GGAATCCATATGAGTATGTTGGTCGTGGT CAC; 5'-GGAATTCTCAAACAGGTAAGAGACAC), respectively. Each PCR product was subcloned into protein expression plasmid pET14b using restriction enzyme *NdeI* and *EcoRI*. EcoCas1 and Cas2 proteins were over-produced individually in strain BL21 AI (Life Technologies, Carlsbad, CA, United States), each with N-terminal (His)₆ tags. Cells were grown at 37°C to OD₆₀₀ 0.5–0.6 in LB broth containing ampicillin (50 µg/ml) and induced using IPTG (1 mM) and arabinose (0.2% wt/vol), with growth continued for 3 hr after induction. Cas1 or Cas2 expressing cells were harvested for re-suspension in buffer H (20 mM Tris.HCl pH7.5, 500 mM NaCl, 5 mM imidazole, 10% glycerol) and flash frozen in liquid nitrogen for storage at –80°C until required. The first purification step was identical for both Cas1 and Cas2: sonicated biomass was clarified by centrifugation (90,000×g, 25 min) and soluble extract was passed into a 5 ml Hi-Trap Nickel chelating column (GE Healthcare) equilibrated with buffer H. Cas1 or Cas2 eluted at 70–100 mM imidazole in a linear imidazole gradient. Sodium chloride was reduced to 50 mM by dialysis against buffer S (20 mM Tris.HCl pH7.5, 50 mM NaCl, 1 mM DTT, 10% glycerol). Cas1 was loaded onto a 5 ml Hi-Trap heparin column and eluted in a gradient of NaCl at 200–300 mM in buffer S. Pooled Cas1 fractions were loaded directly onto a S300 size exclusion column equilibrated in buffer S with 250 mM NaCl. Cas1 fractions were pooled for storage at –80°C in buffer S containing 250 mM NaCl and 40% glycerol. Desalted Cas2 eluted from Ni-NTA was dialyzed into buffer S containing 1.5 M NaCl and applied to a 5 ml Hi-Trap butyl-Sepharose column (GE Healthcare), eluting in the flow through. Cas2 fractions were pooled and loaded directly onto a S300 size exclusion column equilibrated in buffer S with 250 mM NaCl. Following isocratic elution, Cas2 fractions were pooled and stored as for Cas1.

Sequence and preparation of DNA substrates

Substrates were purchased from Integrated DNA Technologies (Coralville, IA, United States) either unlabeled or with a 3'-fluorescein label ([Table 1](#)). Oligonucleotides were 5'-³²P-radiolabelled and gel purified as described previously ([Hutton et al., 2010](#)). Labelled oligonucleotides were annealed with complementary strands by heating with an excess of unlabelled strands at 95°C for 5 min and then slow cooling to room temperature overnight in a heating block. The assembled substrates ([Table 2](#)) were purified by native polyacrylamide (12%) gel electrophoresis with 1× Tris-borate-EDTA (TBE) buffer, followed by band excision, gel extraction and ethanol precipitation before being resuspended in nuclease free water, as described previously ([Hutton et al., 2010](#)). The final substrate concentration was measured using the extinction coefficient and absorbance at 260 nm in a UV-Vis spectrophotometer (Varian Cary, Agilent, Santa Clara, CA, United States) and DNA diluted to ~50 nM or ~200 nM final concentration for use in assays.

Disintegration reactions

Reactions were typically carried out under single turnover conditions. Titration of SsoCas1 ([Figure 3B,C](#)) showed evidence for inhibition at enzyme:substrate ratios above 10:1. For standard assays, 2 µM Cas1 protein was mixed with 200 nM substrate in cleavage buffer (20 mM Tris [pH 7.5], 10 mM NaCl, 1 mM DTT and 5 mM MnCl₂) and incubated at 55°C (for SsoCas1) or 37°C (for EcoCas1). For reactions with Cas1 and Cas2, the proteins were mixed in equimolar concentration and incubated together at either 37 or 55°C for 30 min before being added to the reaction. At indicated times, reactions were quenched by the addition of EDTA to 20 mM final concentration and 1 µl 20 mg/ml Proteinase K (Promega, Madison, WI, United States) and incubation at 37°C for 30 min. The DNA was then separated from the reaction by phenol chloroform extraction. 60 µl neutral phenol:chloroform:isoamyl alcohol (Sigma-Aldrich, St. Louis, MO, United States) was added and the reaction vortexed for 30 s. The sample was then centrifuged (15,000×g, 1 min) and the upper aqueous phase, containing the DNA, collected. Formamide loading dye (100% formamide with 0.25% bromophenol blue, 0.25% xylene cyanol) was added (5 µl) and the sample heated at 95°C for 2 min before being chilled on ice. Reaction products were resolved on a pre-run 20% denaturing (7 M urea) polyacrylamide gel containing 1× TBE in 1× TBE buffer. Gels were run at 80 W, 45°C for 90 min before overnight exposure to a phosphorimaging plate and imaging with a FLA-5000 Imaging System (Fujifilm Life Science, Düsseldorf, Germany).

SacI site repair

Assays with the *SacI* junction substrate were carried out under standard conditions with SsoCas1 for 30 min. FastDigest *SacI* enzyme (1 µl) and 1 µl FastDigest Buffer (Thermo Scientific, Waltham, MA,

Table 1. Sequence of oligonucleotides used for substrate construction

Strand	Sequence 5'→3'	Length
1a	TAGTAAGAGATTAATAAACCCCTCAGATAATCTCTTATAGAATTGAAAGTTCGG	53
1b	TTTTTTTTTTTTTTTTTATTATCTGAGGGTTTATTAATCTCTTACTA	48
1c	CCGAACCTTCAATTCTATAAGAG	23
2a	TAGTAAGAGATTAATAAACCCCTCAGATAACCTCTTATAGAATTGAAAGTTCGG	53
2b	TTTTTTTTTTTTTTTTTGTATCTGAGGGTTTATTAATCTCTTACTA	48
3b	TTTTTTTTTTTTTTTTTAGTTATCTGAGGGTTTATTAATCTCTTACTA	48
4b	TTTTTTTTTTTTTTTTTCGTTATCTGAGGGTTTATTAATCTCTTACTA	48
5b	TTTTTTTTTTTTTTTTTGGTTATCTGAGGGTTTATTAATCTCTTACTA	48
6a	TAGTAAGAGATTAATAAACCCCTCAGATAAGCTCTTATAGAATTGAAAGTTCGG	53
6b	TTTTTTTTTTTTTTTTTACTTATCTGAGGGTTTATTAATCTCTTACTA	48
7a	TAGTAAGAGATTAATAAACCCCTCAGATAAACTCTTATAGAATTGAAAGTTCGG	53
7b	TTTTTTTTTTTTTTTTTATTTATCTGAGGGTTTATTAATCTCTTACTA	48
8b	TTTTTTTTTTTTTTTTTAATTATCTGAGGGTTTATTAATCTCTTACTA	48
9a	TAGTAAGAGATTAATAAACCCCTCAGATAACATCTTATAGAATTGAAAGTTCGG	53
9c	CCGAACCTTCAATTCTATAAGAT	23
10a	TAGTAAGAGATTAATAAACCCCTCAGATAACTCTTATAGAATTGAAAGTTCGG	53
10c	CCGAACCTTCAATTCTATAAGAA	23
11a	TAGTAAGAGATTAATAAACCCCTCAGATAACGTCTTATAGAATTGAAAGTTCGG	53
11c	CCGAACCTTCAATTCTATAAGAC	23
12a	TAGTAAGAGATTAATAAACCCCTCAGATAACTCCTTATAGAATTGAAAGTTCGG	53
12c	CCGAACCTTCAATTCTATAAGGA	23
13a	TAGTAAGAGATTAATAAACCCCTCAGATAACTGCTTATAGAATTGAAAGTTCGG	53
13c	CCGAACCTTCAATTCTATAAGCA	23
14a	TAGTAAGAGATTAATAAACCCCTCAGATAACTACTTATAGAATTGAAAGTTCGG	53
14c	CCGAACCTTCAATTCTATAAGTA	23
SacI-a	TAGTAAGAGATTAATAAACCCCTCAGATGAGCTCTTATAGAATTGAAAGTTCGG	53
SacI-b	TTTTTTTTTTTTTCTCATCTGAGGGTTTATTAATCTCTTACTA	44
1b-3'-FAM	TTTTTTTTTTTTTATTATCTGAGGGTTTATTAATCTCTTACTA-FAM	48
19a	CCTCGAGGGATCCGTCCTAGCAAGCCGCTGCTACCGGAAGCTTCTGGACC	50
19b	GCTCGAGTCTAGACTGCAGTTGAGAGCTTGCTAGGACGGATCCCTCGAGG	50
19c	GGTCCAGAAGCTTCCGGTAGCAGCG	25
20d-10	AGTCTAGACTCGAGC	15
20d-5	ACTGCAGTCTAGACTCGAGC	20
20d	TCTCAACTGCAGTCTAGACTCGAGC	25
25c-d	GGTCCAGAAGCTTCCGGTAGCAGCGTCTCAACTGCAGTCTAGACTCGAGC	50
1c-3'P	CCGAACCTTCAATTCTATAAGAG-phos	25
Sp3-1a	CTGGCGCGGGAACTCTCTAAAAGTATACATTTGTTCTT	39
Sp3-1b	TGTAATTGATAATGTTGAGAGTTCCTCCGCGCCAG	34
Sp3-1c	AAGAACAATGTATACTTTTAGA	23
Sp3-2a	CCAGCGGGGATAAACCGTTTGGATCGGGTCTGGAATTC	39
Sp3-2b	TGTTCCGACAGGGAGCCCGTTTATCCCCGCTGG	34
Sp3-2c	GAAATCCAGACCCGATCCAAAC	23

DOI: [10.7554/eLife.08716.018](https://doi.org/10.7554/eLife.08716.018)

Table 2. DNA constructs used in this study

Substrate	Oligonucleotide components	Junction sequence			
		–2	–1	1	IC
Substrate 1	1a, 1b, 1c	A	G	A	T
Substrate 1-FAM	1a, 1b-3'-FAM, 1c	A	G	A	T
Sacl substrate	Sacl-a, Sacl-b, 1c	A	G	C	T
Substrate 2	2a, 2b, 1c	A	G	G	T
Substrate 3	2a, 3b, 1c	A	G	G	A
3'-phos substrate	2a, 3b, 1c-3'P	A	G	G	A
Substrate 4	2a, 4b, 1c	A	G	G	C
Substrate 5	2a, 5b, 1c	A	G	G	G
Substrate 6	6a, 6b, 1c	A	G	C	A
Substrate 7	7a, 7b, 1c	A	G	T	A
Substrate 8	1a, 8b, 1c	A	G	A	A
Substrate 9	9a, 3b, 9c	A	T	G	A
Substrate 10	10a, 3b, 10c	A	A	G	A
Substrate 11	11a, 3b, 11c	A	C	G	A
Substrate 12	12a, 3b, 12c	G	A	G	A
Substrate 13	13a, 3b, 13c	C	A	G	A
Substrate 14	14a, 3b, 14c	T	A	G	A
Substrate 15	2a, 4b, 1c	A	G	G	C
Substrate 16	11a, 4b, 11c	A	C	G	C
Substrate 17	10a, 4b, 10c	A	A	G	C
Substrate 18	9a, 4b, 9c	A	T	G	C
Substrate 19	19a, 19b, 19c	C	G	G	A
Gap10	19a, 19b, 19c, 20d-10	C	G	G	A
Gap5	19a, 19b, 19c, 20d-5	C	G	G	A
Nick	19a, 19b, 19c, 20d	C	G	G	A
Y-junction	19a, 19b, 20c-d	C	G	G	A
Spacer 3-1 substrate	Sp3-1a, Sp3-1b, Sp3-1c	G	A	G	A
Spacer 3-2 substrate	Sp3-2a, Sp3-2b, Sp3-2c	A	C	G	C

The sequence of the central portion of the junction (positions –2, –1, 1 and the incoming nucleotide (IC)) for each substrate is shown. The oligonucleotides used to assemble the complete substrate are indicated.

DOI: [10.7554/eLife.08716.019](https://doi.org/10.7554/eLife.08716.019)

United States) were then added and the reaction incubated at 37°C for 30 min. Product extraction, separation and visualization was then carried out as described above.

Disintegration reaction time courses

For the time course assays, the concentration of DNA substrates was 50 nM and the concentration of Cas1 protein 50 nM for SsoCas1 or 500 nM for EcoCas1. Reactions were performed as described above with the omission of the Proteinase K step. Following phosphorimaging, substrates and products were quantified using Image Gauge software (Fujifilm) and the reaction course was plotted using Kaleidagraph (Synergy Software, Reading, PA, United States). Experiments were carried out in triplicate and the mean and standard error calculated for each point. For SsoCas1, the data were fitted using a single exponential (Niewoehner *et al.*, 2014). For EcoCas1 the reactions did not go to completion and were therefore fitted with a floating end-point, as described previously (Niewoehner *et al.*, 2014).

Acknowledgements

This work was supported by a grant from the Biotechnology and Biological Sciences Research Council (REF: BB/M000400/1 to MFW). Thanks to Shirley Graham and Kotryna Temcinaite for technical support, and Jing Zhang and Agnes Tello for helpful discussions.

Additional information

Funding

Funder	Grant reference	Author
Biotechnology and Biological Sciences Research Council (BBSRC)	BB/M000400/1	Malcolm F White

The funder had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

CR, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; SS, Acquisition of data; ASB, Acquisition of data, Contributed unpublished essential data or reagents; ELB, Drafting or revising the article, Contributed unpublished essential data or reagents; MFW, Conception and design, Analysis and interpretation of data, Drafting or revising the article

Author ORCIDs

Malcolm F White,  <http://orcid.org/0000-0003-1543-9342>

References

- Arslan Z, Hermanns V, Wurm R, Wagner R, Pul Ü. 2014. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Research* **42**:7884–7893. doi: [10.1093/nar/gku510](https://doi.org/10.1093/nar/gku510).
- Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF. 2011. A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Molecular Microbiology* **79**:484–502. doi: [10.1111/j.1365-2958.2010.07465.x](https://doi.org/10.1111/j.1365-2958.2010.07465.x).
- Barrangou R, Marraffini LA. 2014. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Molecular Cell* **54**:234–244. doi: [10.1016/j.molcel.2014.03.011](https://doi.org/10.1016/j.molcel.2014.03.011).
- Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF. 2008. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *The Journal of Biological Chemistry* **283**:20361–20371. doi: [10.1074/jbc.M803225200](https://doi.org/10.1074/jbc.M803225200).
- Brinkers S, Dietrich HR, de Groote FH, Young IT, Rieger B. 2009. The persistence length of double stranded DNA determined using dark field tethered particle motion. *Journal of Chemical Physics* **130**:215105. doi: [10.1063/1.3142699](https://doi.org/10.1063/1.3142699).
- Chow SA, Vincent KA, Ellison V, Brown PO. 1992. Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science* **255**:723–726. doi: [10.1126/science.1738845](https://doi.org/10.1126/science.1738845).
- Delelis O, Carayon K, Saib A, Deprez E, Mouscadet JF. 2008. Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology* **5**:114. doi: [10.1186/1742-4690-5-114](https://doi.org/10.1186/1742-4690-5-114).
- Diez-Villasenor C, Guzman NM, Almendros C, Garcia-Martinez J, Mojica FJ. 2013. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biology* **10**:792–802. doi: [10.4161/rna.24023](https://doi.org/10.4161/rna.24023).
- Dillingham MS, Kowalczykowski SC. 2008. RecBCD enzyme and the repair of double-stranded DNA breaks. *Microbiology and Molecular Biology Reviews* **72**:642–671. Table of contents. doi: [10.1128/MMBR.00020-08](https://doi.org/10.1128/MMBR.00020-08).
- Fineran PC, Charpentier E. 2012. Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. *Virology* **434**:202–209. doi: [10.1016/j.virol.2012.10.003](https://doi.org/10.1016/j.virol.2012.10.003).
- Gerton JL, Herschlag D, Brown PO. 1999. Stereospecificity of reactions catalyzed by HIV-1 integrase. *The Journal of Biological Chemistry* **274**:33480–33487. doi: [10.1074/jbc.274.47.33480](https://doi.org/10.1074/jbc.274.47.33480).
- Goren M, Yosef I, Edgar R, Qimron U. 2012. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. *RNA biology* **9**:549–554. doi: [10.4161/rna.20177](https://doi.org/10.4161/rna.20177).
- Hagerman PJ. 1988. Flexibility of DNA. *Annual Review of Biophysics and Biophysical Chemistry* **17**:265–286. doi: [10.1146/annurev.bb.17.060188.001405](https://doi.org/10.1146/annurev.bb.17.060188.001405).
- Heler R, Marraffini LA, Bikard D. 2014. Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. *Molecular Microbiology* **93**:1–9. doi: [10.1111/mmi.12640](https://doi.org/10.1111/mmi.12640).

- Hutton RD**, Craggs TD, White MF, Penedo JC. 2010. PCNA and XPF cooperate to distort DNA substrates. *Nucleic Acids Research* **38**:1664–1675. doi: [10.1093/nar/gkp1104](https://doi.org/10.1093/nar/gkp1104).
- Kunin V**, Sorek R, Hugenholtz P. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology* **8**:R61. doi: [10.1186/gb-2007-8-4-r61](https://doi.org/10.1186/gb-2007-8-4-r61).
- Lemak S**, Beloglazova N, Nocek B, Skarina T, Flick R, Brown G, Popovic A, Joachimiak A, Savchenko A, Yakunin AF. 2013. Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from *Sulfolobus solfataricus*. *Journal of the American Chemical Society* **135**:17476–17487. doi: [10.1021/ja408729b](https://doi.org/10.1021/ja408729b).
- Levy A**, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**:505–510. doi: [10.1038/nature14302](https://doi.org/10.1038/nature14302).
- Liu H**, Naismith JH. 2009. A simple and efficient expression and purification system using two newly constructed vectors. *Protein Expression and Purification* **63**:102–111. doi: [10.1016/j.pep.2008.09.008](https://doi.org/10.1016/j.pep.2008.09.008).
- Makarova KS**, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* **1**:7. doi: [10.1186/1745-6150-1-7](https://doi.org/10.1186/1745-6150-1-7).
- Nam KH**, Ding F, Haitjema C, Huang Q, DeLisa MP, KE A. 2012. Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *The Journal of Biological Chemistry* **287**:35943–35952. doi: [10.1074/jbc.M112.382598](https://doi.org/10.1074/jbc.M112.382598).
- Niewoehner O**, Jinek M, Doudna JA. 2014. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Research* **42**:1341–1353. doi: [10.1093/nar/gkt922](https://doi.org/10.1093/nar/gkt922).
- Nuñez JK**, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. 2014. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature Structural & Molecular Biology* **21**:528–534. doi: [10.1038/nsmb.2820](https://doi.org/10.1038/nsmb.2820).
- Nuñez JK**, Lee AS, Engelman A, Doudna JA. 2015. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**:193–198. doi: [10.1038/nature14237](https://doi.org/10.1038/nature14237).
- Ramsden DA**, McBlane JF, van Gent DC, Gellert M. 1996. Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. *The EMBO Journal* **15**:3197–3206.
- Savilahti H**, Rice PA, Mizuuchi K. 1995. The phage Mu transpososome core: DNA requirements for assembly and function. *The EMBO Journal* **14**:4893–4903.
- Shmakov S**, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K. 2014. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Research* **42**:5907–5916. doi: [10.1093/nar/gku226](https://doi.org/10.1093/nar/gku226).
- Sorek R**, Lawrence CM, Wiedenheft B. 2013. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annual Review of Biochemistry* **82**:237–266. doi: [10.1146/annurev-biochem-072911-172315](https://doi.org/10.1146/annurev-biochem-072911-172315).
- Swarts DC**, Mosterd C, van Passel MW, Brouns SJ. 2012. CRISPR interference directs strand specific spacer acquisition. *PLOS ONE* **7**:e35888. doi: [10.1371/journal.pone.0035888](https://doi.org/10.1371/journal.pone.0035888).
- van der Oost J**, Westra ER, Jackson RN, Wiedenheft B. 2014. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Reviews. Microbiology* **12**:479–492. doi: [10.1038/nrmicro3279](https://doi.org/10.1038/nrmicro3279).
- Wei Y**, Terns RM, Terns MP. 2015. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes & Development* **29**:356–361. doi: [10.1101/gad.257550.114](https://doi.org/10.1101/gad.257550.114).
- Wiedenheft B**, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA. 2009. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**:904–912. doi: [10.1016/j.str.2009.03.019](https://doi.org/10.1016/j.str.2009.03.019).
- Yosef I**, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Research* **40**:5569–5576. doi: [10.1093/nar/gks216](https://doi.org/10.1093/nar/gks216).
- Yosef I**, Shitrit D, Goren MG, Burstein D, Pupko T, Qimron U. 2013. DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proceedings of the National Academy of Sciences of USA* **110**:14396–14401. doi: [10.1073/pnas.1300108110](https://doi.org/10.1073/pnas.1300108110).
- Zhang J**, Kasciukovic T, White MF. 2012. The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. *PLOS ONE* **7**:e47232. doi: [10.1371/journal.pone.0047232](https://doi.org/10.1371/journal.pone.0047232).