UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DETAILED ANALYSIS OF PROTEIN-DNA INTERACTIONS DRIVING TYPE II-A CRISPR ADAPTATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

MASON J. VAN ORDEN

Norman, Oklahoma

2020

DETAILED ANALYSIS OF PROTEIN-DNA INTERACTIONS DRIVING TYPE II-A CRISPR ADAPTATION

A DISSERTATION APPROVED FOR THE DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY THE COMMITTEE CONSISTING OF

Dr. Rakhi Rajan, Chair

Dr. Ann West

Dr. Elizabeth Karr

Dr. Elena Zgurskaya

Dr. Christina Bourne

Acknowledgements

I would like to thank the University of Oklahoma Department of Chemistry and Biochemistry for accepting me to their program and giving me this opportunity to learn and grow as a scientist for the past 5 years. I would like to thank Dr. Rakhi Rajan for allowing me to be a part of her laboratory, having confidence in me, and setting an example of hard work and perseverance. I thank my committee members for the support, advice, and constructive criticism that was necessary to get me to this point. I would also like to thank my fellow lab members over the years who have helped run an efficient, friendly lab environment that was easy to come to every day. This includes Dr. S.D. Yogesha, Dr. Peter Klein, Dr. Ramya Sundaresan, Dr. Kesavan Babu, Hari Priya Parameshwaran, and Sydney Newsom, as well as many other graduate and undergraduate students that I worked with over the years.

Most importantly I would like to thank my wife Clarissa and my daughter Layla for supporting me through this process. I would never have been able to accomplish this without their love and support.

Other direct contributions to specific projects are acknowledged at the beginning of each chapter.

Table of Contents

Acknowledgements iv

List of Abbreviations vii

Abstract viii

Chapter 1: CRISPR Adaptation - Background 1

1.1.0 - Discovery of CRISPR-Cas systems 1

1.2.0 - Mechanism of phage inactivation by CRISPR 2

1.3.0 - Classification of CRISPR-Cas systems 3

1.4.0 - Core components of CRISPR adaptation 7

1.5.0 - Sources for new spacers 9

1.6.0 - Role of the protospacer adjacent motif in adaptation 10

1.7.0 - Prespacer processing 11

1.8.0 - Integration site recognition 12

1.9.0 - Integration reaction mechanism 13

1.10.0 - Roles of accessory proteins in CRISPR adaptation 13

1.11.0 - Anti-CRISPR proteins found in phages 16

1.12.0 - Adaptation in type II-A systems 17

1.13.0 - Hypotheses 18

1.14.0 - Significance 19

Chapter 2: Conserved DNA motifs in the type II-A CRISPR leader region 20

2.1.0 - Acknowledgements 20

2.2.0 - Copyright Information 20

2.3.0 - Abstract 20

2.4.0 - Results 21

2.5.0 - Discussion 28

2.6.0 - Conclusion 35

Chapter 3: CRISPR type II-A subgroups exhibit phylogenetically distinct mechanisms for prespacer insertion 36

3.1.0 - Acknowledgements 36

3.2.0 - Copyright Information 36

3.3.0 - Abstract 36

3.4.0 - Results 38

3.5.0 - Discussion 53
Chapter 4: Biophysical characterization of CRISPR adaptation proteins 58
4.1.0 - Acknowledgments 58
4.2.0 - Biophysical characteristics of Cas1-Cas2 58
4.3.0 - Complex formation analysis of various Cas1-Cas2 complexes using SEC-MALS 61
4.4.0 - Structural studies of G3 Cas1 65
Chapter 5: Outlook 73
Methods 76
6.1.0 – Copyright Information 76
6.2.0 - Processing of genomic data 76
6.3.0 - Sequence Alignment 78
6.4.0 - Target and prespacer design 78
6.5.0 - Cloning of cas genes 80
6.6.0 - Protein expression and purification 80
6.7.0 - Integration assays 82
6.8.0 - Analysis of protein complex formation 83
6.9.0 – Crystallization setup and freezing 84
6.10.0 – Crystal Data Analysis 84
References 85
Supporting Tables 93
Supporting Table S1 93
Supporting Table S2 94
Supporting Table S3 99
Supporting Table S4 101
Supporting Table S5 – Nucleotides Used 102
Supporting Figures 111

List of Abbreviations

Acr	anti-crispr
bp	base pair
Cas	CRISPR associated
Cascade	CRISPR associated complex for antiviral defense
CRISPR	clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
E	E. coli target
G1	Group 1
G2	Group 2
G3	Group 3
HP	hairpin
IC	integration complex
IHF	integration host factor
IR	inverted repeat
L	linear
LAS	leader anchoring site
LRH	leader recognition helix
nt	nucleotide
PAM	protospacer adjacent motif
sgRNA	small guide RNA
tracr	trans activating CRISPR

Abstract

CRISPR-Cas is an adaptive immune system that protects prokaryotes against foreign nucleic acids. Prokaryotes gain immunity by acquiring short pieces of the invading nucleic acid, termed prespacers, and inserting them into their CRISPR array using the proteins Cas1 and Cas2. Immediately preceding the CRISPR array is the CRISPR leader region, and prespacers are generally inserted where the leader region meets the CRISPR array (leader-repeat junction). Here, a detailed analysis of the bioinformatic, biochemical, and biophysical characteristics of the DNA and protein elements that govern this site-specific insertion of prespacers is presented. Various sequences of leader-repeat junctions were first analyzed belonging to type II-A, a sub-type of CRISPR systems. Phylogenetic clustering of leader-repeat junctions defined three distinct groups with conserved sequences, G1 with ATTTGAG, G2 with CTRCGAG, and G3 with NNNNNCG. The sequence alignment data showed phylogenetic clustering of Cas proteins and repeat sequences in type II-A systems that mirrored the clustering of leader-repeat junctions. Biochemical characterization of representative Cas1 and Cas2 proteins from each group showed distinct mechanisms in leader-repeat junction recognition and in prespacer insertion. G1 first recognized a 12-bp sequence at the leader–repeat junction and performed leader-side insertion before proceeding to spacer-side insertion. G2 recognized the full repeat sequence and could perform independent leader-side or spacer-side insertions, although the leader-side insertion was faster than spacer-side. G3 showed no sequence specific insertion activity. Protein-DNA complex formation analysis by direct molar mass measurement showed that all three protein complexes form the canonical $Cas1_4$-$Cas2_2$-$prespacer_1$ complex, with the morphology of the prespacer being an essential factor promoting complex formation, at least in the case of G1. These results highlight the intricacy of protein–DNA sequence interactions within the seemingly similar type II-A integration complexes and provide mechanistic insights into prespacer insertion. These insights provide valuable information for the development of a Cas1–Cas2 toolset for inserting small DNAs into diverse DNA targets.

Chapter 1: CRISPR Adaptation - Background

1.1.0 - Discovery of CRISPR-Cas systems

Prokaryotic organisms are in a constant battle against infection from phages. Over time, bacteria have evolved both innate and adaptive immune systems to fight off these infections. Innate immunity, or immunity that a prokaryote starts life with, is generally a prokaryotic cells first response to an infection (4). A variety of mechanisms have been identified as innate immune responses, including receptor mutation, restriction modification systems, and abortive infection (5). These immune responses are genetically programmed and provide protection throughout the cell's lifespan (5). Adaptive immunity, or immune responses that are acquired after the start of life, also plays an important role in the defense against phages (6, 7). In prokaryotes, the only adaptive immune system known to date is comprised of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) proteins (6). A CRISPR system is constituted by a specific region of the prokaryotic genome containing an array of short repetitive sequences called the repeats (usually about 30-40 base pairs (bp)) which are interspaced by short unique sequences (30-40 bp) termed spacers (8). This repeat-spacer array is the defining characteristic of a CRISPR system and is consequently referred to as the CRISPR array. Near the CRISPR array, genes encoding Cas proteins and a promoter containing region termed

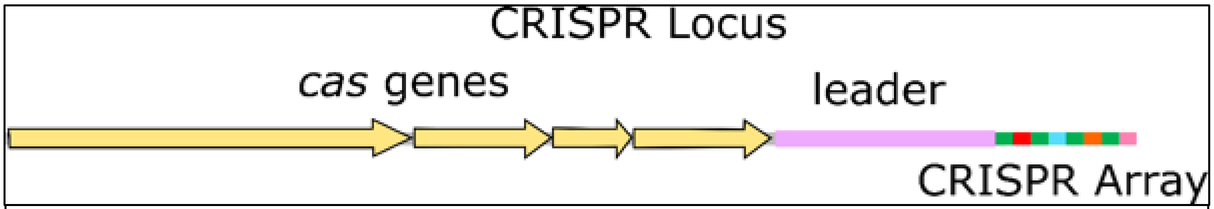


Figure 1: CRISPR locus organization. The CRISPR locus is in the prokaryotic genome and generally consists of *cas* genes, the leader region, and the CRISPR array which contains all the repeats and spacers.

the leader region define the entire CRISPR locus (**Figure 1**). CRISPR systems were first discovered by bioinformaticians who noticed repetitive sequences in prokaryotic genomes (9-13). For years these repetitive sequences went by different names and had unknown functions until in 2002 the name CRISPR was first applied to the repeat-spacer structure (8). In 2005 , bioinformatics studies showed that spacer sequences were derived from foreign DNA, specifically phage DNA (14, 15). This led to the discovery that prokaryotes were able to acquire resistance to phages by gaining a spacer in their CRISPR array which matched the infecting phage (6, 7). New spacers were shown to be acquired near the leader side of the CRISPR array. This polarized integration of new spacers in the CRISPR array creates a chronological memory of past infections, with the newest spacers near the leader region. This order of new spacers is important biologically to prokaryotes as spacers nearest the leader are more expressed, giving the organism the ability to fight recent infections more potently.

1.2.0 - Mechanism of phage inactivation by CRISPR

CRISPR systems can provide immunity to prokaryotes by a three-step process: adaptation, CRISPR RNA processing, and interference (16). Upon phage infection, a short piece of the phage DNA must be excised and integrated as a new spacer in the CRISPR array. This process is called adaptation. Once the new spacer is integrated into the CRISPR array, the array is transcribed into a long piece of RNA containing all the repeats and spacers (7). This long RNA molecule, called pre-CRISPR RNA, is separated into single repeat-spacer segments, either by Cas proteins or other cellular RNAses, which are then termed CRISPR RNAs (crRNAs) (17, 18). This process of crRNA processing is also referred to as crRNA biogenesis. Once crRNAs are formed, they can associate with Cas proteins and serve as guides to aid in targeting nucleic acid segments with complementary sequences to the spacer region of the crRNA (6, 7). The targeting and inactivation of these complementary sequences (usually through a double-stranded break or nucleic acid degradation) is termed interference (**Figure 2**).

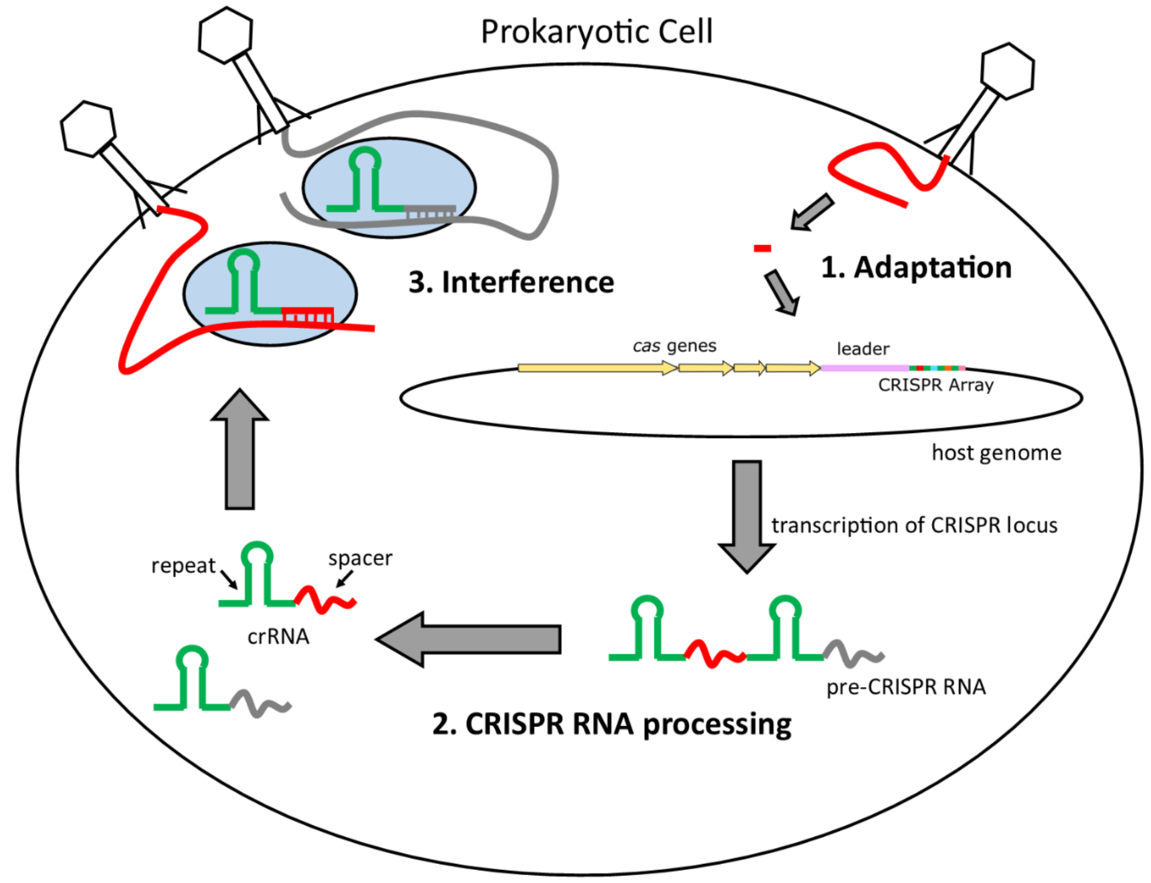


Figure 2: CRISPR immunity. CRISPR systems achieve immunity through adaptation, CRISPR RNA processing, and interference. The CRISPR locus contains genes encoding Cas proteins and a repeat-spacer array, along with a CRISPR leader region.

1.3.0 - Classification of CRISPR-Cas systems

CRISPR systems are classified into classes and types based on the *cas* gene content of the CRISPR locus and the Cas nuclease responsible for the interference step (**Figure 3**). Six different types of CRISPR systems are divided between Class 1 and Class 2, with type I, III, and IV belonging to Class 1 and type II, V, and VI belonging to

Class 2. These types are even further classified into over 40 different sub-types based on the Cas protein content of each CRISPR locus (19).

1.3.1 - Class 1 CRISPR systems

Class I CRISPR systems have an interference complex containing several individual Cas proteins, often containing more than a single copy of certain components. Type I and III interference complexes have been well characterized both biochemically and structurally (20, 21), while type IV remains not well understood. Type I interference complexes are referred to as Cascade (CRISPR associated complex for antiviral defense) complexes (22). Cascade is composed of one copy each of Cas6, Cas5 and Cas8 along with two copies of Cas 11 and six copies of Cas7 (23). Cascade associates with a single crRNA, which guides the complex to the target sequence. Once bound to the target sequence, Cas3, the signature protein for type I systems, is recruited and subsequently degrades the target DNA (21). Type III CRISPR systems are unique in that they target both DNA and RNA (20). Recent work has uncovered that the type III interference complex guided by crRNA binds to complementary RNA sequences (24).

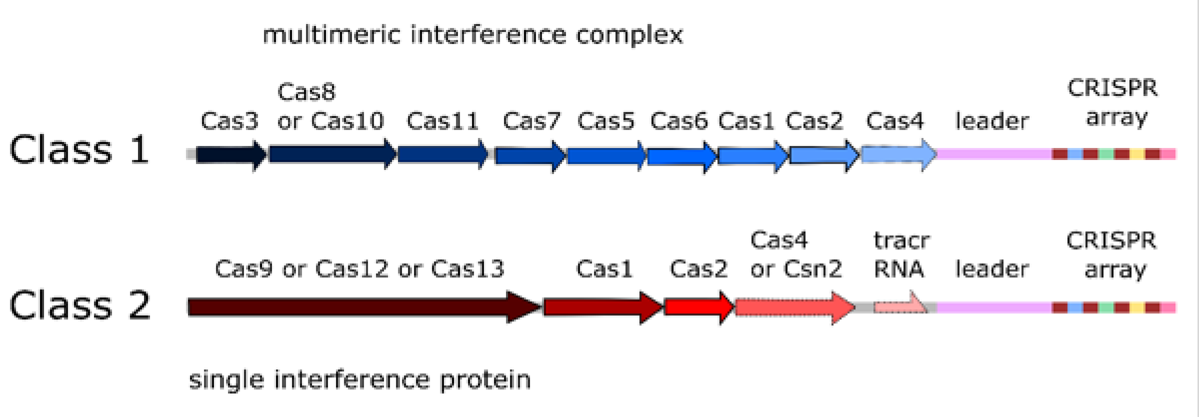


Figure 3: CRISPR Classification. CRISPR systems are classified into two classes and six types. Classes are defined by the presence of a multimeric interference complex composed of multiple Cas proteins or by a single protein responsible for interference.

There are mainly two type III interference complexes, the Csm and Cmr complexes, that are distributed among several subtypes (A through F). They are differentiated by the presence of certain small subunits of the multimeric complex related to the Cas7, either Csm2 (III-A) or Cmr5 (III-B) (25), as well as the presence of an adaptation module of Cas1 and Cas2 in the CRISPR locus. Both Csm and Cmr interference complexes bind to the gene transcript during transcription based on complementarity provided by the crRNA, followed by Cas7-medaited cleavage of the RNA transcript. In addition, type III systems are unique since they have the ability to cleave the DNA component of the transcription bubble using Cas10, another protein component (26-28). Both activities are required for protecting the host (29, 30).

1.3.2 - Class 2 CRISPR systems

Class 2 systems employ a single, multi-domain Cas nuclease to carry out the interference step, which replaces the multiple protein components that is the hallmark of Class 1 systems (**Figure 3**). Types II, V, and VI all use different single effector proteins to target and degrade foreign nucleic acid (19). Cas9, the effector nuclease from type II systems, is the most well-known and studied protein from CRISPR systems. Cas9 requires both a crRNA and a trans activating crRNA (tracr) to successfully target DNA. TracrRNA associates with Cas9 directly and contains a region complementary to the repeat region of crRNA. This complementarity allows the formation of a ternary complex between Cas9, tracrRNA, and crRNA. Studies have shown that fusing the crRNA and tracrRNA through a short loop can create a single guide RNA (sgRNA) that is equally effective as the individual components in targeting and cleaving DNA (31). In 2012, a landmark study showed that Cas9-tracrRNA-crRNA complexes could be programmed to target any location by simply changing the crRNA spacer sequence (31). This discovery has led to the widespread availability of gene editing, with new improvements and possibilities to gene editing being explored at a fast pace. Scientist have recently been able to harness the targeting and nicking power of Cas9 fused to a reverse transcriptase enzyme in a system called Prime editing that will edit genes without the need of double stranded breaks (32). Type V systems employ Cas12, which is mechanistically different from Cas9 in many ways, most notably that Cas9 creates blunt double stranded breaks using two nuclease domains, while Cas12 creates staggered double stranded breaks using a single nuclease domain (33). Certain Cas12 proteins, notably Cas12a, also has no need for a tracrRNA, and only uses a crRNA guide to target and cleave DNA (33). Type VI systems use Cas13, which is an RNA targeting enzyme that utilizes a crRNA as a guide and three RNase domains to process pre-crRNA and degrade viral RNA (34).

1.3.3 - Adaptation in different CRISPR types

CRISPR adaptation, which is the focus of this work and is the process of acquiring new spacers, occurs in a similar fashion throughout the different CRISPR types. Classification is delineated by interference modules because adaptation modules of most CRISPR systems are remarkably similar. Cas1 and Cas2, the core adaptation proteins, are the most well conserved proteins across CRISPR systems (35). Some exceptions do exist, such as a type V system which was shown to not include Cas2, only a Cas1 (36). Type III-B systems, which lack Cas1 and Cas2 genes, likely need other CRISPR systems present in the same host to acquire new spacers. Certain type III systems are also unique in that they have been shown to incorporate new spacers directly from RNA using a reverse transcriptase fused to Cas1 (37). Despite the high level of conservation, certain nuances in mechanisms between CRISPR types have uncovered properties of Cas1 and Cas2 proteins that can be useful to various forms of industry.

1.4.0 - Core components of CRISPR adaptation

CRISPR adaptation occurs in several steps with varying levels of understanding of each step. This process is best characterized in type I CRISPR systems, and general mechanisms described in this chapter will be mostly drawn from studies in type I. Key details that differentiate type I from type II, the focus of this dissertation, will be described later. In most CRISPR systems examined to date, the proteins Cas1 and Cas2 play an important role in adaptation. Generally, Cas1 and Cas2 form a complex consisting of 4 copies of Cas1 and two copies of Cas2 (Cas1-Cas2) that binds foreign DNA and facilitates the site-specific insertion of new spacers into the CRISPR array (**Figure 4**). Potential spacers in the invading DNA are referred to as protospacers, which are then processed into shorter pieces called prespacers that bind to the Cas1-Cas2 complex. The prespacer is integrated into the CRISPR array of the host genome to become a new spacer.

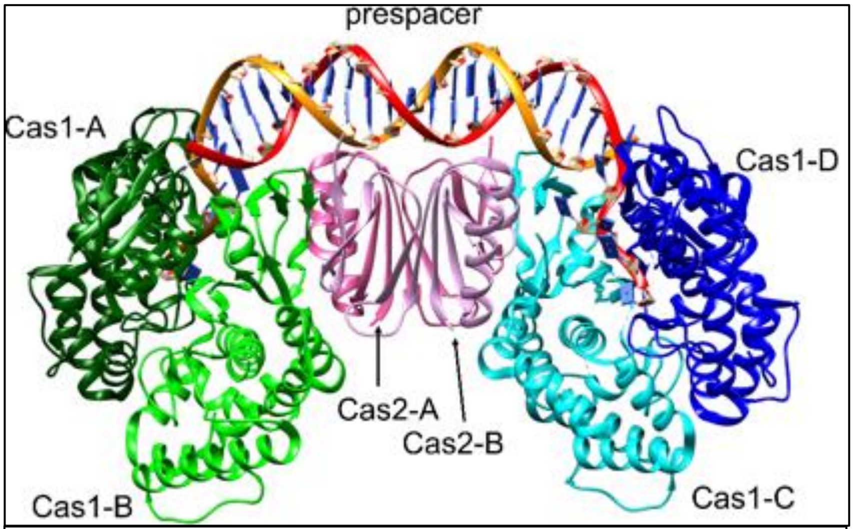


Figure 4: Cas1-Cas2-prespacer Integration complex. Two Cas1 dimers (Cas1: A-D, greens and blues) sandwich a Cas2 dimer (Cas2: A-B, pinks). The prespacer (red and orange) is bound on flat edge and is held in place by interactions with each Cas1 dimer and the Cas2 dimer. PDB id: 5DS4 (2).

This Cas1-Cas2-prespacer complex (hereafter referred to as the integration complex (IC)) consists of two Cas1 dimers sandwiching a Cas2 dimer, binding the prespacer along the flat edge (2) (**Figure 4**). The prespacer is bound by phosphate backbone interactions with the Cas1 dimers on each end of the prespacer and the Cas2 dimer in the middle of the prespacer. The strands of the prespacer are separated into single strands at each end of the complex, with the 3′ end of each strand being routed toward the active site of a Cas1 monomer (1).

Along with the IC, CRISPR leader regions have been shown to be indispensable for adaptation. Leader regions are sections of the CRISPR locus that contain the promoter for transcription of the CRISPR array. They are usually designated as regions between the final *cas* gene in the CRISPR locus and the start of the CRISPR array (38). Different lengths of CRISPR leaders are seen among different CRISPR types, ranging from less than 50 bp in some bacteria to several hundred bp in other organisms (38). Substantial portions of the leader region have been shown to be essential, such as a type I-A system where a full 400 bp of the leader region were shown to be essential for site specific insertion of new spacers (39). Type II-A leader regions were shown to be more dispensable than type I, with only the last 10 bp of the leader directly upstream of the CRISPR array being essential for site specific insertion of new spacers in certain systems (40). Leader regions in some cases have binding sites for accessory proteins to CRISPR adaptation. This is true for integration host factor (IHF) in type I-E systems (41).

1.5.0 - Sources for new spacers

Initially, protospacers need to be chosen and excised from the invading nucleic acid. The selection of protospacers can come from the process of either naïve or primed adaptation.

1.5.1 - Naïve adaptation

Naïve refers to a new protospacer being selected that does not share any sequence homology to spacers already present in the CRISPR array. To select for spacers from a foreign origin, CRISPR adaptation machinery must be biased in some way towards selecting protospacers from viral DNA. In 2015, acquisition of new spacers in *Escherichia coli* was shown to be biased towards free DNA ends (42). This led to the hypothesis that new spacers were selectively taken from actively replicating nucleic acids, or replication forks in general. Since viruses are at much higher copy number than genomic DNA, more replication forks would be present in viral DNA and therefore bias the spacer acquisition to foreign DNA rather than the host genome (42). The RecBCD system has also been implicated in naïve adaptation, with the degradation of free DNA ends being able to produce more substrates for possible spacer selection (43). RecBCD complexes are known to stall degradation at Chi sites found in DNA, which are abundant in prokaryotic genomic DNA but sparse in viral DNA (42). This bias toward degradation of viral DNA even further shifts acquisition bias toward foreign nucleic acid (42). Bias toward free DNA ends has been shown in other systems as well, outside of *E. coli* (44, 45).

Naïve adaptation may also rely on innate immune systems, such as restriction modification systems, for the generation of potential spacers. Restriction modification systems create free DNA ends, which in turn help bias spacer acquisition from foreign DNA rather than self-DNA. Restriction modification systems have been shown to increase spacer acquisition (46), and already employ their own self vs. non-self-discrimination mechanisms that are taken advantage of by the CRISPR systems to bias towards up taking spacers from foreign DNA.

1.5.2 - Primed adaptation

Primed adaptation occurs in conjunction with an already existing CRISPR interference module targeting the foreign nucleic acid. Because of the bias of spacer acquisition machinery toward free DNA ends, CRISPR interference is part of a feedback loop that creates more free DNA ends by cleaving its target, thus providing more possible substrates for spacer acquisition. Primed adaptation serves as a defense mechanism against phages with high mutation rates, where past spacers may not be completely complementary to the phage genome during subsequent infections due to the faster evolution of the phage. In this case, partially matching spacers that are competent to assemble the interference complex on the phage DNA can stimulate spacer acquisition (47). Primed adaptation has been seen in various type I CRISPR systems, as well as one type II system (47-49). Type III CRISPR systems are more tolerant of mismatches in the guide sequence during interference, and therefore have a lower probability for primed adaptation (43).

1.6.0 - Role of the protospacer adjacent motif in adaptation

Along with free DNA ends, protospacer selection is reliant on the presence of a protospacer adjacent motif (PAM) which is a 3-5 nucleotide (nt) sequence recognized directly by Cas proteins (50). The PAM is necessary for discriminating between self and non-self during the interference step (51). Without the PAM requirement in adaptation and interference, CRISPR arrays would be under constant attack by their own interference complexes. PAMs thus provide a mechanism to minimize autoimmunity. During the adaptation phase, PAM recognition can be carried out by different Cas proteins depending on the CRISPR type. The Cas1-Cas2 complex from type I CRISPR systems has been shown to recognize the PAM directly during protospacer selection (52). Cas9, the interference complex, recognizes the PAM during the adaptation phase in type II systems (53, 54). Mutations in the PAM interacting domain of Cas9 caused spacers to be acquired in a PAM independent manner (53). After recognition of the PAM, other Cas proteins (and possibly other cellular factors) are then recruited to the protospacer site where the protospacer is excised from the foreign DNA and bound to Cas1-Cas2. The PAM sequence is also recognized during the interference stage of CRISPR immunity. It is important that the PAM sequence be removed before integration of the prespacer into the CRISPR array, as the presence of the PAM would lead to self-targeting (35). This combination of two separate recognition activities, one form the crRNA sequence and another from the interference complex recognizing the PAM, creates a two-pronged approach that will preferentially target foreign nucleic acid.

1.7.0 - Prespacer processing

Either before or after binding to the Cas1-Cas2 complex, prespacers are known to undergo processing before being integrated into the CRISPR array. As mentioned before, trimming of the PAM sequence is essential to prevent self-targeting. Various studies of *in vitro* activities of Cas1-Cas2 have shown that splayed or trimmed spacers perform better in integration assays (1, 39, 41, 55, 56). Cas4 has been shown to participate in spacer processing in type II-B and various type I systems (56-58). Type II-

A and II-C systems, which lack the Cas4 protein, may rely on other host proteins or Cas4 proteins from other co-existent CRISPR systems (55-58). Host proteins such as DnaQ and ExoT, have been shown to process spacers in a type I system (59). This spacer processing is essential for the integration reaction to occur, as will be shown later in this work.

1.8.0 - Integration site recognition

To be an operational spacer without eliciting host gene disruptions, new spacers need to be inserted at the leader-repeat junction, where the leader region meets the first repeat of the CRISPR array (**Figure 1**). Current knowledge shows that there are distinct mechanisms employed by the different CRISPR types to recognize this site (40, 41, 60). Type I CRISPR systems rely on association of host protein factors with the IC to recognize the leader-repeat junction. *In vitro* assays done in a type I-E system showed that in the absence of the protein IHF, new spacers were inserted throughout the whole genome in a non-specific manner (41). Structural work has shown that IHF binds to the leader region about 25 bp upstream of the CRISPR array, causing a sharp bend in the DNA which allows for upstream DNA sequences in the leader to interact with the IC. Structures of this complex show sequence specific hydrogen bonding interactions between Cas1-Cas2 residues and bases near the leader-repeat junction and the upstream leader region (60). This increase in interactions causes the specificity for prespacer insertion at the leader-repeat junction (60). Type II CRISPR systems rely on intrinsic sequence specificity of Cas1-Cas2 for recognition of the leader-repeat junction rather than depending on other cellular factors (61). *In vivo* experiments showing deep sequencing of spacer acquisition into a plasmid containing the leader-repeat junction showed an ~18 times preference for integration to occur at the leader-repeat junction, as opposed to at a different repeat in the CRISPR array in type II-A systems (61). Another study showed that deletion of certain regions of the leader resulted in ectopic spacer integration, or integration of spacers at repeats other than the first repeat of the CRISPR array (62). Integration site recognition in other CRISPR types is not well studied at the present time.

1.9.0 - Integration reaction mechanism

Structural positioning of the reaction components for spacer integration is a key function of the Cas1-Cas2 complex. Prespacer processing facilitates the proper length of nucleic acid is present for optimal positioning of the 3' OH of each strand of the prespacer. At the leader-repeat junction, a single strand of the prespacer is covalently linked to the host genome by a trans-esterification reaction. A conserved histidine residue in Cas1 acts as a general base to activate the 3' OH of the prespacer to elicit a nucleophilic attack on the phosphate backbone of the host genome. This reaction is magnesium dependent, wherein magnesium facilitates charge stabilization of the reaction intermediate in the active site of Cas1. The IC then bends the repeat DNA to position the second strand of the prespacer for linkage at the repeat-spacer junction by a similar trans-esterification. These two linkages create a new spacer region with two single stranded repeat regions on each side, which are presumably repaired by host DNA repair mechanisms (63) (**Figure 5**).

1.10.0 - Roles of accessory proteins in CRISPR adaptation

Cas1-Cas2 makes up the core integration module in nearly all CRIPSR systems examined to date. However, several systems do employ the use of other proteins to aid in adaptation. Some of these proteins, as mentioned in the previous sections, play roles in protospacer capture, prespacer processing, or integration site recognition. IHF is a non-CRISPR protein which is essential for type I-E systems to recognize the leader-repeat junction (41). The host proteins DnaQ and ExoT process protospacers to the correct length by asymmetrically trimming the prespacers, allowing the insertion reaction to discriminate a functional from a non-functional orientation before inserting. This allows for the non-PAM strand to be inserted first, because of the longer PAM strand, and thus dictates orientation (59). This activity was also seen by another study which showed that a Cas2 which contained a DnaQ domain could perform the same function (64). Cas4 is a protein found in both type I and II CRISPR systems (**Figure 3**). Cas4 has been shown in several studies to help select prespacers via PAM recognition (55, 56). Cas4 has also been shown to be involved in prespacer trimming and orienting prespacers for directional insertion (57, 58). This activity of directional insertion is particularly important because recognition of the PAM during the interference step is crucial to a robust immune response (53). PAM sequences can be found either 5' or 3' of the protospacer sequence, depending on the CRISPR type. Prespacers that are inserted in the wrong orientation will produce crRNAs having PAM on the wrong side, which will produce ineffective interference complexes, essentially incorporating a non-functional spacer. Proteins such as Cas9, Cas4, DnaQ, and ExoT have been shown to facilitate this directional insertion of functional spacers (53, 55-59, 64). Csn2 is a signature protein of type II-A systems and has been shown to be essential for adaptation *in vivo (54)*. Csn2 is known to form a complex with Cas1-Cas2 and bridge the interaction between Cas1-Cas2 and Cas9 (65). Recent structural work has shown that Csn2 may play a role in prespacer protection (66). In this structure, Csn2 is shown

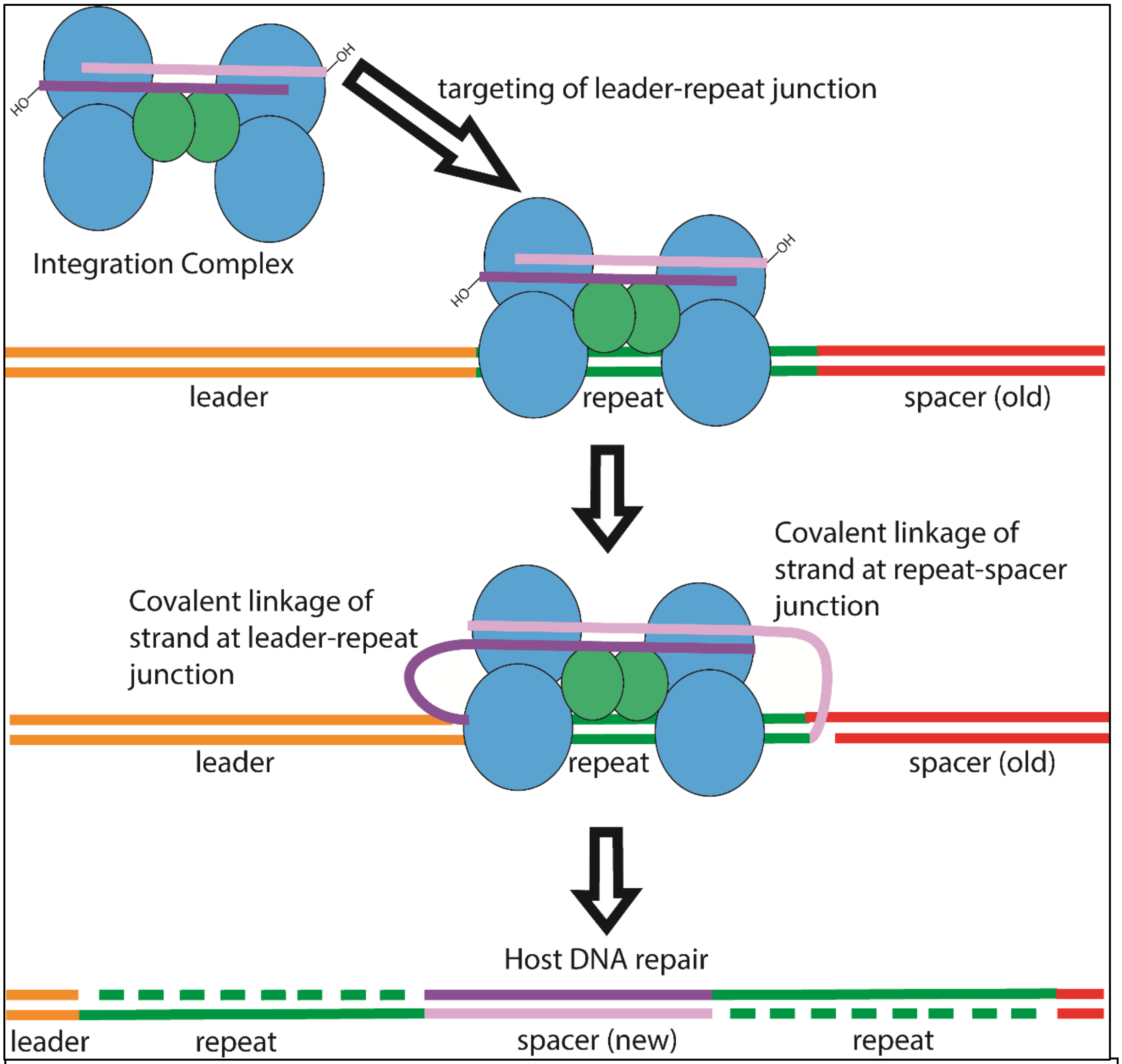

HO
OH
targeting of leader-repeat junction
Integration Complex
HO
OH
leader
repeat
spacer (old)
Covalent linkage of strand at repeat-spacer junction
Covalent linkage of strand at leader-repeat junction
leader
repeat
spacer (old)
Host DNA repair
leader
repeat
spacer (new)
repeat

Figure 5: Prespacer integration mechanism. Prespacers are brought by the integration complex to the leader-repeat junction. Each strand of the prespacer is covalently attached to a strand of the CRISPR locus, one at the leader-repeat junction and one at the repeat-spacer junction. The resulting product is a new spacer region with single stranded repeat regions on each side, which are filled in and repaired by host DNA repair mechanisms.

to form two tetrameric structures surrounding a double stranded DNA molecule. Multiple Cas1-Cas2 complexes bind to the outside of this complex, presumably to catch the prespacer once the Csn2 molecules dissociate (66). The role of Csn2 in this case is hypothesized to be protection of the prespacer from degradation before it can bind to Cas1-Cas2. The mechanisms by which other systems that are devoid of Csn2 or Csn2-like proteins capture, protect, and process prespacers remain unanswered.

1.11.0 - Anti-CRISPR proteins found in phages

In the arms race between prokaryotes and viruses, CRISPR systems are a target for immune escape tactics of many viral systems examined to date (67). Viruses target and deactivate CRISPR systems using anti-CRISPR proteins (Acr proteins) (68). Before the discovery of Acr proteins, the only way viruses were known to evade CRISPR systems was through mutations in the PAM or crRNA guide sequence. Acr proteins were first discovered in different strains of *Pseudomonas aeruginosa* harboring different prophages. Upon infection with different bacteriophages, some strains allowed the replication of the phage while others did not. This eventually led to the discovery of several Acr proteins which directly target and inhibit different aspects of CRISPR immunity. To date, several different Acr proteins have been discovered which inhibit various points of the CRISPR interference step. Acr proteins usually inhibit CRISPR interference by preventing binding of the interference complex to the target nucleic acid or by preventing cleavage by the interference complex. To date, Acr proteins have been found which are specific to type I, II and V CRISPR systems (68). The presence of Acr proteins in bacteriophages also explains reasons for certain organisms to harbor multiple CRISPR systems. Acr proteins can often be specific to a single CRISPR subtype, making the incorporation of multiple CRISPR systems an evolutionary response. Acr proteins are also of great interest to the industrial community, as they provide convenient regulation of CRISPR proteins during popular gene editing strategies. Limiting the amount of time CRISPR proteins are active may limit the off target effects of gene editing with Cas9 (69).

1.12.0 - Adaptation in type II-A systems

Type II-A CRISPR systems, the focus of this work, contain 4 *cas* genes which code for the proteins Cas9, Cas1, Cas2, and Csn2. Cas9 is a large, multidomain interference complex and is responsible for the interference stage of CRISPR immunity in all type II systems. Cas1 and Cas2 are both involved in the adaptation phase of CRISPR immunity and are responsible for the insertion of new spacers (61). The role of Csn2 is known to be in the adaptation phase but is still being studied. Recent structural work has provided some insights and has implicated Csn2 in prespacer capture for adaptation (66).

Protospacers are selected in type II-A systems by recognition of the PAM, with Cas9 being responsible for the PAM specific selection of protospacers (53, 54). Mutation of the PAM interacting domain of Cas9 caused protospacers to be selected non-specifically from both the invading nucleic acid and the host genome (53). Cas9, Cas1, Cas2, and Csn2 are essential for functional new spacers to be inserted under *in vivo* conditions (54). Interestingly, isolated Cas1-Cas2 can insert processed prespacers under *in vitro* conditions, implicating that the integration reaction and selection of integration site requires only cognate Cas1 and Cas2 protein (61). Thus, Cas9 and Csn2 are essential for the selection of a functional spacer. A previous study including several type II-A CRISPR systems from streptococci showed that the last 10 bp of the 3' end of the leader and the first repeat were essential and sufficient to direct spacer insertion *in vivo (40)*. This study also showed the presence of a conserved motif of 5'-ATTTGAG-3' at the 3' end of the leader in the streptococcal type II-A systems that were analyzed (40).

1.13.0 - Hypotheses

An intriguing activity of Cas1 and Cas2 in type II-A systems is their ability to insert prespacers site-specifically without the need of accessory factors. More details on this activity and an investigation into the specific interactions governing it can be found in Chapters 2 and 3. Based on the information available at the time of initiation of the study, a hypothesis that conserved DNA motifs at the leader-repeat junction would be recognized by type II-A Cas1-Cas2 complex was formed. A broad bioinformatics study was completed, collecting sequences of all unique type II-A CRISPR systems available at the time, which found three distinct motifs present at the leader-repeat junction. Phylogenetic analysis of Cas proteins and repeat sequences of these selected type II-A systems showed that the corresponding clusters mirrored the clustering based on leader 3' ends. These findings, that are presented in Chapter 2, led us to hypothesize that the mechanisms by which these three groups of type II-A adaptation proteins interact with their cognate leader-repeat junction will differ. These differences may be due to different catalytic requirements or different molecular assemblies. To address this, a series of *in vitro* protein-DNA assays were carried out that identified the different requirements for integration among the three groups. The details are presented in Chapter 3. To assess the molecular organization and structure of the Cas proteins involved in new spacer acquisition and to distinguish differences in the mechanisms between the different CRISPR sub-types, biophysical analysis of Cas proteins was also carried out by multi-angle light scattering and X-ray crystallography, the details of which are presented in Chapter 4.

1.14.0 - Significance

A greater understanding of CRISPR adaptation provides valuable information useful to a broad range of disciplines. From a biological standpoint, the better we understand how prokaryotes fight off infections, the better we can protect industrially relevant processes that rely on CRISPR containing microbes (such as the dairy industry) (6, 70, 71). The chronological record provided by the CRISPR array can be used to study the evolutionary lineage of prokaryotes and the phages that infect them (72). The processes involved in adaptation also provide useful activities for the development of novel biological tools (72-74). Site-specific spacer insertion by Cas1-Cas2 has been used advantageously by scientists for the development of transcriptional recordings and digital information storage in bacteria (73, 74). With regards to human health, a better understanding of pathogens which contain CRISPR systems will theoretically allow the manipulation of the adaptation mechanisms to make it more susceptible to treatments like phage therapy (75).

Chapter 2: Conserved DNA motifs in the type II-A CRISPR leader region

2.1.0 - Acknowledgements

I would like to acknowledge Dr. Peter Klein, a co-first author on this paper for his contributions in setting up the bioinformatics pipeline and in data collection. I would also like to acknowledge the other authors on this paper, Dr. Kesavan Babu, Dr. Fares Najar, and Dr. Rakhi Rajan for their contributions in data collection, analysis, writing, and resources. My contributions included screening gathered datasets, performing sequence alignments and tree construction, figure making, writing, and editing of the presented manuscript. Funding was provided by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under the grant number P20GM103640.

2.2.0 - Copyright Information

2.3.0 - Abstract

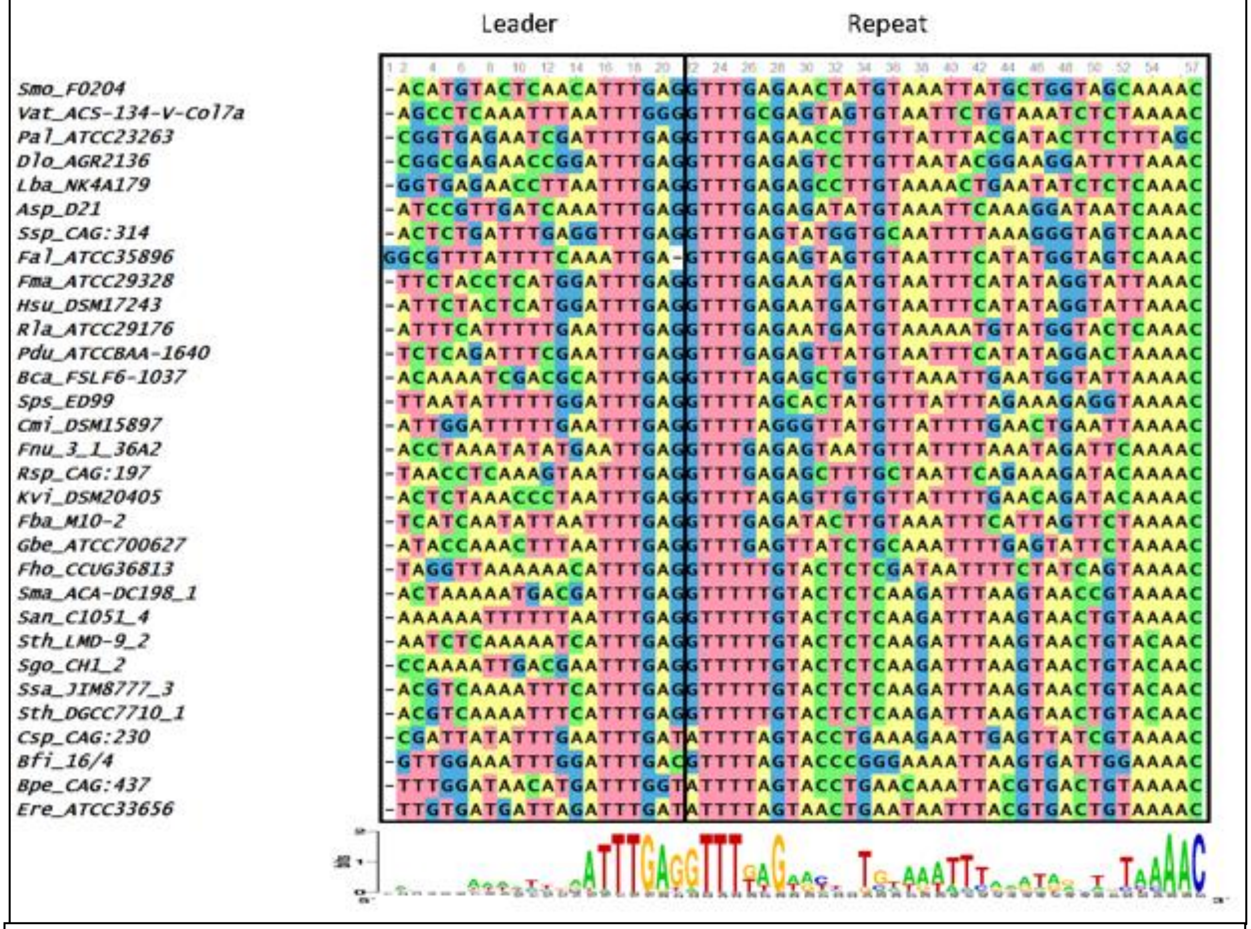
The CRISPR-Cas systems consist of RNA-protein complexes that provide bacteria and archaea with sequence-specific immunity against bacteriophages, plasmids, and other mobile genetic elements. Bacteria and archaea become immune to phage or plasmid infections by inserting short pieces of the intruder DNA (spacer) site-specifically into the leader-repeat junction in a process called adaptation. Previous studies have shown that parts of the leader region, especially the 3' end of the leader, are indispensable for

adaptation. However, a comprehensive analysis of leader ends remains absent. Here, we have analyzed the leader, repeat, and Cas proteins from 167 type II-A CRISPR loci. Our results indicate two distinct conserved DNA motifs at the 3' leader end; ATTTGAG (noted previously in the CRISPR1 locus of *Streptococcus thermophilus* DGCC7710) and a newly defined CTRCGAG, associated with the CRISPR3 locus of *S. thermophilus* DGCC7710. A third group with a very short CG DNA conservation at the 3' leader end is observed mostly in *Lactobacilli*. Analysis of the repeats and Cas proteins revealed clustering of these CRISPR components that mirrors the leader motif clustering, in agreement with the coevolution of CRISPR-Cas components. Based on our analysis of the type II-A CRISPR loci, we implicate leader end sequences that could confer site-specificity for the adaptation-machinery in the different subsets of type II-A CRISPR loci.

2.4.0 - Results

2.4.1 - Analysis of the 3' end of the leader

An initial sequence alignment of the last 20 nucleotides of the leader plus the first repeat showed that the 167 loci clustered into distinct groups. These groups had recognizable conservation at the last 7 nucleotides of the 3' end of the leader and the first 4 nucleotides of the 5' end of the first repeat, or the leader-repeat junction. To obtain an unbiased separation of the different groups, a Cas1 phylogenetic tree was constructed based on protein sequence similarity. The loci belonging to the different clades of the Cas1 tree were grouped together and a sequence alignment of the last 20 nucleotides of the leader along with the first repeat was performed. In order to facilitate interpretation of the trees and alignments, a smaller representative sample of 62 loci was used to generate the main figures and show the relevant relationships. Each of the 3 groups were aligned separately to discern the level of conservation within each group



```
                        1 2  4  6  8  10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54  57
Smo_F0204               -ACATGTACTCAACATTTGAGGTTTGAGAACTATGTAAATTATGCTGGTAGCAAAAC
Vat_ACS-134-V-Col7a     -AGCCTCAAATTTAATTTGGGGTTTGCGAGTAGTGTAATTCTGTAAATCTCTAAAAC
Pal_ATCC23263           -CGGTGAGAATCGATTTTGAGGTTTGAGAACCTTGTTATTTACGATACTTCTTTAGC
Dlo_AGR2136             -CGGCGAGAACCGGATTTGAGGTTTGAGAGTCTTGTTAATACGGAAGGATTTTAAAC
Lba_NK4A179             -GGTGAGAACCTTAATTTGAGGTTTGAGAGCCTTGTAAAACTGAATATCTCTCAAAC
Asp_D21                 -ATCCGTTGATCAAATTTGAGGTTTGAGAGATATGTAAATTCAAAGGATAATCAAAC
Ssp_CAG:314             -ACTCTGATTTGAGGTTTGAGGTTTGAGTATGGTGCAATTTTAAAGGGTAGTCAAAC
Fal_ATCC35896           GGCGTTTATTTTCAAATTGA-GTTTGAGAGTAGTGTAATTTCATATGGTAGTCAAAC
Fma_ATCC29328           -TTCTACCTCATGGATTTGAGGTTTGAGAATGATGTAATTTCATATAGGTATTAAAC
Hsu_DSM17243            -ATTCTACTCATGGATTTGAGGTTTGAGAATGATGTAATTTCATATAGGTATTAAAC
Rla_ATCC29176           -ATTTCATTTTTGAATTTGAGGTTTGAGAATGATGTAAAAATGTATGGTACTCAAAC
Pdu_ATCCBAA-1640        -TCTCAGATTTCGAATTTGAGGTTTGAGAGTTATGTAATTTCATATAGGACTAAAAC
Bca_FSLF6-1037          -ACAAAATCGACGCATTTGAGGTTTTAGAGCTGTGTTAAATTGAATGGTATTAAAAC
Sps_ED99                -TTAATATTTTTGGATTTGAGGTTTTAGCACTATGTTTATTTAGAAAGAGGTAAAAC
Cmi_DSM15897            -ATTGGATTTTTGAATTTGAGGTTTTAGGGTTATGTTATTTTGAACTGAATTAAAAC
Fnu_3_1_36A2            -ACCTAAATATATGAATTGAGGTTTGAGAGTAATGTTATTTTAAATAGATTCAAAAC
Rsp_CAG:197             -TAACCTCAAAGTAATTTGAGGTTTGAGAGCTTTGCTAATTCAGAAAGATACAAAAC
Kvi_DSM20405            -ACTCTAAACCCTAATTTGAGGTTTTAGAGTTGTGTTATTTTGAACAGATACAAAAC
Fba_M10-2               -TCATCAATATTAATTTTGAGGTTTGAGATACTTGTAAATTTCATTAGTTCTAAAAC
Gbe_ATCC700627          -ATACCAAACTTTAATTTGAGGTTTGAGTTATCTGCAAATTTTGAGTATTCTAAAAC
Fho_CCUG36813           -TAGGTTAAAAAACATTTGAGGTTTTTGTACTCTCGATAATTTTCTATCAGTAAAAC
Sma_ACA-DC198_1         -ACTAAAAATGACGATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACCGTAAAAC
San_C1051_4             -AAAAAATTTTTTAATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTAAAAC
Sth_LMD-9_2             -AATCTCAAAAATCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
Sgo_CH1_2               -CCAAAATTGACGAATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
Ssa_JIM8777_3           -ACGTCAAAATTTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
Sth_DGCC7710_1          -ACGTCAAAATTTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
Csp_CAG:230             -CGATTATATTTGAATTTGATATTTTAGTACCTGAAAGAATTGAGTTATCGTAAAAC
Bfi_16/4                -GTTGGAAATTTGGATTTGACGTTTTAGTACCCGGGAAAATTAAGTGATTGGAAAAC
Bpe_CAG:437             -TTTGGATAACATGATTTGGTATTTTAGTACCTGAACAAATTACGTGACTGTAAAAC
Ere_ATCC33656           -TTGTGATGATTAGATTTGATATTTTAGTAACTGAATAATTTACGTGACTGTAAAAC
```

Figure 6: Group 1 sequence alignment. Sequence alignment of the last 20 nucleotides of the 3' end of the leader and the first repeat of selected Group 1 species. Height of the letters in the WebLogo indicates the degree of conservation at specific nucleotide locations. The leader-repeat end is conserved as ATTTGAG-GTTT.

(**Figures 6** and **7**). Strict conservation is seen at the 3' end of the leader as well as at the 5' end of the repeat. Group 1 has the 3' leader end conserved as ATTTGAG (**Figure 6**) and Group 2 has the 3'-leader end conserved as CTRCGAG (where R represents a purine) (**Figure 7A**). Group 3 has a shorter two nucleotide conservation of CG at the 3' leader end (**Figure 7B**). In Groups 1 and 2, the last three nucleotides are conserved as GAG (**Figures 6 and 7A**). An A-rich region is partially conserved adjacent to the CG leader end of Group 3. Interestingly, the CRISPR1 locus of Sth DGCC7710 has the 3' leader end conserved as ATTTGAG while the CRISPR3 locus has the 3’ leader end

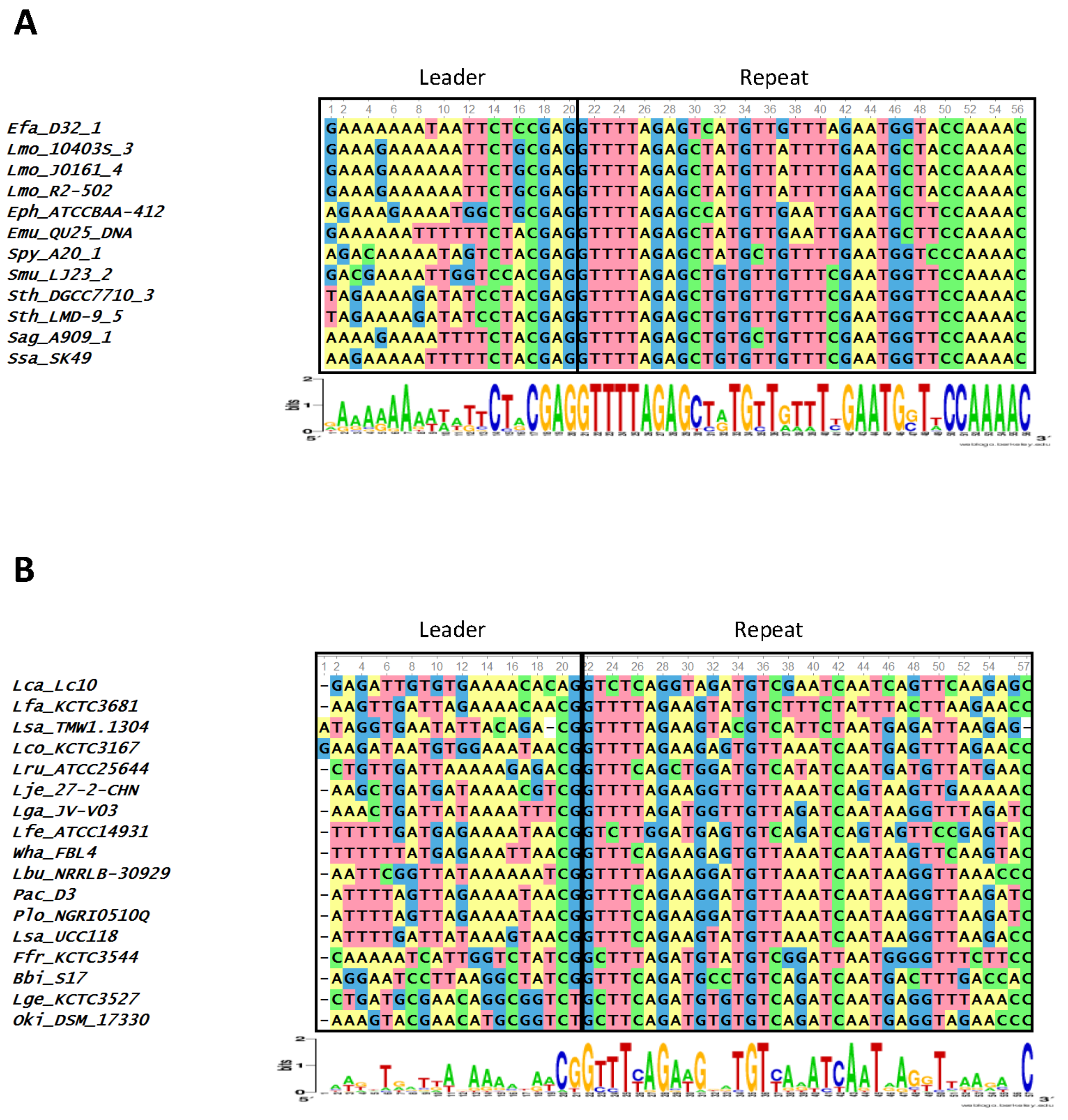


Figure 7: Groups 2 and 3 sequence alignment. Sequence alignment of the last 20 nucleotides at the 3’ end of the leader and the first repeat of selected Group 2(A) and Group 3(B) species. Height of the letters in the WebLogo indicates the degree of conservation at specific nucleotide locations. The leader-repeat of Group 2 loci is conserved as CTRCGAGGTTT, where R represents a purine base. For Group 3 members, this region is conserved as CGGTTT.

conserved as CTACGAG. Of the type II-A CRISPR loci analyzed, 87 belonged to Group1, 55 belonged to Group 2, and 25 belonged to Group 3. Out of the 50 genera analyzed, Group 2 consists of only 5 genera (*Streptococcus, Enterococcus, Listeria,* *Lactobacillus* and *Weissella*) while Group 1 is much more diverse with 42 different genera. Group 3 accounts for 7 genera, but has many loci belonging to the Order Lactobacillales. The leader-repeat junction of Groups 1 and 2 is conserved as GAG-GTTT while in Group 3 it is weakly conserved as CG-GTTT.

2.4.2 - Analysis of the repeat region

The length of the repeat for the type II-A loci analyzed was 36 nucleotides except in 4 cases (*Enterococcus hirae* ATCC 9790 (35 nucleotides long), *Fusobacterium* sp. 1_1_41FAA (37 nucleotides long), *Lactobacillus coryniformis* subsp. coryniformis KCTC 3167 (37 nucleotides long), and *Lactobacillus sanfranciscensis* TMW 1.1304 (35 nucleotides long)). The first repeat sequences of the 3 groups did not possess any distinguishable motifs that corresponded to the segregation of the different groups (**Figure 8**). There is a strong sequence conservation at the 5' end of the repeat as GTTT in all the type II-A loci analyzed (**Figure 8**). Groups 1 and 2 also share a conserved AAAC motif at the 3' end of the repeat. Group 3 members have a conserved C at the 3' end of the repeat, along with a less conserved A-rich region ahead of the C. The repeat sequence belonging to the Group 2 loci is highly conserved across the entire length of the repeat, which may be attributed to the limited number of genera (five) comprising this group compared to Group 1 (forty two). In all the type II-A loci analyzed, the first and last nucleotides of the first repeat are conserved as G and C respectively. A phylogenetic tree was generated using the first repeat sequence of the type II-A loci (**Figure 8**). Even though the reliability of branching is low due to the short length of the sequence, the branches segregate such that members within a clade have similar repeat and leader end conservations. Recently, it was suggested that sequences at the

5’ and 3’ ends of the repeat in *S. pyogenes* type II-A system could be the motifs recognized by Cas1 during spacer acquisition (76). Hence, the conserved 5’ and 3’ repeat ends observed in the first two groups might indicate type II-A specific repeat ends that are essential for adaptation. Further experimental studies will be required to analyze whether the loosely conserved sequences at the 3’ end of the repeat impact effective adaptation in Group 3. The similarity at the 5’ and 3’ ends of the repeat in the different sub-groups of type II-A system and the fact that exchanging leader ends between CRISPR1 and CRISPR3 loci in Sth DGCC7710 (40) impaired adaptation shows that the specificity within the sub-groups of type II-A CRISPR system is most probably attributed by the 3’ leader end and not specified by the repeat ends.

2.4.3 - Analysis of Cas proteins

We extended our analysis to verify whether the different groups of type II-A CRISPR loci observed based on the 3’ leader end conservation relates to Cas proteins. The protein sequences of Cas1 belonging to the selected type II-A loci were aligned by MUSCLE and a phylogenetic tree was generated (**Figure 9**). The loci segregated into 4 main branches, with each branch carrying distinct groups based on the 3’ leader end sequence conservation. A sequence alignment of the leader-repeat junction of the different branches show how the Cas1 sequence is highly correlated with the leader-repeat junction. This confirms previous findings that all the CRISPR-Cas components have coevolved together (77). The phylogenetic tree shows that Group 1 loci are very distant in lineage, which has later evolved into different subsets with specific leader-repeat-Cas1 combinations. Group 2 and Group 3 evolved for very specific genera, while Group 1 has accommodated divergent genera.

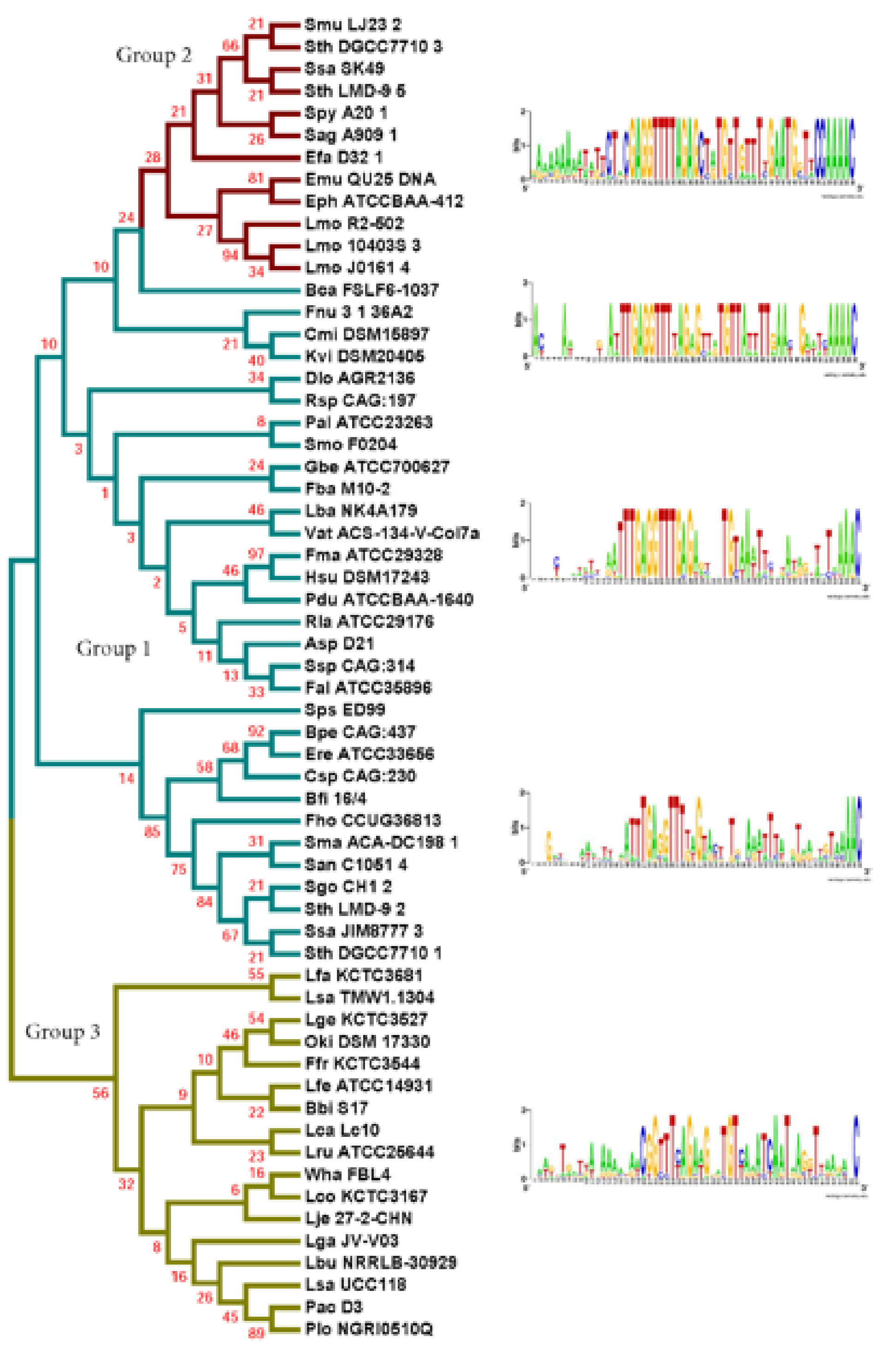


Figure 8: Phylogenetic tree generated from the sequence alignment of the first repeats from selected type II-A species. Groups based on the segregation of the Cas1 tree are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). The tree segregates into 6 main clades and WebLogos were produced with alignments of the last 20 nucleotides at the 3' end of the leader and the first repeat from the loci within each corresponding branch.

A similar analysis was done for the Cas2, Csn2, and Cas9 proteins. The sequence alignments generated using the sequences of the corresponding Cas proteins were used to build phylogenetic trees (**Figures 10 and 11**). All the clades in the different trees have similar 3'-leader ends, except for a few differences in the Cas9 phylogenetic tree where some Group 3 members appeared along with Group 1. A closer analysis of the sequences showed

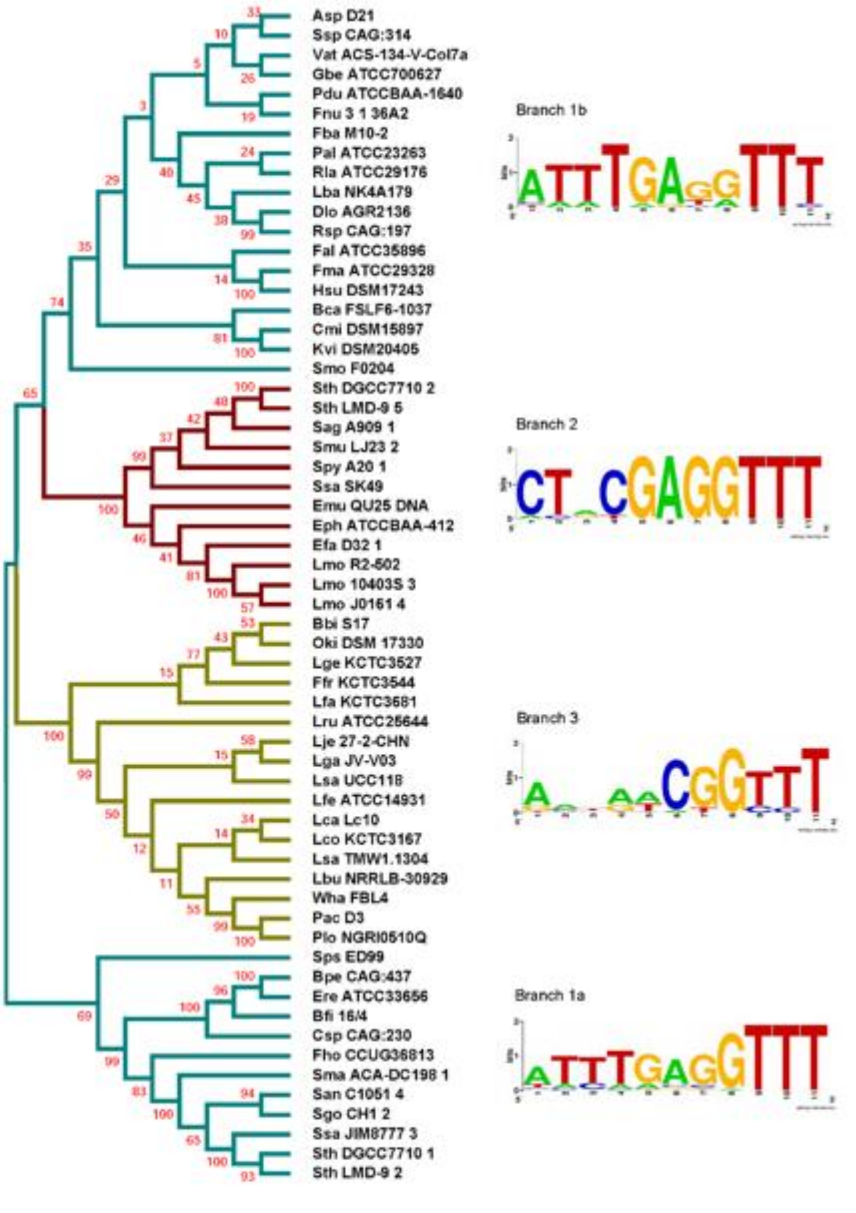


Figure 9: Phylogenetic tree generated from the protein sequence alignment of Cas1. Groups are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). WebLogos were generated by aligning the last 7 nucleotides of the leader and the first 4 nucleotides of the repeat from the loci within each corresponding branch. The tree segregates into 4 branches, two branches showing the Group 1 leader end motif, one branch showing the Group 2 motif, and one branch showing the less-conserved Group 3 leader end. Sps ED99 segregated independently from the final branch but was used in the final branch WebLogo construction based on the leader end and protein length.

high variability in the Cas9 lengths, including an extremely short Cas9 sequence (Plo NGRI0510Q) in the outliers, which may have contributed to the random placement of this Cas9 protein. Cas9 also showed a branch (1b) for Group 1 that did not show prominent leader end conservation as that was observed in branch 1a. Except for the few differences in Cas9, our results indicate the presence of distinct groups within the type II-A CRISPR systems that possess conserved 3' leader ends and group-specific Cas proteins.

It was proposed earlier that the longer version of Csn2 evolved first and the shorter Csn2 proteins evolved from the longer versions (78). Interestingly, our phylogenetic analysis agrees with this and shows a branch that represents the ancestor with an average Csn2 length of 320 amino acids (**Figure 11**). Three main branches evolved from the ancestor and all of them have an average amino acid length of 218-230 amino acids but varying 3' leader ends (**Table S1**). Thus, the ATTTGAG motif is ancestral and universal in the type II-A systems, which later developed to have a similar (ATTTGAG), deviating (CTRCGAG), or less conserved (CG) 3' leader end, with a corresponding change in the protein sequences of all four type II-A Cas proteins. Examining the lengths of Cas1, Cas2, and Cas9 from different groups, we did not observe a strong correlation between the average length of these Cas proteins and the branching group that they belonged.

2.5.0 - Discussion

Though previous studies have shown that the leader-repeat region is important for adaptation, the specific features of the leader-repeat region that may recruit Cas1-Cas2 for adaptation are not clearly defined. We focused on the sequence conservation around the leader-repeat junction and found three distinct DNA motifs at the 3’ leader ends: Group 1 (ATTTGAG), Group 2 (CTRCGAG), and Group 3 (CG). The presence of a conserved 3’ leader end, despite a low sequence conservation in the upstream regions of the leader in bacteria belonging to 50 different unrelated genera, strongly suggests that these DNA motifs play a role in site-specific adaptation. One of the most interesting observations from this analysis is the conservation of GAG-GTTT as the leader-repeat junction in both Group 1 and Group 2 (82%, 117 out of 142 loci) of the type II-A system.

Several studies have implicated the importance of the leader and repeat sequences to drive faithful adaptation. Terns and coworkers reported that streptococci with repeats similar to that present in the CRISPR1 locus (Group 1) of Sth DGCC7710 have the 3’ leader end conserved as ATTTGAG. The accompanying experimental work clearly demonstrated that the 10 nucleotides present at the 3’ end of the leader and the first repeat are essential and sufficient for adaptation, even in a non-CRISPR locus (40). It was concluded that sequences at the leader-repeat junction recruits the adaptation machinery to this region for integration of new spacers (40). In a recent study that analyzed the spacer variation in 126 human isolates of *S. agalactiae*, the 3’ leader end of most of the isolates had a TACGAG sequence (79). Our analysis that focused on many divergent genera uncovered that the DNA motifs that were previously known to be important for streptococcal adaptation is in fact more ubiquitous and conserved across different bacteria.

The importance of the sequences of the leader and the first repeat in driving adaptation is conserved across different CRISPR types. The 60 nucleotides towards the 3’ end of the type I-E CRISPR locus of *Escherichia coli* is essential for adaptation (80). The disruption of the first repeat sequence that left the stem-loop structure intact prevented successful adaptation in a type IE system leading to the conclusion that the cruciform structure of the repeat alone is not sufficient for adaptation (63). Another study showed that the -2 (second last position of leader) and +1 (first nucleotide of repeat) positions of

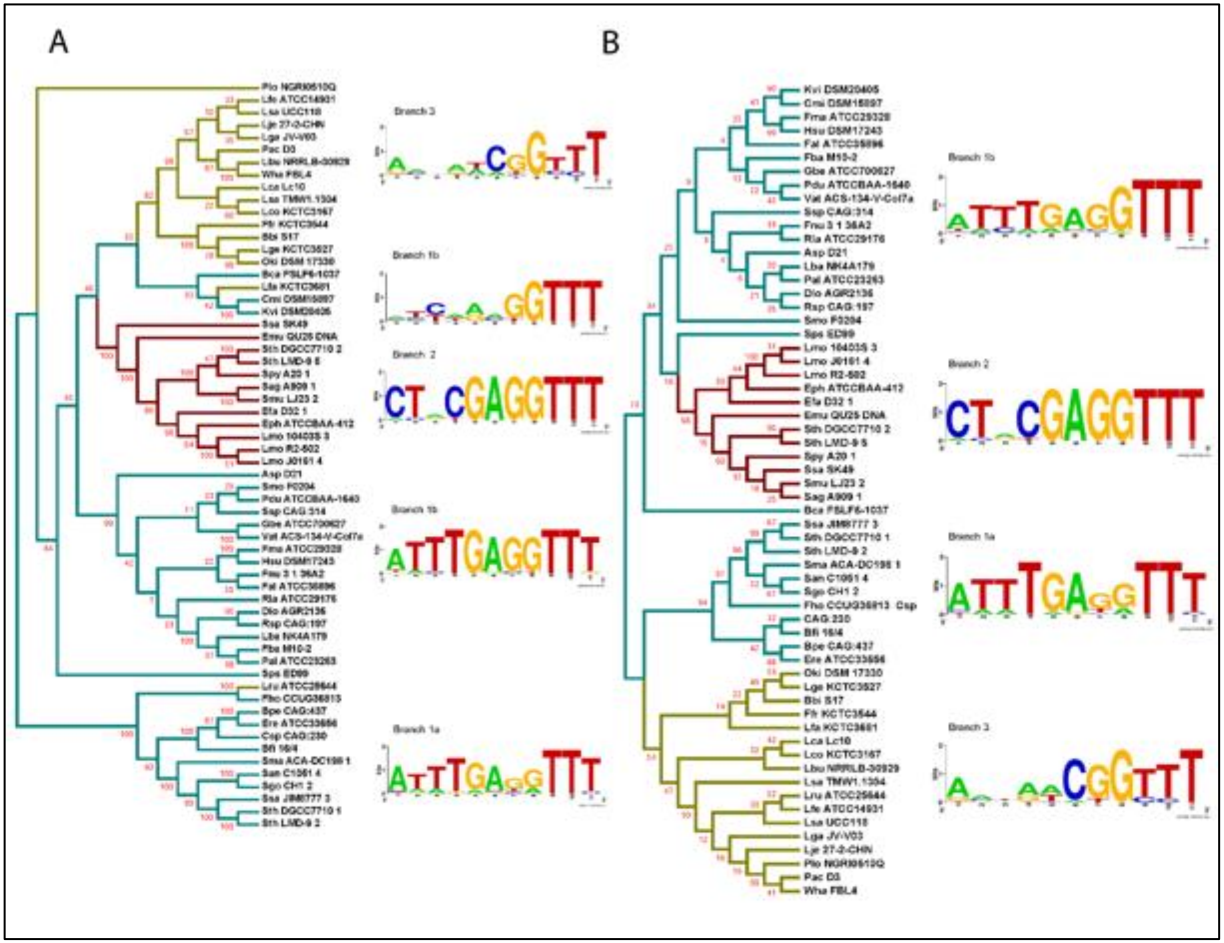


Figure 10: (A) Phylogenetic tree generated from the sequence alignment of Cas9. Groups based on the segregation of the Cas1 tree are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). The tree shows 5 different branches with two branches showing the Group 1 leader end motif, one branch showing the Group 2 motif, and one branch representing the less-conserved Group 3 leader end. One of the branches represent a very loosely conserved Group 1 loci. Three members of Group 3 segregated away from the normal cluster, of which Plo NGRI0510Q has a very short Cas9 sequence. Lru ATCC25644 and Lfa KCTC3681 have normal length Cas9 sequences. **(B) Phylogenetic tree generated from the sequence alignment of Cas2.** All the four branches segregate similarly to those of Cas1 phylogenetic tree. WebLogos for both panels of the figure were generated by aligning the last 7 nucleotides of the leader and the first 4 nucleotides of the repeat from the loci within each corresponding branch.

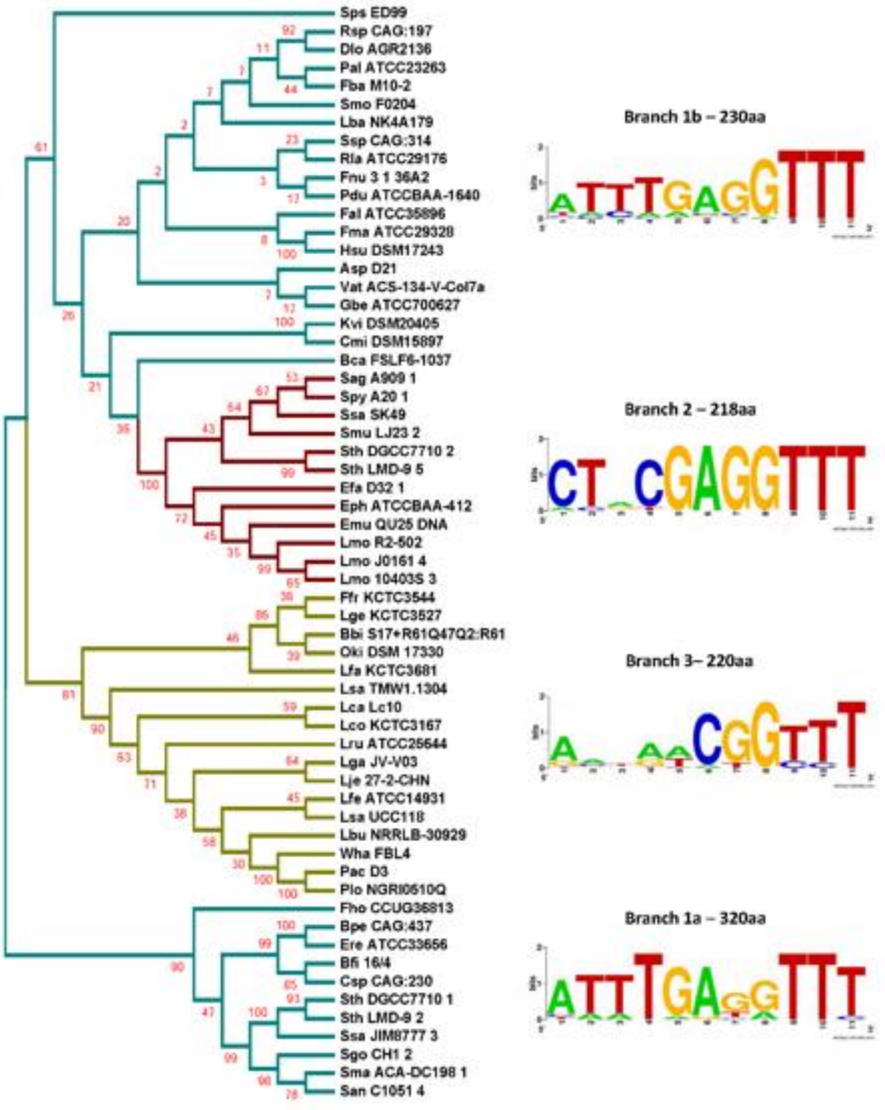


Figure 11: Phylogenetic tree generated from the sequence alignment of Csn2. Groups based on the segregation of the Cas1 tree are shown in cyan (Group 1), red (Group 2), and yellow (Group 3). WebLogos were generated from aligning the last 7 nucleotides of the leader and the first 4 nucleotides of the repeat from the loci within each corresponding branch. Values next to branch labels indicate the average length of the proteins (in amino acids, aa) within the branch. Two branches show the Group 1 leader end motif, one branch shows the Group 2 motif, and one branch shows the less conserved Group 3 leader end.

leader-repeat regions are crucial for adaptation in *E. coli* (type IE) and *Sulfolobus solfataricus* (type I-A) (81, 82). Other studies have experimentally demonstrated that leader and repeat sequences are important for adaptation in streptococcal type II-A systems corresponding to the groups 1 and 2 that we identified in our study(40, 76). Comparing our results with the earlier studies show that leader-repeat sequence conservation that we observed in type II-A sub-groups is relevant for adaptation across diverse bacteria.

There is an interplay between the leader and repeat sequences in adaptation that is CRISPR type specific. For example, in the type I-B system of *Haloarcula hispanica*, inverted repeats (IR) present within the first repeat are essential for recruiting the adaptation machinery to the leader-repeat junction. Once the IRs are located within a repeat, a cut is made by the Cas1-Cas2 complex at the leader-repeat junction and the sequence of the leader is critical for this step. The second cut at the repeat-spacer end is based on a ruler-mechanism and does not depend on the sequence of the repeat (83). Whereas in a type II-A system corresponding to our Group 2, it was shown that the repeat-spacer and repeat-leader ends have the same probability of getting cleaved by Cas1-Cas2, but for a faithful adaptation, the leader-repeat junction is essential (76). In the Group 1 type II-A locus of Sth DGCC7710, mutations in the last 10 nucleotides of the leader abolished adaptation (40). This study also elegantly showed that substitution of the 10 nucleotides at the 3' end of Group 1 leader with that of Group 2 leader abolished adaptation following a phage challenge, further emphasizing the importance of the locus specific leader-repeat junction in adaptation (40).Thus, our observation of the group-specific sequence conservation in type II-A systems at the leader end, along with a lack of distinct group-specific motifs at the 5' and 3' ends of the first repeat, shows that the sub-group specificity in type II-A adaptation arises from the leader sequences that might be specifically distinguished by the Cas1-Cas2 proteins belonging to each sub-group.

Both groups 1 and 2 are active for adaptation and interference (6, 77, 84-89), while Group 3 has been shown to be active in DNA interference (90). Introduction of the type II-A Group 2 locus into *E. coli* protected the bacterium from phage and plasmid infection

(88), demonstrating that intrinsic specificities of protein and DNA components of a CRISPR sub-type are sufficient to drive adaptation and there are no organismal requirements. The three different DNA motifs that we observed at the 3'-end of the leader of the type II-A CRISPR loci may represent three specific functional adaptation units, perhaps guided by leader-sequence specific Cas protein(s). The third group, which consists mostly of lactobacilli, with only two nucleotides conserved instead of seven nucleotides at the 3' leader end in Groups 1 and 2 may represent a more diverse adaptation complex where the protein-DNA sequence interactions are not as tight. It was noted recently that there is considerable variation in the spacer content, even in the ancestral spacers, in *L. gasseri* strains that indicates considerable divergence between the strains (90), thus accounting for the low level of sequence conservation at the 3'-end of the leader. This study also showed that the spacers matched plasmids and temperate phages, though it is not clear how *L. gasseri* acquires spacers from prophages that do not pose threat to bacterial survival (90). These environmental factors may contribute to the low sequence conservation at the 3'-end of the leader in Group 3. Further experiments will be required to assess the adaptation process in Group 3. Group 3 could also be a result of an insufficient amount of genomic data available to completely resolve any more conserved motifs hidden in the different leader end sequences found within the group.

Repeat sequences are specific to a CRISPR locus, even within sub-types (77, 78). The first two nucleotides of the first repeat was shown to be essential for adaptation in the CRISPR1 locus of Sth (40) and the first six nucleotides are essential in adaptation in *S. pyogenes* (76). The importance of G as the first nucleotide in the repeat for efficient disintegration reaction was demonstrated for both *E. coli* and *S. solfataricus* Cas1 proteins (91). We found conservation at the ends of the repeat between groups (**Figure 8**). Only 17/167 loci analyzed did not possess a GTTT at the 5'-end of the first repeat, and 3/167 of the loci did not possess a conserved C at the 3' end. It was previously reported that purified *E. coli* Cas1 possesses nuclease activity against several types of DNA substrates including single stranded DNA, replication forks, Holliday junctions *etc*. without adequate intrinsic sequence specificity and that the four-way DNA junctions recruits Cas1 protein (92). Recently, more studies point to the importance of DNA sequence specificity, especially at the 3'-end of the leader, for driving Cas1 for adaptation (40, 83, 91). The essentiality of IHF for site-specific adaptation in type I-E indicates that even though Cas1 may have the ability of non-sequence specific cleavage in certain CRISPR types, tight regulation by other cellular proteins may enhance site-specific spacer insertion. The position of the IHF site is 9 to 35 nucleotides upstream of the leader-repeat junction in type I systems (93). The 20 nucleotides of the 3' leader end that we analyzed for the type II-A did not possess any similarity to the IHF binding site. It is possible that a cruciform structure formed by leader-repeat or repeat palindromic regions along with specific leader-repeat sequences may recruit the Cas1-Cas2 complex for spacer insertion and that this requirement is critical under *in vivo* conditions.

All four Cas proteins are essential for successful adaptation *in vivo* in type II-A systems (6, 53, 80). Previous studies have shown that the CRISPR components and Cas proteins have coevolved (77). Our analysis showed that all the four type II-A specific Cas proteins and the first repeat clustered into identical groups with similar 3' leader ends. Even though Cas1 protein sequences within type II-A are highly conserved, there are certain differences that segregate them into distinct groups and interestingly these groups have distinct leader sequence conservation. It was previously reported that type II-A CRISPR systems have distinct operon organization that correlates with Csn2 sequence, making Csn2 the signature protein for type II-A systems (78). The longer version of Csn2 originated first and the shorter version evolved from the longer version (78). Our analysis shows that the length of Csn2 is conserved across different clusters (**Figure 11**). Looking at **Figure 11**, branch 1a segregated early from the rest of the tree and consists of the longer version of Csn2, while branches 1b, 2, and 3 all consist of the shorter version of Csn2. Correlating Csn2 branching to the leader end sequences, it is evident that our Group 1 motif of ATTTGAG is present in the ancestral strains, which later evolved to distinct sub-groups possessing either Group 1, Group 2 (CTRCGAG) or Group 3 (CG) leader ends.

2.6.0 - Conclusion

We present an extensive bioinformatic analysis of type II-A CRISPR systems spanning 50 different bacterial genera. We demonstrated the ubiquitous nature of two distinct DNA motifs at the 3’ end of the leader: Group 1 (ATTTGAG) and Group 2 (CTRCGAG) and also discovered a new group (Group 3) with a limited sequence conservation at the 3’-end of the leader. The leader-repeat junction is highly conserved for Groups 1 and 2 as GAG-GTTT. Our work proposes that the Cas proteins of each sub-group within the type II-A system should make sequence-specific association with its cognate DNA region for successful spacer insertion. The observations further strengthen the previous notion that a highly specific interplay between Cas proteins and cognate leader-repeat regions is essential for effective adaptation (40, 63, 80, 81, 94).

Chapter 3: CRISPR type II-A subgroups exhibit phylogenetically distinct mechanisms for prespacer insertion

3.1.0 - Acknowledgements

I would like to acknowledge the co-authors on this paper, Sydney Newsom for her role in performing replications of experiments and Dr. Rakhi Rajan for her contributions in experimental conceptualization, data analysis, writing, and resources. I would also like to thank Dr. Peter Klein for initiating the cloning experiments. My contributions included conceptualization, cloning and purification of proteins, target sequence generation, integration assay development, making figures, writing, and editing. Funding was provided by Oklahoma Center for the Advancement of Science and Technology, Project HR17-124 and in part by National Institutes of Health Grant P20GM103640 and the Office of the Vice President for Research and Partnerships, University of Oklahoma.

3.2.0 - Copyright Information

The original version of this work was published online in the Journal of Biological Chemistry on June 8th, 2020 (95). The work was published under license by the American Society for Biochemistry and Molecular Biology, Inc. https://www.asbmb.org/journals-news/editorial-policies#copyright. Figure numbers and references were changed to be in order with the current work and the introduction was removed, again, to reduce redundancy.

3.3.0 - Abstract

CRISPR-Cas is an adaptive immune system that protects prokaryotes against foreign nucleic acids. Prokaryotes gain immunity by acquiring short pieces of the invading nucleic acid, termed prespacers, and inserting them into their CRISPR array. In type II-A systems, Cas1 and Cas2 proteins insert prespacers always at the leader–repeat

junction of the CRISPR array. Among type II-A CRISPR systems, three distinct groups (G1, G2, and G3) exist according to the extent of DNA sequence conservation at the 3′ end of the leader. However, the mechanisms by which these conserved motifs interact with their cognate Cas1 and Cas2 proteins remain unclear. Here, we performed *in vitro* integration assays, finding that for G1 and G2, the insertion site is recognized through defined mechanisms, at least in members examined to date, whereas G3 exhibits no sequence-specific insertion. G1 first recognized a 12-bp sequence at the leader–repeat junction and performed leader-side insertion before proceeding to spacer-side insertion. G2 recognized the full repeat sequence and could perform independent leader-side or spacer-side insertions, although the leader-side insertion was faster than spacer-side. The prespacer morphology requirements for Cas1–Cas2 varied, with G1 stringently requiring a 5-nucleotide 3′ overhang and G2 being able to insert many forms of prespacers with variable efficiencies. These results highlight the intricacy of protein–DNA sequence interactions within the seemingly similar type II-A integration complexes and provide mechanistic insights into prespacer insertion. These interactions can be fine-tuned to expand the Cas1–Cas2 toolset for inserting small DNAs into diverse DNA targets.

3.4.0 - Results

3.4.1 - G1 and G2 integration complexes support integration preferentially into cognate sequences, while G3 does not.

With the knowledge of conserved (**Figure 12A**) leader 3′ ends within type II-A systems, we set out to characterize the role of these DNA motifs using integration reactions. A representative bacterium was chosen from each group and Cas1-Cas2 proteins were purified from each. We chose to use *in vitro* fluorescence-based integration assays using 5′-FAM-labelled prespacer DNA to monitor prespacer integration, as described

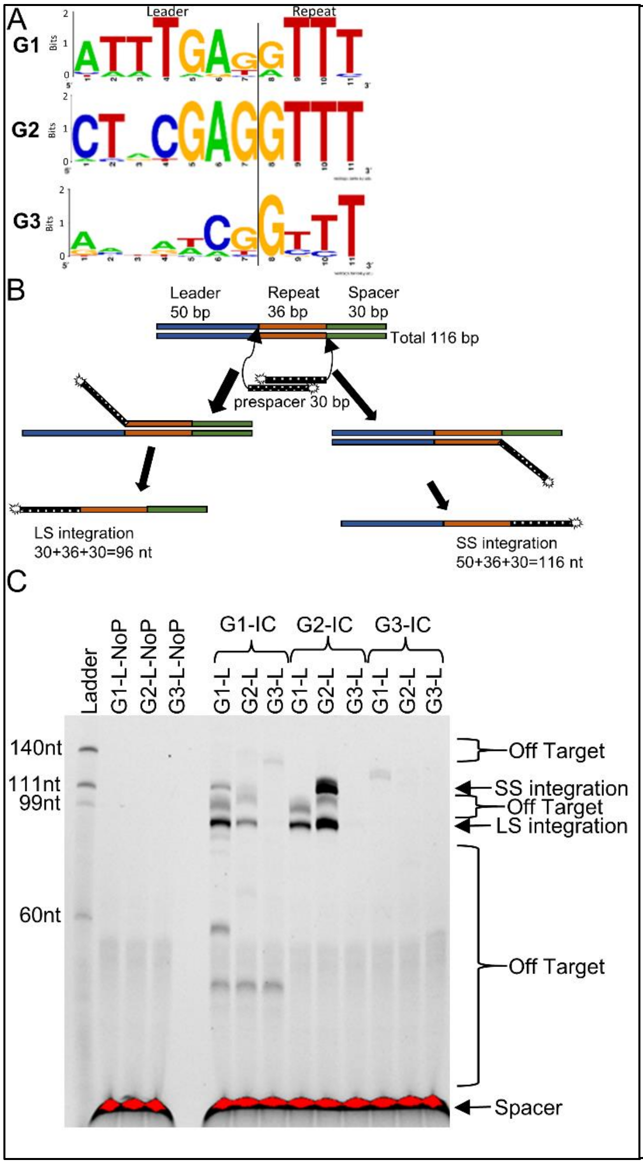


Figure 12: Prespacer DNA integration by the Cas1-Cas2 protein integration complex (IC) into DNA targets that mimic the CRISPR array. A) Sequence logos showing conservation in the last 7 bp of the leader and first 5 bp of the repeat from all three previously identified groups of type II-A CRISPR systems.(3) **B)** Schematic showing integration of FAM labelled prespacer into a 116 bp linear target. Prespacers are integrated either at the leader-side (LS) or spacer-side (SS) of the repeat. After the fluorescent (FAM labelled) strand is separated from the unlabelled strand by denaturing poly-acrylamide gel electrophoresis (PAGE), the LS and SS integration products will appear at 96 nt and 116 nt, respectively. **C)** FAM image of Urea-PAGE integration reactions with combinations of G1-IC, G2-IC, and G3-IC with their cognate linear targets. Total reaction time was 30 minutes. G1-IC integrates both LS and SS into its cognate target, as does G2-IC. G3-IC shows limited activity in the G1-L target. G1-IC and G2-IC can integrate LS into non-cognate targets, if the GAG motif is present at the 3′ leader end. Off-target is when the prespacer integrates somewhere other than LS or SS in the target. No protein (NoP) – IC protein omitted from reaction.

previously (1). Briefly, a 5′-FAM-labelled prespacer with 5 nt 3′ overhangs was mixed with Cas1-Cas2 and a short linear target DNA which contained the cognate leader-repeat region from each representative bacterium. The spacer region in all targets was kept constant to reduce any interference from secondary structure. After a short incubation period, the samples were run on denaturing urea-PAGE to separate individual strands of DNA and images were taken to visualize the FAM label. The leader-side and spacer-side integration reactions cause covalent linkage of one of the strands of the prespacer to the target DNA, which will create distinctly sized ssDNA products on a denaturing gel (**Figure 12B**). The absence of other necessary adaptation proteins (Cas9, Csn2) will result in spacers being inserted in no specific orientation, allowing the sole measurement of intrinsic sequence specificities of Cas1-Cas2 complexes. We used a prespacer having 5 nt 3′ overhangs on both strands unless otherwise noted in the experiments in this study. Divalent metal cations were tested for each Cas1-Cas2 complex, with the metal cation allowing the highest and specific activity being selected for the rest of the study (G1: Mn^{2+}, G2: Mg^{2+}, G3: Mn^{2+}, **Figure S1**).

For a linear target DNA, the two most likely insertion sites can be distinguished on a denaturing gel by following the FAM label (**Figure 12B).** Unincorporated prespacers are also detected at the bottom of each lane. Our results indicate that G1 Cas1-Cas2, which we will call the G1 integration complex (G1-IC), integrates prespacers at both the leader-side and spacer-side of its cognate target sequence (G1-L). We also notice two off-target integration bands, one which is between the leader and spacer-side products (about 99 nt) and another which is around 60 nt (**Figure 12C**). Off-target bands, as will be shown later (**Figure 15**), come from GAG sequence motifs other than the leader-repeat junction that are present in the target sequence. G1-IC also slightly integrates at the leader-side of the cognate G2 target (G2-L, ~10% compared to integration into G1-L along with some off-target integration (~50 nt and ~105 nt)). A very minimal off-target integration occurs with the cognate G3-L (**Figure 12C**). A time course analysis of G1-IC shows faster leader-side integration compared to spacer-side integration (**Figure S2)**. Small amounts of spacer-side and a low molecular weight (~55 nt) off-target band stop increasing in intensity at 5 minutes (**Figure S2**). As time progresses, leader-side integration maximizes at about 10 minutes, with substantially more leader-side than spacer-side integration.

G2 Cas1-Cas2 integration complex (G2-IC) integrates leader-side into both G1-L and G2-L targets, although the efficiency of integration is much higher with the cognate target DNA. For G2-L, there are two bands at the expected sizes corresponding to leader-side (96 nt) and spacer-side (116 nt) integrations. Like G1-IC, G2-IC does not integrate at the spacer-side of non-cognate targets (G1-L and G3-L in this case). Time course analysis of G2-IC (**Figure S3**) shows that at 10 seconds, a prominent leader-side band is present. As time progresses, spacer-side integration increases. Only one off-target band is evident in G2-IC, compared to several off-targets in G1-IC (**Figure S3**)**.** These results also show the robustness of isolated G2-IC compared to isolated G1-IC.

G3 Cas1-Cas2 was promiscuously active in G1-L with integration occurring at a site other than the leader-repeat junction based on the product size (**Figure 12C, Figure S4**). Based on the size of the product in G1-L (~120 nt), the integration can happen either near position 26 in the top strand or near position 90 in the bottom strand (numbering starts at the 5′ end) (**Figure S4**). Since the spacer sequence is conserved in all three linear targets and we do not see this band in all the targets, the most probable integration site is in the top strand. Use of Mn^{2+} instead of Mg^{2+}, shows more off-target integration across multiple targets, compared to a single insertion in G1 -L with Mg^{2+} (**Figure S4**). It is interesting to see an integration complex show no activity against its cognate target sequence but show preferred insertion in another. It should be noted that G3 leader-repeat junction lacks several of the conserved nt based on our bioinformatic analysis (**Fig 1A**).

3.4.2 - Type II-A adaptation shows certain level of cross compatibility between subgroups, but not with a different CRISPR type.

To further understand the integration activities, we designed a variety of mutated target sequences to test the integration tendencies of each protein complex. We started with the three target sequences in **Figure 12C** and added a fourth target DNA containing the leader-repeat region from the type I-E system of *Escherichia coli K-12* (E). We then exchanged the last 7 bp of the leader (at the 3′-end) between the 4 targets in all possible combinations (**Figure 13A**). The exchanged targets and the wild type targets together created 16 different sequence backgrounds to test the integration activity. Integration scores for each group of Cas1-Cas2 proteins are shown in **Figure 13B-C** (gel pictures shown in **Figure S5-S6**). Integration scores are calculated by dividing the intensity of the integration band by the intensity of the same area in the control lane followed by subtracting 1. This value essentially shows the fold change in band intensity compared to background. For G1-IC, leader-side integration is observed in its cognate sequence (G1-L) and in the G2 and G3 backbones where the G1 leader-end motif was present (G2->G1 and G3->G1 respectively). No leader-side integration was observed in any of the E targets, only off-target integration is seen (**Figure S5**). A slight amount of leader-side integration occurred in G2-L, G1->G2, and G3->G2. The G2 leader end motif only differs from G1 by three bp (ATTTGAG in G1 *vs.* CTACGAG in G2), so it is interesting to note that inserting the G2 leader end motif only partially restores leader-side integration activity. This result strongly implies the importance of a continuous 7 bp stretch of DNA at the leader-end for successful leader-side integration (**Figure 12A**). Significant spacer-side integration only occurred in the G1-L target. This shows that spacer-side is much more selective and occurs when a qualifying leader-side integration occurs first. Overall, our data shows that the last 7 bp of the 3′ leader-end helps direct G1-IC to the leader-repeat junction. Absence of leader-side or spacer-side integrations into the E->G1 target, however, shows that more than the 7 bp motif is necessary. It should be noted that there is a strong off-target integration into all four E targets. It is possible that this off-target site is more favourable than the leader-repeat junction of E target, preventing low level insertions at the leader-repeat junction.

G2-IC shows a more robust integration profile into targets containing the cognate upstream leader and repeat sequences. There is a large increase in integration score (4x for LS) for G2-IC compared to G1-IC (**Figure 13C**). Spacer-side integration is present only when the whole G2 backbone is present in the target DNA and appears to be more robust than leader-side integration. This can be substantiated by the fact that two independent insertion events contribute to this intensity (spacer-side from half site and spacer-side from full site integrations), as described previously (61). The occurrence of leader-side integration when G3 targets are mutated to have G2 or G1 leader-end motifs may be attributable to the strong conservation of GAG in both G1 and G2 leader-ends (**Figure 13A**). Similar to G1-IC, G2-IC does not integrate leader-side or spacer-side in E targets, giving evidence that the 3' end of the leader alone is not

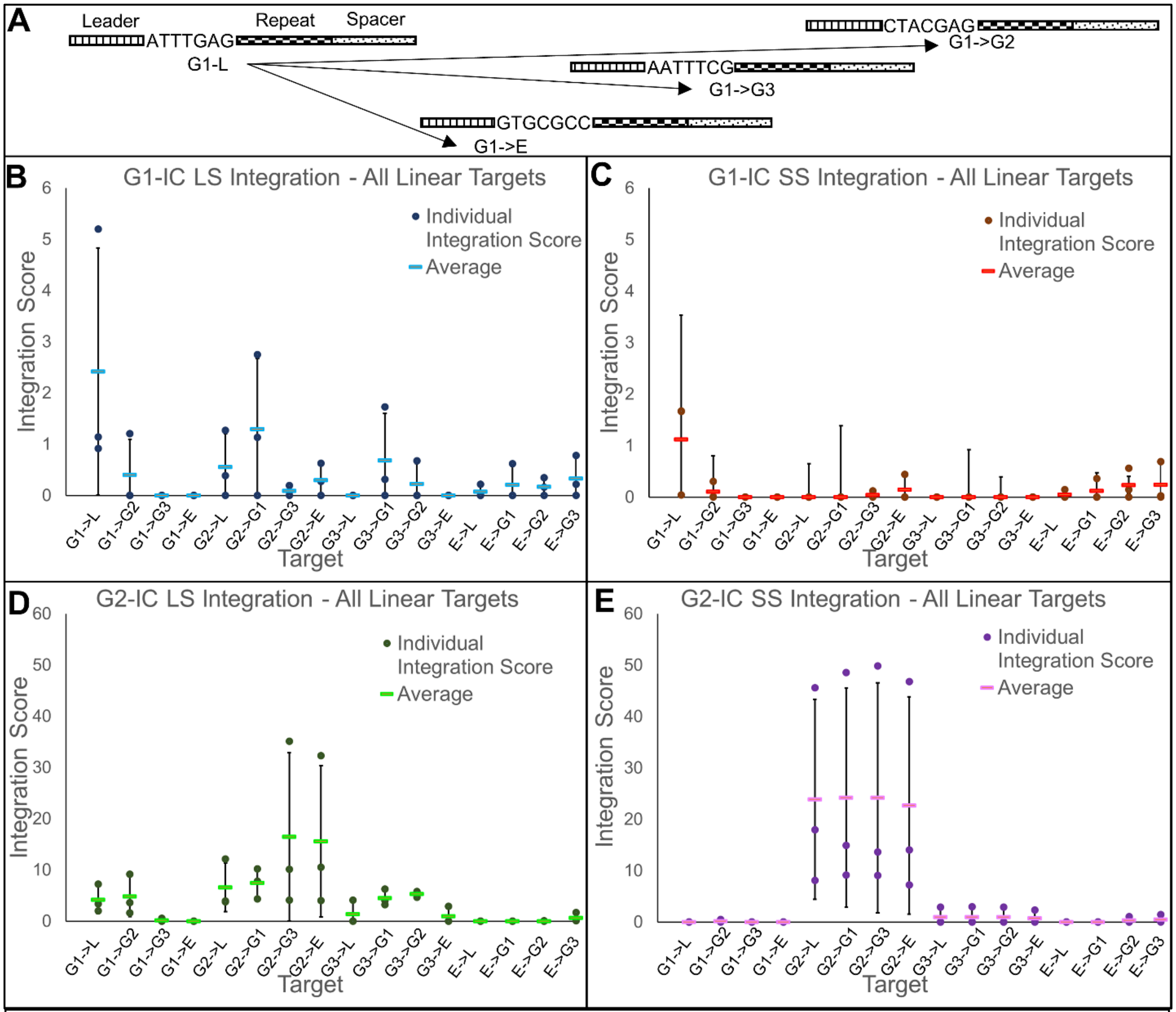


Figure 13: Integration reactions using linear targets with mutated leader 3′ end sequences. A) Schematic of the naming scheme for mutant targets. The naming of the targets shows which group the leader-upstream and repeat sequences came from, followed by which group the 7 bp leader-end motif belongs to (for example G1->G2 consists of the G1 leader-upstream and repeat regions but with the last 7 nt of leader switched with that of G2). **B)** Integration scores of group 1 integration complex (G1-IC) integrating leader-side (LS) into all linear targets. Individual integration scores from each of three replications are shown as dots and the average is shown as a solid bar. **C)** Integration scores of G1-IC integrating spacer-side (SS) into all linear targets. D) Integration scores of the G2 integration complex (G2-IC) integrating LS into all linear targets, and E) G2-IC integrating SS into all linear targets. Solid black error bars (representing standard deviation) are shown.

sufficient. This may be because either the leader-upstream and/or the repeat sequence of type II-A that is absent in the type I-E target design is crucial for prespacer integration. No off-target integration is seen in E targets for G2-IC. One observation is that the intensity of leader side products is higher in non-cognate targets, G2->G3 and G2->E, compared to the cognate G2-L or G2->G1. We attribute this to full site integrations occurring in G2-L and G2->G1, where leader side insertion follows to spacer side insertion. This causes a reduction in the leader side band intensity. Interestingly, this also indicates that only G2-L and G2->G1 can support full site insertion, but not G2->G3 and G2->E.

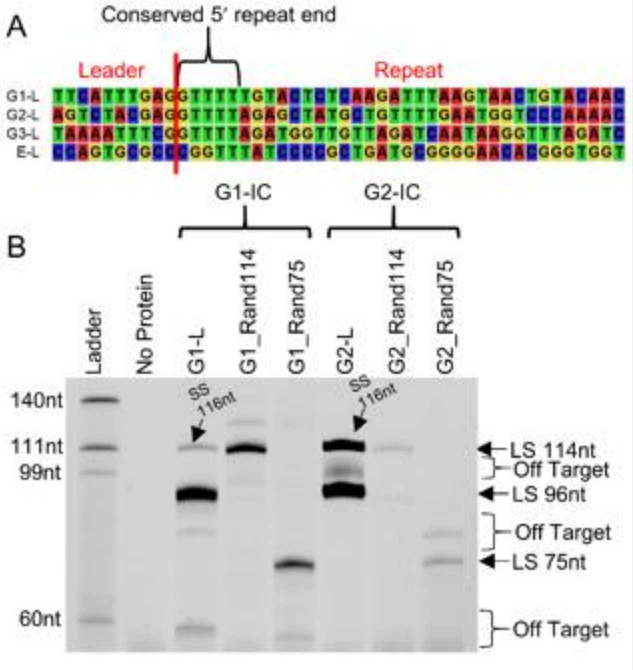

A
Conserved 5′ repeat end
Leader
Repeat
G1-L TTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC
G2-L AGTCTACGAGGTTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC
G3-L TAAAATTTCGGTTTTTAGATGGTTGTTAGATCAATAAGGTTTAGATC
E-L CCAGTGCGCCCGGTTTATCCCCGCTGATGCGGGGAACACGGGTGGT
G1-IC
G2-IC
B
Ladder
No Protein
G1-L
G1_Rand114
G1_Rand75
G2-L
G2_Rand114
G2_Rand75
140nt
111nt
99nt
60nt
SS 116nt
SS 116nt
LS 114nt
Off Target
LS 96nt
Off Target
LS 75nt
Off Target

Figure 14: Testing role of the leader-repeat junction in G1 and G2 integrations. A) Sequence alignment showing the conserved wild-type (WT) leader-repeat region from the three type II-A groups and the *E. coli* CRISPR type I-E system. **B)** FAM image of Urea-PAGE showing G1-IC and G2-IC integrating into linear cognate target as well as randomized linear targets containing the conserved leader-repeat junction at different positions. The G1-L and G2-L integration products are 96 nt (leader-side, LS) and 116 nt (spacer-side, SS), and the Rand114 and Rand75 targets produce LS integration products that are 114 nt and 75 nt, respectively. The results show that while G1 shows reasonable insertion into random DNA (~60%), the ability of G2 to insert into a minimized DNA backbone is significantly reduced.

G3-IC (**Figure S7**) showed only a small amount of off-target integration activity using the G1 derived targets. Since we did not find leader or spacer-side integration products with G3-IC in any of the 16 targets tested, we chose to move forward with only G1-IC and G2-IC for the rest of the study.

3.4.3 - Leader-side integration by G1-IC, but not G2-IC, can be directed by a 12 bp motif

Since no Cas1-Cas2 complex was able to integrate leader or spacer-side into any E targets, we hypothesized that the 5′ end of the type II-A repeat (GTTTT) is a crucial sequence element for prespacer insertion (**Figure 14A**). Even though E targets hold GTTT sequence near the 5′ end of the repeat (two bases into the repeat), they lack GAG conservation at the 3′ end of the leader. (**Figure 14A**). This led us to hypothesize that G1-IC and G2-IC would integrate leader-side using a 12 bp motif mimicking the leader-repeat junction.

To confirm this, we designed randomized DNA targets holding the 12 bp leader-repeat junction. A 104 bp region was randomized by a computer program and was designed to have 50% GC content to avoid other sequence biases (96). To this 104 bp DNA, we inserted the 12 bp leader-repeat junction belonging to either the G1 or G2 CRISPR locus. This 12 bp leader-repeat junction consisted of the last 7 bp of the leader and the first 5 bp of the repeat (**Figure S8**). The 12 bp motif was inserted at two distinct positions in the 104 nt backbone separately, which would result in a 114 nt or a 75 nt product if integration occurred at the leader-repeat junction. Our results show leader-side bands appearing at the predicted sizes only in G1-IC (**Figure 14B)**. G1-IC is efficient in integrating after solely recognizing the 12 bp motif, since the band intensities are comparable to the leader-side integration into the cognate backbone. Slight reduction in these band intensities shows that other sequence elements may be necessary for full insertion activity. G2-IC shows very minimal integration using the 12 bp motif, giving evidence that there is a difference in recognition mechanisms between the two type II-A groups.

3.4.4 - Order of prespacer insertion varies between G1 and G2 systems.

Full site integration is the result of a single prespacer being integrated at both ends into a single target (**Figure S9**). To visualize full site integration, we employed the use of hairpin targets as previously described (61). The hairpin targets used in this study will create a 174 nt leader-side and a 78 nt spacer-side product for half site insertions. A full site integration will create two bands: a 126 nt leader-side product and a 78 nt spacer-side product. G1-IC integrates at all three locations in a cognate hairpin target, as previously shown (97) (**Figure 15**). Leader-side integration was the most intense band, followed by spacer-side and then full site. On a G2 hairpin target (G2-HP), G1-IC integrates mainly at the leader-side, along with some minor off-target integration events (**Figure 15**). Removing the 3′ OH groups from one of the strands of the prespacer yields only a leader-side product, while removing

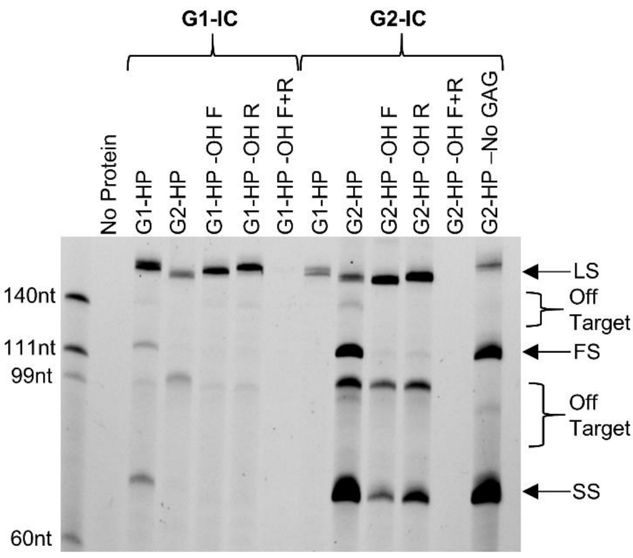


Figure 15: FAM image of Urea-PAGE showing G1-IC and G2-IC integrating prespacers into hairpin (HP) targets to observe full site (FS) integration. Both protein complexes integrate FS only into their cognate target. Removing the 3′ OH abolishes FS integration in both cases as well, showing that the FS band is the product of a single prespacer being inserted at both ends and not two independent integration events. Removing the 3′ OH also abolishes SS integration in the case of G1-IC, but not in G2-IC, which shows a difference in mechanism between the two. Removing all GAGs outside the leader-repeat junction abolishes off-target bands generated by G2-IC (lane: G2-HP-No GAG). The slight variations in the band positions in -OH F and -OH R lanes is most probably due to remaining secondary structures affecting mobility on a gel. [hydroxyl group removed from spacer_Sy_overhangs_5-F (Supporting Table 1, -OH F), hydroxyl group removed from spacer_Sy_overhangs_5-R (Supporting Table 1, -OH R), hydroxyl group removed from both these strands (-OH F+R)].

both 3′ OH groups abolishes all insertions. The abolishment of spacer-side integration in G1-IC when only one 3′ OH group is present shows that leader-side integration is a prerequisite for spacer-side integration.

G2-IC can perform full site integration on cognate target DNA (**Figure 15**) as shown previously (61). Several off-target bands are present in the cognate target lanes. These off-target bands were abolished by changing GAG sequences at positions 19 and 70-74 of the G2-HP target to GCG. When the prespacer possessed only one 3′ OH group, both leader and spacer-side insertions are seen, showing that leader-side and spacer-side integration events can occur independently in this group. A small amount of the full site integration band in lanes where only a single 3′ OH was available shows how often a single target will have two separate prespacers integrating into leader and spacer-sides independently. As expected, there is no integration when 3′ OH is removed from both ends of the spacer. These results indicate an important difference between G1 and G2 integration complexes: G1-IC follows an order while inserting prespacers, G2-IC does not.

3.4.5 - Minimal DNA elements required for prespacer insertions vary for G1 and G2 integration complexes.

To define the minimal sequence requirements for the different insertion events (half site *vs.* full site) in G1 and G2 systems, we performed experiments using randomized hairpin DNA targets. The randomized hairpin target was inserted with either the 36 bp repeat, the 36 bp repeat plus 4 bp of the 3′ leader end, or the 36 bp repeat plus 7 bp of the 3′ leader end (**Figure 16A**). For G1-IC, integration bands are present in all the three expected positions (leader-side half-site – 173 nt, spacer-side half-site – 77 nt, and full site – 126 nt) for targets containing at least the full repeat and 4 bp of the leader 3′ end, with leader-side integration being more prominent (**Figure 16B**). The repeat alone

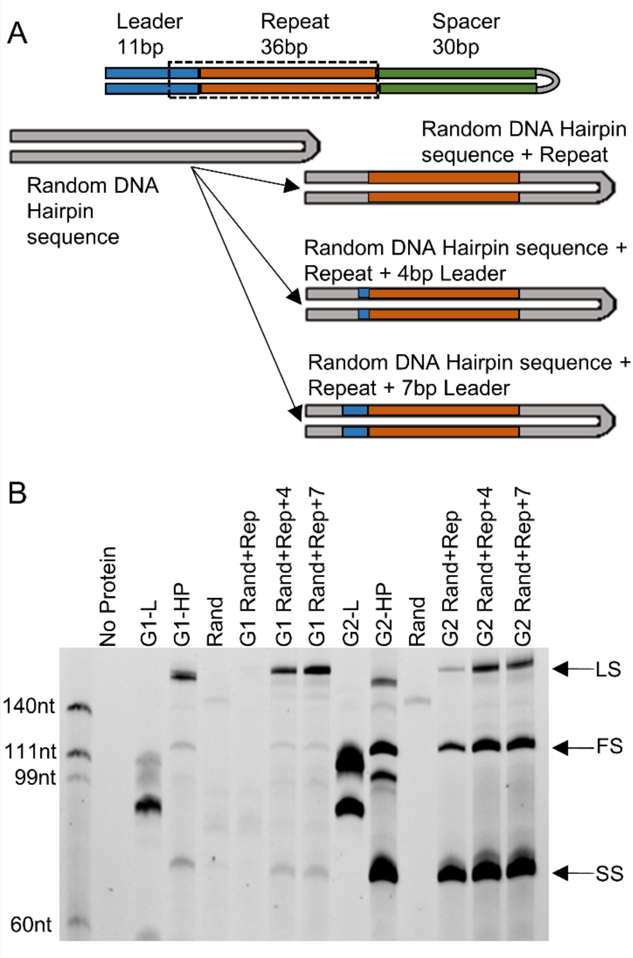


Figure 16: Investigating the role of the whole repeat for integrating prespacers by G1 and G2 into hairpin (HP) targets. A) Construction of the randomized DNA hairpin target (Rand), Rand target holding a 36 nt long repeat belonging to G1 or G2 (Rep), Rep holding a 4 bp region at the 3′ end of the leader (Rep+4), or Rep holding a 7 bp region at the 3′ end of the leader (Rep+7). **B)** FAM image of Urea-PAGE showing G1-IC and G2-IC integrating into cognate linear (-L), cognate hairpin (-HP), and random hairpin (Rand) targets. HP targets produce a full site (FS) band in addition to the leader-side and spacer-side (SS) bands produced by the linear targets. Neither complex integrated significantly into the random DNA target that is devoid of any CRISPR elements. For G1-IC, addition of the repeat and 4 bp of the leader was necessary for LS, SS and FS integration to be present. For G2-IC, only the repeat sequence was necessary. Addition of just 4 bp of the leader increased the activity back to cognate amounts. While Rand and Rand+Rep produced no integration activity in G1-IC, Rand+Rep showed significant activity in G2-IC, showing the differences in the DNA sequence elements required for prespacer insertion in G1 and G2.

showed no integration compared to the G1-HP target or Rand+Rep+4 target. This result is interesting because the G1 Rand+Rep and the G1 Rand+Rep+4 targets (**Figure 16B**) only differ by 4 bp and thus demonstrates the essentiality of the last 4 bp of the leader region in G1 systems. Interestingly, our randomized linear target DNA experiments (**Figure 14**) showed that a 12 bp leader-repeat junction was sufficient to insert prespacer at the leader-side. Thus, the combined results of both linear and hairpin target DNAs show a preference for leader-repeat junction for leader-side insertion, with the necessity of the full repeat for spacer-side insertion by G1-IC.

G2-IC integrates in all three locations when just the repeat is present. Interestingly, repeat alone can drive full site insertion efficiently, since the leader-side band intensity is very low, indicating efficient conversion of leader-side insertions to full site events. Addition of 4 bp of the leader brings the intensity of full site integration back to a comparable amount to G2-HP, showing that the leader end plays some role in integration efficiency but is not a requirement for full site integration. Another interesting observation is that the spacer-side integration intensity is similar across repeat alone, repeat+4 bp, or repeat+7 bp in the case of G2-IC. This again strongly points to the independence in the leader-side and spacer-side insertion in G2, which is distinct from G1. Addition of 7 bp of the leader made no further improvement over the 4 bp leader end in both G1 and G2 systems (**Figure 16B**).

3.4.6 - G1-IC has specific prespacer requirements, while G2-IC is more tolerant.

All the integration assays discussed so far in this work used a prespacer containing 5 nt symmetrical 3′ overhangs on each side as described in a previous study (97). We hypothesized that G1-IC and G2-IC would be capable of integrating prespacers with different types of DNA ends. We created a library of different prespacers mimicking different processing options. The first set of prespacers were made by reducing the length of the spacer_30-F strand (**Supporting Table S5**) from both the 5′ and 3′ ends, simultaneously, before annealing to the spacer_30-R strand (**Figure S10**). These “non-symmetrical” prespacers ranged from 0 to 5 nt overhangs at the 5′ and 3′ ends with all possible combinations, totalling 36 different prespacers. It is important to note that 5′ overhangs in the non-symmetrical prespacers shorten the distance between 3′ OH groups, ranging from 25 nt to 30 nt. In the next set, we reduced the length from the 5′ ends of both the strands of the prespacer, creating different lengths of 3′ overhangs. These “symmetrical” prespacers ranged from 3 to 7 nt long overhangs and maintain the 30 nt distance between 3′ OH groups. We also designed prespacers with varying amounts of splaying, from 0 to 7 nt. The results of G1-IC and G2-IC integrating this unique prespacer library into cognate hairpin targets can be seen in **Figures 17A-B**. G1-IC is very selective about which prespacer is being inserted. Only the symmetrical prespacers with 4 nt and 5 nt 3′ overhangs support full site integration. Looking at the non-symmetrical prespacers, a 3 or 4 nt 3′overhang can result in weak leader-side

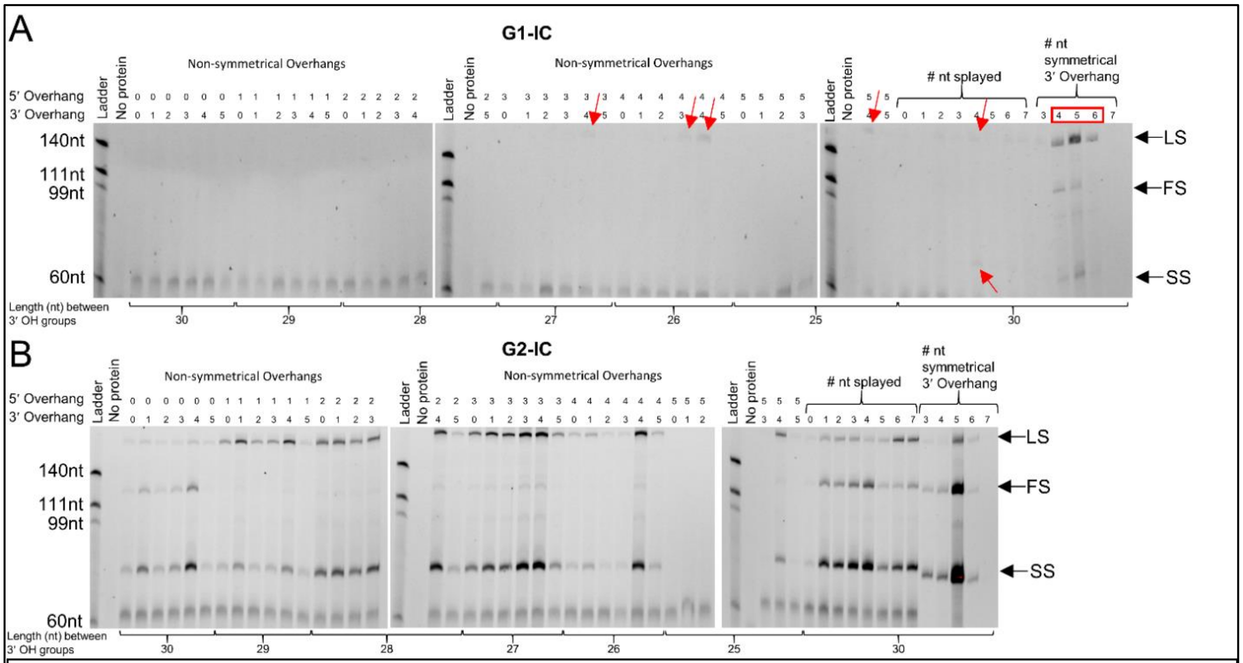

A
G1-IC
Non-symmetrical Overhangs
Non-symmetrical Overhangs
Ladder
No protein
5′ Overhang
3′ Overhang
# nt splayed
# nt symmetrical 3′ Overhang
140nt
111nt
99nt
60nt
LS
FS
SS
Length (nt) between 3′ OH groups
30
29
28
27
26
25
30
B
G2-IC
Non-symmetrical Overhangs
Non-symmetrical Overhangs
Ladder
No protein
5′ Overhang
3′ Overhang
# nt splayed
# nt symmetrical 3′ Overhang
140nt
111nt
99nt
60nt
LS
FS
SS
Length (nt) between 3′ OH groups
30
29
28
27
26
25
30

Figure 17: Investigating prespacer preferences of integration complexes (IC) using integration assays with cognate hairpin targets. A) FAM image of Urea-PAGE showing G1-IC integrating into a cognate hairpin target using various forms of prespacers. The ability of G1-IC to process or deform prespacers for integration is very limited, and only certain prespacer forms are acceptable. 5′ and 3′ 4 nt non-symmetrical overhangs resulted in a small amount of leader-side (LS) integration, as did 4 nt of splaying. 4 nt and 5 nt of symmetrical 3′ overhangs were the only two prespacers to be integrated full site (FS), with 5 nt showing more activity. Red boxes indicate where integration activity was seen. Red arrows indicate positions of faint integration bands. **B)** FAM image showing G2-IC integrating into a cognate hairpin target using various forms of prespacers. Many forms of prespacers were acceptable for integration. In general, a 4 nt 3′ overhang resulted in the best possible integration among non-symmetrical prespacers. Also, decreasing FS integration by using a non-optimal prespacer increases LS and SS integration, showing that non-optimal prespacers can still integrate LS or SS but have trouble proceeding to FS (lanes for splayed 5, 6, and 7). A symmetrical 5 nt 3′ overhang shows the best integration. This shows that G2-IC is more tolerant to different prespacer forms.

insertions (**Figure 17A,** red arrows). The 4 nt splayed prespacer also showed a small amount of leader-side and spacer-side integration. Other studies in G1 and G2 systems have also shown that 3′ overhangs improve prespacer integration activity, indicating structural similarity amongst Cas1-Cas2 complexes (1, 61, 97).

G2-IC supported integration from various forms of prespacers (**Figure 17B**). Most notably, a 4 nt 3′ overhang was generally the best performing prespacer amongst the 36 non-symmetrical spacers and the splayed spacers. The 5 nt 3′ overhang symmetrical spacer, however, outperformed all of them. 5′ overhangs of 1 to 5 nt didn’t support full site integration activity when compared to 0 nt. This may either be due to the 5′ overhangs interfering with complex formation, with the ideal placement of the 3′ OH group in the active site of the complex, or due to less than 30 nt between the two 3′ OH groups of the prespacer. Less than 30-nt between the two 3′ OH groups of the prespacer was shown to reduce integration in previous studies (1, 2, 61, 97). The fact that splayed prespacers possess 30-nt suggests that in addition to the distance between the two 3′ OH groups, 5′ overhangs may inhibit integration as well. A very important finding from using this repertoire of prespacers is that G2-IC is more tolerant and can insert anywhere from a fully double stranded DNA to DNAs with different overhangs, and even those with less than 30-nt in between the two 3′-OH groups. Direct mass measurements using SEC-MALS of G1-IC mixed with a 4 nt splayed prespacer resulted in no complex formation (**Figure S11**), indicating that incorrect prespacer processing hinders integration complex formation. Interestingly, G2-IC does form a complex with the 4 nt splayed prespacer, showing G1-IC is likely not active with many forms of spacers due to an inability to bind rather than being catalytically inactive (**Figure S11**). These results suggest that G2-IC has a more pliable active site/DNA binding region that can accommodate a wide range of prespacers, whereas G1-IC is more restrictive in inserting non-ideal spacers.

3.5.0 - Discussion

3.5.1 - Characterization of multiple group members establishes conservation of mechanisms within groups.

Previous studies used proteins from type II-A CRISPR systems in the *S. thermophilus* DGCC7710, *S. pyogenes* M1GAS and *Enterococcus faecalis* genomes, which belong to the G1, G2, and G2 systems from our analysis, respectively (1, 61, 97). Identities and similarities between the proteins used in our study and the proteins in these previous studies can be seen in **Figure S14**. A comparison of our results with *in vitro* results from the other studies show a mechanistic distinction of prespacer insertion, based on the type II-A groupings that we identified through the bioinformatic work (3). Specifically, the G1 results show a strong dependence on leader-repeat junction for leader-side insertion, followed by a full repeat requirement for spacer-side insertion with an order in prespacer insertion. The G2 loci show dependence on the full repeat sequence for both leader and spacer-side insertions, which are not interdependent. Thus, our study establishes that type II-A systems use different rules for prespacer insertion, based on the members characterized so far, and that these rules have coevolved with leader-repeat and Cas1-Cas2 sequence conservations (3).

Previously published results of an *in vivo* test in a G1 CRISPR system from *S. thermophilus* DGCC7710 showed that the last 10 bp of the leader and the entire 36 bp of the first repeat were enough to direct full site integration *in vivo*, irrespective of whether these DNA sequences were present in or outside of the CRISPR array (40). Our results further extend this finding that, under *in vitro* conditions, a 12 bp leader-repeat junction can promote leader-side integration. Experiments with randomized DNA has further reduced this requirement to a 4 bp 3′ leader end and a full repeat for efficient full site integration. The overall efficiencies of all three integration events (leader-side, spacer-side and full site) were comparable between repeat+4 bp and the full cognate leader-repeat regions (80%, 70%, and 60% respectively). Interestingly, the *in vivo* study in *S. thermophilus* DGCC7710 showed that its G1 system is more efficient in taking spacers compared to its G2 system co-existing in the same bacterium (40). Comparing adaptation efficiency of an isolated Cas1-Cas2 complex from a G1 system from *S. thermophilus CNRZ1066* and a G2 system from *S. pyogenes A20* showed that G2 is more robust than G1 in prespacer insertion. These facts point to inherent differences in prespacer insertion between G1 and G2 proteins. There is one amino acid difference between the G1 Cas1 (Q171K) and Cas2 (I64M) proteins in our present study (*S. thermophilus CNRZ1066*) and the previous *in vivo* study (*S. thermophilus DGCC7710*). We introduced amino acid substitutions in *S. thermophilus* CNRZ1066 G1 Cas1 and Cas2 to produce a 100% amino acid match with that of *S. thermophilus* DGCC7710 G1 system. Integration assays using hairpin targets and the protein variants showed a slightly better efficiency in using different prespacers compared to native proteins, but with significantly lower efficiency in using prespacers with different morphologies similar to the G1 system that was tested in the present study (**Figure S15**). These results support inherent differences in spacer acquisition by different type II-A subgroups.

3.5.2 - Steps in prespacer insertion vary between type II-A groups

Our G2-IC can perform independent leader-side and spacer-side insertions at equal efficiencies, but only a cognate leader-side insertion proceeds to full site insertion (**Figure 15, Figure 16, Figure S2, Figure S3**). The higher intensities for spacer-side insertion for G2-IC lanes (**Figure 15**) is supported by previous studies where leader-side recognition was shown to be essential to perform a full site integration in order to maintain site specificity during prespacer insertions (61). A spacer-side integration, without a leader-side was proposed to induce abortive integrations (61). Thus, in our results, when leader-side integration moves forward to full site integration, it increases the intensity for spacer-side integration band. Our results also shows that when insertion occurs at non-cognate sites that retain some similarities to the cognate leader-repeat junction (G2->G3 and G2->E in **Figure 13C**) it does not proceed to full site insertion, as indicated by the higher intensity for the leader-side insertion band.

In contrast, our G1-IC shows that there is a preference for leader-side insertion as demonstrated by a single 3′-OH prespacer, where insertion is strictly restricted to leader-side, compared to equal insertion events in G2-IC (**Figure 15**). This implicates an inherent, stringent order in the integration events in G1, at least under *in vitro* conditions.

3.5.3 - Absence of conserved leader-repeat junction in G3 may be physiologically relevant

Interestingly, G3 systems have several anomalies compared to G1 and G2 systems based on our bioinformatics analysis (3). There is only a 2 nt 3′ leader end conservation in G3, which differs greatly from the 7 nt conservation found in G1 and G2 (**Figure 12A**). This is significant since 18/25 members of G3 in our previous study belonged to lactobacilli, whereas there is a wider distribution of genera in G1 (~fifty). Physiologically, lactobacilli harbour temperate phages that are maintained in the bacterial genome (71) and may have evolved to not allow spacer uptake to prevent autoimmunity (98). This indicates that the inability to insert a prespacer into the cognate DNA backbone may be the result of an evolutionary event that is beneficial to the bacterium. It is also possible that the inability of G3-IC to perform full site integration *in vitro* may be the result of a missing *in vivo* factor. Further studies are required to conclusively analyse these different scenarios.

3.5.4 - Phylogenetically distinct prespacer insertion mechanisms provide a wide range of applicability.

Our results establish that we cannot generalize the rules for prespacer insertion even within closely related CRISPR subtypes. Even within type II-A, we find two distinct mechanisms (**Figure 18**): G1 with leader-side recognition followed by full site integration (**Figure 15**), G2 with repeat recognition followed by fast leader-side integration and slower spacer-side integration (**Figure 16, Figure S6**), and G3 with promiscuous integration, possibly arising from a missing cellular factor or from a dysfunctional mutation in the protein or DNA sequence elements (**Figure S7**). Thus, even within closely related CRISPR-types, there may be an interplay of cellular factors or mechanistic differences that fine tune the integration reactions. Further studies are essential in deciphering different mechanisms available in nature for prespacer insertion. Interestingly, with the current analyses covering our present study and previously published results (1, 61, 97), we show mechanistic separations that mirror the leader-repeat and Cas1-Cas2 sequence conservations (3).

Potential industrial applications exist for Cas1-Cas2, such as gene tagging, transcriptional recordings, and storage of digital information (73, 74). A better understanding of the functions of Cas proteins in adaptation will make manipulation of the system a possibility, whether that be escalation for better protection of industrially relevant bacteria or combating bacterial pathogens by increasing sensitivity to phages. The results presented here open possibilities of developing several distinct sets of DNA integration tools to match different targeting requirements.

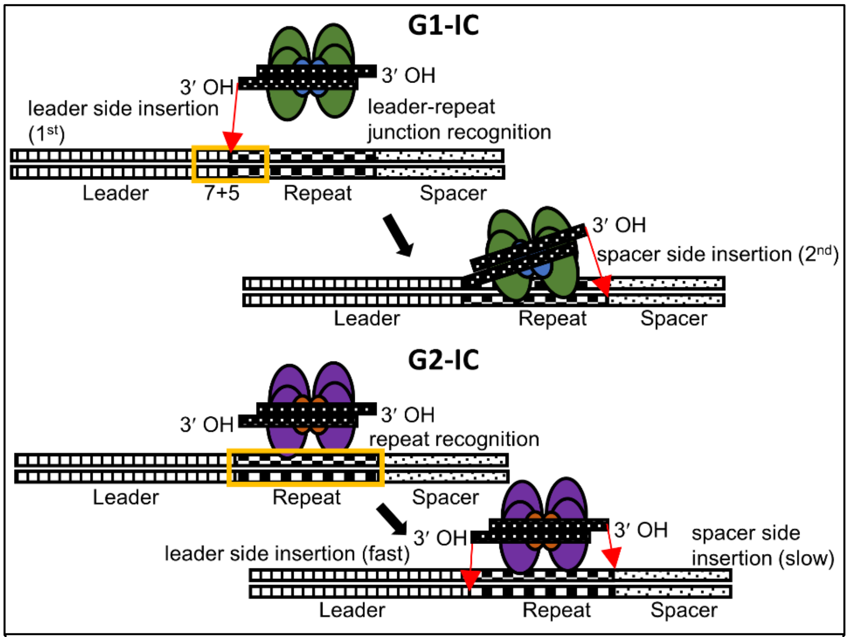

G1-IC
3′ OH
3′ OH
leader side insertion (1st)
leader-repeat junction recognition
Leader
7+5
Repeat
Spacer
3′ OH
spacer side insertion (2nd)
Leader
Repeat
Spacer
G2-IC
3′ OH
3′ OH
repeat recognition
Leader
Repeat
Spacer
3′ OH
3′ OH
leader side insertion (fast)
spacer side insertion (slow)
Leader
Repeat
Spacer

Figure 18: Schematic of the insertion mechanisms presented for G1-IC and G2-IC. G1-IC recognizes a 12 bp sequence at the leader-repeat junction and inserts leader-side first. This is followed by insertion at the spacer-side using a ruler mechanism to define the length. G2-IC recognizes sequences in the repeat. Insertion at the leader-side and spacer-side can happen independently, with leader-side being much faster than spacer-side. Even though the minimal sequence requirement for all prespacer insertions (LS, SS, FS) is a cognate repeat, the efficiency of the process slightly increases (~50%) by adding just 4 bp at the 3′ leader end. This supports the strong 3′ leader end conservation for this group, despite the minimal requirement of a repeat sequence for fulfilling the mechanistic requirements.

Chapter 4: Biophysical characterization of CRISPR adaptation proteins

4.1.0 - Acknowledgments

I would like to thank Dr. Leonard Thomas and Dr. Kesavan Babu for their direct contributions to the crystallography work shown here. Dr. Thomas for his help in screening setup, advice on crystallization, and structural solution work; Dr. Babu for his help in remote data collection and processing. My contributions to this work include conceptualization, cloning and purifying all the proteins, SEC-MALS data collection, screening and optimization crystallization tray setup, and crystal data analysis. Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health, National Institute of General Medical Sciences (P41GM103393). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH.

4.2.0 - Biophysical characteristics of Cas1-Cas2

The Cas1-Cas2 complex, shown in **Figure 19**, consists of two Cas1 dimers sandwiching a Cas2 dimer. Prespacer DNA binds along the long edge and is held by non-specific backbone interactions with various amino acids in two of the Cas1 and the two Cas2 copies. Conserved histidine residues found along the prespacer binding edge of the complex disrupt the base stacking interactions of the double stranded DNA and allow for strand separation to occur. This same effect is seen with tyrosine residues in type I Cas1-Cas2 complexes (2). The relative spacing of the strand-separating residues determines the preferred prespacer length for integration (1, 2). The role of Cas2 in this complex is as a scaffolding protein, as abolishment of its active site had no effect on integration activity (99).

Various structures of CRISPR adaptation proteins (Cas1 and Cas2) are available in the PDB (100). These structures consist of 11 apo-Cas1, 2 apo-Cas2, 19 different Cas1-Cas2 complex with some Cas1-Cas2 bound to different DNAs mimicking either a prespacer or target DNA holding the leader sequence (100). Available structures of Cas1-Cas2 span both type I and type II systems. Even with this abundant structural information, there are still gaps in our knowledge about the molecular mechanisms of type- or group-specific features driving adaptation. For example, all type II Cas1-Cas2 structures belong to type II-A G2 subgroup (details regarding distinction of the different subgroups are presented in chapters 2 and 3). All the Cas1-Cas2 structures solved show a $Cas1_4$-$Cas2_2$ stoichiometry, implying that this is most likely conserved across

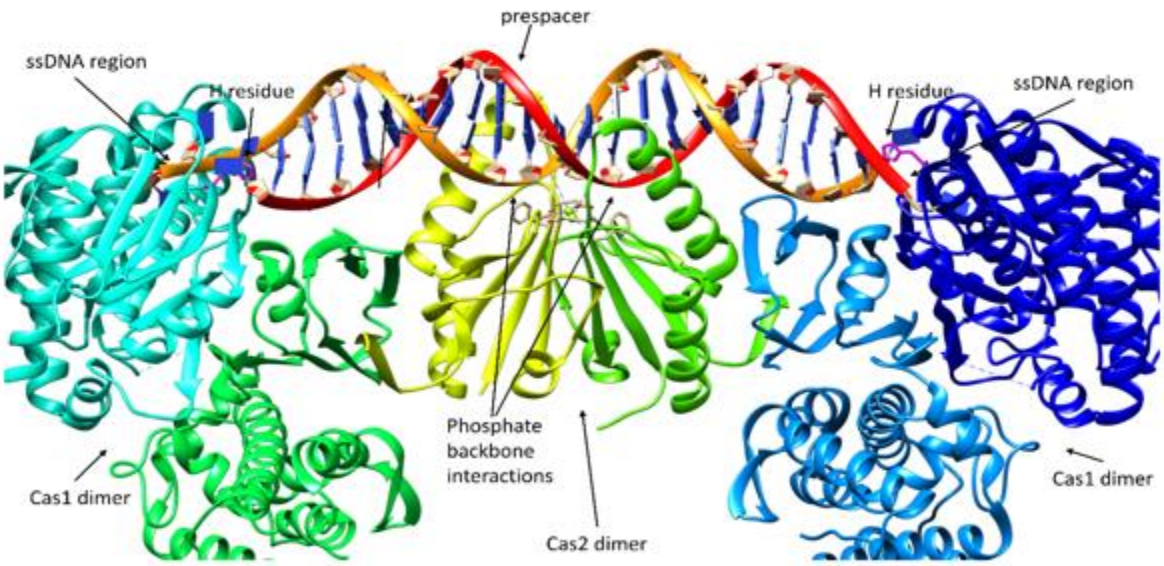


Figure 19: Structural features of G2 Cas1-Cas2-prespacer. Overall structure and important features are labeled. Histidine residues which interrupt the base stacking interactions of the double-stranded prespacer are shown in magenta. PDB id 5XVN (1).

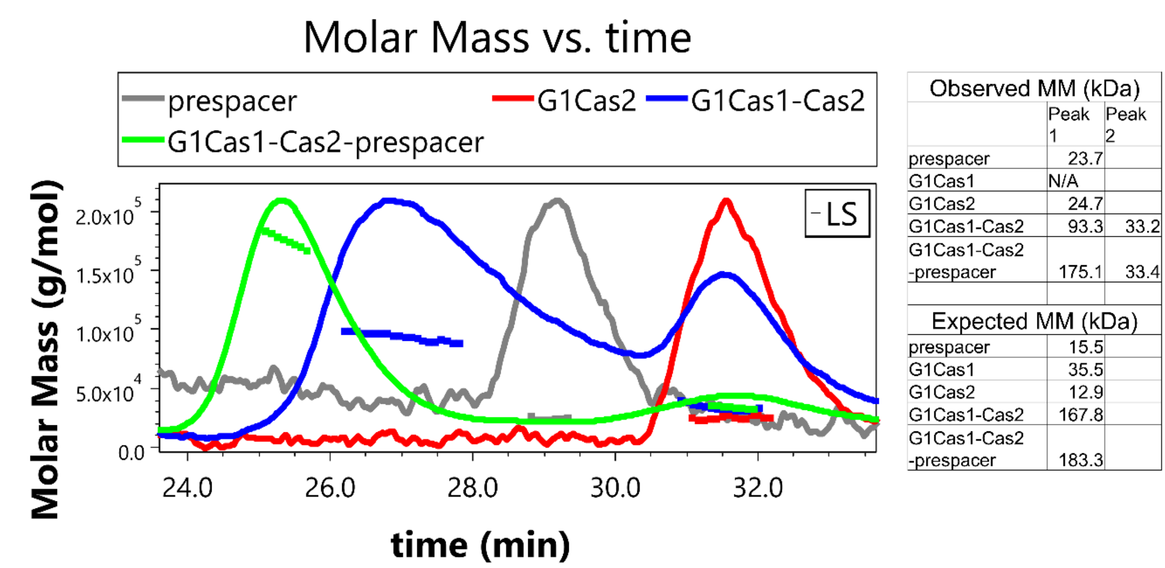
different CRISPR types. Exceptions to this rule do exist, such as the Cas1 tetramer found in a type V system capable of site specific integration and the Cas1-reverse transcriptase fusion protein from a type III systems which integrates spacers taken from RNA (36, 74).

As stated previously, bioinformatics and biochemical characterization shows clear differences in the mechanisms even between seemingly similar type II-A subgroups.

From available structures, several interactions have been identified in Cas1-Cas2 integration complexes from G2 systems which may explain the intrinsic sequence specificity shown by this group (1, 65, 66). **Figure 24B** shows key hydrogen bonding interactions between four amino acids found in the Leader Recognition Helix (LRH) of Cas1 and the bases of the leader DNA. This LRH is not present in Cas1-Cas2 complexes in type I CRISPR systems. This presents an intriguing hypothesis, where the LRH region and its interactions with the target DNA may differ between the different sub-groups accounting for the differences in the mechanisms of prespacer insertions as was observed in our biochemical analyses (Chapter 3). Our goal is to determine the molecular differences that account for group-specific prespacer insertion. An advantage of identifying these interactions is that mutating the LRH may result in altered specificities for Cas1-Cas2 insertion, which will enable the development of biotechnological tools for site-specific tagging. To derive sub-group specific molecular mechanisms, protein complex formation of Cas1-Cas2 complexes was characterized from all three groups and the role of prespacer morphology for proper complex formation was determined. We also attempted to crystalize and solve the structure of G3 Cas1, which shows no intrinsic sequence specificity during prespacer insertion and may provide key details into interactions governing specificity. In addition, protein-DNA complex crystallography with G1 and G3 sub-groups was initiated.

4.3.0 - Complex formation analysis of various Cas1-Cas2 complexes using SEC-MALS

We first tested whether canonical proteins from G1, G2, and G3 formed Cas1-Cas2 complexes as well as canonical protein-prespacer complex. A prespacer having symmetrical 5 nt 3′ overhangs were used for this analysis. We analyzed complex formation and calculated stoichiometries using size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS).



Observed MM (kDa)	Peak 1	Peak 2
prespacer	23.7	
G1Cas1	N/A	
G1Cas2	24.7	
G1Cas1-Cas2	93.3	33.2
G1Cas1-Cas2 -prespacer	175.1	33.4

Expected MM (kDa)	
prespacer	15.5
G1Cas1	35.5
G1Cas2	12.9
G1Cas1-Cas2	167.8
G1Cas1-Cas2 -prespacer	183.3

Figure 20: Molecular weight analysis of G1 Cas1-Cas2-prespacer. Chromatograms showing light scattering signal and molar mass calculations vs. time. The table shows observed average molar mass calculations of each peak and actual molar masses of selected components.

4.3.1 - Complex formation analysis of G1 Cas1-Cas2 with prespacer DNA

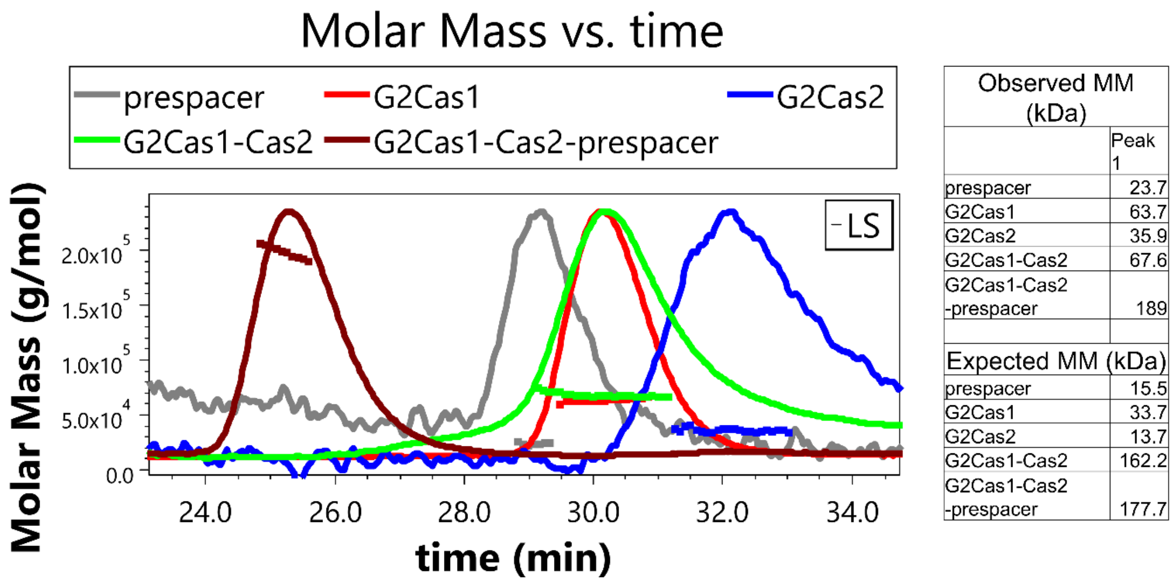
G1 Cas1, G1 Cas2, and 5 nt 3' OH overhang prespacer were subjected to SEC either individually or combined before being measured by the light scattering detector. Light scattering signal can be seen in the chromatogram in **Figure 20**, with absolute molar mass calculations shown as small squares. G1 Cas1 was not able to be tested in the apo-form because of solubility issues in the absence of binding partners. Interestingly, G1 Cas1-Cas2 formed a smaller than expected complex with a molar mass of 93.3 kDa. The expected molar mass for the Cas1-Cas2 complex with a stoichiometry of 4:2 is 168 kDa, and hence the 93 kDa complex is most probably an association of Cas1-Cas2 proteins in a non-canonical stoichiometry. Expected molar masses matched up well with observed values G1 Cas2 alone (which forms a dimer), and the full G1 Cas1-Cas2-prespacer complex. This confirmed the expected stoichiometry of the Cas1-Cas2-prespacer complex to be 4:2:1. Analysis of the fraction for Cas1-Cas2-prespacer complex showed that both protein and prespacer were present in the fraction (**Figure S12**). These results indicate that the prespacer DNA was necessary for Cas1-Cas2 to interact together in the canonical 4:2 ratio and form a stable complex, which is different from type I-E system where the proteins can associate and form a stable complex in a 4:2 ratio even in the absence of a prespacer (99).

4.3.2 - Complex formation analysis of G2 Cas1-Cas2 with prespacer DNA

Similar to G1 proteins, G2 Cas1-Cas2 was not able to associate into a complex without the presence of prespacer DNA (**Figure 21**). Interestingly, the peak similar to the 93 kDa complex, which represents non-canonical complex formation between Cas1 and Cas2, did not form with the G2 proteins. This is evident by the presence of a clear Cas1 alone peak in the G2-Cas1-Cas2 chromatogram. G2 Cas1 alone matched up well to a dimer molar mass while G2 Cas2 alone was closer to a trimer molar mass. The full complex was confirmed to be a 4:2:1 stoichiometry with fractions showing both proteins and prespacer on appropriate gels (**Figure S12**). Thus, G2 systems are different from G1 in that Cas1-Cas2 truly assemble only in the presence of prespacer since the non-canonical Cas1-Cas2 complex formation does not occur in this group.

4.3.3 - Complex formation analysis of G3 Cas1-Cas2 with prespacer DNA

G3 proteins behaved similarly to G1 Cas1-Cas2-prespacer (**Figure 22)**. Lacking the prespacer, G3 Cas1-Cas2 formed an averaged 94.7 kDa complex, showing the proteins have some non-canonical stoichiometric association with each other in the absence of the prespacer. Full complex formation with the canonical 4:2:1 stoichiometry required the presence of DNA. The slightly increased observed molar mass may indicate a second DNA molecule associating with the complex. Acceptable molar mass measurements are seen for the prespacer, a Cas1 dimer, and a Cas2 dimer.



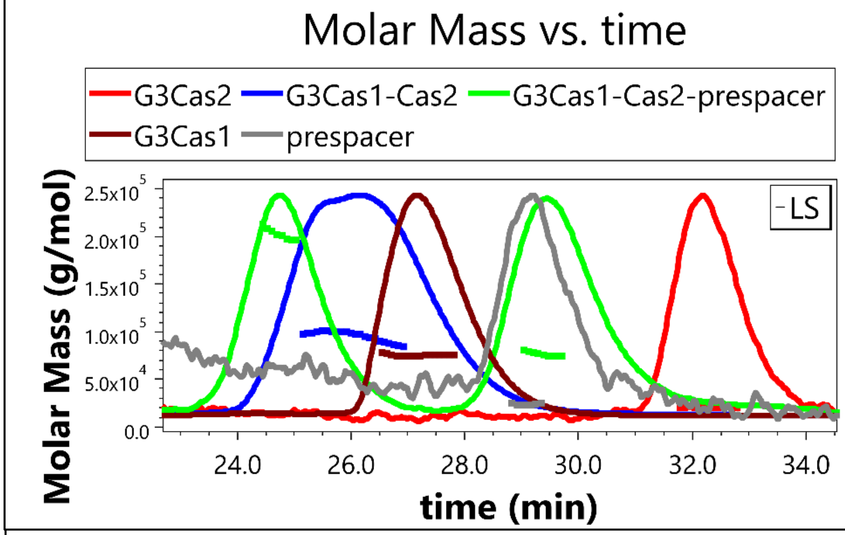
Observed MM (kDa)	
	Peak 1
prespacer	23.7
G2Cas1	63.7
G2Cas2	35.9
G2Cas1-Cas2	67.6
G2Cas1-Cas2 -prespacer	189

Expected MM (kDa)	
prespacer	15.5
G2Cas1	33.7
G2Cas2	13.7
G2Cas1-Cas2	162.2
G2Cas1-Cas2 -prespacer	177.7

Figure 21: Molar mass analysis of G2 Cas1-Cas2-prespacer. Chromatograms showing light scattering signal and molar mass calculations vs. time. The table shows observed average molar mass calculations of each peak and actual molar masses of selected components.

4.3.4 - Improper spacer morphology affects Cas1-Cas2 complex formation

Seeing that in all three cases, the presence of a prespacer was necessary to allow complex formation to occur, we decided to test the effect that prespacer morphology would have on complex formation. **Figure 23** shows the effect in G1 when using a 4 nt splayed spacer, which still showed ~10% activity in integration assays (**Figure 17**), instead of the 5 nt 3' overhang prespacer used for integration assays. Addition of the 4 nt prespacer resulted in the formation of a 90.3 kDa complex, like what was seen with no prespacer added in **Figure 20**. This indicates that a prespacer needs to be of a certain morphology for efficient complex formation to occur. The inefficiency of stable complex formation in G1 with the splayed spacer explains the decrease in integration activity seen in **Figure 17**. When done with G2, complete complex formation occurred using the splayed spacer (**Figure 23**). This contrast shows that some integration complexes are more robust in their ability to bind to different types of prespacers.

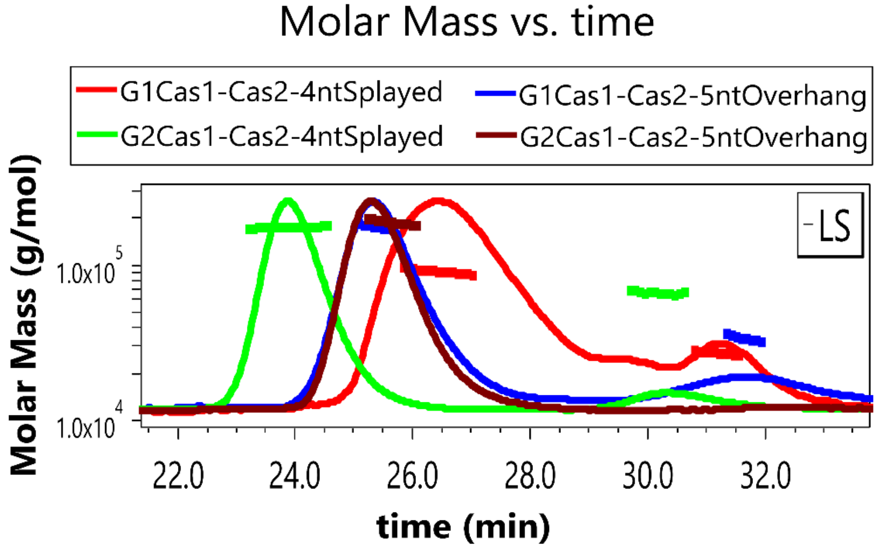


Observed MM (kDa)	Peak 1	Peak 2
prespacer	23.7	
G3Cas1	75.3	
G3Cas2	20.8	
G3Cas1-Cas2	94.7	
G3Cas1-Cas2 -prespacer	202	76.1
Expected MM (kDa)		
prespacer	15.5	
G3Cas1	35.1	
G3Cas2	12.3	
G3Cas1-Cas2	165	
G3Cas1-Cas2 -prespacer	180.5	

Figure 22: Molar mass analysis of G3 Cas1-Cas2-prespacer. Chromatograms showing light scattering signal and molar mass calculations vs. time. The table shows observed average molar mass calculations of each peak and actual molar masses of selected components.

4.4.0 - Structural studies of G3 Cas1

As stated previously, differences in structure of the LRH in Cas1 proteins may be a contributing factor to the differences in integration activity between G1, G2, and G3. We initially performed sequence analysis of selected Cas1 proteins from G1, G2, and G3 (**Figure 24**). The sequence alignment showed a conserved histidine residue, H153, which is present in G1 and G2 but not in G3. This result is interesting because in the crystal structure of G2 Cas1-Cas2-prespacer-leader (PDB 5XVN, (1)) H153 makes sequence specific contacts with the LAS. This gave us reason to believe that this absence of H153 in G3 may be responsible for the lack of sequence specific insertion activity. Another interesting fact is the lactobacilli harboring type II-A G3 CRISPR system has lysogens integrated into the genome and we hypothesize that substitutions in crucial active site residues and absence of sequence conservation at the leader-repeat junction (**Figure 12A**) in the native context is physiologically relevant to lactobacilli to maintain an inactive CRISPR adaptation. To gain insights into these



Observed MM (kDa)		
	Peak 1	Peak 2
G1Cas1-Cas2-5ntOverhang	175	33.4
G1Cas1-Cas2-4ntSplayed	90.3	27.1
G2Cas1-Cas2-5ntOverhang	187	
G2Cas1-Cas2-4ntSplayed	175	65.8
Expected MM (kDa)		
G1Cas1-Cas2-5ntOverhang	183	
G1Cas1-Cas2-4ntSplayed	186	
G2Cas1-Cas2-5ntOverhang	178	
G2Cas1-Cas2-4ntSplayed	181	

Figure 23: Effect of non-ideal prespacers on protein-DNA complex formation. Chromatograms showing light scattering signal and molar mass calculations vs. time. The table shows observed average molar mass calculations of each peak and expected molar masses of selected components. The observed shift in retention time for the G2Cas1-Cas2-4ntSplayed came from a difference in flow rate during the experiment, not from size or shape differences.

questions that are distinct from active CRISPR adaptation that have been previously characterized, we set out to crystallize and solve the structures of Cas1 and Cas1-Cas2-prespacer complex from G3. Detailed structural analysis of the LRH region of G3, along with other differing regions if found, will likely provide evidence for key protein-DNA interactions that drive the differences in integration activity seen in G1, G2, and G3.

4.4.1 - Initial crystal screening, hit identification, and optimization

We initially set out to crystallize G3 Cas1 proteins from *Lactobacillus gasseri* in the apo

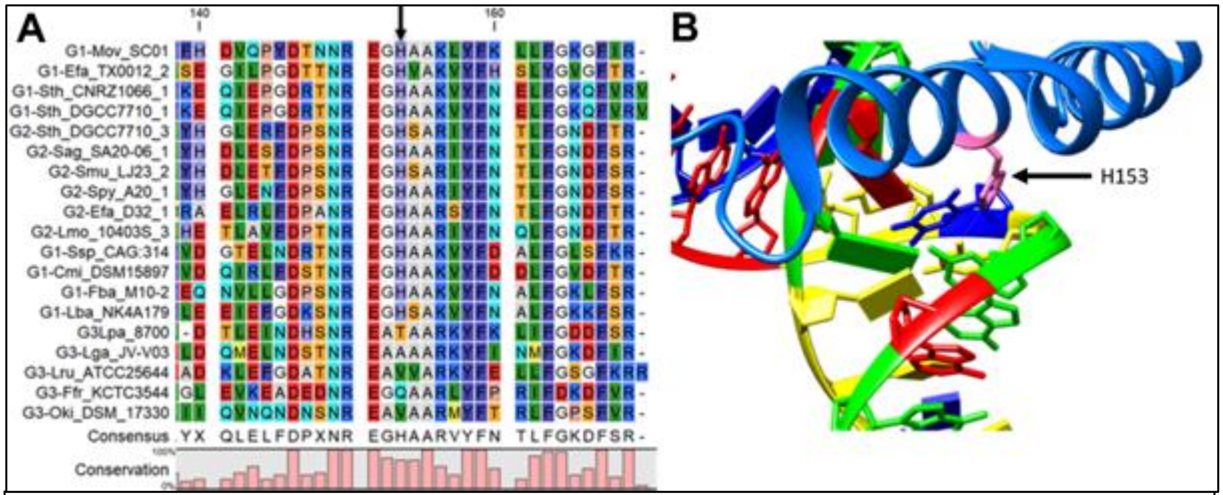
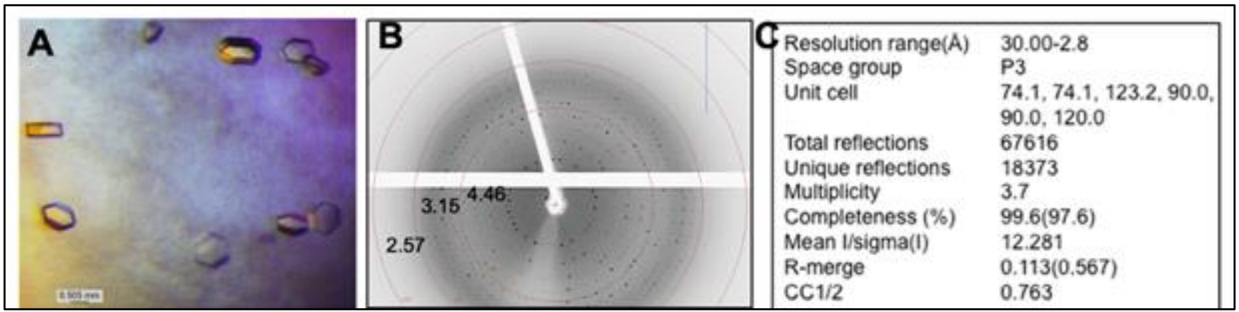


Figure 24: A) Sequence alignment of various Cas1 proteins from G1, G2, and G3 systems. H153 is completely conserved in G1 and G2 but absent in G3. **B) Crystal structure of Cas1 from G2 (PDB 5XVN (1)) bound to leader and spacer DNA.** H153 (magenta) makes a sequence specific contact with the opposite strand of the GAG sequence motif conserved at the leader-repeat junction in G1 and G2.

form as well as the Cas1-Cas2 complex. This was before detailed molar mass analysis showed that prespacer DNA was necessary for canonical stoichiometric stable complex formation. Initial crystals were seen for G3 Cas1 in the Wizard Classic crystallization screen from Rigaku, well D9 which contained 20% PEG 3000, 100 mM sodium acetate/acetic acid pH 4.5. This condition was optimized for PEG concentration and pH and yielded hexagonal crystals of manageable size for data collection which can be seen in **Figure 25A**. For Cas1-Cas2 complex, we obtained a crystal condition that yielded complex crystals as per analysis on a gel (**Figure 26A-B**). This crystal diffracted to a low resolution (~8Å) (**Figure 26C**). Since the SEC-MALS studies showed that the complex association in non-canonical stoichiometries, we have not optimized this condition further.

4.4.2 - Initial data analysis of hexagonal G3 Cas1 crystals

Crystals were initially screened in house using a Dectris Pilatus P200 hybrid pixal detector coupled to a Rigaku MicroMax 007HF microfocus X-ray generator, using CuKa**α** radiation. Initial diffraction resolution was to about 3.5 Å. Indexing of spots appeared to be difficult, as diffraction patterns were noisy, and spots were not clearly defined above the background. Several attempts at data collection were tried, with and without cryo-protectants. Higher resolution datasets using 20% glycerol as a cryoprotectant were collected at synchrotron radiation sources under cryo conditions (2.8Å)(**Figure 25B-C**). Indexing and integration were carried out using the program HKL2000 from HKL Research, Inc. Analysis of the generated datasets was carried out in the Phenix suite of programs. Results of twinning analysis in Xtriage, a data quality analysis software in Phenix, show strong evidence for twinning (101). All the statistics



Resolution range(Å)	30.00-2.8
Space group	P3
Unit cell	74.1, 74.1, 123.2, 90.0, 90.0, 120.0
Total reflections	67616
Unique reflections	18373
Multiplicity	3.7
Completeness (%)	99.6(97.6)
Mean I/sigma(I)	12.281
R-merge	0.113(0.567)
CC1/2	0.763

Figure 25: Crystals of G3 Cas1. A) Hexagonal crystals obtained from optimization of the Wizard Classic screen condition. **B)** Diffraction pattern collected at the synchrotron radiation source. Spots observed clearly go out beyond 3 Å. **C)** Diffraction statistics from data collection at the synchrotron source.

independent of twin laws listed in the file agree with the data being a near perfect twin (102). This twinning problem caused great difficulty in molecular replacement attempts and in experimental phasing attempts using a seleno-methionine version of these crystals. Eventually a new crystal form was deemed necessary for the project to continue.

4.4.3 - Re-screening, hit identification, and structure determination

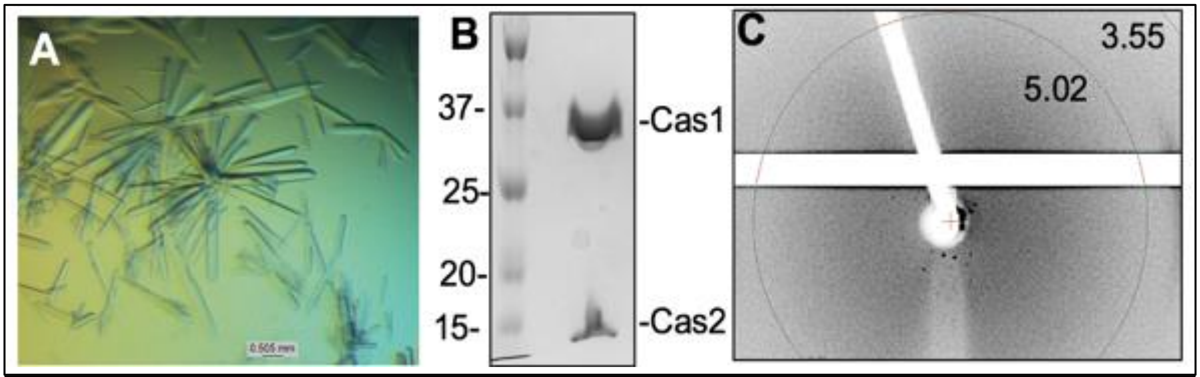


Figure 26: Crystals of G3 Cas1-Cas2. A) Rod crystals obtained from optimization. **B)** SDS-PAGE of washed crystals loaded directly into the well. Ladder positions are in kilodaltons. **C)** Diffraction pattern of rod crystals shows very low-resolution scattering.

Re-screening, with the addition of a seed stock made from the hexagonal crystals, brought about another hit in the Berkley screen from Rigaku. Well D11 (100 mM BisTris pH 6.0, 100 mM $LiSO_4$, 100 mM NaCl, 10% hexanediol, 20% PEG 3350) showed a large single crystal, seen in **Figure 27A**, which was harvested and used to collect a full data set on home source X-rays. We will denote this as crystal form 2. From this single crystal we collected a dataset that diffracted to 3.2Å at home source. Data analysis showed no evidence of twinning, and hence structure determination was attempted for this crystal. Molecular replacement was performed using an averaged Cas1 structure generated by the Ensembler program in Phenix using 4 different Cas1 structures available in the PDB (PDB IDs: 4ZKJ, 5XVN, 6QXF, 5XVP). The Mathews Coefficient predicted four full chain monomers in the asymmetric unit. Molecular replacement gave a partial model and this model was subjected to model building using Phenix Autobuild to build four monomers into the asymmetric unit. Only one of the chains had good density covering the whole protein region (pink monomer in **Figure 28**). This monomer was superimposed to the other three partial monomers, which gave the model shown in **Figure 28**. The R_{work} and R_{free} at this stage of refinement were 27.09 and 35.07, respectively. Further refinement of the model did not reduce the R factors. In addition, a crystal packing analysis showed that the monomers are placed non-ideally with large water channels between the set of dimers (**Figure 28**). Statistics shown in the table in **Figure 28**, such as an R_{free} of 0.35 and Ramachandran outliers of 6.99%, show that the model does not completely explain the diffraction pattern observed. Due to these reasons, we are not confident in the current model and pursued further optimization of this condition to collect more datasets, with improved resolution to ease the structure determination process.

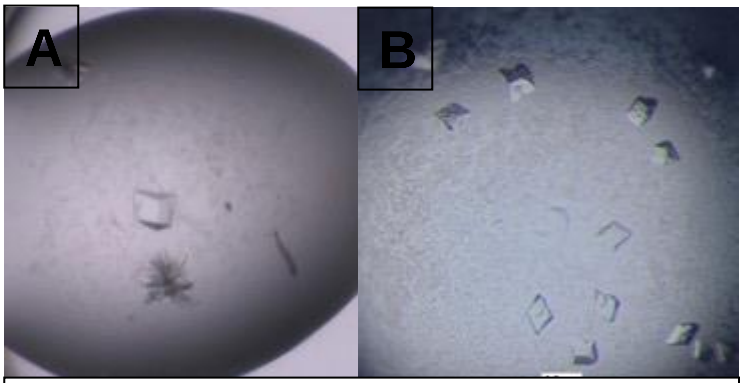


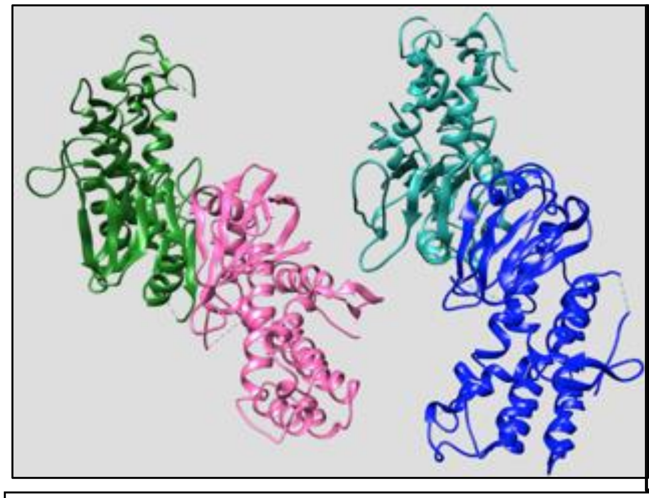
Figure 27: Re-crystallization of G3 Cas1.A) Single large crystal hit in Berkley broad screen using G3 Cas1, seeded with hexagonal crystals. **B)** Optimization of condition in A producing many diffracting crystals.

4.4.4 – Optimization of crystal form 2

Replicating the single crystal condition proved to be difficult at first, creating fused multiple crystals despite several optimization steps. We identified the reason for this being the change in the seed stock we have been using for optimization of this condition. We used the small crystals present in the crystallization drop of the initial form 2 crystal (**Figure 27A**) for further optimization of this condition with the logic that this will provide a better seed since they originated from the same condition. Since we were not successful with this strategy, we used the hexagonal seed stocks for exact replication of crystal in **Figure 27A**. Interestingly, this produced the original crystal back which is shown in **Figure 27B**. We setup optimization trays following this strategy with both native and selenomethionine proteins. Several of these crystals were used for data collection at SSRL on beamline 9-2, which gave a slightly higher resolution for the native dataset (3.2Å) and a selenomethionine dataset at 4.0Å.

4.4.5 - Data analysis and results of molecular replacement

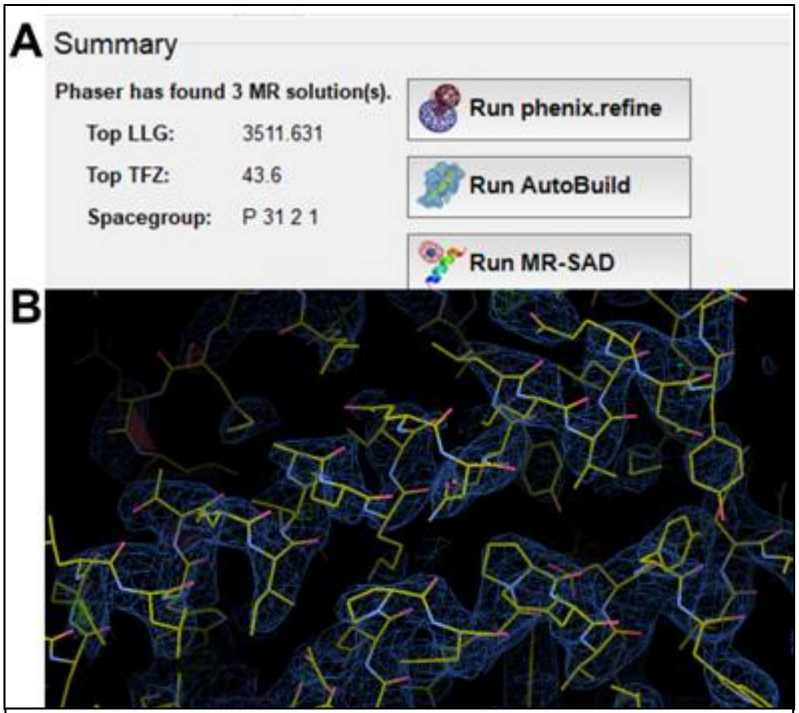
Datasets from native and selenmethionine version of G3 Cas1 were collected at synchrotron radiation sources and analyzed for twinning using Xtriage. No evidence of twinning was found, which was an encouraging sign for phasing. Molecular replacement using a split search model of the alpha helical region of an already published crystal structure of Cas1 (PDB ID 4ZKJ) separated from the beta sheet region, resulted in the scores shown in **Figure 29A**. A TFZ score above 8.0 and an LLG score above 1000



G3 Cas1	
Space group	P 32 2 1
Unit cell	157.057, 157.057, 111.508, 90, 90, 120
Unique reflections	26436 (2557)
Completeness (%)	97.36 (98.46)
Wilson B-factor	70.62
Reflections used in refinement	25829 (2557)
Reflections used for R-free	1952 (197)
R-work	0.2709 (0.3526)
R-free	0.3507 (0.4749)
Number of non-hydrogen atoms	8512
macromolecules	8512
Protein residues	1082
RMS(bonds)	0.012
RMS(angles)	1.47
Ramachandran favored (%)	76.53
Ramachandran allowed (%)	16.48
Ramachandran outliers (%)	6.99
Rotamer outliers (%)	0
Clashscore	23.86
Average B-factor	76.15
macromolecules	76.15

Figure 28: G3 Cas1 model generated from molecular replacement. Two copies of the Cas1 dimer are shown in the asymmetric unit. The table shows selected statistics for the shown model.

indicates a correct solution (103). Careful inspection of the model generated show clear agreement from the model and calculated electron density (**Figure 29B**). With a Mathews coefficient indicating 2821 residues in the asymmetric unit, it was calculated to have about 8 copies, which differs from the model obtained from home source. Several regions of the model and electron density show bad agreement, one problem being molecular replacement solution placing too many copies of the alpha helical region and too few copies of the beta sheet region (8 alpha helical regions and 2 beta sheet regions).



Phaser has found 3 MR solution(s).

Top LLG: 3511.631

Top TFZ: 43.6

Spacegroup: P 31 2 1

Figure 29: Molecular replacement results of most recent G3 Cas1 data collection. A) Molecular replacement statistics (LLG and TFZ scores) indicate a good solution was found. **B)** Electron density map and molecular model shown in Coot, generated by the molecular replacement shown in A. Good agreement is shown thorough some areas of the structure, with other regions showing much less agreement than others.

4.4.6 - Future directions

Newly collected datasets will be carefully processed, with an emphasis on experimental phasing from the selenomethionine datasets as molecular replacement hasn't given good results in the past. Also, experimental phasing needs to be carried out on multiple datasets, as only one of the several datasets has been used to this point. Careful attention to the space group and unit cell dimensions is essential as some variability has been seen from the autoindexing done by the synchrotron software.

Along with solving the crystal structure of G3 Cas1, further structural studies involving Cas1-Cas2-prespacer complexes will provide necessary information for the development of Cas1-Cas2 complexes with designed insertion sites. Current screening has resulted in at least one crystal hit for the G3 Cas1-Cas2-prespacer complex (15% Reagent Alcohol, 100mM HEPES pH 7.5, 200mM $MgCl_2$) which is currently being optimized (**Figure 30**). More screening with the G1 Cas1-Cas2-prespacer complex will be necessary to provide the structural data from all three groups. A complete structural profile of Cas1-Cas2-prespacer complexes will allow for targeted approaches to be used in developing Cas1-Cas2 as a biological tool.

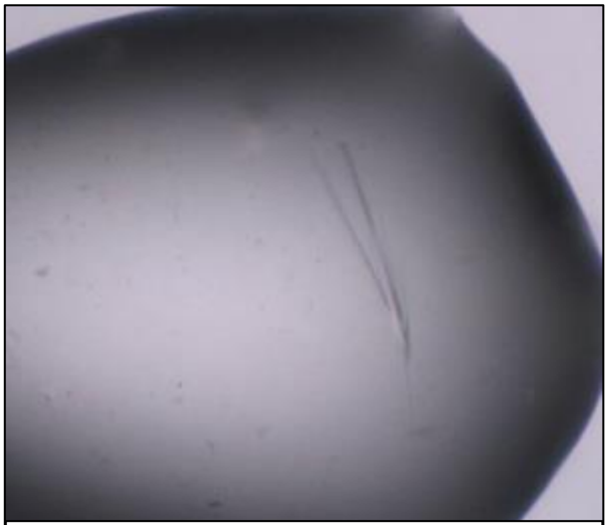


Figure 30: Crystal hit using G3 Cas1-Cas2-prespacer. Long needle crystals present after a two-week incubation at room temperature using purified G3-Cas1-Cas2-prespacer complex.

Chapter 5: Outlook

With the information set forth in this work and in combination with work done by other groups, a solid knowledge base has been established for the development of tunable Cas1-Cas2 complexes for the site-specific insertion of short DNAs. Bioinformatic work has established the importance of the leader-repeat junction sequence and the co-evolution of CRISPR genes with the leader-repeat sequence. Phylogenetic analysis of different type II-A CRISPR loci appear to show several ancestral strains of type II-A which were similar to G1, from which both a deviant leader-repeat junction (G2) and a non-functional (G3) system evolved. The purpose of CRISPR systems which evolve away from CRISPR systems and deactivate their Cas1-Cas2 machinery remains an important question in the field to study. The conservation shown at leader ends in type II-A systems appears to be correlated with the intrinsic site-specific activity seen by integration complexes in type II-A. It will be interesting to expand leader-repeat junction analysis to other CRISPR types and see how leaders are conserved across a wider range of CRISPR systems.

The subsequent biochemical work has shown that different leader-repeat junction sequences correspond with differing mechanisms of prespacer insertion, at least in members of each group assessed so far. Work done by other groups has established the mechanisms presented in this work for G1 and G2. More biochemical work is needed in other CRISPR types to assess the site-specific insertion activity of Cas1-Cas2. To date, type II contains the only integration complex capable of site-specifically inserting prespacers without the need of accessory proteins. Lack of work done in types IV, V, and VI leaves a large gap in knowledge and a possible untapped group of integration proteins to be used in a similar fashion to the potential tools shown in this work. Until now, the G2-IC appears to be the better choice for further development as a biological tool, with its higher activity *in vitro* and its ability to form complexes and insert various forms of prespacers. Another major point drawn from this work is the spacer processing requirements of G1, of which a detailed mechanism is missing. This gap in knowledge provides a possible direction for future studies in G1.

A large amount of structural information will be needed to bring the idea of tunable Cas1-Cas2 complexes to fruition. With only Cas1-Cas2 complexes available from G2, a large gap in data provides potential for future structural studies to identify key areas that affect site-specific spacer insertion. Comparison of these structures to non-site-specific insertion proteins (such as those from G3) will be essential in development of biological tools.

The work presented in this dissertation has opened several future directions of research in the Rajan laboratory including:

- A bioinformatics pipeline has been established for the use in analyzing leader sequences outside of type II-A CRISPR systems.
- The initial crystal hits and optimized conditions for producing Cas1-Cas2-prespacer complex will enable characterization of all type II-A sub-groups to determine molecular determinants of differences in prespacer integration mechanisms.
- Crystal structure of G3 Cas1 along with site directed mutagenesis of G3 Cas1 to restore conserved amino acids in G1 and G2 are being pursued to identify molecular determinants of inactive native versions of G3 type II-A system.

- A major direction in this study is directed evolution approaches where G2 Cas1 and Cas2 are being engineered to enable site-specific DNA insertion with varied sequence specificities. Success of this approach will directly link the fundamental mechanistic information gleaned from this thesis to biotechnology applications.

Methods

6.1.0 – Copyright Information

The methods shown in this chapter were taken directly from the publications listed for Chapters 2 and 3. Methods were published alongside the corresponding results under the same licensing listed at the beginning of Chapters 2 and 3.

6.2.0 - Processing of genomic data

In this study, the type II-A loci were collected by multiple ways. Initially, Bacterial Generic Feature Format (GFF) and accessioned protein product FASTA files were downloaded from NCBI and scanned for II-A specific Cas protein names (Cas9/Csn1 and Csn2) in the annotation field. The genomes containing Cas9/Csn1 and/or Csn2 annotation entries were downloaded from NCBI in GenBank format. The datasets were screened manually for the presence of *cas1*, *cas2*, *cas9*, and *csn2*, and only the loci with all four type II-A specific *cas* genes were used for further analysis. The genomic region flanking downstream of the *csn2* gene was further processed to extract the leader and the first repeat of the CRISPR array. The protein sequences of Cas9, Cas1, Cas2, and Csn2 that were coded by the upstream region flanking the *csn2* gene were extracted from NCBI. The presence of all four proteins limits our dataset to strictly type II-A loci. A total of 129 loci were identified based on Cas9/Csn1 and/or Csn2 annotation search. Previously, Chylinski *et al* reported type II-A loci based on a Cas9 sequence search (78, 104). A total of 32 type II-A loci that represented species and genera that were absent in our initial dataset were selected from the Chylinski list for our study. In addition, we performed protein sequence homology search by DELTA-BLAST (105) using a representative Csn2 sequence from each subfamily as mentioned in Chylinski *et al* 2014 (78) ((NCBI protein accession number: 116101487 for subfamily I, 116100822 for subfamily II, 389815356 for subfamily III, 385326557 for subfamily IV, 315659845 for subfamily V as mentioned in Chylinski *et al* 2014) (78). By this search a total of 6 loci were identified from bacterial genera *Weissella, Globicatella, Nosocomiicoccus, Caryophanon* and *Virgibacillus*. The final dataset consisted of 167 type II-A loci with a wide representation based on the current knowledge of type II-A diversity. A total of 50 different bacterial genera were present in our dataset. (**Table S2** and **S3**).

The orientation of the Cas proteins was used in assessing the transcription direction of the leader-repeat units. To analyze leader and repeat sequences, an approximately 400-nucleotide stretch of sequences downstream of *csn2* gene were examined using CRISPR finder tool (106), and an in-house script to locate the tandem repeats. Since there were differences in the repeat length as it exists in the genomic locus and as reported in the CRISPRdb (106), we used the in-house program to locate the repeats (**Table S4**). The accuracy of the repeat extracted by our script was validated manually by checking the genomic data for the length and sequence of the repeat within a CRISPR array. The loci that lacked predicted repeats or Cas protein(s) were omitted from further analysis. In the case of bacteria with multiple CRISPR types, the components belonging to a type II-A locus were taken as one dataset. For example, Sth DGCC7710 has four CRISPR loci. Only loci 1 and 3 that correspond to type II-A were selected for our analysis. The Cas proteins and leader-repeat elements of CRISPR1 were kept as one unit, while that belonging to CRISPR3 represented another unit. Recently, several bioinformatics tools for the identification and analysis of leader and repeat regions have been developed (107, 108). For a selected subset, we compared the orientation of leader sequences and repeats as predicted by CRISPRDetect tool (108) and our results, and saw agreement between the methods.

6.3.0 - Sequence Alignment

We used **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation (MUSCLE) with its default settings (109) to perform all the sequence alignments in this study. The MUSCLE output was used to generate phylogenetic trees with MEGA6 (110) using the Maximum Likelihood Tree option and Jones-Taylor-Thornton (JTT) model. Additionally, MUSCLE alignments were used to generate alignment figures in UGENE (111) and sequence logos with WebLogo (112).

6.4.0 - Target and prespacer design

Target DNA sequences for each group were designed to mimic the leader-repeat junction and contained the last 50 base pairs of the leader and the first repeat from the genomic DNA of each representative bacteria, as shown in the CRISPR database (106). For each group, this sequence was followed by the first spacer sequence taken from the CRISPR array in the *Streptococcus pyogenes A20* type II-A genome (**Figure 12A**). The *S. pyogenes A20* type II-A spacer sequence was used in all linear targets because initial experiments using cognate native first spacer from each bacterium gave low activity, indicating a role of spacer secondary structure affecting the integration process, at least under the *in vitro* settings. To avoid such variations in our analysis, we used a spacer sequence with the lowest predicted secondary structure, which was G2 (**Figure S16**). G1 came from *Streptococcus thermophilus CNRZ1066,* G2 from *Streptococcus pyogenes A20* (type II-A), G3 from *Lactobacillus paragasseri JV-V03*, and E, type I-E system, from *Escherichia coli str. K-12 substr. MG1655*. The four original linear target sequences (G1-L, G2-L, G3-L, and E-L) were synthesized as double stranded gene fragments by Integrated DNA Technologies (IDT), PCR amplified, and cloned between BamHI and EcoRI (New England Biolabs) sites of pUC19. Sequence verified plasmids were used as templates for Site Directed Mutagenesis (113) using overlapping primers, having the last 7 bp of the leader 3′ ends exchanged in all possible combinations to create twelve mutant substrates (**Figure S17**). The sequences for the gene fragments and mutated targets, as well as PCR primers, can be found in **Table S1**. Sequence verified plasmids were used as PCR templates to create the 12 mutant 116 bp targets for integration assays. The hairpin targets could not be synthesized as a single molecule because of its secondary structure (77 bp stem, 6 nt loop). To accommodate this, a 67 nt fragment was annealed and ligated to a long template strand containing one strand of the stem (77 nt), the loop (6 nt), and 10 nt of the returning stem. The 67 nt fragment was phosphorylated by Calf Intestinal Phosphatase to facilitate ligation to the 93 nt DNA by T4 DNA ligase. The prespacer used in the bulk of this study was taken from a recently published paper and contained 5 nt 3′ overhangs (97). Each strand of the prespacer was ordered from IDT with a 5′ 6-fluorescein amidite (6-FAM) label and was annealed at an equimolar ratio to create double stranded, doubly labelled prespacers. For construction of the prespacer library, a single 6-FAM labelled reverse strand was used to create all non-symmetrical overhangs. This was done by annealing a shortened non-labelled forward strand to create the desired overhangs. For spacers that are splayed or with the symmetrical 3′ overhang spacers, unlabelled forward and reverse strands were ordered and were manually labelled using the 5′ EndTag^{TM} Labeling DNA/RNA Kit from Vector Laboratories before annealing complementary strands at an equimolar ratio.

6.5.0 - Cloning of cas genes

The *cas1* and *cas2* gene sequences from each group were codon optimized by IDT for protein expression in *Escherichia coli* Bl21-DE3. Synthetic gene fragments were made for all 6 Cas genes. G1 Cas1, G1 Cas2, G3 Cas1, and G3 Cas2 were cloned using sequence and ligation independent cloning (SLIC) (114) into the pET His6 SUMO TEV LIC cloning vector for expression with a SUMO His6 tag (115). G2 Cas1 and G2 Cas2 were cloned into pMCSG9 using ligation independent cloning (LIC) for expression with an N-terminal Maltose Binding Protein (MBP) with a His6X tag (116). Solubility tags were necessary for initial protein solubility. Primers used for each reaction can be seen in **Table S5**.

6.6.0 - Protein expression and purification

Plasmids containing the correct gene inserts were transformed into BL21-DE3 cells for protein expression. Cells were grown in 2xYT medium containing the appropriate antibiotic (ampicillin 100 µg/mL, kanamycin 50 µg/mL) at 37°C and at OD_{600} (optical density at 600 nm) of 0.6-0.8 protein was induced with 0.5 mM Isopropyl β-D-1-thiogalactopyranoside at 18°C overnight. The cell pellets were re-suspended in lysis buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 20 mM Imidazole, 1 mM TCEP, and 10% glycerol) and lysed by sonication, clarified at 38000 rcf for 30 minutes, and the supernatant was loaded onto a Histrap Crude column (GE). Proteins were eluted from the nickel column using a gradient of nickel elution buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 500 mM Imidazole, 1 mM TCEP, and 10% glycerol). From this point different protocols were used for each protein.

G1 Cas1 and G1 Cas2 were dialyzed overnight into 50 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM TCEP, and 10% glycerol with the simultaneous addition of tobacco etch virus (TEV) protease to cleave off the SUMO His6X solubility tag. The protein was then loaded back onto a Histrap Crude column to separate out the SUMO tag and collect the pure protein in the flow through. Pure protein fractions were checked by SDS-PAGE before being concentrated and loaded onto a HiLoad 16/600 Superdex S-75 pg column (GE) size exclusion column, using fresh dialysis buffer as the gel filtration buffer.

G2 Cas1 was dialyzed overnight with TEV in 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 1 mM TCEP, and 10% glycerol. This removed the MBP-His6X tag. The protein was further purified on a HiTrap Heparin HP (GE) column and pure fractions were concentrated and run on a HiLoad 16/600 Superdex S-75 pg column (GE). G2 Cas2 was purified similarly, but 500 mM NaCl was in all buffers, except the 1 M elution buffer for the heparin column, which enhanced protein solubility.

G3 Cas1 was dialyzed overnight in 50 mM Tris-HCl pH 8.0, 300 mM NaCl, 1 mM TCEP, and 10% glycerol with TEV to remove the SUMO His6X tag. The protein was then loaded back onto a Histrap Crude column and pure protein separated from the SUMO tag was collected in the flow through. The flow through was concentrated and run on a HiLoad 16/600 Superdex S-75 pg (GE). G3 Cas2 was purified similarly except for the use of 500 mM NaCl in all buffers to keep the protein soluble. All proteins were concentrated to ~5 mg/mL for Cas1 and ~1 mg/mL for Cas2. Protein concentrations were measured using the absorbance at 280 nm on a NanoDrop (ThermoFisher) and extinction coefficients predicted by ProtParam (117). Small aliquots for integration assays (3 µL-10 µL) were flash frozen in liquid nitrogen and stored at -80°C. The concentrations of protein dilutions made for integration assays were measured again using the NanoDrop for accuracy in experiments. All purified protein samples were analysed by high resolution intact protein mass spectrometry at the Laboratory for Molecular Biology and Cytometry Research at the University of Oklahoma Health Sciences Center. Each protein that was compatible was measured within one Dalton of the expected mass, confirming full length protein samples (**Figure S18**). SDS-PAGE confirms that each protein sample is >95% pure (**Figure S19**). NCBI Protein accession numbers are G1 Cas1: WP_011227029, G1 Cas2: WP_011227030, G2 Cas1: AFV38400, G2 Cas2: AFV38399, G3 Cas1: EFJ70028, G3 Cas2: EFJ70029.

6.7.0 - Integration assays

Integration assays were performed similarly as before (1) with slight modifications. Final reaction conditions contained 50 ng of target DNA, 200 nM prespacer, and 500 nM Cas1-Cas2 complex (since the complex is four Cas1 molecules to two Cas2 molecules, we mixed equal volumes of 40 µM Cas1 and 20 µM Cas2 to make a 10X stock solution of 5 µM). Optimum protein concentration was determined by titration assay (**Figure S20**). The 30-minute reaction time was determined by time courses for each protein complex (**Figure S2-S3**). Reaction buffer differed for each protein complex. G1-IC and G3-IC reaction buffer contained 20 mM HEPES pH 8.0, 25 mM NaCl, 1 mM TCEP, 10 mM $MnCl_2$, and 10% DMSO. G2-IC reaction buffer contained 20 mM HEPES pH 8.0, 25 mM NaCl, 1 mM TCEP, 10 mM $MgCl_2$, and 10% DMSO. Different metals in the buffers for each protein complex were based on optimization of reactions using various metals (**Figure S1**). 10 µL reactions were run at room temperature for 30 minutes before being quenched with 10 µL of 95% formamide, 50 mM EDTA, 0.025% bromophenol blue, and 0.025% xylene cyanol. Quenched reactions were heated to 95°C for 5 minutes before resolving on Urea-PAGE. The gel composition was 12.5% acrylamide, 6 M urea, and 20% formamide. A total of three replications were done for each experiment, including proteins from two independent protein preparations. Representative gels used for quantifications can be seen in **Figure S5-S6**. FAM images and ethidium bromide images were taken using a ChemiDoc MP imaging system (Bio-Rad). Quantifications of FAM bands were done using ImageJ (118). Integration scores were calculated using the following equation:

$$Integration\ Score = \left(\frac{[intensity\ of\ integration\ band]}{[intensity\ of\ same\ position\ in\ the\ No\ Protein\ lane]}\right) - 1$$

Standard deviation (σ) was calculated using the following equation:

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

While this quantification method does not account for loading differences between lanes, this was the only reasonable way we would get a quantitative information from this data. The amount of free prespacer left was in amounts causing saturation during quantification and hence a quantification within each lane was not possible. Nevertheless, this method showed its ability to compare between the different experimental conditions in this study.

6.8.0 - Analysis of protein complex formation

Protein complex formation was analyzed using size exclusion chromatography coupled with multi angle light scattering (SEC-MALS). Running buffer for all the SEC-MALS runs consisted of 20 mM HEPES pH 8.0, 200 mM NaCl, 1 mM TCEP, 10 mM $MgCl_2$, and 1% glycerol. Individual samples were prepared by diluting 50 µL of 5 µM complex + 5 µM prespacer (in gel filtration buffer from the protein purification) with 450 µL of 20 mM HEPES pH 8.0, 150 mM NaCl, 1 mM TCEP, and 10 mM $MgCl_2$. The samples were incubated overnight at 4ºC and centrifuged at 16000 rcf. The samples were then injected onto a Superdex 200 Increase column 10/300 GL (GE) with an inline miniDAWN TREOS (Wyatt Technology Co.) multi-angle light scattering detector and Optilab T-rEX (Wyatt Technology Co.) differential refractometer. Molecular weight analysis was carried out using ASTRA (v 7.3.0) software (Wyatt Technology Co.) and other data analysis was done in Microsoft Excel.

6.9.0 – Crystallization setup and freezing

G3 Cas1 crystals used for collecting the non-twinned data were obtained first by sitting drop vapor diffusion using 50% well solution and 50% G3 Cas1 at 4.8 mg/mL. Well solution consisted of 100 mM BisTris pH 6.0, 100 mM $LiSO_4$, 100 mM NaCl, 10% hexanediol, 20% PEG 3350. 0.3 µL of 1:100000 diluted seed stock was added as well. Optimization was carried out using hanging drop vapor diffusion with similar results. Crystals were allowed to grow for at least 4 days at room temperature before harvesting. Crystals were frozen in well solution supplemented with 20% glycerol as a cryoprotectant.

6.10.0 – Crystal Data Analysis

Data was collected at home source or at synchrotron radiation, as stated in the text. Frames were indexed, scaled, and merged using either iMosFlm or XDS (119, 120). Resulting datafiles were analyzed further in Phenix (103), using a previously solved Cas1 protein as a molecular replacement model (PDB ID 4ZKJ). Resulting solutions were further analyzed in Coot and refined using phenix.refine (103, 121).

References

1. Xiao Y, Ng S, Nam KH, Ke A. How type ii crispr-cas establish immunity through cas1-cas2-mediated spacer integration. Nature. **2017**;550(7674):137-41.
2. Nunez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. Foreign DNA capture during crispr-cas adaptive immunity. Nature. **2015**;527(7579):535-8.
3. Van Orden MJ, Klein P, Babu K, Najar FZ, Rajan R. Conserved DNA motifs in the type ii-a crispr leader region. PeerJ. **2017**;5:e3161.
4. Fineran PC. Resistance is not futile: Bacterial 'innate' and crispr-cas 'adaptive' immune systems. Microbiology (Reading, England). **2019**;165(8):834-41.
5. Shabbir MA, Hao H, Shabbir MZ, Wu Q, Sattar A, Yuan Z. Bacteria vs. Bacteriophages: Parallel evolution of immune arsenals. Front Microbiol. **2016**;7:1292.
6. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. Crispr provides acquired resistance against viruses in prokaryotes. Science. **2007**;315(5819):1709-12.
7. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small crispr rnas guide antiviral defense in prokaryotes. Science. **2008**;321(5891):960-4.
8. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Molecular microbiology. **2002**;43(6):1565-75.
9. Ishino M, Sawada Y, Yaegashi T, Demura M, Fujinaga K. Nucleotide sequence of the adenovirus type 40 inverted terminal repeat: Close relation to that of adenovirus type 5. Virology. **1987**;156(2):414-6.
10. Mojica FJ, Juez G, Rodríguez-Valera F. Transcription at different salinities of haloferax mediterranei sequences adjacent to partially modified psti sites. Molecular microbiology. **1993**;9(3):613-21.
11. van Belkum A, Scherer S, van Alphen L, Verbrugh H. Short-sequence DNA repeats in prokaryotic genomes. Microbiology and molecular biology reviews : MMBR. **1998**;62(2):275-93.
12. van Belkum A, van Leeuwen W, Scherer S, Verbrugh H. Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. Research in microbiology. **1999**;150(9-10):617-26.
13. Mojica FJ, Díez-Villaseñor C, Soria E, Juez G. Biological significance of a family of regularly spaced repeats in the genomes of archaea, bacteria and mitochondria. Molecular microbiology. **2000**;36(1):244-6.
14. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (crisprs) have spacers of extrachromosomal origin. Microbiology (Reading, England). **2005**;151(Pt 8):2551-61.
15. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of molecular evolution. **2005**;60(2):174-82.
16. Marraffini LA. Crispr-cas immunity in prokaryotes. Nature. **2015**;526(7571):55-61.

17. Haurwitz RE, Jinek M Fau - Wiedenheft B, Wiedenheft B Fau - Zhou K, Zhou K Fau - Doudna JA, Doudna JA. Sequence- and structure-specific rna processing by a crispr endonuclease. **2010**(1095-9203 (Electronic)).
18. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide rnas for invader defense in prokaryotes. Genes & development. **2008**;22(24):3489-96.
19. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, Moineau S, Mojica FJM, Scott D, Shah SA, Siksnys V, Terns MP, Venclovas C, White MF, Yakunin AF, Yan W, Zhang F, Garrett RA, Backofen R, van der Oost J, Barrangou R, Koonin EV. Evolutionary classification of crispr-cas systems: A burst of class 2 and derived variants. Nat Rev Microbiol. **2020**;18(2):67-83.
20. Pyenson NC, Marraffini LA. Type iii crispr-cas systems: When DNA cleavage just isn't enough. Current opinion in microbiology. **2017**;37:150-4.
21. Zheng Y, Li J, Wang B, Han J, Hao Y, Wang S, Ma X, Yang S, Ma L, Yi L, Peng W. Endogenous type i crispr-cas: From foreign DNA defense to prokaryotic engineering. Frontiers in bioengineering and biotechnology. **2020**;8:62.
22. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and atp-dependent helicase in the crispr/cas immune system. The EMBO journal. **2011**;30(7):1335-42.
23. Hayes RP, Xiao Y, Ding F, van Erp PB, Rajashankar K, Bailey S, Wiedenheft B, Ke A. Structural basis for promiscuous pam recognition in type i-e cascade from e. Coli. Nature. **2016**;530(7591):499-503.
24. Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. Rna-guided rna cleavage by a crispr rna-cas protein complex. Cell. **2009**;139(5):945-56.
25. Tamulaitis G, Venclovas Č, Siksnys V. Type iii crispr-cas immunity: Major differences brushed aside. Trends in microbiology. **2017**;25(1):49-61.
26. Elmore JR, Sheppard NF, Ramia N, Deighan T, Li H, Terns RM, Terns MP. Bipartite recognition of target rnas activates DNA cleavage by the type iii-b crispr-cas system. Genes & development. **2016**;30(4):447-59.
27. Estrella MA, Kuo FT, Bailey S. Rna-activated DNA cleavage by the type iii-b crispr-cas effector complex. Genes & development. **2016**;30(4):460-70.
28. Kazlauskiene M, Tamulaitis G, Kostiuk G, Venclovas Č, Siksnys V. Spatiotemporal control of type iii-a crispr-cas immunity: Coupling DNA degradation with the target rna recognition. Molecular cell. **2016**;62(2):295-306.
29. Jiang W, Samai P, Marraffini LA. Degradation of phage transcripts by crispr-associated rnases enables type iii crispr-cas immunity. Cell. **2016**;164(4):710-21.
30. Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA. Co-transcriptional DNA and rna cleavage during type iii crispr-cas immunity. Cell. **2015**;161(5):1164-74.
31. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-rna-guided DNA endonuclease in adaptive bacterial immunity. Science. **2012**;337(6096):816-21.
32. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, Liu DR. Search-and-replace genome

editing without double-strand breaks or donor DNA. Nature. **2019**;576(7785):149-57.

33. Garcia-Doval C, Jinek M. Molecular architectures and mechanisms of class 2 crispr-associated nucleases. Curr Opin Struct Biol. **2017**;47:157-66.

34. O'Connell MR. Molecular mechanisms of rna targeting by cas13-containing type vi crispr-cas systems. J Mol Biol. **2019**;431(1):66-87.

35. Sasnauskas G, Siksnys V. Crispr adaptation from a structural perspective. Curr Opin Struct Biol. **2020**;65:17-25.

36. Wright AV, Wang JY, Burstein D, Harrington LB, Paez-Espino D, Kyrpides NC, Iavarone AT, Banfield JF, Doudna JA. A functional mini-integrase in a two-protein-type v-c crispr system. Molecular cell. **2019**;73(4):727-37.e3.

37. Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, Lambowitz AM, Fire AZ. Direct crispr spacer acquisition from rna by a natural reverse transcriptase-cas1 fusion protein. Science. **2016**;351(6276):aad4234.

38. Alkhnbashi OS, Shah SA, Garrett RA, Saunders SJ, Costa F, Backofen R. Characterizing leader sequences of crispr loci. Bioinformatics (Oxford, England). **2016**;32(17):i576-i85.

39. Rollie C, Graham S, Rouillon C, White MF. Prespacer processing and specific integration in a type i-a crispr system. Nucleic acids research. **2018**;46(3):1007-20.

40. Wei Y, Chesne MT, Terns RM, Terns MP. Sequences spanning the leader-repeat junction mediate crispr adaptation to phage in streptococcus thermophilus. Nucleic Acids Res. **2015**;43(3):1749-58.

41. Nunez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. Crispr immunological memory requires a host factor for specificity. Molecular cell. **2016**;62(6):824-33.

42. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. Crispr adaptation biases explain preference for acquisition of foreign DNA. Nature. **2015**;520(7548):505-10.

43. Weissman JL, Stoltzfus A, Westra ER, Johnson PLF. Avoidance of self during crispr immunization. Trends in microbiology. **2020**;28(7):543-53.

44. Modell JW, Jiang W, Marraffini LA. Crispr-cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature. **2017**;544(7648):101-4.

45. Shiimori M, Garrett SC, Chambers DP, Glover CVC, 3rd, Graveley BR, Terns MP. Role of free DNA ends and protospacer adjacent motifs for crispr DNA uptake in pyrococcus furiosus. Nucleic acids research. **2017**;45(19):11281-94.

46. Dupuis M, Villion M, Magadán AH, Moineau S. Crispr-cas and restriction-modification systems are compatible and increase phage resistance. Nature communications. **2013**;4:2087.

47. Nussenzweig PM, McGinn J, Marraffini LA. Cas9 cleavage of viral genomes primes the acquisition of new immunological memories. Cell Host Microbe. **2019**;26(4):515-26.e6.

48. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the crispr/cas adaptive bacterial immunity system. Nature communications. **2012**;3:945.

49. Swarts DC, Mosterd C, van Passel MW, Brouns SJ. Crispr interference directs strand specific spacer acquisition. PloS one. **2012**;7(4):e35888.

50. Mosterd C, Rousseau GM, Moineau S. A short overview of the crispr-cas adaptation stage. Canadian journal of microbiology. **2020**.
51. Marraffini LA, Sontheimer EJ. Self versus non-self discrimination during crispr rna-directed immunity. Nature. **2010**;463(7280):568-71.
52. Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. Structural and mechanistic basis of pam-dependent spacer acquisition in crispr-cas systems. Cell. **2015**;163(4):840-53.
53. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. Cas9 specifies functional viral targets during crispr-cas adaptation. Nature. **2015**;519(7542):199-202.
54. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in type ii-a crispr-cas adaptation. Genes & development. **2015**;29(4):356-61.
55. Kieper SN, Almendros C, Behler J, McKenzie RE, Nobrega FL, Haagsma AC, Vink JNA, Hess WR, Brouns SJJ. Cas4 facilitates pam-compatible spacer selection during crispr adaptation. Cell reports. **2018**;22(13):3377-84.
56. Lee H, Zhou Y, Taylor DW, Sashital DG. Cas4-dependent prespacer processing ensures high-fidelity programming of crispr arrays. Molecular cell. **2018**;70(1):48-59 e5.
57. Shiimori M, Garrett SC, Graveley BR, Terns MP. Cas4 nucleases define the pam, length, and orientation of DNA fragments integrated at crispr loci. Molecular cell. **2018**;70(5):814-24 e6.
58. Lee H, Dhingra Y, Sashital DG. The cas4-cas1-cas2 complex mediates precise prespacer processing during crispr adaptation. eLife. **2019**;8.
59. Ramachandran A, Summerville L, Learn BA, DeBell L, Bailey S. Processing and integration of functionally oriented prespacers in the escherichia coli crispr system depends on bacterial host exonucleases. The Journal of biological chemistry. **2020**;295(11):3403-14.
60. Wright AV, Liu JJ, Knott GJ, Doxzen KW, Nogales E, Doudna JA. Structures of the crispr genome integration complex. Science. **2017**;357(6356):1113-8.
61. Wright AV, Doudna JA. Protecting genome integrity during crispr immune adaptation. Nat Struct Mol Biol. **2016**;23(10):876-83.
62. McGinn J, Marraffini LA. Crispr-cas systems optimize their immune response by specifying the site of spacer integration. Molecular cell. **2016**;64(3):616-23.
63. Arslan Z, Hermanns V, Wurm R, Wagner R, Pul U. Detection and characterization of spacer integration intermediates in type i-e crispr-cas system. Nucleic Acids Res. **2014**;42(12):7884-93.
64. Drabavicius G, Sinkunas T, Silanskas A, Gasiunas G, Venclovas C, Siksnys V. Dnaq exonuclease-like domain of cas2 promotes spacer integration in a type i-e crispr-cas system. EMBO reports. **2018**;19(7).
65. Ka D, Jang DM, Han BW, Bae E. Molecular organization of the type ii-a crispr adaptation module and its interaction with cas9 via csn2. Nucleic acids research. **2018**;46(18):9805-15.
66. Wilkinson M, Drabavicius G, Silanskas A, Gasiunas G, Siksnys V, Wigley DB. Structure of the DNA-bound spacer capture complex of a type ii crispr-cas system. Molecular cell. **2019**;75(1):90-101 e5.

67. Bondy-Denomy J, Pawluk A, Maxwell KL, Davidson AR. Bacteriophage genes that inactivate the crispr/cas bacterial immune system. Nature. **2013**;493(7432):429-32.
68. Hwang S, Maxwell KL. Meet the anti-crisprs: Widespread protein inhibitors of crispr-cas systems. The CRISPR journal. **2019**;2(1):23-30.
69. Nakamura M, Srinivasan P, Chavez M, Carter MA, Dominguez AA, La Russa M, Lau MB, Abbott TR, Xu X, Zhao D, Gao Y, Kipniss NH, Smolke CD, Bondy-Denomy J, Qi LS. Anti-crispr-mediated control of gene editing and synthetic circuits in eukaryotic cells. Nature communications. **2019**;10(1):194.
70. Ibrahim SA, Alazzeh AY, Awaisheh SS, Song D, Shahbazi A, AbuGhazaleh AA. Enhancement of alpha- and beta-galactosidase activity in lactobacillus reuteri by different metal ions. Biological trace element research. **2010**;136(1):106-16.
71. Sanozky-Dawes R, Selle K, O'Flaherty S, Klaenhammer T, Barrangou R. Occurrence and activity of a type ii crispr-cas system in lactobacillus gasseri. Microbiology (Reading, England). **2015**;161(9):1752-61.
72. Sheth RU, Yim SS, Wu FL, Wang HH. Multiplex recording of cellular events over time on crispr biological tape. Science. **2017**;358(6369):1457-61.
73. Shipman SL, Nivala J, Macklis JD, Church GM. Crispr-cas encoding of a digital movie into the genomes of a population of living bacteria. Nature. **2017**;547(7663):345-9.
74. Schmidt F, Cherepkova MY, Platt RJ. Transcriptional recording by crispr spacer acquisition from rna. Nature. **2018**;562(7727):380-5.
75. Rehman S, Ali Z, Khan M, Bostan N, Naseem S. The dawn of phage therapy. Reviews in medical virology. **2019**;29(4):e2041.
76. Wright AV, Doudna JA. Protecting genome integrity during crispr immune adaptation. Lid - 10.1038/nsmb.3289 [doi]. Nat Struct Mol Biol. **2016**(1545-9985 (Electronic)).
77. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. Diversity, activity, and evolution of crispr loci in streptococcus thermophilus. Journal of bacteriology. **2008**;190(4):1401-12.
78. Chylinski K, Makarova KS, Charpentier E, Koonin EV. Classification and evolution of type ii crispr-cas systems. Nucleic acids research. **2014**;42(10):6091-105.
79. Lier C, Baticle E, Horvath P, Haguenoer E, Valentin A-S, Glaser P, Mereghetti L, Lanotte P. Analysis of the type ii-a crispr-cas system of streptococcus agalactiae reveals distinctive features according to genetic lineages. Frontiers in Genetics. **2015**;6:214.
80. Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the crispr adaptation process in escherichia coli. Nucleic Acids Res. **2012**;40(12):5569-76.
81. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. Intrinsic sequence specificity of the cas1 integrase directs new spacer acquisition. Elife. **2015**;4.
82. Nunez JK, Lee AS, Engelman A, Doudna JA. Integrase-mediated spacer acquisition during crispr-cas adaptive immunity. Nature. **2015**;519(7542):193-8.

83. Wang R, Li M, Gong L, Hu S, Xiang H. DNA motifs determining the accuracy of repeat duplication during crispr adaptation in haloarcula hispanica. Nucleic Acids Res. **2016**(1362-4962 (Electronic)).

84. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. Phage response to crispr-encoded resistance in streptococcus thermophilus. Journal of bacteriology. **2008**;190(4):1390-400.

85. Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S. The crispr/cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. **2010**;468(7320):67-71.

86. Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, Moineau S, Glover CV, 3rd, Graveley BR, Terns RM, Terns MP. The three major types of crispr-cas systems function independently in crispr rna biogenesis in streptococcus thermophilus. Molecular microbiology. **2014**;93(1):98-112.

87. Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Ratsima Hariniaina E, Gonzalez-Zorn B, Poyart C, Rosinski-Chupin I, Glaser P. The highly dynamic crispr1 system of streptococcus agalactiae controls the diversity of its mobilome. Molecular microbiology. **2012**;85(6):1057-71.

88. Sapranauskas R, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. The streptococcus thermophilus crispr/cas system provides immunity in escherichia coli. Nucleic Acids Res. **2011**;39(21):9275-82.

89. Paez-Espino D, Morovic W, Sun CL, Thomas BC, Ueda K, Stahl B, Barrangou R, Banfield JF. Strong bias in the bacterial crispr elements that confer immunity to phage. Nature communications. **2013**;4:1430.

90. Sanozky-Dawes R, Selle K, O'Flaherty S, Klaenhammer T, Barrangou R. Occurrence and activity of a type ii crispr-cas system in lactobacillus gasseri. Microbiology. **2015**(1465-2080 (Electronic)).

91. Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MFA-Ohoo. Intrinsic sequence specificity of the cas1 integrase directs new spacer acquisition. Lid - 10.7554/elife.08716 [doi]. Elife. **2015**(2050-084X (Electronic)).

92. 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. A dual function of the crispr-cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol. **2011**;79(2):484-502.

93. Nunez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. Crispr immunological memory requires a host factor for specificity. Mol Cell. **2016**(1097-4164 (Electronic)).

94. Diez-Villasenor C, Guzman NM, Almendros C, Garcia-Martinez J, Mojica FJ. Crispr-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among crispr-cas i-e variants of escherichia coli. RNA biology. **2013**;10(5):792-802.

95. Van Orden MJ, Newsom S, Rajan R. Crispr type ii-a subgroups exhibit phylogenetically distinct mechanisms for prespacer insertion. The Journal of biological chemistry. **2020**.

96. Maduro M. Random DNA sequence generator [Available from: https://faculty.ucr.edu/~mmaduro/random.htm.
97. Kim JG, Garrett S, Wei Y, Graveley BR, Terns MP. Crispr DNA elements controlling site-specific spacer integration and proper repeat length by a type ii crispr-cas system. Nucleic acids research. **2019**;47(16):8632-48.
98. Rollie C, Chevallereau A, Watson BNJ, Chyou TY, Fradet O, McLeod I, Fineran PC, Brown CM, Gandon S, Westra ER. Targeting of temperate phages drives loss of type i crispr-cas systems. Nature. **2020**;578(7793):149-53.
99. Nunez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. Cas1-cas2 complex formation mediates spacer acquisition during crispr-cas adaptive immunity. Nat Struct Mol Biol. **2014**;21(6):528-34.
100. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic acids research. **2000**;28(1):235-42.
101. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Echols N, Headd JJ, Hung LW, Jain S, Kapral GJ, Grosse Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner RD, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. The phenix software for automated determination of macromolecular structures. Methods (San Diego, Calif). **2011**;55(1):94-106.
102. Yeates TO, Fam BC. Protein crystals and their evil twins. Structure (London, England : 1993). **1999**;7(2):R25-9.
103. Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB, Croll TI, Hintze B, Hung LW, Jain S, McCoy AJ, Moriarty NW, Oeffner RD, Poon BK, Prisant MG, Read RJ, Richardson JS, Richardson DC, Sammito MD, Sobolev OV, Stockwell DH, Terwilliger TC, Urzhumtsev AG, Videau LL, Williams CJ, Adams PD. Macromolecular structure determination using x-rays, neutrons and electrons: Recent developments in phenix. Acta crystallographica Section D, Structural biology. **2019**;75(Pt 10):861-77.
104. Fonfara I, Le Rhun A, Chylinski K, Makarova KS, Lecrivain AL, Bzdrenga J, Koonin EV, Charpentier E. Phylogeny of cas9 determines functional exchangeability of dual-rna and cas9 among orthologous type ii crispr-cas systems. Nucleic Acids Res. **2014**;42(4):2577-90.
105. Boratyn GM, Schaffer Aa Fau - Agarwala R, Agarwala R Fau - Altschul SF, Altschul Sf Fau - Lipman DJ, Lipman Dj Fau - Madden TL, Madden TL. Domain enhanced lookup time accelerated blast. Biol Direct. **2012**(1745-6150 (Electronic)).
106. Grissa I, Vergnaud G, Pourcel C. The crisprdb database and tools to display crisprs and to generate dictionaries of spacers and repeats. BMC bioinformatics. **2007**;8:172.
107. Alkhnbashi OS, Shah SA, Garrett RA, Saunders SJ, Costa F, Backofen R. Characterizing leader sequences of crispr loci. Bioinformatics. **2016**(1367-4811 (Electronic)).
108. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM. Crisprdetect: A flexible algorithm to define crispr arrays. BMC Genomics. **2016**(1471-2164 (Electronic)).

109. Edgar RC. Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **2004**;32(5):1792-7.

110. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. Mega6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution. **2013**;30:2725-9.

111. Okonechnikov K, Golosova O Fau - Fursov M, Fursov M. Unipro ugene: A unified bioinformatics toolkit. Bioinformatics. **2012**;8(1367-4811 (Electronic)).

112. Crooks GE, Hon G, Chandonia JM, Brenner SE. Weblogo: A sequence logo generator. Genome Res. **2004**;14(6):1188-90.

113. Carey MF, Peterson CL, Smale ST. Pcr-mediated site-directed mutagenesis. Cold Spring Harbor protocols. **2013**;2013(8):738-42.

114. Li MZ, Elledge SJ. Slic: A method for sequence- and ligation-independent cloning. Methods in molecular biology (Clifton, NJ). **2012**;852:51-9.

115. Pet his6 sumo tev lic cloning vector (1s) was a gift from scott gradia (addgene plasmid # 29659 ; rrid:Addgene_29659.

116. Aslanidis C, de Jong PJ. Ligation-independent cloning of pcr products (lic-pcr). Nucleic acids research. **1990**;18(20):6069-74.

117. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. Protein identification and analysis tools in the expasy server. Methods in molecular biology (Clifton, NJ). **1999**;112:531-52.

118. Rasband WS. Imagej, u. S. National institutes of health, bethesda, maryland, USA. **1997-2018**.

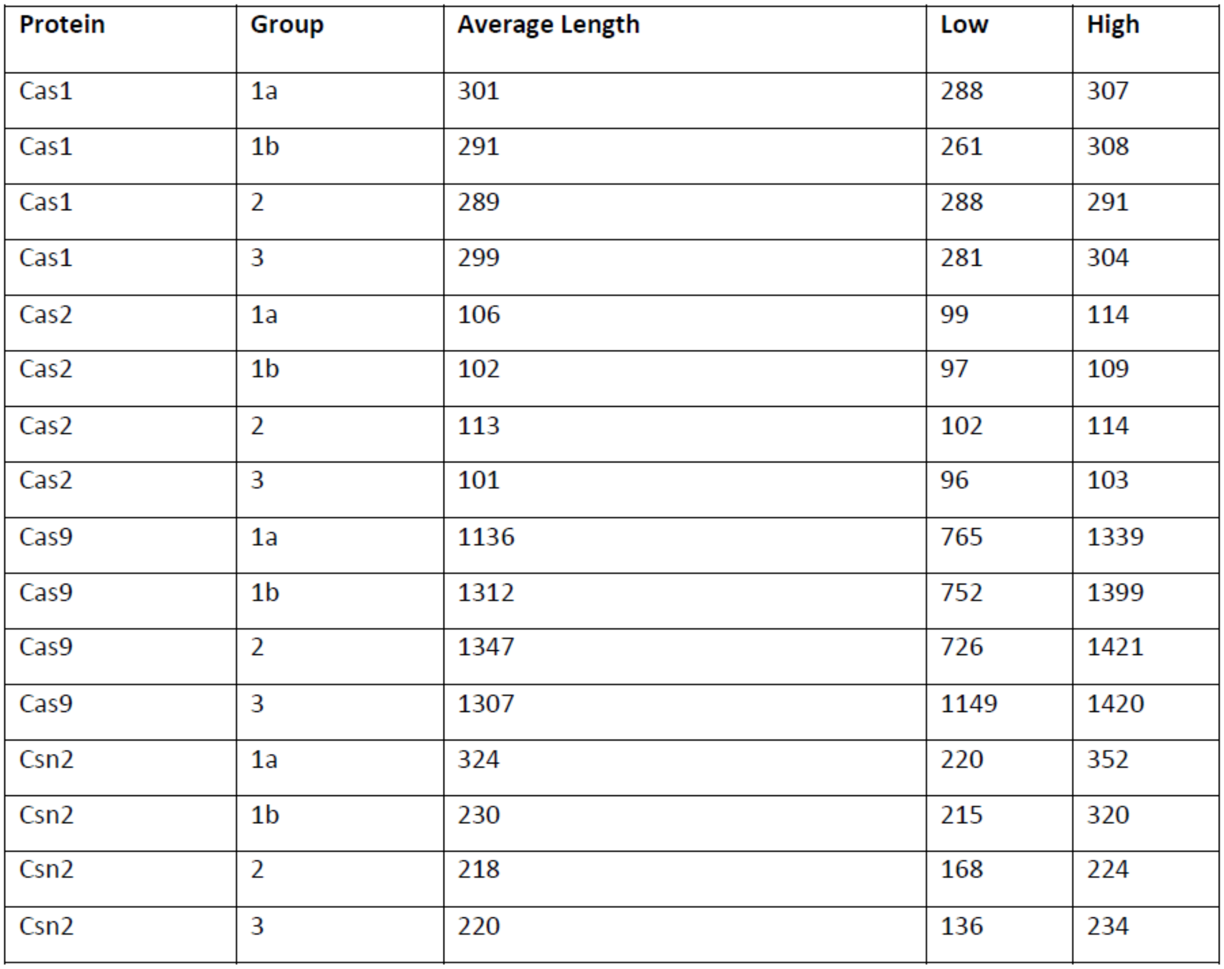
119. Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AG. Imosflm: A new graphical interface for diffraction-image processing with mosflm. Acta crystallographica Section D, Biological crystallography. **2011**;67(Pt 4):271-81.

120. Kabsch W. Xds. Acta crystallographica Section D, Biological crystallography. **2010**;66(Pt 2):125-32.

121. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of coot. Acta crystallographica Section D, Biological crystallography. **2010**;66(Pt 4):486-501.

Supporting Tables

Supporting Table S1



Protein	Group	Average Length	Low	High
Cas1	1a	301	288	307
Cas1	1b	291	261	308
Cas1	2	289	288	291
Cas1	3	299	281	304
Cas2	1a	106	99	114
Cas2	1b	102	97	109
Cas2	2	113	102	114
Cas2	3	101	96	103
Cas9	1a	1136	765	1339
Cas9	1b	1312	752	1399
Cas9	2	1347	726	1421
Cas9	3	1307	1149	1420
Csn2	1a	324	220	352
Csn2	1b	230	215	320
Csn2	2	218	168	224
Csn2	3	220	136	234

Supporting Table S2

Strain	Abbreviation	Nucleotide Accession Number
Acidaminococcus intestini RyC-MR95	Ain_RyC-MR95	NC_016077
Acidaminococcus sp. D21	Asp_D21	NZ_ACGB01000044
Bifidobacterium bifidum S17	Bbi_S17	NC_014616
Brochothrix campestris FSL F6-1037	Bca_FSLF6-1037	NZ_AODH01000013
Butyrivibrio fibrisolvens 16/4	Bfi_16/4	FP929036
Bacteroides pectinophilus CAG:437	Bpe_CAG:437	FR894965
Butyrivibrio sp. AC2005	Bsp_AC2005	NZ_AUJI01000018
Brochothrix thermosphacta DSM 20171	Bth_DSM20171	NZ_JHZM01000008
Coriobacterium glomerans PW2	Cgl_PW2	NC_015389
Caryophanon latum DSM 14151	Cla_DSM_14151	NZ_MATO01000048
Catellicoccus marimammalium M35/04/3	Cma_M35/04/3	NZ_AMYT01000017
Catenibacterium mitsuokai DSM 15897	Cmi_DSM15897	NZ_ACCK01000419
Clostridium sp. CAG:230	Csp_CAG:230	FR881840
Carnobacterium sp. ZWU0011	Csp_ZWU0011	NZ_JTLM01000044
Dorea longicatena AGR2136	Dlo_AGR2136	NZ_AUJS01000026
Eubacterium dolichum DSM 3991	Edo_DSM3991	NZ_DS483477
Enterococcus faecium 1,141,733	Efa_1141733	NZ_GG688466.1
*Enterococcus_faecalis*_D32_1	Efa_D32_1	NC_018221
*Enterococcus_faecalis*_OG1RF_1	Efa_OG1RF_1	NC_017316
Enterococcus faecalis TX0012	Efa_TX0012	NZ_GL456502
Enterococcus faecalis TX0012_2	Efa_TX0012_2	NZ_GL456502
Enterococcus hirae ATCC 9790	Ehi_ATCC9790	NZ_KB946228.1
Enterococcus italicus DSM 15952	Eit_DSM15952	NZ_AIJM01000004.1
Enterococcus mundtii QU 25 DNA	Emu_QU25_DNA	NC_022878
Enterococcus phoeniculicola ATCC BAA-412	Eph_ATCCBAA-412	NZ_ASWE01000001
Eubacterium ramulus atcc 29099	Era_ATCC29099	NZ_KI271077
Eubacterium rectale ATCC 33656	Ere_ATCC33656	NC_012781
Eggerthella sp. YY7918	Esp_YY7918	NC_015738
Eubacterium ventriosum ATCC 27560	Eve_ATCC27560	NZ_DS264278.1
Eubacterium yurii subsp. margaretiae ATCC 43715	Eyu_ATCC43715	AEES01000067
Filifactor alocis ATCC 35896	Fal_ATCC35896	NC_016630
Firmicutes bacterium M10-2	Fba_M10-2	NZ_KE159700
Fructobacillus fructosus KCTC 3544	Ffr_KCTC3544	NZ_AEOP01000035.1
Facklamia hominis CCUG 36813	Fho_CCUG36813	AGZD01000007
Finegoldia magna ATCC 29328	Fma_ATCC29328	NC_010376
Finegoldia magna SY403409CC001050417	Fma_SY403409CC001050417	NZ_AFUI01000017
Fusobacterium nucleatum subsp. vincentii 3_1_36A2	Fnu_3_1_36A2	NC_022196
Fusobacterium sp. 1_1_41FAA	Fus_1_1_41FAA	NZ_GG770381.1
Fusobacterium sp. 3_1_36A2	Fus_3_1_36A2	NC_022196.1
*Gemella_bergeri*_ATCC_700627	Gbe_ATCC700627	NZ_KI271806
*Gemella_haemolysans*_M341	Gha_M341	GL883582

Gordonibacter pamelaeae 7-10-1-b	Gpa_7-10-1-b	NC_021021
Globicatella sanguinis NBRC 15551	Gsa_NBRC15551	NZ_BCQX01000009
Helcococcus sueciensis DSM 17243	Hsu_DSM17243	NZ_AUHK01000002
Kandleria vitulina DSM 20405	Kvi_DSM20405	NZ_KL370857
Lactobacillus animalis KCTC 3501	Lan_KCTC3501	NZ_GL573157.1
Lachnospiraceae bacterium NK4A179	Lba_NK4A179	NZ_ATWC01000025
Lactobacillus brevis subsp. gravesensis ATCC 27305	Lbr_ATCC27305	NZ_GG669606.1
Lactobacillus buchneri NRRL B-30929	Lbu_NRRLB-30929	CP002652
Lactobacillus casei Lc-10	Lca_Lc10	NZ_AFYT01000017.1
Lactobacillus ceti DSM 22408	Lce_DSM22408	NZ_KE383993
Lactobacillus coryniformis subsp. coryniformis KCTC 3167	Lco_KCTC3167	NZ_GL544592.1
Lactobacillus curvatus CRL 705	Lcu_CRL705	NZ_AGBU01000042.1
Lactobacillus farciminis KCTC 3681	Lfa_KCTC3681	NZ_GL575017
Lactobacillus fermentum ATCC 14931	Lfe_ATCC14931	NZ_GG669901.1
Lactobacillus gasseri JV-V03	Lga_JV-V03	NZ_GL379580
Leuconostoc gelidum KCTC 3527	Lge_KCTC3527	NZ_AEMI01000021.1
Lactobacillus hominis CRBIP 24.179	Lho_CRBIP24.179	NZ_CAKE01000018.1
Lactobacillus iners LactinV 11V1-d	Lin_11V1-d	NZ_AEHN01000016.1
Lactobacillus jensenii 27-2-CHN	Lje_27-2-CHN	NZ_KI411428.1
Lactobacillus johnsonii DPC 6026	Ljo_DPC6026	NC_017477.1
Listeria monocytogenes 10403S_3	Lmo_10403S_3	NC_018586
Listeria monocytogenes EGD	Lmo_EGD	NC_017544
Listeria monocytogenes J0161_4	Lmo_J0161_4	NC_018591
Listeria monocytogenes strain J2-031	Lmo_J2-031	NC_022568
Listeria monocytogenes strain N1-011A	Lmo_N1-011A	NC_017545
Listeria monocytogenes strain R2-502	Lmo_R2-502	NC_021837
Listeria monocytogenes serotype 7 str. SLCC2482_3	Lmo_SLCC2482_3	NC_021826
Listeria monocytogenes SLCC2540_3	Lmo_SLCC2540_3	NC_021838
Listeria monocytogenes SLCC5850_3	Lmo_SLCC5850_3	NC_018592
Lactobacillus paracasei subsp. paracasei 8700	Lpa_8700	NC_022112
Lactobacillus pentosus KCA1	Lpe_KCA1	NZ_CM001538.1
Lactobacillus plantarum ZJ316	Lpl_ZJ316	NC_020229
Lactobacillus rhamnosus GG	Lrh_GG	NC_017482
Lactobacillus ruminis ATCC 25644	Lru_ATCC25644	NZ_AFYE01000073.1
Lactobacillus sanfranciscensis TMW 1.1304	Lsa_TMW1.1304	NC_015978.1
Lactobacillus salivarius UCC118	Lsa_UCC118	NC_007929
Mycoplasma arginini HAZ145_1	Mar_HAZ145_1	NZ_AP014657.1
Mycoplasma canis PG 14	Mca_PG14	NZ_AJFQ01000005
Mycoplasma canis UFG4	Mca_UFG4	NZ_AJFU01000005.1
Mycoplasma cynos C142	Mcy_C142	NC_019949.1
Mycoplasma gallisepticum str. F	Mga_str._F	NC_017503.1
Mycoplasma mobile 163K	Mmo_163K	NC_006908
Mycoplasma ovipneumoniae SC01	Mov_SC01	NZ_AFHO01000003

Megasphaera sp. UPII 135-E	Msp_UPII135-E	NZ_AFUG01000024.1
Mycoplasma synoviae 53	Msy_53	NC_007294
Nosocomiicoccus ampullae strain LUREC	Nam_LUREC	NZ_MBFG01000013
Oenococcus kitaharae DSM 17330	Oki_DSM_17330	NZ_CM001398
Olsenella uli DSM 7084	Oul_DSM7084	NC_014363
Pediococcus acidilactici D3	Pac_D3	NZ_KB889550
Pseudoramibacter alactolyticus ATCC 23263	Pal_ATCC23263	NZ_GL622359
Peptostreptococcus anaerobius CAG:621	Pan_CAG:621	NZ_CP016534.1
Planococcus antarcticus DSM 14505	Pan_DSM14505	CAYH010000043
Peptoniphilus duerdenii ATCC BAA-1640	Pdu_ATCCBAA-1640	NZ_GL397071
Pediococcus lolii NGRI 0510Q	Plo_NGRI0510Q	NZ_BANK01000034.1
Ruminococcus lactaris ATCC 29176	Rla_ATCC29176	NZ_DS990175
Roseburia sp. CAG:197	Rsp_CAG:197	HF999864
Streptococcus agalactiae 09mas018884	Sag_09mas018884	NC_021485
*Streptococcus agalactiae*_2603V/R_1	Sag_2603V/R_1	NC_004116
*Streptococcus agalactiae*_A909_1	Sag_A909_1	NC_007432
*Streptococcus agalactiae*_GD201008-001_1	Sag_GD201008-001_1	NC_018646
Streptococcus agalactiae ILRI005	Sag_ILRI005	NC_021486
*Streptococcus agalactiae*_NEM316_1	Sag_NEM316_1	NC_004368
*Streptococcus agalactiae*_SA20-06_1	Sag_SA20-06_1	NC_019048
*Streptococcus anginosus*_C1051_4	San_C1051_4	NC_022244
*Streptococcus dysgalactiae*_subsp.equisimilis_ATCC_12394_2	Sdy_ATCC12394_2	NC_017567
*Streptococcus dysgalactiae*_subsp.equisimilis_GGS_124_1_2	Sdy_GGS124_2	NC_012891
*Streptococcus dysgalactiae*_subsp.equisimilis_AC-2713_2	Sdy_AC-2713_2	NC_019042
*Streptococcus dysgalactiae*_subsp.equisimilis_RE378_2	Sdy_RE378_2	NC_018712
*Streptococcus equi*_subsp.zooepidemicus_str.MGCS10565_4	Seq_MGCS10565_4	NC_011134
*Streptococcus gallolyticus*_subsp.gallolyticus_ATCC_43143_1	Sga_ATCC43143_1	NC_017576
*Streptococcus gallolyticus*_subsp.gallolyticus_ATCC_43143_2	Sga_ATCC43143_2	NC_017576
*Streptococcus gallolyticus*_subsp.gallolyticus_ATCC_BAA-2069_1	Sga_ATCCBAA-2069_1	NC_015215
*Streptococcus gallolyticus*_subsp.gallolyticus_ATCC_BAA-2069_2	Sga_ATCCBAA-2069_2	NC_015215
*Streptococcus gallolyticus*_UCN34_1	Sga_UCN34_1	NC_013798
*Streptococcus gallolyticus*_UCN34_2	Sga_UCN34_2	NC_013798
*Streptococcus gordonii*_str.Challis_substr.CH1_2	Sgo_CH1_2	NC_009785

*Streptococcus intermedius*_B196_1	Sin_B196_1	NC_022246
*Streptococcus infantarius*_subsp.infantarius_CJ18_1	Sin_CJ18_1	NC_016826
*Streptococcus lutetiensis*_033_1	Slu_033_1	NC_021900
*Streptococcus lutetiensis*_033_2	Slu_033_2	NC_021900
*Streptococcus macedonicus*_ACA-DC_198_1	Sma_ACA-DC198_1	NC_016749
Solobacterium moorei F0204	Smo_F0204	GL637674
*Streptococcus mutans*_GS-5_1	Smu_GS-5_1	NC_018089
*Streptococcus mutans*_LJ23_2	Smu_LJ23_2	NC_017768
*Streptococcus mutans*_NN2025_2	Smu_NN2025_2	NC_013928
*Streptococcus mutans*_UA159_1	Smu_UA159_1	NC_004350
*Streptococcus pasteurianus*_ATCC_43144_1	Spa_ATCC43144_1	NC_015600
Streptococcus parasanguinis F0449	Spa_F0449	NZ_AJMV01000063.1
Staphylococcus pseudintermedius ED99	Sps_ED99	NC_017568
Streptococcus pseudoporcinus LQ 940-04	Sps_LQ940-04	NZ_AEUY02000005.1
*Streptococcus pyogenes*_A20_1	Spy_A20_1	NC_018936
*Streptococcus pyogenes*_M1_GAS_1	Spy_M1GAS_1	NC_002737
*Streptococcus pyogenes*_MGAS10270_1	Spy_MGAS10270_1	NC_008022
*Streptococcus pyogenes*_MGAS15252_1	Spy_MGAS15252_1	NC_017040
*Streptococcus pyogenes*_MGAS1882_1	Spy_MGAS1882_1	NC_017053
*Streptococcus pyogenes*_MGAS5005_1	Spy_MGAS5005_1	NC_007297
*Streptococcus pyogenes*_MGAS6180_2	Spy_MGAS6180_2	NC_007296
*Streptococcus pyogenes*_NZ131_1	Spy_NZ131_1	NC_011375
*Streptococcus salivarius*_JIM8777_3	Ssa_JIM8777_3	NC_017595
Streptococcus sanguinis SK49	Ssa_SK49	NZ_GL890985
Subdoligranulum sp. CAG:314	Ssp_CAG:314	FR900985
Streptococcus sp.I-G2_3	Ssp_I-G2_3	NC_022584
Streptococcus suis D9	Ssu_D9	NC_017620
*Streptococcus suis*_ST3_2	Ssu_ST3_2	NC_015433
Streptococcus suis YB51	Ssu_YB51	NC_022516
*Streptococcus thermophilus*_LMG_18311_1	Sth_LMG18311_1	NC_006448
*Streptococcus thermophilus*_JIM_8232_1	Sth_JIM8232_1	NC_017581
*Streptococcus thermophilus*_CNRZ1066_1	Sth_CNRZ1066_1	NC_006449
*Streptococcus thermophilus*_DGCC7710_1	Sth_DGCC7710_1	AWVZ01000002
Streptococcus thermophilus DGCC7710_3	Sth_DGCC7710_3	AWVZ01000001
*Streptococcus thermophilus*_LMD-9_2	Sth_LMD-9_2	NC_008532
*Streptococcus thermophilus*_LMD-9_5	Sth_LMD-9_5	NC_008532
*Streptococcus thermophilus*_MN-ZLW-002_1	Sth_MN-ZLW-002_1	NC_017927
*Streptococcus thermophilus*_MN-ZLW-002_3	Sth_MN-ZLW-002_3	NC_017927
*Streptococcus thermophilus*_ND03_1	Sth_ND03_1	NC_017563
*Streptococcus thermophilus*_ND03_3	Sth_ND03_3	NC_017563
Treponema denticola ATCC 33521	Tde_ATCC33521	NZ_KB445539
Treponema denticola ATCC 35405	Tde_ATCC35405	NC_002967
Treponema putidum OMZ 758	Tpu_OMZ758	NZ_CP009228
Veillonella atypica ACS-134-V-Col7a	Vat_ACS-134-V-Col7a	NZ_AEDS01000047

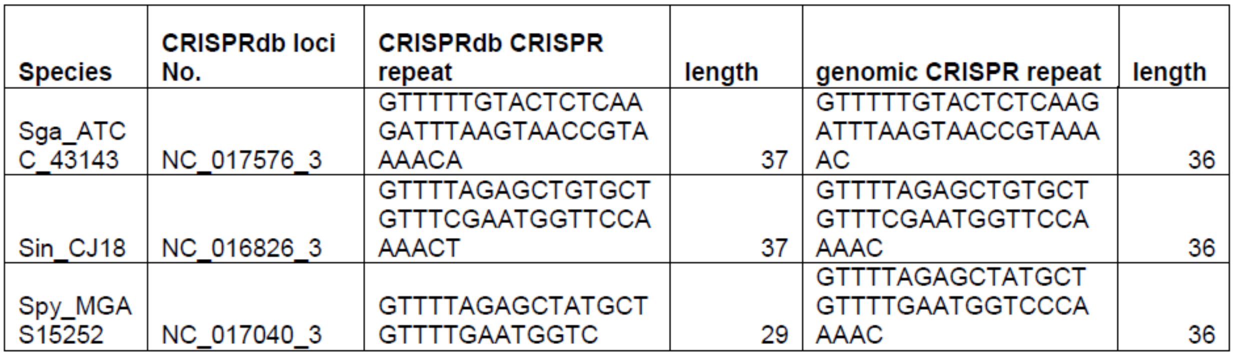
Veillonella parvula ATCC 17745	Vpa_ATCC17745	NZ_ADFU01000012.1
Veillonella sp. 6_1_27	Vsp_6_1_27	NZ_GG770216.1
Veillonella sp. oral taxon 780 str. F0422	Vsp_780strF0422	NZ_AFUJ01000012.1
Virgibacillus sp. SK-1	Vsp_SK1	NZ_CCXU01000007
Weissella cibaria strain FBL5	Wci_FBL5	NZ_LVYB01000027
Weissella halotolerans FBL4	Wha_FBL4	NZ_LVVN01000018.1

Supporting Table S3

Phylum	*Class*	*Order*	*Family*	*Genus*
Firmicutes	Negativicutes	Acidaminococcales	Acidaminococcaceae	Acidaminococcus
Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
Firmicutes	Bacilli	Bacillales	Listeriaceae	Brochothrix
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Butyrivibrio
Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Catellicoccus
Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Clostridioides
Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Carnobacterium
Firmicutes	Bacilli	Bacillales	Planococcaceae	Caryophanon
Actinobacteria	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Coriobacterium
Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Catenibacterium
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Dorea
Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Eubacterium
Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Eggerthella
Firmicutes				
Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Filifactor
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Fructobacillus
Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Facklamia
Firmicutes	Tissierellia	Tissierellales	Peptoniphilaceae	Finegoldia
Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium
Firmicutes	Bacilli	Bacillales		Gemella
Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Gordonibacter
Firmicutes	Bacilli	Lactobacillales	Aerococcaceae	Globicatella
Firmicutes	Tissierellia	Tissierellales	Peptoniphilaceae	Helcococcus
Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Kandleria
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Leuconostoc
Firmicutes	Bacilli	Bacillales	Listeriaceae	Listeria
Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	Mycoplasma
Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	Megasphaera
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Nosocomiicoccus
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Oenococcus

Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Olsenella
Firmicutes	Tissierellia	Tissierellales	Peptoniphilaceae	Peptoniphilus
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Pediococcus
Firmicutes	Clostridia	Clostridiales	Eubacteriaceae	Pseudoramibacter
Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Peptostreptococcus
Firmicutes	Bacilli	Bacillales	Planococcaceae	Planococcus
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Solobacterium
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
Spirochaetes	Spirochaetia	Spirochaetales	Spirochaetaceae	Treponema
Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	Veillonella
Firmicutes	Bacilli	Bacillales	Bacillaceae	Virgibacillus
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella

Supporting Table S4



Species	CRISPRdb loci No.	CRISPRdb CRISPR repeat	length	genomic CRISPR repeat	length
Sga_ATCC_43143	NC_017576_3	GTTTTTGTACTCTCAAGATTTAAGTAACCGTAAAACA	37	GTTTTTGTACTCTCAAGATTTAAGTAACCGTAAAAC	36
Sin_CJ18	NC_016826_3	GTTTTAGAGCTGTGCTGTTTCGAATGGTTCCAAAACT	37	GTTTTAGAGCTGTGCTGTTTCGAATGGTTCCAAAAC	36
Spy_MGAS15252	NC_017040_3	GTTTTAGAGCTATGCTGTTTTGAATGGTC	29	GTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAAC	36

Supporting Table S5 – Nucleotides Used

Description	Sequence (5´->3´)
G1-Cas1_LIC-F	TACTTCCAATCCAATGCCATGACTTGGCGCGTTGTGCAC
G1-Cas1_LIC-R	TTATCCACTTCCAATGTTACTTGCGCCATTCCAGGGAAGA
G1-Cas2_LIC-F	TACTTCCAATCCAATGCCATGCGCTACGAGGCGCTGCG
G1-Cas2_LIC-R	TTATCCACTTCCAATGTTATTAAATCACCACCAGTTTAT
G2-Cas1_LIC-F	TACTTCCAATCCAATGCCATGGCGGGTTGGCGCACAGTG
G2-Cas1_LIC-R	TTATCCACTTCCAATGTTAGATACGAAATTCAGGCACACC
G2-Cas2_LIC-F	TACTTCCAATCCAATGCCATGTCTTACCGGTATATGCG
G2-Cas2_LIC-R	TTATCCACTTCCAATGTTATTAAGATTCATCAAACGCCT
G3-Cas1_LIC-F	TACTTCCAATCCAATGCAATGGGTTGGCGCTCCGTAATC
G3-Cas1_LIC-R	TTATCCACTTCCAATGTTATTATCAGACGTTGTCATTTATCGC

G3-Cas2_LIC-F	TACTTCCAATCCAATGCAATGCGGTTGATGATTATGTTCG
G3-Cas2_LIC-R	TTATCCACTTCCAATGTTATTATCATAAAATGACGGTCCGATC
spacer_30-F	TCAGCTACTCCGATGGCCCATATGCGGATC
spacer_30-R	GATCCGCATATGGGCCATCGGAGTAGCTGA
spacer_splayed_28+1-R	CATCCGCATATGGGCCATCGGAGTAGCTGT
spacer_splayed_26+2-R	CTTCCGCATATGGGCCATCGGAGTAGCTCT
spacer_splayed_24+3-R	CTACCGCATATGGGCCATCGGAGTAGCACT
spacer_splayed_20+25-R	CTAGGGCATATGGGCCATCGGAGTACGACT
spacer_splayed_18+6-R	CTAGGCCATATGGGCCATCGGAGTTCGACT
spacer_splayed_16+7-R	CTAGGCGATATGGGCCATCGGAGATCGACT
spacer_NS_overhangs_1-5_0-3	ATCCGCATATGGGCCATCGGAGTAGCTGA
spacer_NS_overhangs_0-5_1-3	GATCCGCATATGGGCCATCGGAGTAGCTG
spacer_NS_overhangs_1-5_1-3	ATCCGCATATGGGCCATCGGAGTAGCTG
spacer_NS_overhangs_1-5_2-3	ATCCGCATATGGGCCATCGGAGTAGCT
spacer_NS_overhangs_2-5_1-3	TCCGCATATGGGCCATCGGAGTAGCTG
spacer_NS_overhangs_2-5_2-3	TCCGCATATGGGCCATCGGAGTAGCT
spacer_NS_overhangs_0-5_2-3	GATCCGCATATGGGCCATCGGAGTAGCT
spacer_NS_overhangs_2-5_0-3	TCCGCATATGGGCCATCGGAGTAGCTGA
spacer_NS_overhangs_3-5_1-3	CCGCATATGGGCCATCGGAGTAGCTG
spacer_NS_overhangs_1-5_3-3	ATCCGCATATGGGCCATCGGAGTAGC
spacer_NS_overhangs_3-5_2-3	CCGCATATGGGCCATCGGAGTAGCT
spacer_NS_overhangs_2-5_3-3	TCCGCATATGGGCCATCGGAGTAGC
spacer_NS_overhangs_3-5_3-3	CCGCATATGGGCCATCGGAGTAGC

spacer_NS_overhangs_0-5_3-3	GATCCGCATATGGGCCATCGGAGTAGC
spacer_NS_overhangs_3-5_0-3	CCGCATATGGGCCATCGGAGTAGCTGA
spacer_NS_overhangs_4-5_1-3	CGCATATGGGCCATCGGAGTAGCTG
spacer_NS_overhangs_1-5_4-3	ATCCGCATATGGGCCATCGGAGTAG
spacer_NS_overhangs_4-5_2-3	CGCATATGGGCCATCGGAGTAGCT
spacer_NS_overhangs_2-5_4-3	TCCGCATATGGGCCATCGGAGTAG
spacer_NS_overhangs_4-5_3-3	CGCATATGGGCCATCGGAGTAGC
spacer_NS_overhangs_3-5_4-3	CCGCATATGGGCCATCGGAGTAG
spacer_NS_overhangs_4-5_4-3	CGCATATGGGCCATCGGAGTAG
spacer_NS_overhangs_4-5_0-3	CGCATATGGGCCATCGGAGTAGCTGA
spacer_NS_overhangs_0-5_4-3	GATCCGCATATGGGCCATCGGAGTAG
spacer_NS_overhangs_5-5_1-3	GCATATGGGCCATCGGAGTAGCTG
spacer_NS_overhangs_1-5_5-3	ATCCGCATATGGGCCATCGGAGTA
spacer_NS_overhangs_5-5_2-3	GCATATGGGCCATCGGAGTAGCT
spacer_NS_overhangs_2-5_5-3	TCCGCATATGGGCCATCGGAGTA
spacer_NS_overhangs_5-5_3-3	GCATATGGGCCATCGGAGTAGC
spacer_NS_overhangs_3-5_5-3	CCGCATATGGGCCATCGGAGTA
spacer_NS_overhangs_4-5_5-3	CGCATATGGGCCATCGGAGTA
spacer_NS_overhangs_5-5_4-3	GCATATGGGCCATCGGAGTAG
spacer_NS_overhangs_5-5_5-3	GCATATGGGCCATCGGAGTA
spacer_NS_overhangs_0-5_5-3	GATCCGCATATGGGCCATCGGAGTA
spacer_NS_overhangs_5-5_0-3	GCATATGGGCCATCGGAGTAGCTGA
spacer_Sy_overhangs_3-F	AGTCGTTACTGGTGAACCAGTTTCAAT
spacer_Sy_overhangs_3-R	GAAACTGGTTCACCAGTAACGACTGAG
spacer_Sy_overhangs_4-F	GTCGTTACTGGTGAACCAGTTTCAAT

spacer_Sy_overhangs_4-R	AAACTGGTTCACCAGTAACGACTGAG
spacer_Sy_overhangs_5-F	TCGTTACTGGTGAACCAGTTTCAAT
spacer_Sy_overhangs_5-R	AACTGGTTCACCAGTAACGACTGAG
spacer_Sy_overhangs_6-F	CGTTACTGGTGAACCAGTTTCAAT
spacer_Sy_overhangs_6-R	ACTGGTTCACCAGTAACGACTGAG
spacer_Sy_overhangs_7-F	GTTACTGGTGAACCAGTTTCAAT
spacer_Sy_overhangs_7-R	CTGGTTCACCAGTAACGACTGAG
G1-L-Target	TGATTTTATAATCACTATGTGGGTATAAAAACGT CAAAATTTCATTTGAGGTTTTTGTACTCTCAAGA TTTAAGTAACTGTACAACGGGTGGTTGGCTGAC GCATCGCAATATTAA
G1->G2_Target	TGATTTTATAATCACTATGTGGGTATAAAAACGT CAAAATTTCCTACGAGGTTTTTGTACTCTCAAGA TTTAAGTAACTGTACAACGGGTGGTTGGCTGAC GCATCGCAATATTAA
G1->G3_Target	TGATTTTATAATCACTATGTGGGTATAAAAACGT CAAAATTTCAATTTCGGTTTTTGTACTCTCAAGA TTTAAGTAACTGTACAACGGGTGGTTGGCTGAC GCATCGCAATATTAA
G1->E_Target	TGATTTTATAATCACTATGTGGGTATAAAAACGT CAAAATTTCGTGCGCCGTTTTTGTACTCTCAAGA TTTAAGTAACTGTACAACGGGTGGTTGGCTGAC GCATCGCAATATTAA
G2-L_Target	GAGACAAATAGTGCGATTACGAAATTTTTTAGAC AAAAATAGTCTACGAGGTTTTAGAGCTATGCTG

	TTTTGAATGGTCCCAAAACGGGTGGTTGGCTGACGCATCGCAATATTAA
G2->G1_Target	GAGACAAATAGTGCGATTACGAAATTTTTTAGACAAAAATAGTATTTGAGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACGGGTGGTTGGCTGACGCATCGCAATATTAA
G2->G3_Target	GAGACAAATAGTGCGATTACGAAATTTTTTAGACAAAAATAGTAATTTCGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACGGGTGGTTGGCTGACGCATCGCAATATTAA
G2->E_Target	GAGACAAATAGTGCGATTACGAAATTTTTTAGACAAAAATAGTGTGCGCCGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACGGGTGGTTGGCTGACGCATCGCAATATTAA
G3-L_Target	CTTGTTGATTGGTACTAACTGTCCGATTAAAAACTGATTATAAAATTTCGGTTTTAGATGGTTGTTAGATCAATAAGGTTTAGATCGGGTGGTTGGCTGACGCATCGCAATATTAA
G3->G2_Target	CTTGTTGATTGGTACTAACTGTCCGATTAAAAACTGATTATAACTACGAGGTTTTAGATGGTTGTTAGATCAATAAGGTTTAGATCGGGTGGTTGGCTGACGCATCGCAATATTAA
G3->G1_Target	CTTGTTGATTGGTACTAACTGTCCGATTAAAAACTGATTATAAATTTGAGGTTTTAGATGGTTGTTAG

	ATCAATAAGGTTTAGATCGGGTGGTTGGCTGACGCATCGCAATATTAA
G3->E_Target	CTTGTTGATTGGTACTAACTGTCCGATTAAAAACTGATTATAAGTGCGCCGTTTTAGATGGTTGTTAGATCAATAAGGTTTAGATCGGGTGGTTGGCTGACGCATCGCAATATTAA
E-L_Target	GCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCAGTGCGCCCGGTTTATCCCCGCTGATGCGGGGAACACGGGTGGTTGGCTGACGCATCGCAATATTAA
E->G2_Target	GCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCACTACGAGCGGTTTATCCCCGCTGATGCGGGGAACACGGGTGGTTGGCTGACGCATCGCAATATTAA
E->G1_Target	GCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCAATTTGAGCGGTTTATCCCCGCTGATGCGGGGAACACGGGTGGTTGGCTGACGCATCGCAATATTAA
E->G3_Target	GCAGAGGCGGGGGAACTCCAAGTGATATCCATCATCGCATCCAAATTTCGCGGTTTATCCCCGCTGATGCGGGGAACACGGGTGGTTGGCTGACGCATCGCAATATTAA
G1Rand114-Target	TTAGATAACATGATTAGCCGAAGTTATTTGAGGTTTTATACGGGATATTGACCGTAAACTCCTCCTC

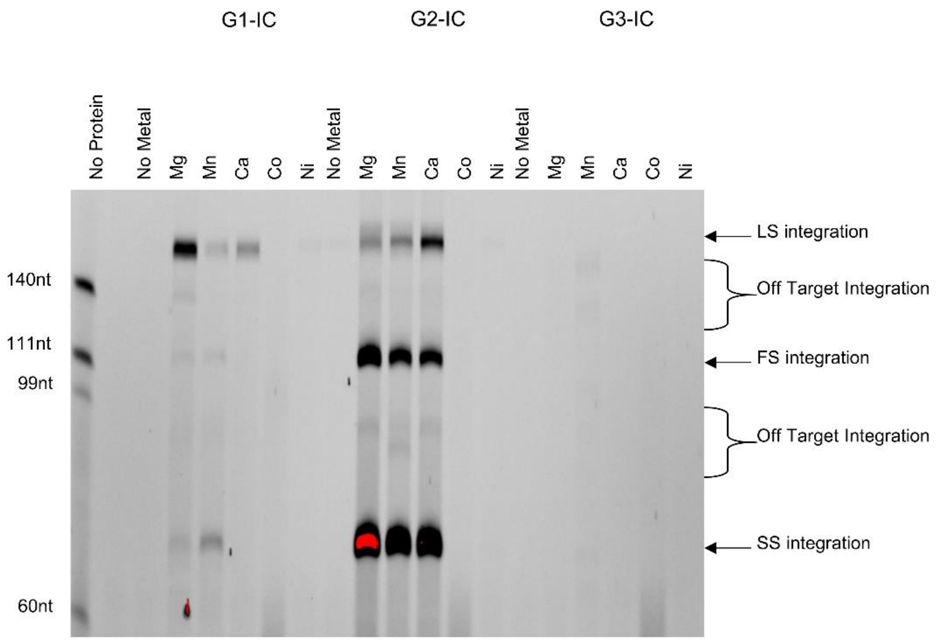
	GGGTGTGGTTCCTTTATTTGATAATATGCAACCGCTACCATTATTGATT
G1Rand75-Target	TTAGATAACATGATTAGCCGAAGTTATACGGGATATTGACCGTAAACTCCTCCTCGGGTGTGGTATTTGAGGTTTTTCCTTTATTTGATAATATGCAACCGCTACCATTATTGATT
G2Rand114-Target	TTAGATAACATGATTAGCCGAAGTTCTACGAGGTTTTATACGGGATATTGACCGTAAACTCCTCCTCGGGTGTGGTTCCTTTATTTGATAATATGCAACCGCTACCATTATTGATT
G2Rand75-Target	TTAGATAACATGATTAGCCGAAGTTATACGGGATATTGACCGTAAACTCCTCCTCGGGTGTGGTCTACGAGGTTTTTCCTTTATTTGATAATATGCAACCGCTACCATTATTGATT
G1-HP-Target	ATTTCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACTGCGCTGGTTGATTTACATGTCTCTCTCGATAGAGAGAGACATGTAAATCAACCAGCGCAGTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGAAAT
G2-HP-Target	TAGTCTACGAGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACTGCGCTGGTTGATTTACATGTCTCTCTcgatagAGAGAGACATGTAAATCAACCAGCGCAGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACCTCGTAGACTA

G3-HP-Target	TAAAATTTCGGTTTTAGATGGTTGTTAGATCAATAAGGTTTAGATCTGCGCTGGTTGATTTACATGTCTCTCTCGATAGAGAGAGACATGTAAATCAACCAGCGCAGATCTAAACCTTATTGATCTAACAACCATCTAAAACCGAAATTTTA
G2-HP-Target_NoGAG	TAGTCTACGAGGTTTTAGCGCTATGCTGTTTTGAATGGTCCCAAAACTGCGCTGGTTGATTTACATGTCGCGCGCGATAGCGCGCGACATGTAAATCAACCAGCGCAGTTTTGGGACCATTCAAAACAGCATAGCGCTAAAACCTCGTAGACTA
Random Haripin Target	TAGCAAGGCTTCAGTCGCGCGTCCGAATCTAGCTCTACTTTAGAGGCATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGCCTCTAAAGTAGAGCTAGATTCGGACGCGCGACTGAAGCCTTGCTA
G1-Rand+Rep-HP-Target	TAGCGCGGCTTGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACAAGCCGCGCTA
G1-Rand+Rep+4-HP-Target	TAGCGCGTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCACGCGCTA

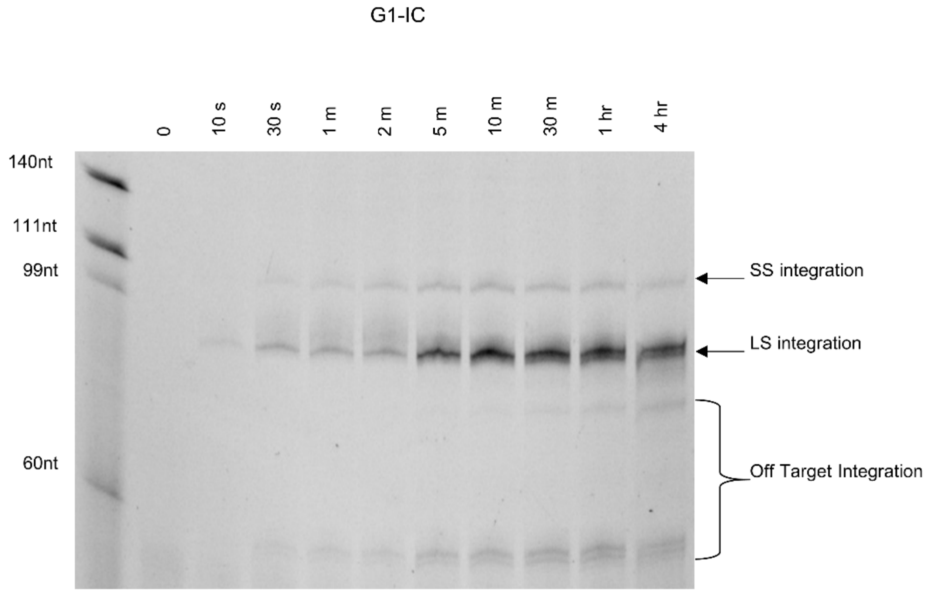
G1-Rand+Rep+7-HP-Target	TAGCATTTGAGGTTTTTGTACTCTCAAGATTTAAGTAACTGTACAACATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGTTGTACAGTTACTTAAATCTTGAGAGTACAAAAACCTCAAATGCTA
G2-Rand+Rep-HP-Target	TAGCAAGGCTTGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACAAGCCTTGCTA
G2-Rand+Rep+4-HP-Target	TAGCAAGCGAGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACCTCGCTTGCTA
G2-Rand+Rep+7-HP-Target	TAGCCTACGAGGTTTTAGAGCTATGCTGTTTTGAATGGTCCCAAAACATAAGTAACACCACCACTGCGACCCTACGATAGTAGGGTCGCAGTGGTGGTGTTACTTATGTTTTGGGACCATTCAAAACAGCATAGCTCTAAAACCTCGTAGGCTA

Supporting Figures

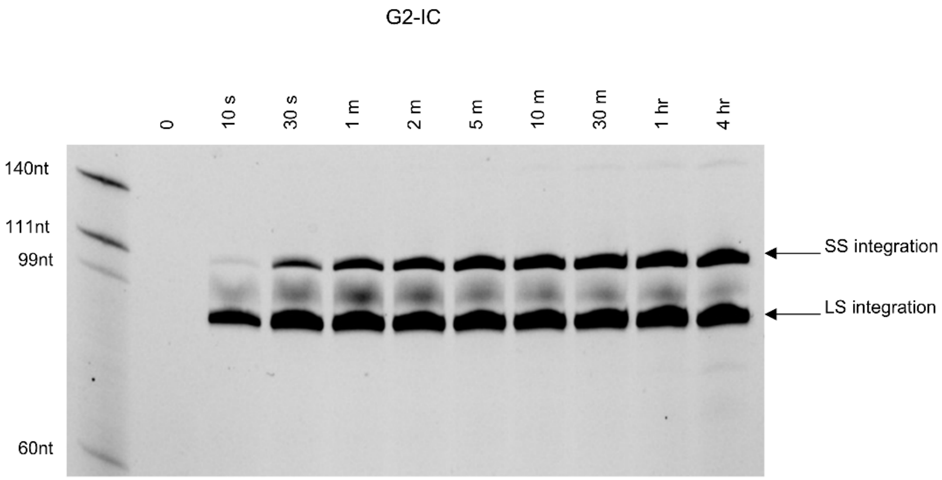
Supporting figures were taken directly from the online supporting information published alongside Chapter 3 which can be found at https://www.jbc.org/content/early/2020/06/08/jbc.RA120.013554/suppl/DC1.



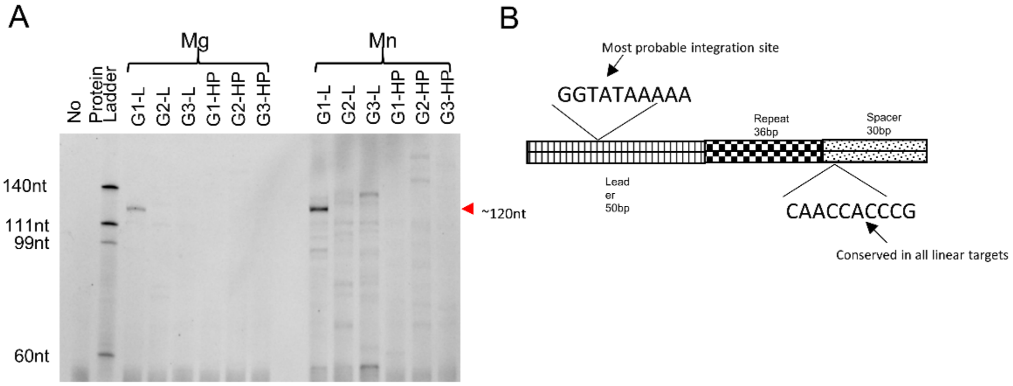
Supporting Figure S1 - FAM image of Urea-Formamide PAGE showing integration by all three groups of proteins into their cognate hairpin targets using different divalent metal cations. G1-IC shows LS activity with Mg, Mn and Ca. While Mg shows the best LS intensity, the similar amounts of LS, FS, and SS produced by Mn, along with less of the off-target band just below 140nt, make it the metal of choice because of increased specificity. G2-IC shows good activity with Mg, Mn, and Ca, with Mg being the most. G3-IC shows little activity against the G3-HP target; however, Mn shows some very light bands compared to the other blank lanes.



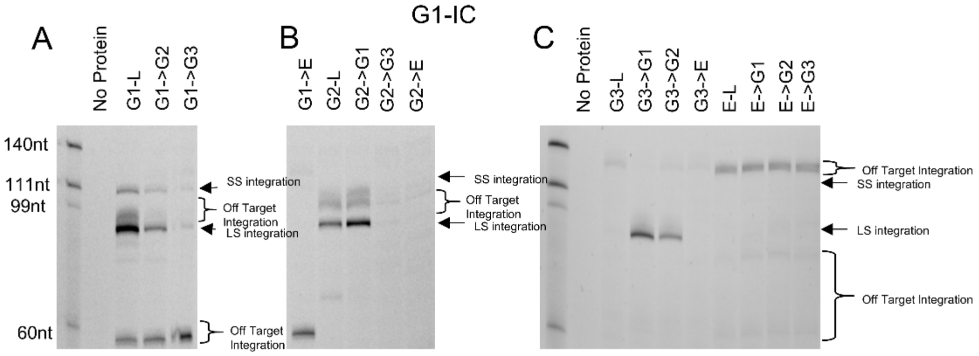
Supporting Figure S2 - FAM image of Urea-Formamide PAGE showing a time course of integration (both leader side (LS) and spacer side (SS)) by G1-IC into G1-L. LS integration is seen very quickly, at 10 seconds (s), while SS integration builds up to a maximum at about 5 minutes (m). Several off-target bands accumulate over 4 hours (hr), however, the delayed intensities show how LS integration is the preferred site. This also shows that in G1, LS is preferred over SS since the intensities are lower for SS even at longer time points. This is different from G2 (see figure S3). G1-IC was at 500 nM.



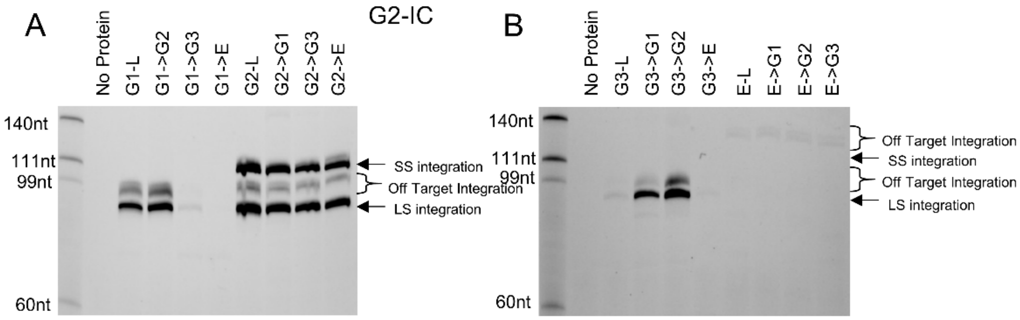
Supporting Figure S3 - FAM image of Urea-Formamide PAGE showing a time course of integration (both leader side (LS) and spacer side (SS)) by G2-IC into G2-L. Similar to G1, intense LS integration is seen immediately at 10 seconds (s) while SS integration maximizes later, at about 10 minutes (m) for G2. In contrast to G1, an intense off target band is seen between LS and SS, which slowly deteriorates over time. This could be due to disintegration activity of G2-IC. No other off target integrations are seen. Compared to G1, the amount of SS compared to LS is higher in G2 at 4 hours (hr), which may indicate a more efficient FS integration reaction. G2-IC was at 500 nM.



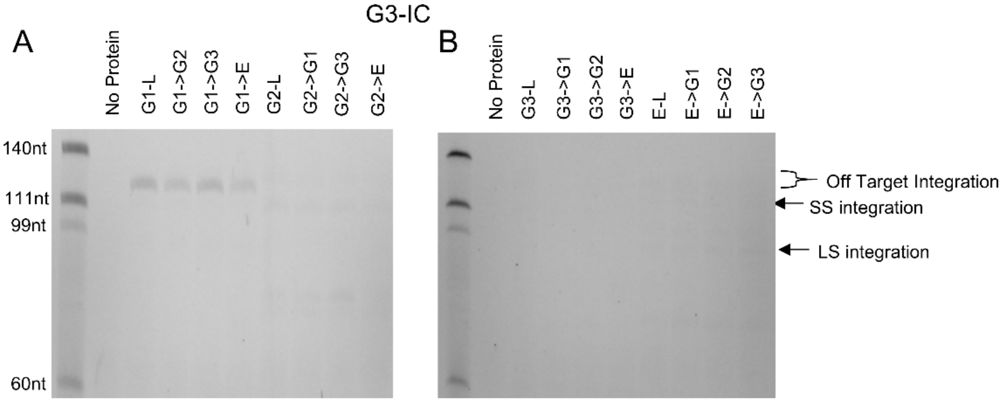
Supporting Figure S4 – A) FAM image of Urea-Formamide PAGE showing integration of G3-IC into various targets with either Mg or Mn. Much greater activity is seen when using Mn, however, the activity is very promiscuous. B) Schematic of mapped integration sites from the band positions in G1-L. Since the spacer sequence is conserved in all linear targets, and the band is not found in all linear targets, the most likely site of integration is in the leader.



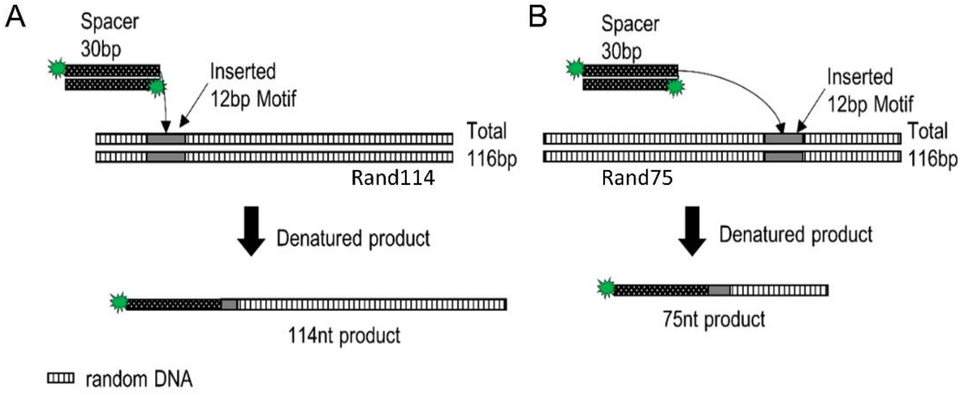
Supporting Figure S5 - Representative FAM images of Urea-Formamide PAGE gels for G1 integration assays used for integration score quantifications in Figure 2B. Each lane has a different target sequence, the first part of the name indicating which backbone sequence was used to make the target and the second part indicating the 7 bp leader-end motif present.



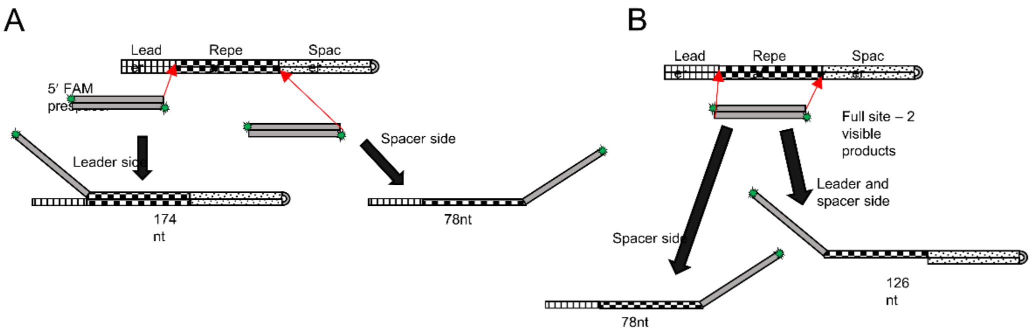
Supporting Figure S6 - Representative FAM images of Urea-Formamide PAGE gels for G2 integration assays used for integration score quantifications in Figure 2C. Each lane has a different target sequence, the first part of the name for all mutant sequences indicating which backbone sequence was used to make the target and the second part indicating the 7 bp leader-end motif present.



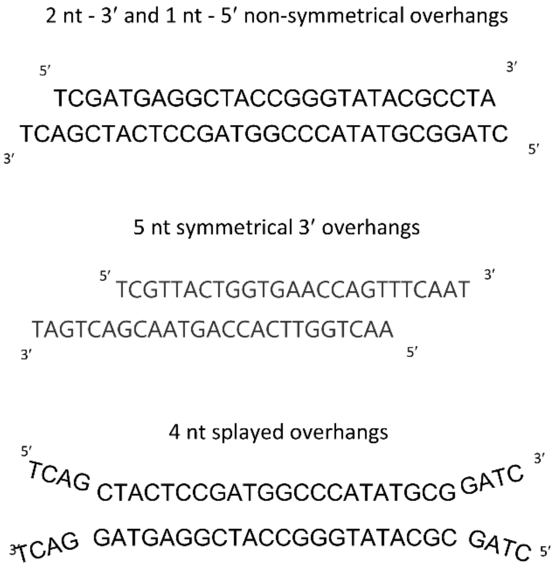
Supporting Figure S7 - Representative FAM images of Urea-Formamide PAGE gels for G3 integration assays. Each lane has a different target sequence, the first part of the name indicating which backbone sequence was used to make the target and the second part indicating the 7 bp leader-end motif present. G3-IC integrates promiscuously into targets containing the G1 backbone.



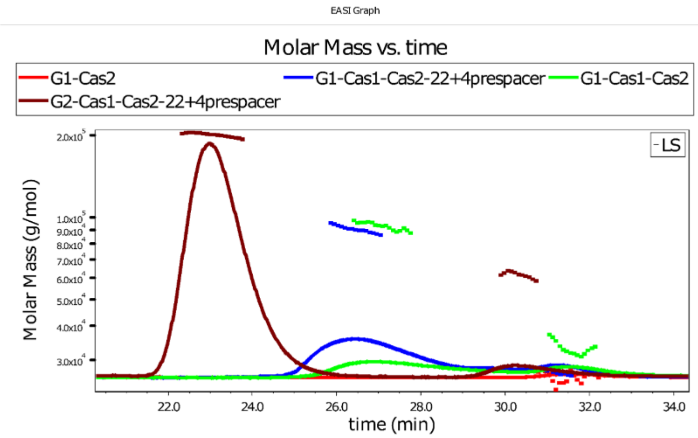
Supporting Figure S8 - Schematic explaining the design of the randomized linear DNA targets for integration assays in Fig 3.



Supporting Figure S9 - Schematic of prespacer integration into hairpin targets. A single 5′ FAM labelled prespacer can integrate at either the leader side, spacer side, or both sides of the repeat. This results in the three differently sized products shown, which are distinguishable on a denaturing polyacrylamide-urea gel.

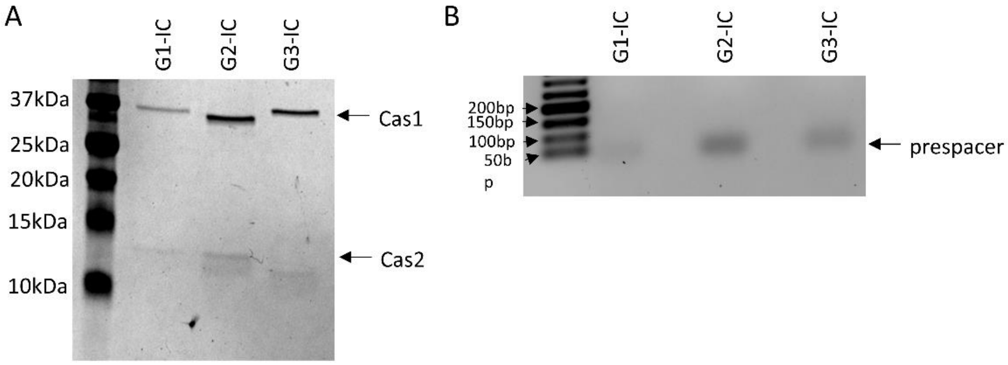

2 nt - 3′ and 1 nt - 5′ non-symmetrical overhangs
5′
3′
TCGATGAGGCTACCGGGTATACGCCTA
TCAGCTACTCCGATGGCCCATATGCGGATC
3′
5′
5 nt symmetrical 3′ overhangs
5′
TCGTTACTGGTGAACCAGTTTCAAT
3′
TAGTCAGCAATGACCACTTGGTCAA
3′
5′
4 nt splayed overhangs
5′
TCAG
CTACTCCGATGGCCCATATGCG
GATC
3′
3′
TCAG
GATGAGGCTACCGGGTATACGC
GATC
5′

Supporting Figure S10 – Prespacer design to check for subsrate preferences. Non-symmetrical prespacers were designed by annealing the full reverse strand to one of the forward strands, as shown above. In total 36 different non-symmetrical prespacers were used. Symmetrical prespacers had the same length of overhang on each side and were only designed for 3′ overhangs. Splayed prespacers had the same number of nts in each strand but were mismatched at the ends to create splaying.

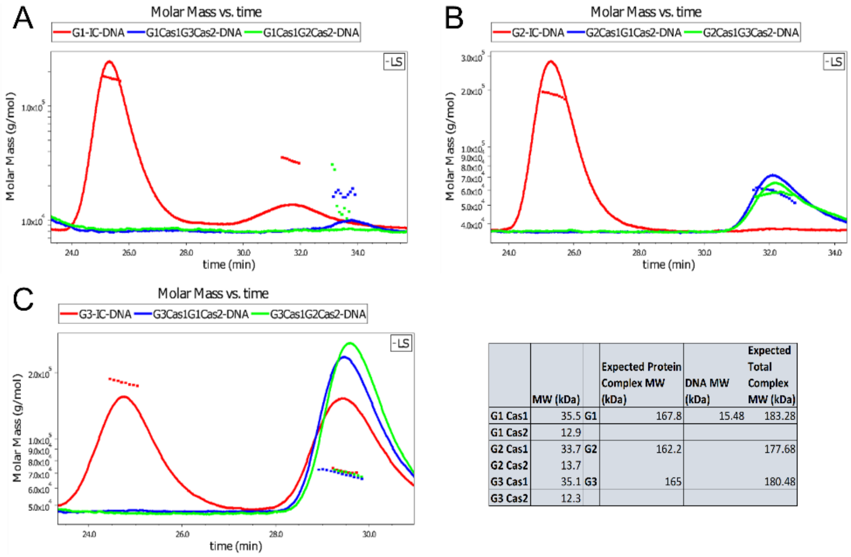


	Expected MW	Observed MW
G1 Cas2	12927	26030
G1 Cas1-Cas2	167822	93090
G1 Cas1-Cas2-prespacer	186395	90330
G2 Cas1-Cas2-prespacer	180501	200300

Supporting Figure S11 - Graph showing the light scattering signal and the molecular weight calculations for G1 Cas2, G1 Cas1-Cas2, G1 Cas1-Cas2-prespacer, and G2 Cas1-Cas2-prespacer analyzed by SEC-MALS. The DNA used in this graph was the 4 nt splayed prespacer (22+4 prespacer), in contrast to the 5 nt symmetrical 3′ overhangs prespacer used in other MALS experiments (Fig. 7, Fig. S12). The calculated MW for each peak is shown in the table along with the expected MW of each monomer. Both Cas1 and Cas2 form dimers when isolated on a gel filtration column, so we would expect peaks corresponding to dimer MWs when not in an integration complex. Based on the light scattering signal, the prespacer does not associate with the Cas1-Cas2 complex and the full integration complex does not form in G1, while it does in G2.



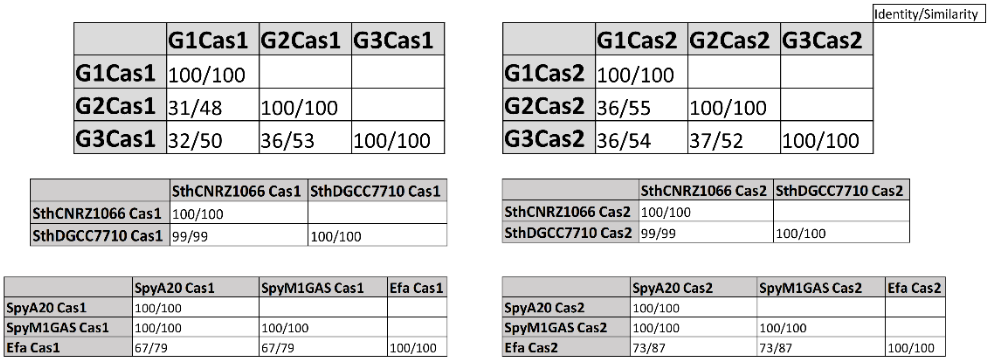
Supporting Figure S12 – A) SDS PAGE of fractions taken from the G1-IC, G2-IC, and G3-IC peaks during SEC-MALS analysis. The gel shows the presence of both Cas1 and Cas2 in the complex containing peak. B) Agarose gel analysis of the same fractions shown in A. All three fractions show the presence of prespacer DNA (5 nt 3′ overhangs) as well, showing that indeed the SEC-MALS peaks analyzed contain all three components of the complex.



	MW (kDa)		Expected Protein Complex MW (kDa)	DNA MW (kDa)	Expected Total Complex MW (kDa)
G1 Cas1	35.5	G1	167.8	15.48	183.28
G1 Cas2	12.9				
G2 Cas1	33.7	G2	162.2		177.68
G2 Cas2	13.7				
G3 Cas1	35.1	G3	165		180.48
G3 Cas2	12.3				

	Measured MW (kDa)
G1-IC (Peak 1)	175
G1-IC (Peak 2)	34
G1Cas1-G2-Cas2-DNA	N/A
G1Cas1-G3-Cas2-DNA	N/A
G2-IC	189
G2Cas1-G1-Cas2-DNA	58
G2Cas1-G3-Cas2-DNA	58
G3-IC (Peak 1)	181
G3-IC (Peak 2)	72
G3Cas1-G1-Cas2-DNA	69
G3Cas1-G2-Cas2-DNA	70

Supporting Figure S13 – Graphs showing the light scattering signal and the molecular weight calculations for various Cas1-Cas2-DNA complexes analyzed by SEC-MALS. Only cognate Cas1 and Cas2 complexes were able to associate with each other and form the canonical Cas1$_4$-Cas2$_2$-DNA$_1$ complex. Interchanging of Cas1 and Cas2 complexes between groups resulted in no complex formation. The suspected G1 Cas1 dimer peak only measured at 34 kDa, which means it is a monomer rather than a dimer. This contrasts the other peaks for the G2 Cas1 dimer and the G3 Cas1 dimer, which measure closer to dimer MWs. G1 Cas1 appeared to oligomerize into a large complex when paired with a non-cognate Cas2. Cas2 protein peaks did not show up well by themselves because of their low MW and low concentration during these experiments.



Identity/Similarity

	G1Cas1	G2Cas1	G3Cas1
G1Cas1	100/100		
G2Cas1	31/48	100/100	
G3Cas1	32/50	36/53	100/100

	G1Cas2	G2Cas2	G3Cas2
G1Cas2	100/100		
G2Cas2	36/55	100/100	
G3Cas2	36/54	37/52	100/100

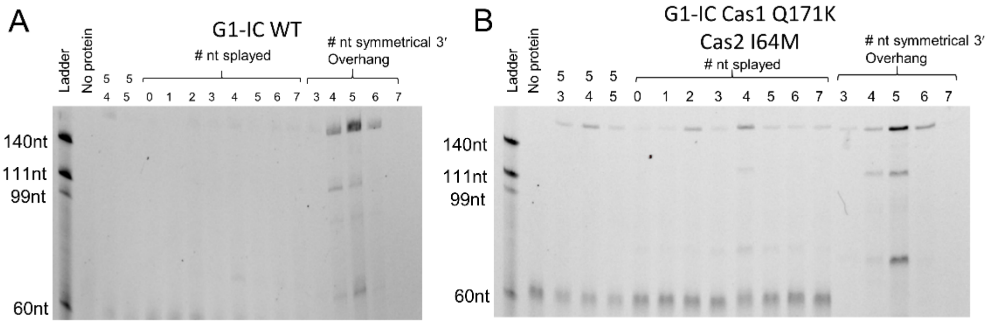
	SthCNRZ1066 Cas1	SthDGCC7710 Cas1
SthCNRZ1066 Cas1	100/100	
SthDGCC7710 Cas1	99/99	100/100

	SthCNRZ1066 Cas2	SthDGCC7710 Cas2
SthCNRZ1066 Cas2	100/100	
SthDGCC7710 Cas2	99/99	100/100

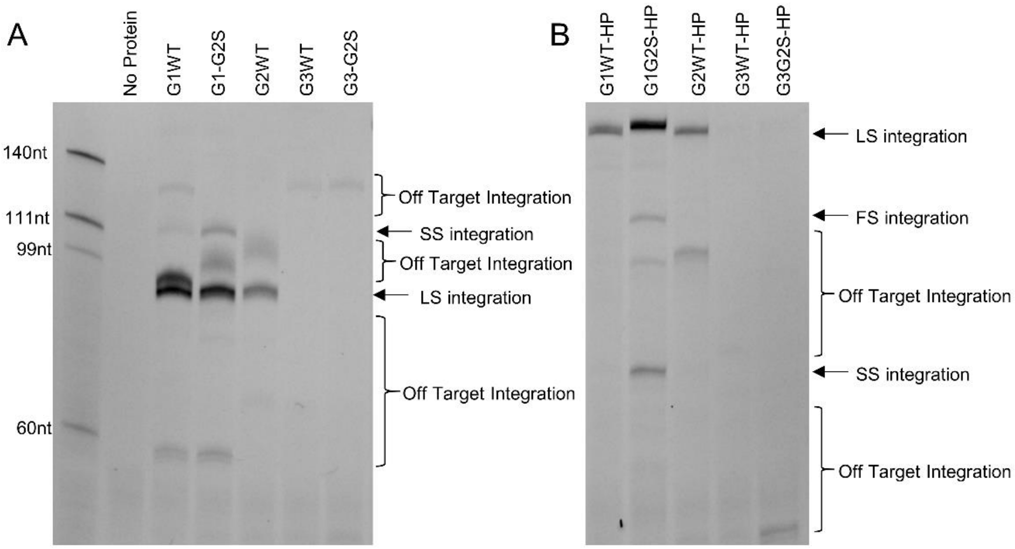
	SpyA20 Cas1	SpyM1GAS Cas1	Efa Cas1
SpyA20 Cas1	100/100		
SpyM1GAS Cas1	100/100	100/100	
Efa Cas1	67/79	67/79	100/100

	SpyA20 Cas2	SpyM1GAS Cas2	Efa Cas2
SpyA20 Cas2	100/100		
SpyM1GAS Cas2	100/100	100/100	
Efa Cas2	73/87	73/87	100/100

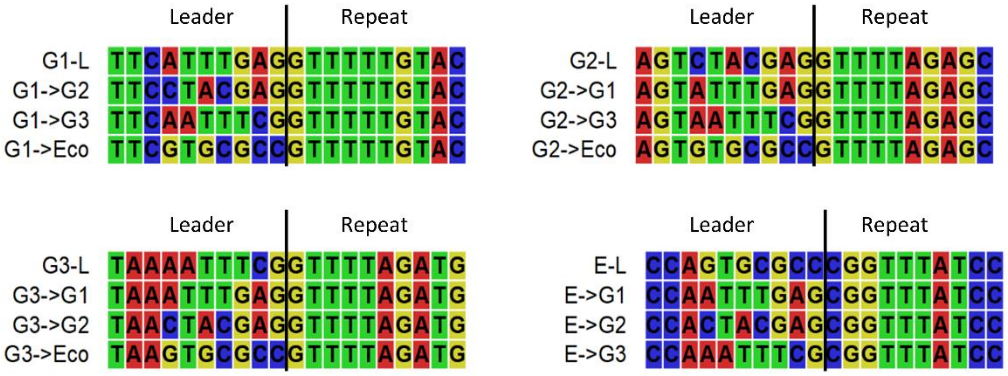
Supporting Figure S14 – Protein sequence identity/similarity matrices for various types of Cas1 and Cas2 proteins discussed in this study.



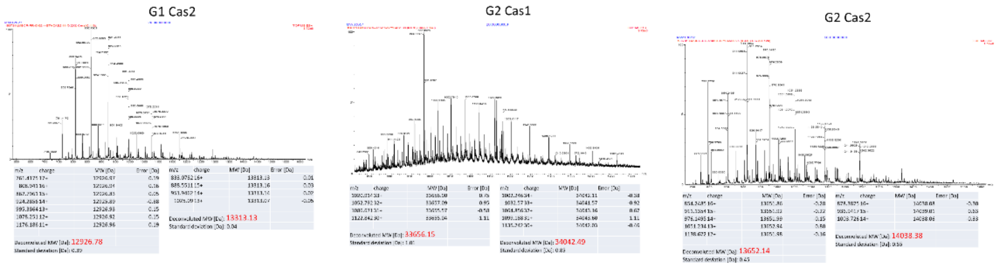
Supporting Figure S15– Comparison of activity of G1-IC with the mutant G1-IC Cas1 Q171K and Cas2 I64M, which is identical to Cas1-Cas2 from another G1 member . The gels are similar to Gel 3 from figure 6A.

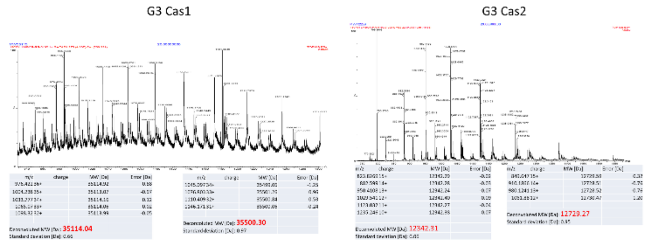


Supporting Figure S16 – Effect of spacer secondary structure in integration assays. Each lane has a different target sequence, either containing the first spacer sequence from the type II-A locus of S. thermophilus (WT) or the first spacer sequence from the type II-A locus of S. pyogenes (G2S). Integration activity, most notably in SS, increased when G2S was inserted into G1 (G1-G2S). Hairpin targets showed an increase in both SS and FS integration with the G1G2S-HP target when compared to G1WT-HP. These increases necessitated the use of G2S targets throughout the study, and nomenclature throughout the main figures (G1-L, G2-L, G3-L).


Leader
Repeat
G1-L TTCATTTGAGGTTTTTGTAC
G1->G2 TTCCTACGAGGTTTTTGTAC
G1->G3 TTCAATTTCGGTTTTTGTAC
G1->Eco TTCGTGCGCCGTTTTTGTAC
Leader
Repeat
G2-L AGTCTACGAGGTTTTAGAGC
G2->G1 AGTATTTGAGGTTTTAGAGC
G2->G3 AGTAATTTCGGTTTTAGAGC
G2->Eco AGTGTGCGCCGTTTTAGAGC
Leader
Repeat
G3-L TAAAATTTCGGTTTTAGATG
G3->G1 TAAATTTGAGGTTTTAGATG
G3->G2 TAACTACGAGGTTTTAGATG
G3->Eco TAAGTGCGCCGTTTTAGATG
Leader
Repeat
E-L CCAGTGCGCCCGGTTTATCC
E->G1 CCAATTTGAGCGGTTTATCC
E->G2 CCACTACGAGCGGTTTATCC
E->G3 CCAAATTTCGCGGTTTATCC

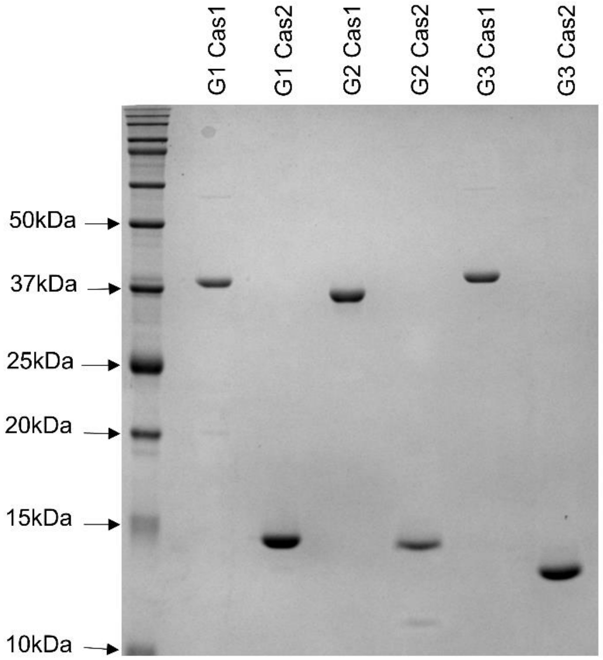
Supporting Figure S17 - Sequence alignment of all linear targets. Each group of sequences shows how the last 7 nt of the leader were changed to form each mutant target.





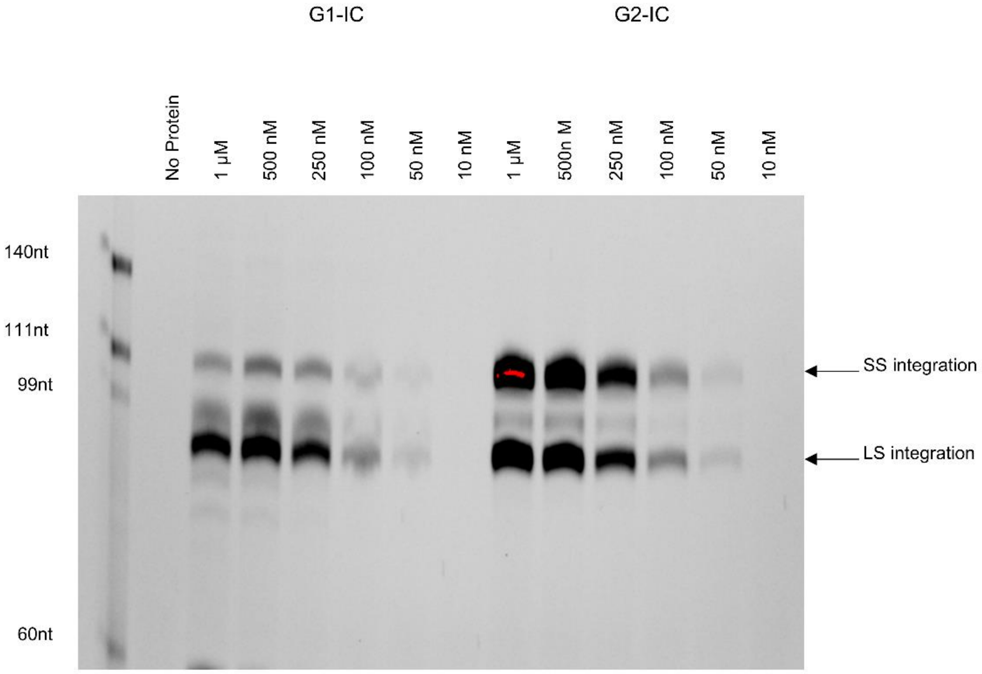
	Expected MW (Da)
G1 Cas2	12927
G2 Cas1	33657
G2 Cas2	13652
G3 Cas1	35115
G3 Cas2	12343

Supporting Figure S18 – Mass spectrometry analysis of 5/6 of the Cas proteins used in this study. All masses are within 1 Da of the expected full length MW. G1 Cas1 was not compatible with the mass spectrometry buffer and was verified to be present at the correct size by SDS-PAGE.



Expected Sizes (Da):
G1Cas1 – 35493 G1Cas2 – 12926
G2Cas1 – 33656
G2Cas2 – 13652
G3Cas1 – 35114
G3Cas2 – 12342

Supporting Figure S19 - SDS-PAGE showing purity of Cas1 and Cas2 proteins from the three groups. Each protein shows >95% purity.



Supporting Figure S20 - FAM image of Urea-Formamide PAGE showing a protein titration of G1-IC and G2-IC into cognate linear targets, meaning G1-IC with G1-L and G2-IC with G2-L. Both sets of proteins showed maximum activity at 500 nM. Reaction times were 30 minutes. A final concentration of 500 nM was used for all integration assays.