Microbial Warfare: Illuminating CRISPR adaptive immunity using single-molecule fluorescence

Loeff, Luuk

DOI
10.4233/uuid:08c08aec-53f0-4419-ba97-11fbb5a3dd49

Publication date
2017

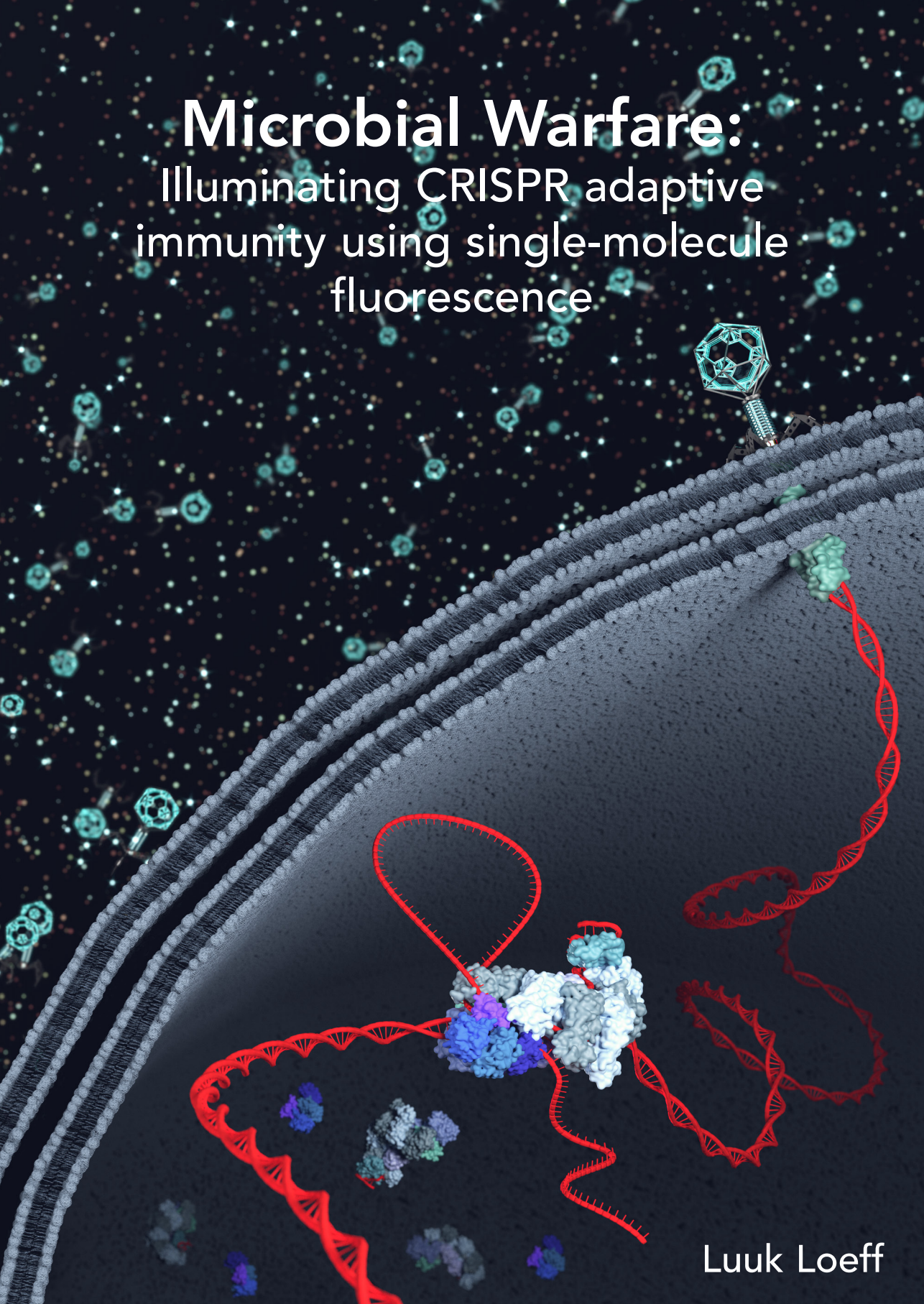
Document Version
Final published version

Citation (APA)
Loeff, L. (2017). *Microbial Warfare: Illuminating CRISPR adaptive immunity using single-molecule fluorescence*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:08c08aec-53f0-4419-ba97-11fbb5a3dd49

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Microbial Warfare:

Illuminating CRISPR adaptive immunity using single-molecule fluorescence



Luuk Loeff

Microbial Warfare:

Illuminating CRISPR adaptive immunity using single-molecule fluorescence

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar op vrijdag 6 oktober 2017 om 12:30 uur

door

Luuk LOEFF

Master of Science in de Biomoleculaire Wetenschappen
Vrije Universiteit Amsterdam, Nederland
geboren te Den Haag, Nederland

This dissertation has been approved by the

promotor:	Prof. Dr. C. Dekker &
copromotor:	Dr. C. Joo

Composition of the doctoral committee:

Rector Magnificus	Chairman	
Prof. Dr. C. Dekker	Promotor	Delft University of Technology
Dr. C. Joo	Copromotor	Delft University of Technology

Independent members:

Prof. Dr. B. Wiedenheft	Montana State University
Prof. Dr. E. Woo	Korea Research Institute of Bioscience & Biotechnology
Prof. Dr. M. Dogterom	Delft University of Technology
Dr. J. Hohlbein	Wageningen University
Dr. S.J.J. Brouns	Delft University of Technology

Reserve member:

Prof. Dr. A. Engel	Delft University of Technology

Printed by: Gildeprint
Cover Image: L. Loeff

Casimir PhD series: 2017-30
ISBN: 978-90-8593-314-4

An electronic version of this dissertation is available at http://repository.tudelft.nl/

Table of Contents

1

How prokaryotes mediate CRISPR adaptive immunity

In preparation
Springer Book: "Biophysics of RNA-Protein Interactions"

Luuk Loeff & Chirlmin Joo**

** Corresponding author

Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, 2628 CJ, Delft, The Netherlands

1.1 Abstract

Prokaryotes are constantly threatened by a large array of viruses and other mobile genetic elements. The evolutionary arms race between these prokaryotes and their invaders has resulted in a wide arsenal of defense mechanism, that enable the host to fight off the invaders. Among these defense mechanisms is an adaptive and inheritable immune system that is conveyed through Clustered regularly interspaced short palindromic repeats (CRISPR) and their CRISPR associated proteins (Cas). Immunity relies on the integration of short stretches of invasive nucleic acids (spacers) into the genome of the host. Subsequent, transcription and processing of these spacers result in small crRNA molecules that guide Cas proteins for sequence specific target degradation. In this chapter, we will review the molecular mechanisms of CRISPR immunity, with a main focus on the *E. coli* type I-E CRISPR-Cas system.

1.2 Introduction

Living systems have to constantly adapt to the ever-changing environment in order to survive. As a consequence, evolution has driven each species to have diverse survival strategies. For example, prokaryotic viruses (bacteriophages) are ten times more abundant than their prokaryotic hosts [1–4]. Yet, despite this shear abundance of bacteriophages, prokaryotes are one of the most abundant life forms on planet earth [5, 6]. To cope with this high load of invaders, prokaryotes have evolved numerous defense mechanisms that act on various stages of the bacteriophages life cycle. The combination of these defense mechanisms has allowed prokaryotes to fight off the invading bacteriophages and thrive in a wide variety hostile and competitive of environments.

Analogous to immune systems in humans, defense mechanisms in prokaryotes can be divided into innate (Figure 1.1A, Figure 1.1B & Figure 1.1C) and adaptive immune systems (Figure 1.1D). Innate immune systems are non-specific defense mechanisms that respond to invaders in a generic way, whereas adaptive immune systems are tuned towards one specific invader. In prokaryotes, innate immunity is comprised of several mechanisms that include: abortive infection mechanisms in which the host cell undergoes programmed cell death to prevent phage propagation (Figure 1.1A) [7, 8]; surface modifications that block phage uptake (Figure 1.1B) [7–9]; and restriction-modification systems that target invading DNA elements (Figure 1.1C) [7, 10]. Together these innate defense mechanisms provide the first line of defense against invading bacteriophages.

Until recently, it was thought that adaptive immune systems were exclusively found in eukaryotes. However, the perception changed with the discovery of Clustered regularly interspaced short palindromic repeats (CRISPR) and their CRISPR associated proteins (Cas). CRISPR-Cas loci are widely spread throughout prokaryotic genomes and provide an inheritable RNA-guided adaptive immune system against invading DNA or RNA [11–13]. The CRISPR loci consist of an array of repeat sequences that are separated by unique sequences called spacers. These spacers are often derived from bacteriophages or other mobile genetic elements (MGE) [11] and facilitate the recognition and destruction of MGE [12].

The CRISPR immune response is conveyed by the *cas* genes, which are usually found adjacent to the CRISPR-array. CRISPR systems function in three distinct stages, namely; (I) The adaptation stage, where Cas proteins integrate small fragments of foreign nucleic acids (spacers) into the CRISPR locus; (II) The CRISPR RNA (crRNA) biogenesis stage, in which the CRISPR locus is transcribed and processed into small interfering crRNAs by the Cas proteins; (III) The interference stage, where the crRNAs guide Cas effector complexes to complementary target sites for degradation (Figure 1.1D) [14]. In this chapter, we will provide an overview of the molecular mechanisms that underlie CRISPR-mediated defense in *E. coli*.

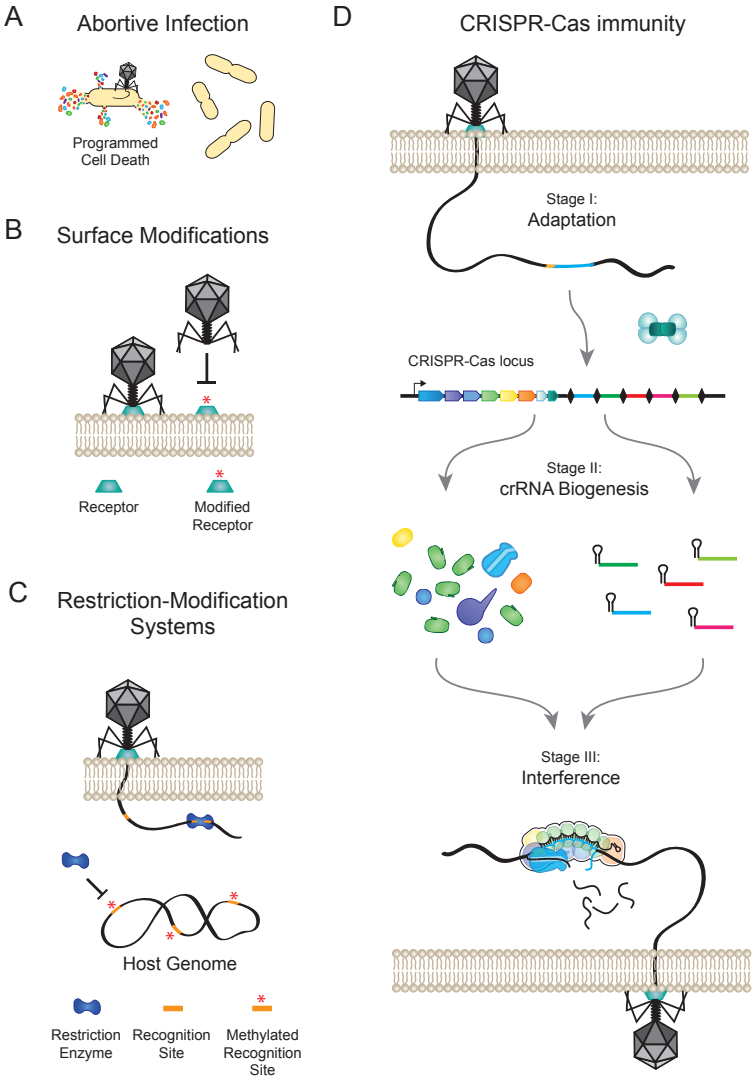


Figure 1.1: Overview of independent defense mechanisms that are found in prokaryotes

(A) Schematic representation of the innate defense mechanism called abortive infection. During an abortive infection, the infected host cell undergoes programmed cell death, to prevent phage propagation [7, 8]. **(B)** Schematic of the innate defense mechanism called surface modification. The modification of surface receptors or the expression of polysaccharides can block the absorption of bacteriophages, rending the invader harmless [7–9]. **(C)** A schematic representation of the innate restriction-modification defense system. Restriction enzymes can target and cut specific DNA sequences in the viral genome. The host genome is protected from the cleavage activity by the restriction enzyme through methylation of the genomic DNA [7, 10]. **(D)** A schematic overview of the CRISPR-Cas adaptive immune system. CRISPR immunity is conveyed in three distinct stages. During the adaptation stage, small fragments of invading DNA are incorporated into the CRISPR locus. The second stage of CRISPR immunity is crRNA biogenesis, in which the CRISPR locus is transcribed and processed into small guide RNA molecules. The last stage of CRISPR immunity is interference, where the invading DNA located and destroyed by the CRISPR-associated proteins [14].

1.3 History of CRISPR-Cas

In the early 90s, Mojica et al. set out to study salt-associated DNA modifications in the halophilic archaea *Haloferax mediterranei* [15]. Their initial DNA sequencing experiments uncovered a locus in a non-coding area of the genome of *H. mediterranei* with a peculiar pattern. The locus consisted of partially palindromic DNA repeats [16], that were separated by unknown, seemingly random, DNA sequences (Figure 1.2A). Moreover, the repeats were flanked by an open reading frame of an unknown set of proteins (Figure 1.2B). Because the function of this locus remained elusive, the term Tandem REPeats (TREP) was coined, describing the repetitive nature of the repeats. Although TERP were reported in several other species of prokaryotes (e.g. *E. coli* & *Mycobaterium spp.*) [17, 18], its importance remained unknown for more than a decade.

Advances in DNA sequencing techniques allowed for genome wide sequence analysis of prokaryotes. These experiments revealed that TREP sequences were widely distributed among archaea and bacteria, and led to an abundance of acronyms for these loci (e.g. TERP, DR, LTRR, SRSR and SPIDR) [19–25]. To reduce abundance of acronyms had accumulated over time, the term CRISPR was coined [26], which led to the recognition and definition of a new family of prokaryotic repeats. In addition, these genome sequencing experiments revealed the presence of conserved *cas* genes that were located nearby the repeat loci (Figure 1.2B). Based on the location of the Cas proteins and their sequence similarity with other proteins, it was presumed that their activity was linked to the repeats sequences [26].

Critical insights in the function of CRISPR-Cas came when it the origin of the spacer sequences was found. BLAST (Basic Local Alignment Search Tool) searches in the rapidly expanding genome databases revealed that spacers were derived from bacteriophages and conjugative plasmids [27–29]. This breakthrough led to the hypothesis that CRISPR loci may encode an adaptive immune system that protected prokaryotes against infections through RNA interference [30]. Two years later the first experimental evidence came from a study that characterized phage resistance

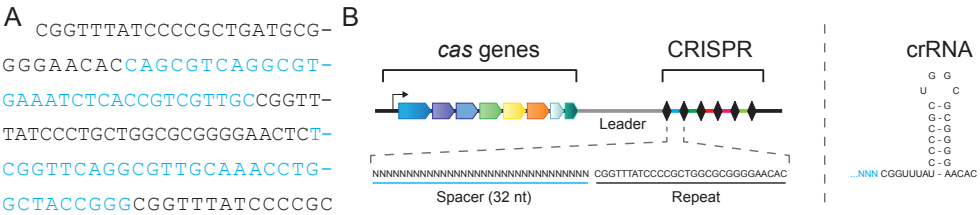


Figure 1.2: Organisation of the CRISPR locus

(A) An example of partially palindromic repeat sequences, located within the CRISPR-Array. Black sequences indicate the repeats, whereas the blue sequences indicate spacer sequences. **(B)** Schematic representation of the CRISPR locus. The CRISPR locus is comprised of a CRISPR array and a set of Cas genes. The CRISPR array holds the repeat sequences and virus derived spacers. Expression of the CRISPR array yields crRNA molecules that act as guides for the Cas proteins. The *Cas* genes encode for a set of proteins that is able to confer the immune response.

of *Streptococcus thermophilus* strains. The authors found that phage sensitive *S. thermophilus* strains could acquire resistance after being exposed to bacteriophages. Remarkably, resistance coincided with insertion of new spacers in the CRISPR array. Moreover, sequence analysis of the CRISPR array showed that the newly acquired spacers were highly similar to the genome of the phage's they were challenged with [11]. This was the first time that CRISPR-Cas adaptive immunity was caught in action.

Scientists soon began to understand the details of CRISPR-Cas based immunity. Biochemical characterization an *E. coli* CRISPR system showed that the repeats are processed by the Cas proteins into small guide RNAs, so called crRNAs (Figure 1.1D & Figure 1.2B) [12]. These crRNAs retain the virus derived spacer and is used to guide a complex of Cas proteins to target foreign DNA sequences [12, 13]. Later, it became clear that Cas proteins interfere with mobile genetic elements (MGE) through DNA cleavage [31]. These pioneering experiments established the CRISPR-Cas field and led to the characterisation of many other Cas proteins [32–35]. The discovery that CRISPR-Cas systems can be re-purposed as programmable restriction enzymes for genome engineering [36, 37], fast tracked the characterisation of CRISPR-Cas systems and shaped the field as we know it to date.

1.4 Classification of CRISPR-Cas systems

The constant evolutionary arms race between prokaryotes and their invaders has resulted in an extreme diversity of CRISPR-Cas systems [32, 33, 38, 39]. Since its discovery (see 1.3 on page 5) numerous Cas proteins have been identified and characterized, yet, new systems with novel activities are still being found [34, 40] (e.g. C2c2 a CRISPR system that targets single stranded RNA [41]). The diversity of CRISPR-Cas systems poses a challenge when it comes to annotation and classification of these systems [32, 33]. To date, CRISPR systems are classified using a two-step classification system that consists of 2 classes, 6 types and 19 subtypes [33, 34]. Despite the wealth in diversity, CRISPR systems share a common architecture: an array of alternating repeat and spacer sequences and a set of *cas* genes that convey immunity (Figure 1.2 & Figure 1.3).

Most CRISPR systems contain the two universal core proteins Cas1 and Cas2, which are responsible for the insertion of new spacers in the CRISPR array (so called adaptation, see 1.5.1 on page 12) [32, 33, 42–44]. Cas1 is the most highly conserved Cas protein making it a good maker for annotation and classification [32]. However, some functionally active CRISPR systems rely on adaptation modules from other CRISPR loci, and are therefore not equipped with an adaptation module [32]. To overcome this hurdle, a two-step classification system is used. First, CRISPR-Cas immune systems are divided into two broad classes: Class I and Class II [33], Class I CRISPR systems are characterized by the presence of multi-subunit crRNA effector complexes [12, 45] (e.g. Cascade, see 1.4.1 on page 7), whereas Class II systems carry out immunity though a single-protein (e.g. Cas9, see 1.4.2.1 on page 10) [46]. These classes are further divided into types based on the presence of signature proteins (Figure 1.3).

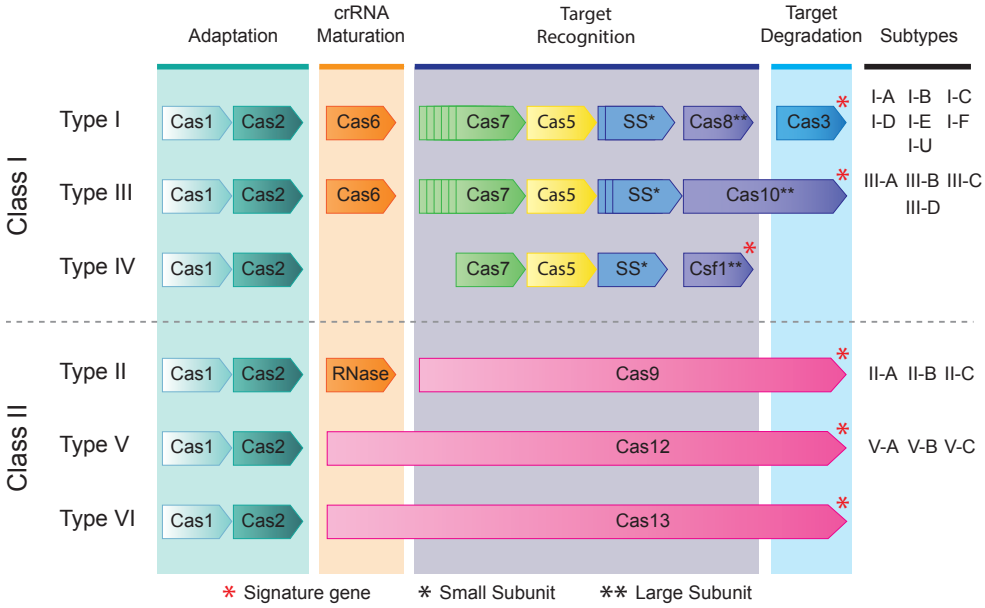


Figure 1.3: Classification of CRISPR-Cas systems

CRISPR systems can be classified using a two-step classification system. First, CRISPR systems are divided into two broad classes based on the presence of multi-subunit or single protein crRNA effector complexes. The systems are further divided into types and subtypes based on the presence of signature genes. As a result, CRISPR systems are divided into two classed, 6 types and 19 subtypes. * indicates the signature gene for the specific type. * indicates the small subunit (e.g. Cse2 of Cascade). ** indicates the large subunit (e.g. Cse1 of Cascade).

1.4.1 Class I CRISPR-Cas systems

1.4.1.1 Type I CRISPR-Cas systems

All type I CRISPR loci contain the signature gene *cas3* (or *cas3'*), which encodes a large protein with separate helicase and nuclease activities (Figure 1.3). The Cas3 helicase is highly conserved and belongs to the super family two (SF2) helicases (see Chapter 3 on page 73) [33, 47]. In most type I systems, this SF2 helicase domain is fused to a metal-dependent histidine aspartate (HD) domain with endonuclease activity (Figure 1.3) [33, 48, 49]. However, in some CRISPR loci the HD nuclease is encoded by a separate gene (*cas3''*) that is usually located adjacent to the *cas3'* gene (Cas3 helicase). Together, these proteins are responsible for target degradation during the CRISPR-interference immune response (see Chapter 3 on page 73).

Apart from the Cas3 protein, type I systems share another feature: the formation of crRNA guided Cascade (CRISPR associated complex for anti-viral defence) like complexes that are responsible for target recognition (Figure 1.4A) [12, 32, 45]. Based on the composition of these complexes, type I systems can be divided into seven subtypes: Type I-A to I-F and I-U (Figure 1.3) [33]. Each subtype has an unique combination of Cas proteins and distinct features of the operon organization. For example, in type I-C, I-D, I-E and I-F all cas genes are encoded by a single operon, whereas for type I-A and I-B the cas genes seem to be clustered in two or more operons [33].

The most extensively studied type I CRISPR system, is the I-E subtype from the model strain *E. coli* K12. The I-E subtype harbours an 11 subunit Cascade complex that is comprised of $Cas8_1$ (Cse1), SS_2 (Cse2), $Cas7_6$, $Cas5_1$ and $Cas6_1$ [50–54] (Figure 1.4A) and a Cas3 nuclease-helicase fusion (Figure 1.3). Together with the type I-F system, another well characterized type I CRISPR system from Pseudomonas aeruginosa, the type I-E system has descended from a single ancestor making these systems monophyletic [33]. Despite their similarities between these two types, the type I-F system differs in its Cas protein architecture, for example a Cas3-Cas2 fusion [55] and a 9 subunit Cascade complex: $Csy1_1$, $Csy2_1$, $Csy3_6$, and $Cas6_1$ [56, 57].

Compared to the type I-E and I-F systems, the remaining subtypes (I-A to I-D) are less well characterized. Yet, there is an increasing effort in understanding these remaining subtypes. For example, recent cryoelectron microscopy reconstructions of the type I-C Cascade revealed that this system contains a large Cas8 subunit that resembles a fusion of the Cse1 and Cse2 subunits of *E. coli* Cascade (subtype I-E) (Figure 1.4A) [58]. These continuous efforts in characterizing the type I subtypes will broaden our understanding of CRISPR immunity and may shed more light on how these CRISPR systems have diverged.

1.4.1.2 Type III CRISPR-Cas systems

Type III CRISPR loci contain the signature gene *cas10* (Figure 1.3), and encode the multi-subunit crRNA guided effector complexes: Csm (subtypes III-A and III-D) or Cmr (subtypes III-B and III-C) [21, 22]. The subtypes III-A and III-B are distinguished based on the small subunit of their effector complexes, type III-A loci contain the *csm2* gene whereas type III-B loci contain the *cmr5* gene. Moreover, for type III-B loci usually lack the *cas1*, *cas2*, and *cas6* genes and therefore require other CRISPR systems to provide this functional module [22]. The absence of some functional modules in certain subtypes provides strong evidence CRISPR-Cas systems are highly modular.

The signature protein Cas10 is the largest subunit of the type III effector complexes, which can be divided into four domains: HD domain, two palm domains, and a C-terminal α-helical domain (D4) [46]. The domain features of the Cas10 protein is what distinguishes the III-C (Cmr) and III-D (Csm) subtypes [47]. For example, in type III-C systems one of the palm domains appears to be inactive, whereas type III-D loci typically encode a Cas10 protein that lacks the HD domain [22]. Interestingly, the HD domain of Cas10 contains conserved structural motifs that are shared with the HD domain of Cas3 in the type I system [37, 48, 49].

Apart from the homology between the HD domain of Cas10 and Cas3, the type I (Cascade) and type III (Csm/ Cmr) effector complexes also share a common architecture (Figure 1.4A & Figure 1.4B) [50]. For example, in both Cascade and Csm/Cmr effector complexes the crRNA is held by proteins from the Cas7 family (e.g. Cas7 and Cmr4, Figure 1.4A & Figure 1.4B) to form a helical backbone (Figure 1.4A & Figure 1.4B) [39–43]. Even though the amino-acid sequence among Cas7 proteins from these complexes are different, the proteins share a common hand

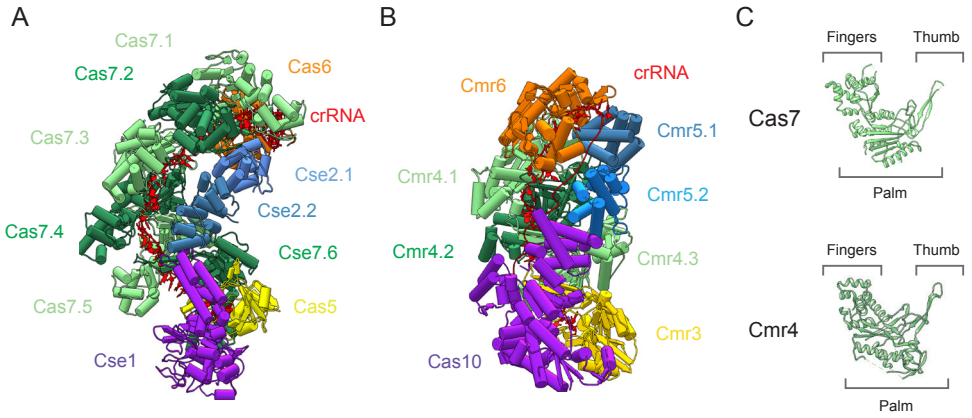


Figure 1.4: Comparison of type I and type III CRISPR systems

(A) Crystal structure of a type I Cascade effector complex at a 3.2 Å resolution [53]. The crRNA guided (Red) Cascade complex comprised of five subunits with an uneven stochiometry: $Cse1_1$ (purple), $Cse2_2$ (blue), $Cas7_6$ (green), $Cas5_1$ (yellow) and $Cas6_1$ (orange). **(B)** Crystal structure of a type III effector complex at a 2.1 Å resolution [59]. The crRNA guided (Red) CMR complex comprised of five subunits with an uneven stochiometry: $Cas10_1$ (purple), $Cmr5_2$ (blue), $Cmr4_3$ (green), $Cmr3_1$ (yellow) and $Cmr6_1$ (orange). **(C)** Comparison of the Cas7 and Cmr4 proteins of type I and type III CRISPR systems, respectively. Both Cas7 and Cmr4 protein have a hand-like fold with: palm, thumb and finger domains. In each complex multiple copies of the protein form the backbone of the respective effector complex (see [A] and [B]).

like morphology with a thumb, finger and palm domains (Figure 1.4C) [41–43]. The conserved structural similarity between type I and type III systems suggests that these systems have descended from a common ancestor [22].

Unlike type I systems that target double stranded DNA, type III systems have been shown to target single stranded (ss) RNA and ssDNA. The exact mechanism for targeting by type III systems has remained unclear for a long time [47]. For example, Csm complexes were shown to degrade DNA *in vivo* [19, 51], whereas Cmr complexes were shown to target single stranded (ss) RNA *in vitro* [34, 52, 53]. Recent reports have shed more light on how these systems work. It was shown that both Csm and Cmr complexes can bind to ssRNA transcripts, which triggers two distinct enzymatic activities: sequence specific single stranded ribonuclease activity (ssRNase) and single stranded deoxyribonuclease (ssDNase) activity [54, 55]. This leads to a transcript-activated degradation model, in which transcripts of RNA polymerase II are bound by the Csm/Cmr effector complexes with ssRNase activity [56, 57]. Such mechanism regulates the ssDNase activity ensuring that foreign nucleic acids are destroyed.

1.4.2 Class II CRISPR-Cas systems

1.4.2.1 Type II CRISPR-Cas systems

All type II CRISPR loci contain the signature gene *cas9* and are exclusively found in bacteria. The *cas9* gene encodes a single protein effector complex (Figure 1.3) with multiple domains that is responsible for both target recognition and target cleavage (RuvC and HNH domain) (Figure 1.5A) [46, 71–73]. Apart from target recognition and cleavage, the Cas9 effector protein also coordinates the incorporation of new spacers in type II systems [74]. Unlike the type I and type III systems, that use a single crRNA, Cas9 requires an additional tracrRNA for the activation of the crRNA guided effector complex [75]. The tracrRNA is partially complementary to the repeats within the respective CRISPR-array and is usually encoded in the CRISPR locus [75, 76]. The multi-functionality of the Cas9 protein makes type II systems among the most compact CRISPR systems, and has therefore been harnessed as genome engineering tool [77–79].

Based on the locus organisation, type II systems can be further divided into three distinct subtypes: type II-A to II-C [33, 35]. The subtype II-A is characterized by the presence of the signature gene *csn2*, which is involved in the integration of new spacers but is not required for target degradation [74]. By contrast, the type II-B system lacks the *csn2* gene but is characterized by the presence of the *cas4* gene that is also found in some type I systems [33]. It was shown that Cas4 exhibits 3′ to 5′ exonuclease activity [80, 81], and is likely playing a role in spacer acquisition [80]. The type II-C systems have the most minimalistic architecture, encompassing only three genes (*cas1, cas2, cas9*) [35]. The absence of the *csn2* and *cas4* genes in these loci suggests that spacer adaptation occurs through a distinct mechanism that may require additional factors [35].

1.4.2.2 Type V CRISPR-Cas systems

Type V systems are characterized by the presence of the *cas12* gene (Figure 1.3). Like type II systems, the *cas12* gene encodes a large multi-domain protein (Cpf1) that is required for both target recognition and target cleavage [82]. However, Cpf1 has some distinct features that distinguish it from Cas9 proteins. For example, where type II systems require a tracrRNA for activation, Cpf1 requires a single guide RNA (crRNA) [42, 82, 83]. Moreover, Cpf1 lacks the HNH domain that is present in type II systems. Recent crystal structures have revealed that Cpf1 depends a RuvC domain (also found in type II systems) and a Nuc domain for DNA cleavage (Figure 1.5B) [42, 83]. The distinct domain organisation of type V effector complexes make these loci different from the established type II systems [84].

A recent computational prediction has divided type V systems into three putative subtypes: V-A to V-C [34]. Each subtype is predicted to have a domain organisation that is like Cpf1 with a RuvC like nuclease domain (Figure 1.5B). It was shown that one of these subtypes (V-B, C2c1) requires a tracrRNA, which contrasts with Cpf1 [34]. Further biochemical characterisation and structural studies of these putative subtypes, will aid in understanding their functions and will help in uncovering their unique features.

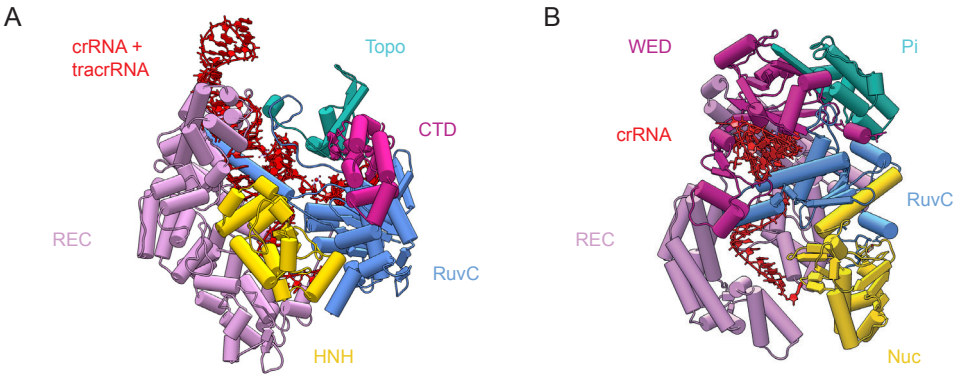


Figure 1.5: Comparison of type II and type V CRISPR systems

(A) Crystal structure of the type II Cas9 effector protein at a 2.9 Å resolution [72]. The RNA-guided (red) effector protein Cas9 can be divided into multiple domains: REC lobe (pink), HNH domain (yellow), RuvC domain (blue), CTD domain (magenta) and topo domain (green). **(B)** Crystal structure of the type II Cpf1 effector protein at a 2.8 Å resolution [42]. The RNA-guided (red) effector protein Cpf1 can be divided into multiple domains: REC lobe (pink), Nuc domain (yellow), RuvC domain (blue), WED domain (magenta) and Pi domain (green).

1.4.2.3 Type IV CRISPR-Cas systems

Type VI CRISPR loci contain the signature gene *cas13* that encodes a single protein effector complex called C2c2 (Figure 1.3). The C2c2 protein lacks homology to any of the known proteins. However, C2c2 does contain two HEPN motifs that are also found in type III CRISPR systems and higher eukaryotes. Therefore, it was predicted that C2c2 may target RNA instead of DNA [34]. Recent, biochemical characterisation of a *Leptotrichia shahii* C2c2 protein revealed that this protein can cleave ssRNA targets by using a single guide RNA [85]. Another study has shown that C2c2 exhibits two distinct RNase activities, that allows it to generate mature crRNAs and cleave ssRNA targets [86]. Further characterization and exploration of this system is required to establish if there are subtypes of this system and to establish how the function these subtypes differ.

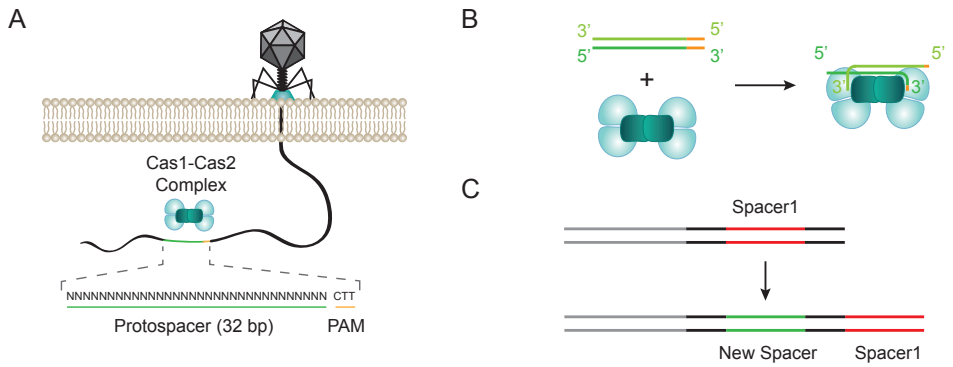


Figure 1.6: Three steps in CRISPR adaptation

(A) The first step of adaptation is the identification of viral DNA fragments (protospacers) that can be integrated by the Cas1-Cas2 complex into the CRISPR-array. New protospacers are identified based on the presence of a three nucleotide sequence motif called PAM. The PAM is located immediately adjecent to the protospacer. **(B)** In the second step is processing of the viral fragment, yielding a 33 bp protospacer that is eligable for integration. During this process the Cas1-Cas2 complex undergoes a conformational change. **(C)** The last step of adaptation is the integration of the protospacer intro the CRISPR array.

1.5 The type I-E CRISPR-Cas system

1.5.1 Adaptation

The first step in CRISPR immunity is called adaptation or naïve adaptation, which refers to the uptake of new foreign DNA fragments (spacers) in the CRISPR locus on the hosts genome (Figure 1.1). Adaptation is a complex procedure that requires at least three steps (Figure 1.6). The first step is to identify the invading DNA as a target for spacer acquisition (Figure 1.6A). Second, a DNA fragment of 33 base pairs (bp) in length (protospacer) is obtained from the foreign DNA (Figure 1.6B). Finally, the obtained spacer is integrated in the hosts CRISPR array to serve as a molecular memory against future invasions of mobile genetic elements (Figure 1.6C). The molecular basis for the adaptation process has only recently been uncovered and there is a continuous effort to obtain a comprehensive mechanistic understanding of the steps that lead to adaptation.

In the *E. coli* type I-E system, naïve spacer acquisition solely depends on two Cas proteins, Cas1 and Cas2 (Figure 1.6 & Figure 1.7) [43, 87–89], which are dispensable for later steps in CRISPR immunity, such as crRNA biogenesis (see 1.5.3 on page 17) and CRISPR interference (see 1.5.4 on page 19) [12, 90–92]. Through electrostatic and hydrophobic interactions these the Cas1 and Cas2 proteins form a stable heterohexameric complex, which is composed of two dimers of Cas1 and a bridging dimer of Cas2 (Figure 1.7) [87–89]. In its DNA-free state the complex adopts a 'wings-up' configuration, in which each Cas1 dimer represents a wing (Figure 1.7A) [87]. Upon binding a protospacer, the complex undergoes a conformational change in which the Cas1 dimers rotate downwards in the 'wings-down' configuration (Figure 1.7B). This conformational rearrangement of the Cas1-Cas2 complex likely facilitates

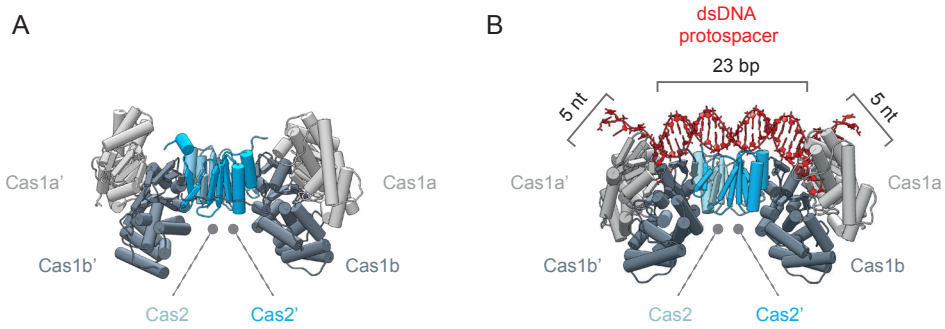


Figure 1.7: Two conformations of the Cas1-Cas2 complex

Crystal structures of the Cas1-Cas2 complex in apo and a DNA bound state. **(A)** Crystal structure of the apo Cas1-Cas2 complex at a 2.9 Å resolution [87]. The Cas1 and Cas2 proteins form a stable heterohexameric complex in which Cas2 dimer (blue) is sandwitched between two dimers of Cas1 (grey). The Cas1-Cas2 complex adops a wings-up conformation in its apo state. **(B)** Crystal structure of the Cas1-Cas2 complex bound to dsDNA at a 4.5 Å resolution [89]. The Cas1-Cas2 complex houses 23 bp dsDNA core (red), with two splayed ends of 5 nt each (red). Moreover, susbtrate binding introduces a conformational change (wings down), which likely facilitates spacer itegration.

spacer integration in the CRISPR locus [88, 89]. Single-molecule techniques such as single-molecule FRET could reveal how the conformational changes of the Cas1-2 complex coordinate spacer integration process.

For the first step of adaptation foreign DNA needs to be recognized and processed by the Cas1-Cas2 complex. The Cas1-Cas2 complex identifies suitable protospacers based on the presence of a 3 bp protospacer adjacent motif (PAM), which is also a prerequisite for the CRISPR-interference stage of immunity (see 1.5.4 on page 19) [43, 93, 94]. The absence of PAMs in the spacer flanking repeat sequences prevents self-recognition and thereby inhibits autoimmunity. Moreover, it was shown that the Cas1-Cas2 complex preferentially acquires new spacers from plasmids despite the large excess of chromosomal DNA in the cell [43].

A recent genome wide study on the origin of spacers shed light on the mechanism that drives the preference for foreign DNA [94]. It was shown that the Cas1-Cas2 complex derives new spacers from DNA degradation intermediates that are formed during the repair of double stranded DNA breaks (DSB). In *E. coli* DSB are repaired by the RecBCD complex, which is recruited to the DBS and then rapidly unwinds and degrades the DNA until it encounters a Chi site (Figure 1.8A) [91, 95, 96]. It was found that most newly acquired spacers were derived from DNA that was located between replication fork stalling sites, a common source of DSB, and the nearest Chi site [94].

The use of degradation intermediates of RecBCD generates a bias for foreign DNA by means of two mechanisms. First, the genome of *E. coli* is highly enriched for Chi sites compared to plasmid DNA, resulting in relatively small amounts self DNA for spacer integration (Figure 1.8A) [94]. In contrast, the lack of Chi sites in foreign DNA results in an excess of degradation products that can be repurposed for spacer integration by the Cas1-Cas2 complex. Second, plasmids or viral DNA are

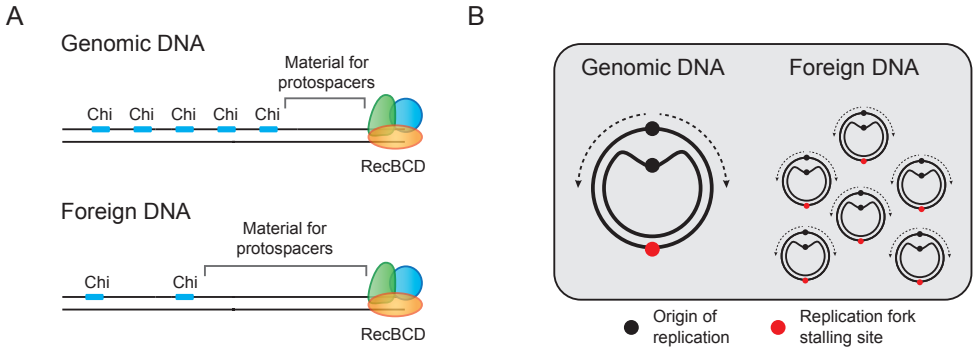


Figure 1.8: Mechanisms for self vs non-self spacer aqcuisition

(A) The Cas1-Cas2 complex re-purposes degradation products of the RecBCD complex. The *E. coli* genome is highly enriched for chi sites, which stalls degradation by RecBCD. Thereby, only a small amount of genomic DNA becomes available for spacer integration. Foreign DNA is deficient in chi sites and is thereby more extensivly processed by RecBCD. **(B)** Replication stall sites are a common source of double stranded breaks and a hotspot for new spacers. Plasmids are commonly found in high copy number, generating a bias towards foreign DNA.

commonly present in high-copy numbers. Each of these mobile genetic elements can stall the replication fork, which result in degradation intermediates that can be repurposed by the Cas1-Cas2 complex (Figure 1.8B). This suggests that acquisition by the Cas1-Cas2 complex exhibits a strong preference for high-copy DNA and thereby it limits acquisition of self DNA.

For the second step of acquisition, the Cas1-Cas2 complex captures a protospacer of 33 bp in length to integrate it into the CRISPR-array (Figure 1.6B). Recent crystal structures of the Cas1-Cas2 complex bound to a 33 bp protospacer revealed the mechanism by which Cas1-Cas2 determines the size of the protospacers [88, 89]. The complex binds a dual forked DNA substrates in which the Cas2 dimer houses 23 bp dsDNA core (Figure 1.7B) [88, 89]. The end of the substrate is bracketed by a tyrosine residue (Y22) in the Cas1 monomers, threading single stranded DNA (ssDNA) into the active site of Cas1 (Figure 1.7B) [88, 89]. One of the Cas1 monomers recognizes the 5′-CTT-3′ PAM (a PAM for type I-E systems), which positions the ssDNA overhang such that it can be cleaved within the on the C-T junction [89]. Trimming of the ssDNA overhangs on both sides results in a protospacer of 33 bp that is comprised of 32 bp of foreign DNA and the first nucleotide of the PAM (Figure 1.9A) [97]. Notably, the degradation products of RecBCD are single-stranded DNA [94], whereas the substrate for the Cas1-Cas2 has been shown to be double stranded DNA [44]. How re-annealing of the DNA strands occurs remains to be explored.

The final step of acquisition is integration of the protospacer in the CRISPR array. Directly upstream of the CRISPR array, an AT rich leader sequence is found that spans 100 to 300 bp (Figure 1.2B & Figure 1.9B) [21]. New spacers are preferentially integrated at the junction this leader sequence and the first repeat (Figure 1.6C & Figure 1.9C) [43, 98, 99]. Integration at this location results in a chronological record

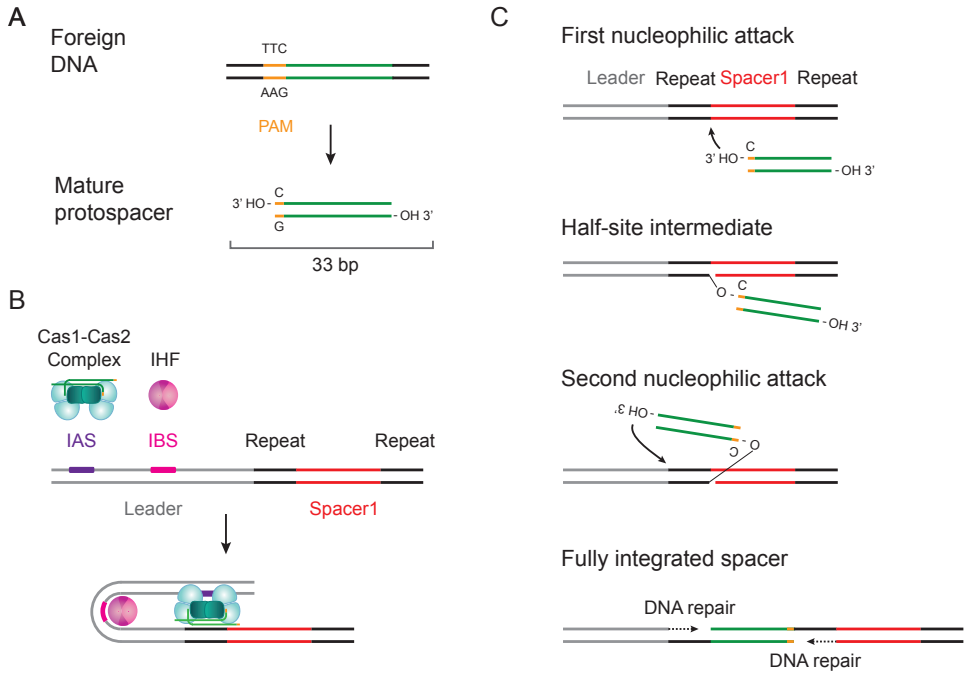


Figure 1.9: Mechanism for spacer integration by Cas1-Cas2

(A) Foreign DNA is recognized by the Cas1-Cas2 complex by means of a PAM sequence that is immediately adject to the protospacer. When the Cas1-Cas2 complex locates a pre-spacer, its is processed by the complex to yield a mature spacer. This mature spacer is comprised of 32 bp of foreign DNA and the first nucleotide of the PAM. **(B)** New spacers are consistenly inegrated at the leader end of the CRISPR-array. The AT-rich leader sequence harbors two binding sites, an IAS site that docks the Cas1-Cas2 complex and a IBS site that binds the IHF protein. When IHF binds the DNA it introduces a sharp ~160° bend that positions the Cas1-Cas2 complex on the first repeat for integration. **(C)** Spacer integration is a multi step process that requires two nucleophillic attacks. The first nucleophillic attack occurs at the minus strand of the CRISPR-array, on the junction of the first repeat and the first spacer. This attack is facillitated by the 3'-OH group of the first nucleotide of the PAM. Thereby the nucleotide of the PAM also determines the orientation of the spacer. The second nucleophillic attack occurs at the plus strand of the CRISPR array, on the junction of the leader sequence and the first repeat. Next, the resulting gaps are closed by an unknown repair mechanism to complete the integration of the new spacer.

of the invaders that have been encountered by the cell or its ancestors [43, 99]. Two sequence motifs in the leader sequence guide the integration of new spacers at this specific location [99, 100]. The integrase anchoring site (IAS) is located furthest upstream of the CRISPR array (Figure 1.8B). This motif thought to recruits the Cas1-Cas2 complex to the leader sequence [100]. The second motif encodes an integration host factor binding site (IBS) [99]. Integration host factor (IHF) is a heterodimer belonging to a family of bacterial histone-like proteins [99]. When IHF binds DNA, it introduces a sharp ~160° bend [100, 101]. Bending of the leader sequence has been suggested to position the Cas1-Cas2 complex such that the complex is located at the first repeat for spacer integration (Figure 1.9B) [100].

Next, integration at the first repeat sequence is mediated though a two-step nucleophilic attack, in which the 3′-OH ends of the protospacer are essential for integration [44]. First, the Cas1-Cas2 complex catalyses a nucleophilic attack between the 3′-OH group of the protospacer and the minus strand of the CRISPR array, resulting in a half site integration intermediate (Figure 1.9C) [44]. Second, the Cas1-Cas2 complex catalyses another nucleophilic attack between the first repeat and the leader sequence (Figure 1.9C). This results in an integrated protospacer with on either side a ssDNA gap. It is hypothesized that both ligase and polymerase activity is required to complete the integration reaction (Figure 1.9C). Notably, the Cas1-Cas2 complex determines the orientation of the new spacer based on the presence of a 3′-OH C nucleotide that origins from the 5′-CTT-3′ PAM sequence (Figure 1.9A & Figure 1.9C) [44, 97].

In recent years, substantial progress has been made in understanding the adaptation process. Yet, some outstanding questions remain unsolved, such as how does the Cas1-Cas2 complex process DNA precursors to form protospacers? What is the role of the catalytic activity of Cas2? Further biochemical, structural and single-molecule studies could greatly enhance our understanding of this process.

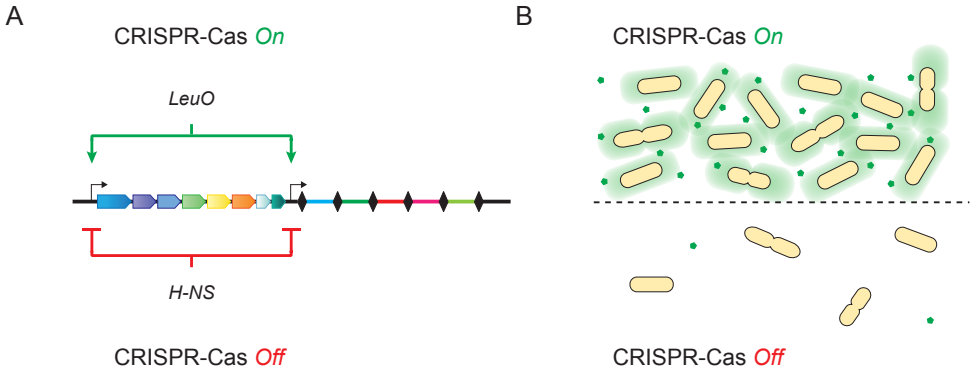


Figure 1.10: Regulation of the CRISPR-locus

(A) The CRISPR locus of *E. coli* contains multiple promotors that are tightly regulated. The heat stable H-NS protein, represses the expression of the CRISPR machinery and crRNAs. By constrast, LeuO is able to alleviate this repression. **(B)** Prokaryotes tightly regulate CRISPR expression though quorum sensing. When the concentration of autoinducers (green hexagons) is high, as a result of high cell density, CRISPR is turned on. In contrast when the level of autoinducers is low, as a result of high cell density, CRISPR is turned off.

1.5.2 Regulation of the CRISPR locus

The second stage of CRISPR immunity is crRNA biogenesis, which involves transcription of the CRISPR locus, maturation of the crRNAs and assembly of the crRNA-guided effector complex. The type I-E CRISPR locus of *E. coli* is tightly controlled by the heat-stable nucleoid structuring protein H-NS [102]. H-NS inhibits the expression of the Cas proteins and the CRISPR array, rendering the immune system inactive (Figure 1.10A) [98, 103, 104]. The H-NS binding sites are flanked by binding sites of the transcriptional activator LeuO, which can alleviate the repression of H-NS (Figure 1.10A) [102]. Apart from LeuO it has been shown that bacterial stress (e.g. envelope stress) can also activate the expression of CRISPR locus [105], suggesting that immunity in *E. coli* is thightly tuned based on its cellular state.

Tight regulation of gene expression allows bacteria to reduce the energy costs that are associated with the CRISPR immune system. For example, constitutive expression of the CRISPR locus is a costly process and would be disadvantageous when thread of a bacteriophage is absent [106]. Two recent reports, using two distinct model organisms, have shown that CRISPR immunity is modulated by sensing the cell density through quorum sensing (QS) [107, 108]. At low cell densities, when the thread of a spreading phage is low [109], CRISPR immunity is repressed (Figure 1.10B). However, at high cell densities, when the thread of a spreading phage is high [109], the cells start to produce auto-inducers that act as a transcriptional activator for CRISPR systems (Figure 1.10B). By using QS, the cells limit the costs that are associated with CRISPR immunity, and thereby increase their fitness [106].

1.5.3 CRISPR RNA biogenesis

Transcription of the CRISPR locus yields a set of Cas proteins and a long precursor crRNA (pre-crRNA) that encompasses the repeats and viral fragments (Figure 1.11A). Given the pseudo-palindromic nature of the repeat sequences, the pre-crRNA adopts a secondary stem-loop structure. Both the sequence and the shape of the stem loop, act as a hall marks for processing by the metal-independent endoribonuclease Cas6e [110, 111]. Subsequently, the Cas6e protein binds the stem loop and cleaves the pre-crRNA within the repeat. This yields a mature crRNA that is comprised of an 8 nt 5′ handle, a 32 nt spacer and a 21 nt 3′ handle with a stem loop structure (Figure 1.11B) [12]. After cleavage, the Cas6e protein remains associated to the 3′ stem loop and assembles into an effector complex with other Cas proteins [50, 51].

In *E. coli*, the Cas proteins assemble into a multi-subunit effector complex that is commonly referred to as Cascade (CRISPR associated complex for anti-viral defence) (Figure 1.12A & Figure 1.12B) [12]. The Cascade complex consists of five Cas proteins with an uneven subunit stoichiometry: $Cse1_1$, $Cse2_2$, $Cas5e_1$, $Cas6e_1$ and $Cas7_6$ (Figure 1.12A) [50–54]. These eleven subunits assemble, together with the crRNA, in a sea-horse shaped effector complex that encompasses a head, backbone, belly and tail (Figure 1.12B) [50, 51]. The head of the complex is formed the Cas6e subunit, which provides a binding site for the helical backbone of Cascade. The backbone of Cascade consists of six Cas7 subunits (Cas7.1 to Cas7.6) with a hand like shape (Figure

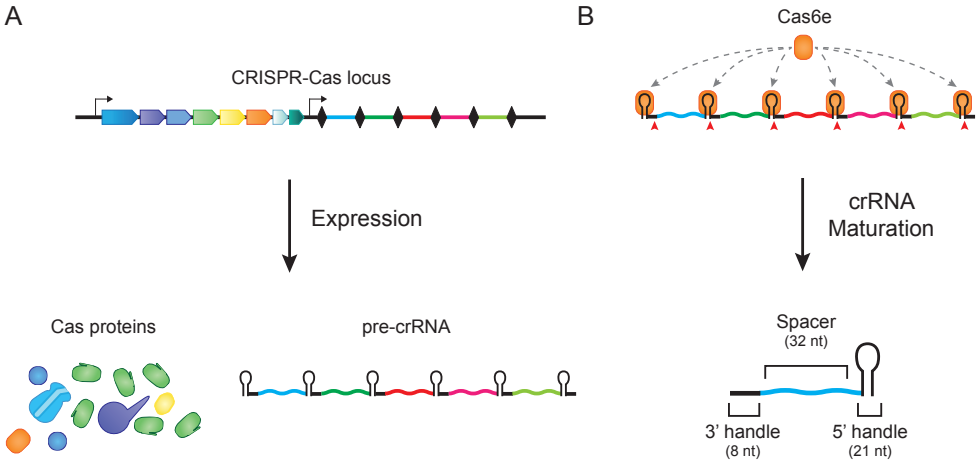


Figure 1.11: Transcription the CRISPR locus

(A) Transcription of the CRISPR locus, results in a pool of Cas proteins and pre-crRNA molecules. Given the palindromic nature of the repeats, the pre-crRNA adopts a secondary hairpin structure. **(B)** The pre-crRNA molecule is processed by the Cas6e protein that tightly binds the hairpin structures in the repeats. Subsequent cleavage by the Cas6e protein, results in mature crRNA molecule. These molecules consist of a 8 nt 3′ handle, a 32 nt spacer and a 21 nt 5′ handle with stemloop structure.

1.4C & Figure 1.12) [52–54]. The thumb of each Cas7 subunit holds and positions the crRNA at six nucleotide intervals, forming an interwoven architecture. Consequently, every sixth nucleotide of the crRNA is flipped out of plane and is unable to interact with the target DNA [52–54, 112, 113].

After assembly of the Cas7 backbone, the conserved 5′ handle of the crRNA (Figure 1.11C) is capped by the Cas5e subunit. When Cas5e binds to the 5′ handle, it forms sequence specific interactions with the RNA and it introduces a conformational change in Cas7.6. This conformational change is thought to prevent filament formation of Cas7 [53]. Strikingly, Cas5e also adopts a hand-like architecture with thumb and palm domain, suggesting that Cas5e is structurally related to Cas7 and its homologs [52–54, 59]. Additionally, the Cas5e protein serves as a binding site for the Cse1 of Cascade, which is also known as the large subunit. Together with Cas5e, Cse1 forms the tail of the complex (Figure 1.12) and is responsible for discriminating foreign DNA from genomic DNA (see 1.5.4 on page 19).

The belly of the Cascade complex is formed by the two Cse2 subunits (Cse2.1 and Cse2.2) (Figure 1.12) [50–54]. These two subunits bridge the head and the tail of the complex and are held in place by the Cas7 backbone [114]. The Cse2 dimer has two positively charged faces that are located on either side of the dimer. The charged faces are thought to stabilize the interactions with the two strands of the target DNA and are therefore, extremely important in the target binding reaction [52–54, 114].

1.5.4 CRISPR interference

The last step of CRISPR immunity is interference, which is a step wise process that involves crRNA guided cleavage of double stranded DNA (dsDNA) targets [12]. In *E. coli*, CRISPR interference relies on multiple proteins, that locate, identify and degrade target DNA substrates (called protospacers). The Cascade surveillance complex has the daunting task of locating invading protospacers amongst the vast amount of DNA in the cell. To distinguish protospacers from genomic DNA of the host, Cascade searches for a trinucleotide protospacer adjacent motif (PAM) that is located immediately upstream of the protospacer (Figure 1.13A & Figure 1.13B) [115, 116]. This PAM sequence is absent in the CRISPR locus and thereby provides a robust mechanism for discriminating "self" from "non-self" [112, 116].

Cascade exhibits a rather stringent regime when it comes to PAM recognition, triggering robust interference for only five of the 64 possible PAM sequences (5′-CTT-3′, 5′-CTA-3′, 5′-CCT-3′, 5′-CTC-3′ and 5′-CAT-3′, on the target strand) [112, 115–118]. To identify PAM sequences, Cascade uses three structural features of its Cse1 subunit, a glutamine wedge, a glycine loop and a lysine finger, that probe the minor groove of the DNA. These structural features only tolerate specific nucleotides at the different positions of the PAM, giving rise to the five PAMs that trigger interference [112, 113, 115–118]. Strikingly, all the spacers in the CRISPR-locus are flanked by a 5′-CGG-3′ PAM from the repeat sequence [117]. This PAM is the combination of the least favoured nucleotides at each position, inhibiting Cascade targeting reaction despite the perfect match with the crRNA [113].

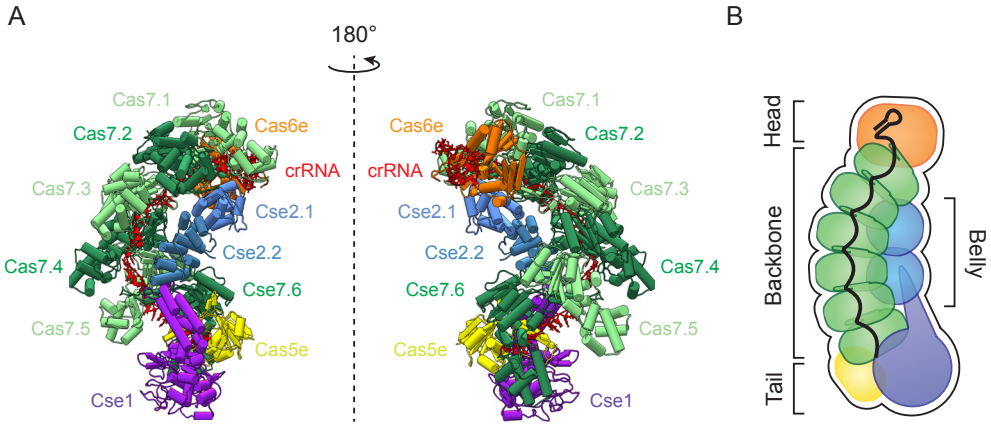


Figure 1.12: Assembly of the Cascade complex

(A) Orthogonal views of the Cascade effector complex [53]. The Cascade adopts a seahorse shaped structure with an uneven stochiometry: $Cse1_1$ (purple), $Cse2_2$ (blue), $Cas7_6$ (green), $Cas5_1$ (yellow) $Cas6_1$ (orange) and crRNA (red). **(B)** Schematic representation of the Cascade complex. Color coding for the subunits is as desribed in [A]. Cascade forms a seahore shaped complex with a head, backbone, belly and tail.

Besides the hallmark for foreign DNA, the PAM sequence carries an additional benefit. Recent studies have shown that the target search mechanism of Cascade is largely depended on 3D diffusion, finding protospacers through random collisions with the DNA [119]. If Cascade would probe the DNA for a match over the full length of its crRNA at every collision with the DNA, it would spend a substantial amount of time on the DNA. To reduce the complexity of its target search, Cascade initially screens the DNA for PAM sequences (Figure 1.13A) [119], allowing Cascade to avoid a large fraction the DNA. Thereby, Cascade can greatly reduce its search time [119–121].

Once Cascade locates a PAM sequence, it uses the glutamine wedge of the Cse1 subunit to locally melt the DNA and probe the DNA for a match with its crRNA [113]. Ultimately, a match results in hybridization between the complementary target strand and a displaced (nontarget) strand (called a R-loop) (Figure 1.13A). R-loop formation initiated at the "seed" sequence that is defined as the first eight nucleotides (with exception of the 6th nucleotide) of the protospacer (Figure 1.13B) [122]. A matching seed sequence is crucial for propagation of the R-loop [123] and is therefore a prerequisite for CRISPR interference [122]. Upon hybridization of the seed and the crRNA, the displaced nontarget strand is stabilized by positively charged residues within the Cse1 subunit, resulting in the formation of an initial recognition complex (Figure 1.13C) [113, 124].

After formation of initial recognition complex, the R-loop propagates in a directional manner, towards the PAM distal end of the protospacer (Figure 1.13D) [123–125]. Because the crRNA is held at every 6th nucleotide by the Cas7 backbone, the crRNA-DNA hybridization reaction occurs in 5 nucleotide segments [52–54, 112] (Figure 1.13D). If Cascade encounters a mismatch in one of these segments, the formation of the R-loop will stall and interference will likely be aborted [123, 125]. This directional R-loop formation, serves as a proofreading mechanism, allowing Cascade to rapidly reject off-targets without probing the remaining DNA. By using such proofreading mechanism, Cascade limits the time it spends on off targets that do not meet the requirements for interference.

When the R-loop reaches the end of the protospacer, Cascade undergoes a conformational change that involves movement of several subunits, including the head (Cas6e), tail (Cse1) and belly (Cse2) of the complex [50–54, 123, 124]. For example, the Cas6e subunit of the complex moves down towards the PAM proximal site of the protospacer. Consequently, the Cse2 dimer slides down along the backbone, rotating the Cse1 subunit of the Cascade complex [50–54, 113]. This global conformational change, positions the Cse1 and Cse2 subunits such that they can contact the displaced nucleotides that are located at each pinch point of Cas7 backbone (nucleotides 6, 12, 18, 24 and 30) [52] (Figure 1.13D). The interactions of Cse1 and Cse2 with the flipped out bases, "lock" the R-loop and result in a stable protein-DNA complex [123, 124, 126] (Figure 1.13D).

The locked R-loop licences DNA degradation by the trans-acting Cas3 protein with helicase and nuclease activities [113, 127, 128]. The *E. coli* Cas3 protein is comprised of an N-terminal metal-dependent histidine-aspartate (HD) endonuclease domain and a C-terminal superfamily 2 helicase domain (Figure 1.14) [47, 129, 130]. The Cas3

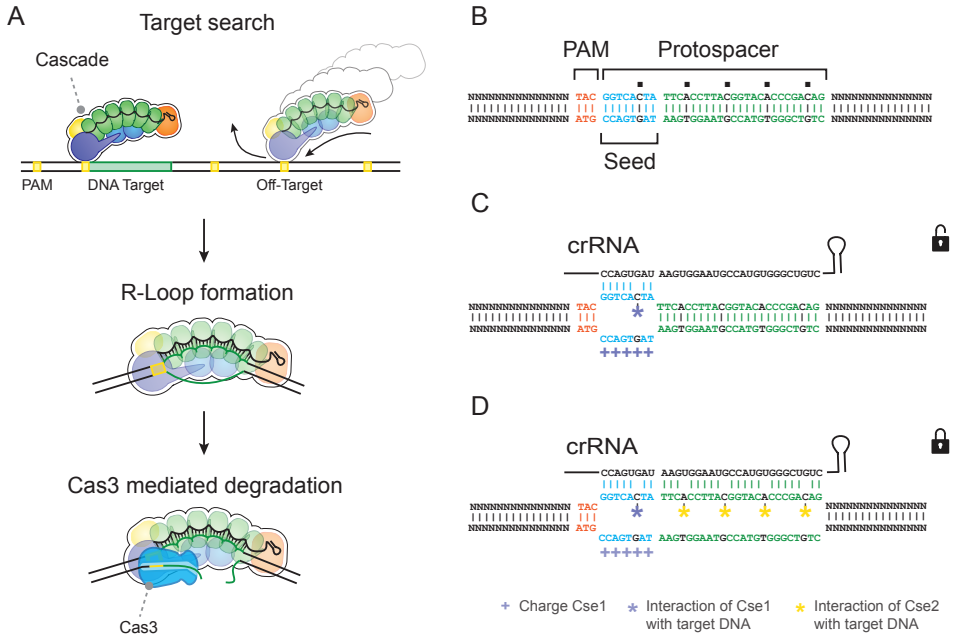


Figure 1.13: R-loop formation by the Cascade complex

(A) Steps towards target recognition by the Cascade complex. Cascade starts by screening the DNA for PAM sequences. Once Cascade has found a PAM sequence, it locally melts the DNA and probes the neighbouring DNA for complementarity with its crRNA. Once a target site is found, Cascade forms a R-loop, in which the crRNA is paired with the complementary target-site whereas the nontarget strand is displaced. The formation of the R-loop triggers the recruitment of Cas3 for subsequent target degradation. **(B)** Schematic representation of sequence elements required for R-loop formation. Protospacer adjacent motif (PAM) is highlighted in orange, the seed sequence is highlighted in blue and the protospacer is highlighted in blue and green. The black dots and nucleotides represent the location of the pinch points of the Cas7 backbone. **(C)** Schematic representation of seed bubble formation by Cascade. Cascade initially probes the seed sequence by forming a so-called seed bubble. The seed bubble is stabilized by a positively charged groove on the Cse1 subunit of Cascade. **(D)** Schematic representation of full R-loop formation. Once Cascade reaches the end of the target site it locks the R-loop through interactions between the target DNA and the Cse1 and Cse2 subunits.

helicase has a typical arrangement of SF2 helicases [47, 129, 130], with two RecA domains that are separated by a cleft in which ATP driven conformational changes take place [129, 131]. Additionally, the helicase contains a Cas3-specific accessory C-terminal domain (CTD) (Figure 1.14A), which is connected to RecA2 domain through a long linker [129, 130]. This domain may be involved in the interaction with the Cse1 subunit of Cascade that recruits Cas3 [127, 129].

When Cas3 is recruited to the Cascade marked R-loop, Cas3 nicks the foreign DNA ~11 nucleotides away from the PAM in the displaced nontarget strand (Figure 1.14B) [132–134]. This initial nick, generates a single-stranded DNA (ssDNA) overhang which facilitates loading of the helicase domain (Figure 1.14) [129] (see Chapter 3 on page 73). Once the helicase domain is loaded, the interaction between Cascade

and Cas3 becomes stable (see Chapter 3 on page 73). Upon ATP hydrolysis, Cas3 unwinds the DNA along the nontarget strand in a 3′ to 5′ direction, while remaining in tight contact with the Cascade complex [49, 129, 132, 135] (see Chapter 3 on page 73). As a result, loops are formed in the target strand (Figure 1.14) [119]. This mechanism acts as a fail-safe to ensure that Cas3 is only active on DNA that is flagged for degradation by Cascade. Thereby, limiting the potential toxic effect of off-target degradation.

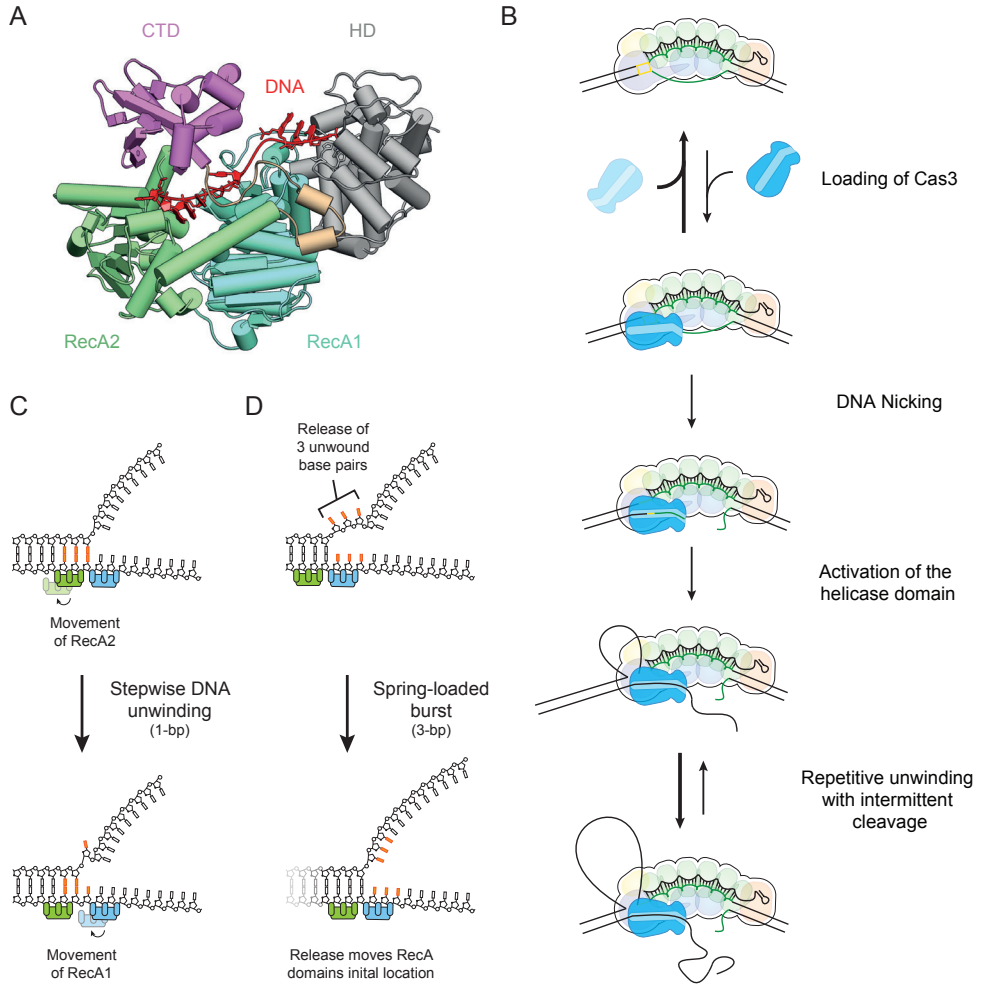


Figure 1.14: CRISPR interference by the trans-acting Cas3 protein

(A) Crystal structure of *Thermobifida fusca* Cas3 at a 2.65 Å resolution [128]. Cas3 is comprised of a histidine-aspartate (HD) endonuclease domain (grey), two RecA domains (green and cyan) and a C-terminal domain (CTD) (purple). DNA is highlighted in red. **(B)** Steps towards CRISPR mediated DNA degradation in type I systems. These steps include, recruitment of Cas3, nicking of the R-loop, loading of the helicase domain and repetitive DNA unwinding with intermittent cleavage. **(C)** To break open the dsDNA helix, Cas3 undergoes successive conformational changes, in which the RecA domains open 1-bp at a time. This process repeats until 3-bp are opened. **(D)** Cas3 holds the opened base-pairs until the third base pair is opened. This third step triggers the release of the DNA, resulting in a 3-bp spring loaded burst that moves the RecA domains to their initial location.

To break open the dsDNA helix the Cas3 helicase uses an 'inchworm-like' unwinding mechanism (see Chapter 3 on page 73). The unwinding cycle by Cas3 begins with the RecA1 monomer tightly bound to the DNA and the RecA2 monomer weakly associated with the DNA (Figure 1.14C). Upon ATP hydrolysis, the RecA2 monomer dissociates from the DNA and moves forward to tightly bind a position one base pair a head (Figure 1.14C). Next, the RecA1 monomer that was initially tightly bound, becomes weakly associated and undergoes a similar cycle (Figure 1.14C). Cas3 undergoes three of such cycles, followed by a spring-loaded burst that moves the enzyme forward by 3 base pairs (Figure 1.14D) (see Chapter 3 on page 73). This returns the helicase in its original conformation and allows Cas3 to unwind the next three base pairs in a similar fashion (Figure 1.14D). Given the conserved features of the Cas3 helicase [47], it is likely that the Cas3 helicases from other types I systems (e.g. type-IF) also unwind DNA through an 'inchworm-like' mechanism.

The 'inchworm-like' unwinding mechanism of Cas3 requires careful coordination of its RecA domains. However, the two RecA domains frequently miscoordinate, allowing the DNA to slip back and re-anneal over short and long distances (see Chapter 3 on page 73). These slipping events limit the translocation distance and allow Cas3 to repeatedly present its intrinsically deficient HD nuclease domain with single-stranded DNA (see Chapter 3 on page 73). This provides a mechanism to ensure DNA cleavage even when the nuclease domain exhibits sparse activity. As a result, Cas3 generates a distribution of degradation products that are close to spacer length. Subsequently, the fragments, of about 90 nucleotides in size, can be repurposed by the Cas1-Cas2 spacer integration complex to serve as precursors for new spacers (see Chapter 3 on page 73). This allows the CRISPR-system to update the CRISPR memory and amplify the CRISPR immune response during CRISPR interference [136]. Taken together, the stepwise recognition of the Cascade complex and the repetitive discontinuous unwinding behavior of Cas3, provides *E. coli* with an immune system that exhibits high-fidelity target detection and robust cleavage activity whilst limiting deleterious off-target effects.

1.5.5 Primed spacer acquisition

Driven by an evolutionary arms race, bacteriophages and other MGE have developed various strategies to escape CRISPR-Cas immunity. These strategies include mutations that abrogate CRISPR-Cas targeting ("escape mutants") and small proteins that bind and inhibit the CRISPR-Cas machinery (Figure 1.15) [39, 122, 137]. These small inhibiting proteins, so called anti-CRISPRs, have only recently been discovered. Therefore, it remains unclear which strategies hosts have acquired to overcome the detrimental effects of these inhibitory proteins. However, it is easy to speculate that the presence of anti-CRISPR proteins has driven the wide diversity of CRISPR-Cas systems (see 1.4 on page 6) and the existence of multiple CRISPR-Cas systems in single bacterial and archaeal strains.

Compared to anti-CRISPRs, much more is known about the hosts response against escape mutants. Evolutionary mutations in the PAM and/ or protospacer regions can result in a loss of target recognition by the immune system, allowing the MGE

to escape CRISPR immunity [64]. Typically, these escape mutations arise in the PAM and/ or the seed sequence (Figure 1.13B) that are both essential for target recognition by the Cascade complex [112, 117, 122]. Apart from these mutations, multiple mutations in the PAM distal part of the protospacer may also be detrimental to CRISPR immunity [112, 117].

To avoid the lethal effects of escape mutants, type I CRISPR systems have evolved a mechanism that is commonly referred to as primed adaptation. During the primed adaptation response, the host uses the pre-existing spacers to acquire a new set of spacers at a much higher rate from the same foreign DNA [97, 112, 117, 124, 136, 138–142]. For example, in the type I-E system the priming response enhances the acquisition rate by 10- to 20- fold over naïve acquisition [138, 139]. This memory update, allows the host to maintain immunity and keep pace with the rapidly evolving MGE. Interestingly, primed adaptation even occurs in the absence of escape mutations [97, 136, 142], suggesting that CRISPR immunity is actively maintained by the host.

The primed adaptation response requires tight coordination of the all the components involved in CRISPR immunity, including the Cascade targeting complex, Cas3 degradation module and the Cas1-Cas2 adaptation complex [97, 138–141]. This contrasts with naïve adaptation, which solely depends on the Cas1 and Cas2 proteins [43, 88]. Primed adaptation response starts with target recognition by the Cascade complex. Therefore, the position and the number of mutations in the protospacer, strongly affect the efficiency of primed spacer acquisition [112, 117]. A recent high-throughput screen of escape mutants revealed that depending on the position, primed adaptation may tolerate up to 13 mismatches, suggesting that priming is

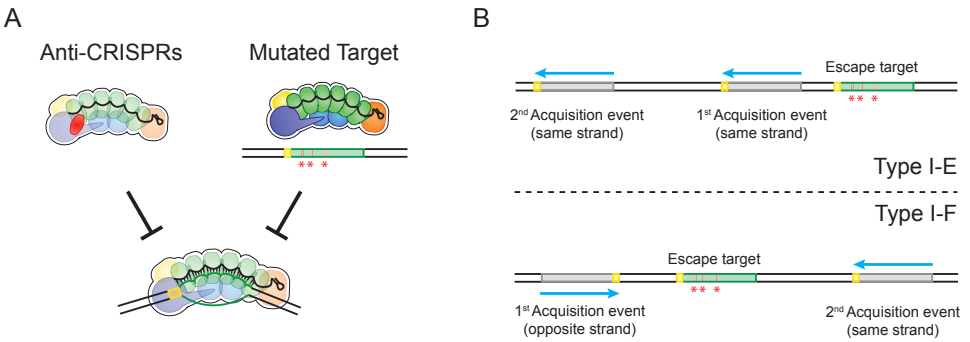


Figure 1.15: Primed adaptation in response to escape mutants

(A) Schematic of the mechanisms that allow bacteriophages to escape the CRISPR-Cas adaptive immune system. **(B)** Schematic representation of the primed adaptation in the type I-E and type I-F systems. In reponse to escape mutants, cells induce primed adaptation that results in an enhanced rate of spacer acquisition. In the type I-E system, new spacers are derrived from the same strand that was initially targeted. In contrast the type I-E system, priming in the type I-F system results in bi-directional spacer acquisition. First a spacer is acquired from the opposite strand, which fuels a secondairy primed adaptation reponse that results in spacer acquisition from the same strand.

an astonishingly robust response [112]. Consequently, these mutations trigger two distinct conformational states of the Cascade complex (see Chapter 2 on page 39) [124, 128], resulting in either Cas1-Cas2 dependent or independent recruitment of Cas3 [119, 124, 128].

Upon recruitment, Cas3 starts unwinding and degrading the invading DNA. This generates ssDNA fragments with a size of 30 to 100 nucleotides, which re-anneal to form precursors for spacer integration (pre-spacers) by the Cas1-Cas2 complex [136]. The precise mechanism for re-annealing if these ssDNA fragments remains unclear. However, the complex formation of Cascade and Cas3 during the interference response (see Chapter 3 on page 73) and the Cas1-Cas2 dependent recruitment of Cas3 to mutated targets [119, 128], suggests that Cascade, Cas3 and Cas1-Cas2 form a complex upon primed adaptation. Such complex formation allows for direct transfer of the newly generated pre-spacers to the integration complex, resulting in a fast and efficient adaptation response.

Interestingly, primed adaptation among the various type I subtypes show a distinct integration pattern (Figure 1.15B). For example, the type I-E system displays a strong bias for spacers that have been acquired from the same strand as the priming protospacer [97, 112, 117, 138], whereas in the type I-F and type I-B systems primed adaptation occurs from both strands (Figure 1.15B) [55, 140–142]. Several models have been suggested for the distinct behaviour amongst the different subtypes. In the type I-F system, the first new spacer that is typically acquired from the opposite strand of the primed protospacer, which coincides with the directional translocation of Cas3 (Figure 1.15B) (see Chapter 3 on page 73). This may fuel a secondary primed adaptation response in the opposite direction (Figure 1.15B). Alternatively, it was recently shown that two Cas2-Cas3 fusion proteins of the type I-F system form a four-lobed propeller shaped complex with four Cas1 molecules (Cas2-Cas3$_2$, Cas1$_4$) [55]. It has been suggested that the stoichiometry of this complex with two Cas3 molecules, may be responsible for the bi-directional acquisition. Even though, these results provide some mechanistic basis for primed acquisition, more extensive characterisation of priming in the different subtypes is needed to uncover the mechanistic details and order of events during primed adaptation.

1.6 Thesis outline

Chapter 2 on page 39: "Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex".

In chapter 2 we show that Cascade distinguishes bona fide targets and mutated targets using disparate binding modes. Using single-molecule FRET, we observe that the recognition of bona fide targets by Cascade is an ordered process, starting with seed bubble formation after which a complete R-loop is established. This tightly controlled and ordered process allows Cascade to recognize targets with high-fidelity. In contrast, mutated targets are recognized with low fidelity, displaying short-lived seed-independent binding that can occur from any segment of the crRNA. These dual roles of Cascade in immunity with distinct fidelities underpin CRISPR-Cas robustness, allowing for efficient degradation of bona fide targets and priming of mutated DNA targets.

Chapter 3 on page 73: "The CRISPR associated Cas3 protein repetitively probes the target DNA with a 1-nt step size".

In chapter 3 we investigate the mechanism of CRISPR interference using single-molecule FRET. We show that Cascade and Cas3 remain in tight contract while Cas3 unwinds the DNA, resulting in loops in the target strand. Cas3 unwinds DNA in distinct bursts of three base pairs that underlies three one base pair steps. Miscoordination within the helicase domain of Cas3 results in slipping, which allows Cas3 to repeatedly present the intrinsically deficient nuclease domain with ssDNA. This generates a distribution degradation products with an average size of ~90 nucleotides. Our study reveals an unanticipated level of complexity, in which the discontinuous and burst-like helicase properties of Cas3 are the driving force behind CRISR interference.

Chapter 4 on page 103: "TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms".

In chapter 4 we investigate how terminal uridylation transferases recognize and uridylate precursor microRNAs (pre-miRNA). We find that the overhang of a pre-miRNA is the key structural element that is recognized by TUT7 and its paralogues, TUT4 and TUT2. While TUT7 mono-uridylates the 1-nt overhang of group II pre-miRNAs to its canonical end structure, it generates an oligo-U tail for pre-miRNAs where the 3' end is further recessed into the stem. The oligo-U tails on the trimmed pre-miRNAs may promote rapid degradation of non-functional pre-miRNA species. In contrast processive oligo-uridylation in the presence of Lin28, both mono- and oligo-uridylation by TUT7 is conveyed through a distributive mode of uridylation. Our study reveals dual roles and mechanisms of uridylation in repair and removal of defective pre-miRNAs.

Chapter 5 on page 153: "Single-molecule pull-down for investigating protein–nucleic acid interactions".

In chapter 5 we combine single-molecule fluorescence with various protein complex pull-down techniques. We describe several different strategies and list the challenges that have to be overcome for the development of these techniques. As a proof-of-concept, we highlight three examples of protein complexes involved in small RNA biogenesis (Drosha-DGCR8, human Dicer-TRBP, Drosophila Dicer 2-Loqs-PD, and a TUT4 complex) and illustrate how we elucidate the molecular bases of their functions. With this protocol, single-molecule fluorescence can be widely used to study nucleoprotein complexes.

Chapter 6 on page 177: "A fast and automated step detection method for analysing single-molecule trajectories".

In chapter 6 we describe an automated step detection method to analyse single-molecule trajectories. The algorithm is based on chi-squared minimization, capable of detecting steps in single-molecule trajectories without any prior knowledge on their size or location. We first describe how the step finding procedure is performed and how the optimal number of fitted steps is determined. Next, we provide a description on how the selection criteria for the optimal number of steps change when data exhibits a wide variety of step-sizes and plateau lengths. Finally, these considerations lead to a robust "hands-off" fitting procedure that is suitable for various types of single-molecule trajectories.

1.7 References

1 O. Bergh, K. Y. Børsheim, G. Bratbak, M. Heldal, High abundance of viruses found in aquatic environments. *Nature*. **340**, 467–468 (1989).

2 S. Chibani-chennoufi, A. Bruttin, M. Dillmann, H. Bru, Phage-Host Interaction : an Ecological Perspective. *J. Bacteriol.* **186**, 3677–3686 (2004).

3 M. G. Weinbauer, Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).

4 M. Breitbart, F. Rohwer, Here a virus , there a virus , everywhere the same virus ? *Trends Genet.* **13**, 278–284 (2005).

5 N. R. Pace, A Molecular View of Microbial Diversity and the Biosphere. *Science*. **276**, 734–740 (1997).

6 T. P. Curtis, W. T. Sloan, J. W. Scannell, Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci.* **99**, 10494–10499 (2002).

7 S. J. Labrie, J. E. Samson, S. Moineau, Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).

8 E. V Koonin, M. Krupovic, Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nature.* **16**, 184–192 (2015).

9 E. R. Westra *et al.*, The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339 (2012).

10 K. Vasu, V. Nagaraja, Diverse Functions of Restriction-Modification Systems in Addition to. *Microbiol. Mol. Biol. Rev.* **77**, 53–72 (2013).

11 R. Barrangou *et al.*, CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*. **315**, 1709–1712 (2007).

12 S. J. J. Brouns *et al.*, Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*. **340**, 216–219 (2008).

13 L. A. Marraffini, E. J. Sontheimer, CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*. **322**, 1843–1845 (2008).

14 L. A. Marraffini, CRISPR-Cas immunity in prokaryotes. *Nature*. **526**, 55–61 (2015).

15 G. Juez, F. Rodriguez-Valera, N. Herrero, F. J. M. Mojica, Evidence for salt-associated restriction pattern modifications in the archaeobacterium Haloferax mediterranei. *J. Bacteriol.* **172**, 7278–7281 (1990).

16 F. J. M. Mojica, G. Juez, F. Rodriguez-Valera, Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol. Microbiol.* **9**, 613–621 (1993).

17 Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, A. Nakata, Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169**, 5429–33 (1987).

18 P. M. A. Groenen, A. E. Bunschoten, D. van Soolingen, J. D. A. van Errtbden, Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol. Microbiol.* **10**, 1057–1065 (1993).

19 Y. Kawarabayasi *et al.*, Complete sequence and gene organization of the genome of a hyperthermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76 (1998).

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

21 R. Jansen, J. D. A. Van Embden, W. Gaastra, L. M. Schouls, Identification of a novel family of sequence repeats among prokaryotes. **6**, 23–33 (2002).

22 H. P. Klenk *et al.*, The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature.* **390**, 364–70 (1997).

23 D. R. Smith *et al.*, Complete genome sequence of Methanobacterium thermoautotrophicum DH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155 (1997).

24 Q. She *et al.*, The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7835–7840 (2001).

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

26 R. Jansen, J. D. A. Van Embden, W. Gaastra, L. M. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).

27 F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, E. Soria, Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).

28 C. Pourcel, G. Salvignol, G. Vergnaud, CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology.* **151**, 653–663 (2005).

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

30 K. S. Makarova, N. V Grishin, S. A. Shabalina, Y. I. Wolf, E. V Koonin, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct.* **1**, 7 (2006).

31 J. E. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature.* **468**, 67–71 (2010).

32 K. S. Makarova *et al.*, Evolution and Classification of the CRISPR–Cas Systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2011).

33 K. S. Makarova *et al.*, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

34 S. Shmakov *et al.*, Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol. Cell.* **60**, 385–397 (2015).

35 K. Chylinski, K. S. Makarova, E. Charpentier, E. V. Koonin, Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* **42**, 6091–6105 (2014).

36 S. W. Cho, S. Kim, J. M. Kim, J.-S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 230–2 (2013).

37 M. Jinek *et al.*, RNA-programmed genome editing in human cells. *Elife.* **20**, 1–9 (2013).

38 A. Pawluk *et al.*, Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell.* **167**, 1829–1838.e9 (2016).

39 J. Bondy-Denomy, A. Pawluk, K. L. Maxwell, A. R. Davidson, Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*. **493**, 429–32 (2013).

40 D. Burstein, L. B. Harrington, S. C. Strutt, A. J. Probst, New CRISPR-Cas systems from uncultivated microbes. *Nature*, 1–20 (2016).

41 P. Mohanraju *et al.*, Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science*, **353**, 6299 (2016).

42 T. Yamano *et al.*, Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell*. **165**, 949–962 (2016).

43 I. Yosef, M. G. Goren, U. Qimron, Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**, 5569–5576 (2012).

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

45 C. R. Hale *et al.*, RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell*. **139**, 945–956 (2009).

46 M. Jinek *et al.*, A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. **337**, 816–822 (2012).

47 R. N. Jackson, M. Lavin, J. Carter, B. Wiedenheft, Fitting CRISPR-associated Cas3 into the Helicase Family Tree. *Curr. Opin. Struct. Biol.* **24**, 106–114 (2014).

48 S. Mulepati, S. Bailey, Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J. Biol. Chem.* **286**, 31896–31903 (2011).

49 T. Sinkunas *et al.*, Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).

50 M. M. Jore *et al.*, Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18**, 529–536 (2011).

51 B. Wiedenheft *et al.*, Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*. **477**, 486–489 (2011).

52 S. Mulepati, A. Héroux, S. Bailey, Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science*. **345**, 1479–84 (2014).

53 R. N. Jackson *et al.*, Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science.* **345**, 1473–9 (2014).

54 H. Zhao *et al.*, Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. *Nature.* **515**, 147–50 (2014).

55 M. F. Rollins *et al.*, Cas1 and the Csy complex are opposing regulators of Cas2 / 3 nuclease activity. *Proc. Natl. Acad. Sci.* **1** (2017), doi:10.1073/pnas.1616395114.

56 B. Wiedenheft *et al.*, RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci.* **108**, 10092–10097 (2011).

57 S. Chowdhury *et al.*, Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. *Cell.* **169**, 47–51.e11 (2017).

58 M. L. Hochstrasser, D. W. Taylor, J. E. Kornfeld, E. Nogales, J. A. Doudna, DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. *Mol. Cell.* **63**, 840–851 (2016).

59 T. Osawa, H. Inanaga, C. Sato, T. Numata, Crystal structure of the crispr-cas RNA silencing cmr complex bound to a target analog. *Mol. Cell.* **58**, 418–430 (2015).

60 G. Tamulaitis, Č. Venclovas, V. Siksnys, Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. *Trends Microbiol.* **xx**, 1–13 (2016).

61 C. Benda *et al.*, Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol. Cell.* **56**, 43–54 (2014).

62 T. Y. Jung *et al.*, Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure.* **23**, 782–790 (2015).

63 R. N. Jackson, B. Wiedenheft, A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. *Mol. Cell.* **58**, 722–728 (2015).

64 L. A. Marraffini, E. J. Sontheimer, Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature.* **463**, 568–71 (2010).

65 C. Rouillon *et al.*, Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell.* **52**, 124–134 (2013).

66 R. H. J. Staals *et al.*, Structure and Activity of the RNA-Targeting Type III-B CRISPR-Cas Complex of Thermus thermophilus. *Mol. Cell.* **52**, 135–145 (2013).

67 J. R. Elmore *et al.*, Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev.* **30**, 447–459 (2016).

68 M. A. Estrella, F. T. Kuo, S. Bailey, RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. *Genes Dev.* **30**, 460–470 (2016).

69 M. Kazlauskiene, G. Tamulaitis, G. Kostiuk, Č. Venclovas, V. Siksnys, Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Mol. Cell.* **62**, 295–306 (2016).

70 P. Samai *et al.*, Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. *Cell.* **161**, 1164–1174 (2015).

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

72 M. Jinek *et al.*, Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*. **343**, 1247997 (2014).

73 F. Jiang, K. Zhou, L. Ma, S. Gressel, J. A. Doudna, A Cas9 – guide RNA complex preorganized for target DNA recognition. *Science.* **348**, 1477–1481 (2015).

74 R. Heler *et al.*, Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature*. **519**, 1–16 (2015).

75 E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature.* **471**, 602–607 (2011).

76 K. Chylinski, A. Le Rhun, E. Charpentier, The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol.* **10**, 726–37 (2013).

77 P. Mali, K. M. Esvelt, G. M. Church, Cas9 as a versatile tool for engineering biology. *Nat. Methods*. **10**, 957–963 (2013).

78 J. A. Doudna, E. Charpentier, The new frontier of genome engineering with CRISPR-Cas9. *Science* (2014), doi:10.1126/science.1258096.

79 H. Wang, M. La Russa, L. S. Qi, CRISPR / Cas9 in Genome Editing and Beyond. *Annu. Rev. Biochem.* (2016), doi:10.1146/annurev-biochem-060815-014607.

80 J. Zhang, T. Kasciukovic, M. F. White, The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. *PLoS One.* **7**, 1–8 (2012).

81 S. Lemak *et al.*, Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from sulfolobus solfataricus. *J. Am. Chem. Soc.* **135**, 17476–17487 (2013).

82 B. Zetsche *et al.*, Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell.* **163**, 759–771 (2015).

83 D. Dong *et al.*, The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature.* **532**, 1–16 (2016).

84 K. S. Makarova, F. Zhang, E. V. Koonin, SnapShot: Class 2 CRISPR-Cas Systems. *Cell.* **168**, 328–328.e1 (2017).

85 O. O. Abudayyeh *et al.*, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, (2016).

86 A. East-Seletsky *et al.*, Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature.* **538**, 270–273 (2016).

87 J. K. Nuñez *et al.*, Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–34 (2014).

88 J. K. Nuñez, L. B. Harrington, P. J. Kranzusch, A. N. Engelman, J. A. Doudna, Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature.* **527**, 535–538 (2015).

89 J. Wang *et al.*, Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell.* **163**, 840–853 (2015).

90 E. V Koonin, K. S. Makarova, CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10**, 679–86 (2013).

91 G. Amitai, R. Sorek, CRISPR–Cas adaptation: insights into the mechanism of action. *Nat. Rev. Microbiol.* **advance on**, 67–76 (2016).

92 S. A. Jackson *et al.*, CRISPR-Cas : Adapting to change. *Science*, 1–9 (2017).

93 C. Díez-Villaseñor, N. M. Guzmán, C. Almendros, J. García-Martínez, F. J. M. Mojica, CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. *RNA Biol.* **10**, 792–802 (2013).

94 A. Levy *et al.*, CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature.* **520**, 505–510 (2015).

95 A. F. Taylor, G. R. Smith, RecBCD enzyme is altered upon cutting DNA at a chi recombination hotspot. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5226–5230 (1992).

96 S. K. Amundsen, A. F. Taylor, M. Reddy, G. R. Smith, Intersubunit signaling in RecBCD enzyme, a complex protein machine regulated by Chi hot spots. *Genes Dev.* **21**, 3296–3307 (2007).

97 D. C. Swarts, C. Mosterd, M. W. J. van Passel, S. J. J. Brouns, CRISPR interference directs strand specific spacer acquisition. *PLoS One.* **7**, 1–7 (2012).

98 K. Pougach *et al.*, Transcription, processing and function of CRISPR cassettes in Escherichia coli. *Mol. Microbiol.* **77**, 1367–1379 (2010).

99 J. K. Nuñez, L. Bai, L. B. Harrington, T. L. Hinder, J. A. Doudna, CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell.* **62**, 824–833 (2016).

100 K. N. R. Yoganand, R. Sivathanu, S. Nimkar, B. Anand, Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Genome Biol. Evol.*, 1–15 (2016).

101 B. M. J. Ali *et al.*, Compaction of single DNA molecules induced by binding of integration host factor (IHF). *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10658–10663 (2001).

102 E. R. Westra *et al.*, H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. *Mol. Microbiol.* **77**, 1380–1393 (2010).

103 I. Yosef, M. G. Goren, R. Kiro, R. Edgar, U. Qimron, High-temperature protein G is essential for activity of the Escherichia coli clustered regularly interspaced short palindromic repeats (CRISPR)/Cas system. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20136–20141 (2011).

104 A. Plagens, B. Tjaden, A. Hagemann, L. Randau, R. Hensel, Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. *J. Bacteriol.* **194**, 2491–2500 (2012).

105 R. Perez-Rodriguez *et al.*, Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in Escherichia coli. *Mol. Microbiol.* **79**, 584–599 (2011).

106 E. R. Westra *et al.*, Parasite exposure drives selective evolution of constitutive versus inducible defense. *Curr. Biol.* **25**, 1043–1049 (2015).

107 N. M. Høyland-Kroghsbo *et al.*, Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system. *Proc. Natl. Acad. Sci. U. S. A.*, 201617415 (2016).

108 A. G. Patterson *et al.*, Quorum Sensing Controls Adaptive Immunity through the Regulation of Multiple CRISPR-Cas Systems. *Mol. Cell.* **64**, 1–7 (2016).

109 Z. Erez *et al.*, Communication between viruses guides lysis–lysogeny decisions. *Nature.* **541**, 488–493 (2017).

110 J. Carte, R. Wang, H. Li, R. M. Terns, M. P. Terns, Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489–3496 (2008).

111 R. E. Haurwitz, M. Jinek, B. Wiedenheft, K. Zhou, J. A. Doudna, Sequence- and Structure-Specific RNA Processing by a CRISPR Endonuclease. *Science.* **329**, 1355–1358 (2010).

112 P. C. Fineran *et al.*, Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1629-38 (2014).

113 R. P. Hayes *et al.*, Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. *Nature.* **530**, 499–503 (2016).

114 P. B. G. Van Erp *et al.*, Mechanism of CRISPR-RNA guided recognition of DNA targets in Escherichia coli. *Nucleic Acids Res.* **43**, 8381–8391 (2015).

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

116 E. R. Westra *et al.*, Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. *PLoS Genet.* **9** (2013).

117 C. Xue *et al.*, CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Res.* **43**, 10831–10847 (2015).

118 R. T. Leenay *et al.*, Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol. Cell.* **62**, 137–147 (2015).

119 S. Redding *et al.*, Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. *Cell.* **163**, 854–865 (2015).

120 S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature.* **507**, 62–67 (2014).

121 M. Klein, S. D. Chandradoss, M. Depken, C. Joo, Why Argonaute is needed to make microRNA target search fast and reliable Misha. *Semin. Cell Dev. Biol.*, 1–9 (2016).

122 E. Semenova *et al.*, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 10098–10103 (2011).

123 M. Rutkauskas *et al.*, Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep.* **10**, 1534–1543 (2015).

124 T. R. Blosser *et al.*, Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex. *Mol. Cell.* **58**, 60–70 (2015).

125 M. Klein, B. Eslami-Mossallam, D. Gonzalez Arroyo, M. Depken, The Kinetic Basis Of CRISPR-Cas Off-Targeting Rules. *bioRxiv* (2017)

126 M. D. Szczelkun *et al.*, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9798–803 (2014).

127 M. L. Hochstrasser *et al.*, CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6618–23 (2014).

128 C. Xue *et al.*, Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity Short Article Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Mol. Cell.* **64**, 1–9 (2016).

129 Y. Huo *et al.*, Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**, 771–7 (2014).

130 B. Gong *et al.*, Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16359–64 (2014).

131 M. R. Singleton, M. S. Dillingham, D. B. Wigley, Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* **76**, 23–50 (2007).

132 S. Mulepati, S. Bailey, In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA Target. *J. Biol. Chem.* **288**, 22184–22192 (2013).

133 T. Sinkunas *et al.*, In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *EMBO J.* **32**, 385–394 (2013).

134 Y. Xiao *et al.*, Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR- Cas System Article Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell.* **170**, 48–60.e11 (2017).

135 E. R. Westra *et al.*, CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Mol. Cell.* **46**, 595–605 (2012).

136 T. Künne *et al.*, Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Mol. Cell.* **63**, 1–13 (2016).

137 J. Bondy-denomy *et al.*, Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature.* **526**, 136–139 (2015).

138 K. A. Datsenko *et al.*, Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).

139 E. Savitskaya, E. Semenova, V. Dedkov, A. Metlitskaya, K. Severinov, High-throughput analysis of type I-E CRISPR / Cas spacer acquisition in E . coli. *RNA Biol.* **10**, 716–725 (2013).

140 M. Li, R. Wang, D. Zhao, H. Xiang, Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42**, 2483–2492 (2014).

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

142 R. H. J. Staals et al., Interference dominates and amplifies spacer acquisition in a native CRISPR-Cas system. Nat. Commun. 23, 127–135 (2016).Acids Res. **42**, 2483–2492 (2014).

2

Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex

Molecular Cell
2015 Apr 2;58(1):60-70. doi: 10.1016/j.molcel.2015.01.028.

Timothy R. Blosser* [1, 3], **Luuk Loeff* [1]**, Edze R. Westra [2, 4], Marnix Vlot [2], Tim Künne [2], Małgorzata Sobota [2], Cees Dekker [1], Stan J.J. Brouns** [2] & Chirlmin Joo** [1]

* These authors have contributed equally to this work

** Co-corresponding authors

[1] Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, 2628 CJ, Delft, The Netherlands

[2] Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, 6703 HB, Wageningen, The Netherlands

[3] Present address: Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, Massachusetts, USA

[4] Present address: The School of Biosciences, University of Exeter, TR10 9EZ, United Kingdom

2.1 Abstract

Small RNA-guided protein complexes play an essential role in CRISPR-mediated immunity in prokaryotes. While these complexes initiate interference by flagging cognate invader DNA for destruction, recent evidence has implicated their involvement in new CRISPR memory formation, called priming, against mutated invader sequences. The mechanism by which the target recognition complex mediates these disparate responses interference and priming remains poorly understood. Using single-molecule FRET, we visualize how bona fide and mutated targets are differentially probed by *E. coli* Cascade. We observe that the recognition of bona fide targets is an ordered process that is tightly controlled for high fidelity. Mutated targets are recognized with low fidelity, which is featured by short-lived and PAM- and seed-independent binding by any segment of the crRNA. These dual roles of Cascade in immunity with distinct fidelities underpin CRISPR-Cas robustness, allowing for efficient degradation of bona fide targets and priming of mutated DNA targets.

2.2 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) loci are widely spread throughout prokaryotic genomes and provide an inheritable RNA-guided adaptive immune system against bacteriophages and mobile genetic elements [1–7]. In response to invading phages or mobile genetic elements, CRISPR-associated (Cas) proteins integrate small fragments of foreign DNA into the CRISPR array, which are subsequently processed into mature CRISPR RNAs (crRNAs). crRNAs form a complex with one Cas protein (Cas9 from Type II, see 1.4.2 on page 10) or multiple Cas proteins (Types I and III, see 1.4.1 on page 7), which utilizes the crRNA as a guide to trigger degradation of cognate invading nucleic acids. While it is DNA that is targeted in Types I and II [8], recent studies suggest that both DNA and RNA are targeted in Type III [9–13]. Among the target recognition complexes, Cas9 has been widely applied as a versatile tool for genome engineering in a broad spectrum of organisms [14, 15].

In the CRISPR-Cas/I-E system of *Escherichia coli*, mature crRNAs are incorporated into Cascade (CRISPR-associated complex for antiviral defense), an eleven-subunit complex comprised of five different Cas proteins ($Cse1_1$, $Cse2_2$, $Cas7_6$, $Cas5_1$ and $Cas6_1$) [16] (Figure 2.1A). In the CRISPR interference pathway, Cascade generates an R-loop between the crRNA and its double-stranded DNA (dsDNA) target (protospacer), which subsequently leads to target degradation by the nuclease-helicase Cas3 [17–19]. The first 8 nt (with exception of the 6th nt) of the protospacer, or "seed" region, must be a perfect match for efficient R-loop formation [20]. Additionally, R-loop formation requires an immediately neighboring tri-nucleotide protospacer adjacent motif (PAM). This conserved PAM sequence at the seed end of the protospacer is recognized by the Cse1 subunit and is essential for the discrimination between targets and non-targets [21, 22].

The mechanism by which Cascade finds its target among the vast amount of DNA in the cell remains elusive. It has been hypothesized that Cascade transiently associates with PAM sequences, interrogating neighboring sequences for a complementary seed, followed by directional R-loop formation [20]. A recent single-molecule study has visualized the transient interactions of Cas9 with PAM-rich sequences in real time [23]. Another study with Cascade and Cas9 has shown directional R-loop formation and how PAM and protospacer complementarity influence its stability [24]. However, it is yet to be shown how the stepwise interaction between PAM, seed and protospacer is coordinated and how off-targeting is avoided during target recognition.

Recent in vivo studies have revealed an additional functionality of CRISPR-Cas immunity. When facing "escape mutants", previously targeted sequences that bear mutations in their PAM and/or protospacer, Cascade initiates a response called priming wherein the CRISPR-Cas system acquires new spacer sequences from the mutant at an elevated rate to restore immunity [25–28]. High-throughput plasmid loss assays of a randomized PAM and protospacer library have revealed that priming is a robust process, tolerating up to 13 mutations in the PAM and protospacer

sequence [26]. Even though Cascade is essential for priming, its role in this process is poorly understood. Intriguingly, biochemical studies have shown that a single point mutation in the PAM or seed sequence leads to a drastic decrease in the binding affinity of Cascade [29]. Therefore, it is puzzling how Cascade can associate with these mutated substrates despite its low affinity and further, how it distinguishes these mutated substrates from bona fide targets to initiate priming.

Single-molecule fluorescence is a powerful tool for elucidating the intricate mechanistic details of complex protein-nucleic acid interactions [30–34]. To dissect Cascade's two distinct functional roles, we developed a single-molecule FRET assay to monitor the interaction of Cascade with bona fide and mutated substrates. Real-time observation of Cascade-target interactions revealed that an initial recognition complex proceeds to a stable R-loop only if the crRNA makes an extensive match with the target. In addition to this "canonical binding mode", we identified an alternative binding mode of Cascade that is triggered by partial complementarity to a target. Using an in vivo assay, we validated that this binding mode enables Cascade to probe mutated DNA substrates and consequently initiate priming.

2.3 Results

2.3.1 Single-molecule observation of Cascade target binding

For single-molecule measurements, Cascade was labeled with a biotin on the N-terminus of its Cse1 subunit (Figure S2.1A) and immobilized to the surface of a microscope slide via a biotin-streptavidin linkage (Figure 2.1A). Dye-labeled dsDNA targets were added to the slide, and individual binding events were imaged in real time with a total-internal-reflection-fluorescence (TIRF) microscope (Figure 2.1A). DNA constructs consisted of a protospacer, a PAM, and an additional 15 base pair flank (Figure 2.1B). The target strand (complementary to the crRNA) was labeled with an acceptor dye (Cy7) at protospacer position +9, whereas the non-target strand was labeled with a donor dye (Cy3) at protospacer position +17. These labeling positions yielded a FRET value of ~0.65 (named E_C for a FRET state which represents a closed conformation of dsDNA between nt 9 and 17) (Figure 2.1E) as measured by immobilization of the DNA alone (see 2.5 on page 53 and Table S2.1). Control experiments showed that dye labeling of the DNA at protospacer positions +9 and +17 did not appreciably affect the target binding reaction of Cascade (Figure S2.1F).

We first explored Cascade's interaction with a bona fide target DNA, a substrate that triggers interference in vivo. This substrate contains a protospacer with perfect complementarity to the crRNA and an interference-permissive PAM (named 'interfering PAM') (Figure 2.1B) [19, 26]. After equilibration of the DNA with the immobilized Cascade, the measured FRET distribution exhibited one major peak centered at 0.44 (named E_O for a FRET state which represents an open conformation of dsDNA between nt 9 and 17), a decrease from the starting value of E_C (0.65) (Figure 2.1E). This decrease in FRET is consistent with the expected open DNA conformation resulting from R-loop formation upon Cascade binding. A similar decrease in FRET was observed upon exchanging the position of the donor and acceptor dyes (Figure

S2.1C) or when Cascade was pre-bound to the DNA prior to immobilization (Figure S2.1D), indicating that the observed decrease in FRET was not due to a protein- or surface-induced photophysical effects.

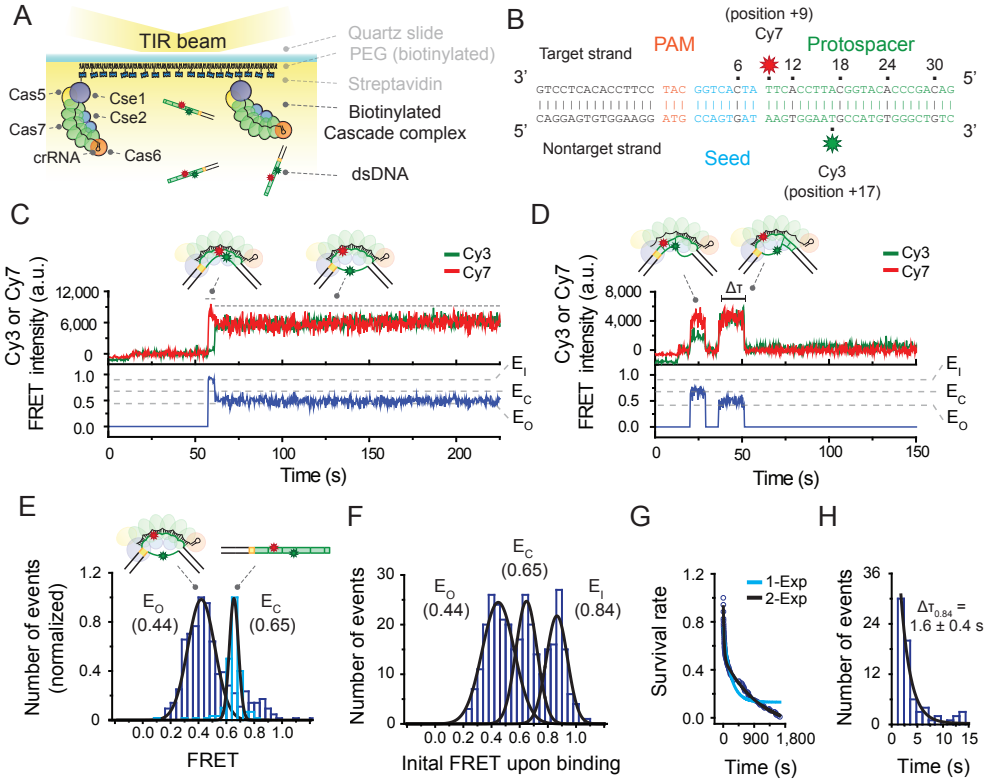


Figure 2.1: Two binding modes of Cascade revealed by a single-molecule FRET assay

(A) Schematic of a single-molecule FRET experiment used to monitor binding of Cascade to target DNA substrates. **(B)** The bona fide target construct consists of a 15 bp flank (black), a PAM (orange), and a protospacer (green), with its seed highlighted in blue. Cy7 (red star) was attached to position +9 of the target strand and Cy3 (green star) to position +17 of the non-target strand. **(C)** A representative time trace of donor (Cy3, green) and acceptor (Cy7, red) fluorescence and corresponding FRET (blue) exhibiting the long-lived binding of the bona fide target. High FRET (~0.84, named E_I for FRET efficiency of an intermediate state) exhibited upon binding is followed by low FRET (~0.44, named E_O for FRET efficiency of an open state). DNA was added at time 10 sec. **(D)** A representative time trace exhibiting the short-lived binding of the bona fide target exhibits two FRET states (E_O ~0.44 and E_C ~0.65). E_C is for FRET efficiency of a closed state). The duration of each state is measured as the dwell time (Δτ). DNA was added at time 10 sec. **(E)** The FRET distribution of the bona fide target DNA alone (light blue) or after equilibration with immobilized Cascade (purple) with peaks at E_C (0.65) and E_O (0.44), respectively (derived from Gaussian fit, black line). Data obtained from 5 fields of view each. **(F)** A histogram of the initial FRET upon binding (average of first 1.5 sec of each event) of the bona fide target exhibits three peaks at FRET = E_O (0.44), E_C (0.65), E_I (0.84) (derived from Gaussian fit, black line). **(G)** The survival rate of events that start at E_I (0.84) was fitted using a single (light blue color) and a double (black color) exponential curve. The double exponential fit resulted in two characteristic times (25.9 and 1040 sec). **(H)** The dwell time distribution of E_I (0.84) state of bona fide target binding with mean $\Delta\tau_{E0.84}$ (derived from single exponential fit, black line). Error represent standard deviation (3 individual data sets). See also Figure S2.1 and Table S2.1.

Next we characterized the kinetics and structural dynamics of Cascade binding in real time by adding a bona fide target substrate to immobilized Cascade during data acquisition. Interestingly, time trajectories exhibited disparate binding events that varied in their dwell time and FRET value. The dwell time distribution followed a double-exponential decay curve (Figure 2.1G, a fit in black), suggesting heterogeneity in binding. A histogram of the initial FRET of binding events exhibited three distinct peaks (centered at E_O (0.44), E_C (0.65), and 0.84) (Figure 2.1F), which, combined with dwell time analysis, led us to divide the events into two distinct types.

2.3.2 Two distinct binding modes of Cascade

The first type of binding event initiated at a FRET of 0.84, and persisted over the entire duration of our observation time (30 minutes) (Figure 2.1C) and was therefore considered to be irreversible over the time scale of our experiment (Figure 2.1G). Interestingly, events of this type did not remain at their initial FRET of 0.84, but exhibited a transition after 1.6 ± 0.4 seconds (Figure 2.1H) to a final FRET of 0.44 (Figure 2.1C). This observation is consistent with the single FRET peak centered at 0.44 (E_O) observed at equilibrium (Figure 2.1E). The initial transient state (0.84, named E_I for an initial transient state) may represent a target-recognition complex wherein the crRNA interacts with the dsDNA before full displacement of the non-target strand (schematic, Figure 2.1C). Notably, the FRET of the initial state is higher than that of the DNA alone (E_C, 0.65, Figure 2.1E), likely arising from a subtle conformational change of the dsDNA upon target recognition (e.g. twisting or bending) [19, 35].

The observed transition ($E_I \rightarrow E_O$) may represent a previously hypothesized locking process, wherein Cascade slides its Cse2 dimer toward its Cse1 subunit upon target recognition [24, 36], ultimately resulting in the displacement of the non-target strand and stable R-loop formation (schematic, Figure 2.1C). Taken together, considering

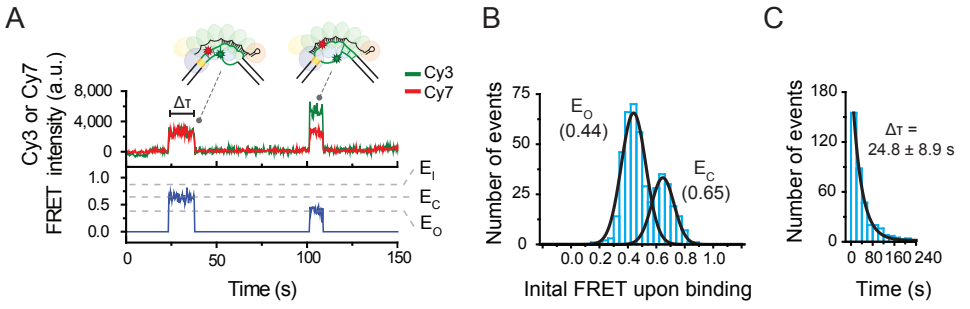


Figure 2.2: Short-binding of Cascade to PAM-mutated targets

(A) A representative time trace exhibiting the short-lived binding of the PAM-mutated target exhibits two FRET states, E_O (0.44) and E_O (0.65). The duration of each state is measured as the dwell time (Δτ). DNA was added at time 12 sec. **(B)** A histogram of the initial FRET upon binding (average of first 1.5 sec of each event) of Mut[PAM] exhibits peaks at E_O (0.44) and E_C (0.65) (derived from Gaussian fit, black line). **(C)** The dwell time distribution of Mut[PAM] binding events with mean Δτ (derived from single exponential fit, black line). Error represents std (3 individual data sets).

Cascade's strong target association and observed conformational change [24, 29, 36], we interpret the first type of binding event to correspond to Cascade's canonical mode of target binding that leads to interference in vivo. We therefore refer to this event type as Cascade's interference mode of binding.

Unlike the interference mode, the second type of binding event was short-lived (25.9 sec, Figure 2.1G) and exhibited an initial FRET of either E_O or E_C (Figure 2.1D). These states were further distinguished from the interference mode as they did not exhibit any kinetic intermediates, nor did they show transitions to other FRET states. As a substrate containing no complementarity (Mut[S1-6]) to the crRNA showed negligible binding (Figure S2.1E), we speculate that these short binding events (named "non-canonical mode") arise from sequence-specific interactions wherein the probed region of the target DNA is either opened in a locally formed R-loop (E_O) or remains closed (E_C).

To explore the origin of Cascade's disparate binding interactions, we first focused on the role of the PAM. We repeated our assay with a DNA substrate containing a point mutation in the PAM (Mut[PAM], Table S2.2) that represents one of the dominant mutant phenotypes of bacteriophages that escape CRISPR interference [29] and subsequently trigger priming in vivo [25, 26]. Notably, while Cascade was still able to interact with Mut[PAM], only binding events characteristic of its non-canonical mode were observed (Figure 2.2A). A histogram of the initial FRET of each event exhibited only two peaks, centered at E_C and E_O (Figure 2.2B), identical to the peak positions observed for the non-canonical binding mode (Figure 2.1F). In addition, the binding events observed for Mut[PAM] were short-lived, exhibiting a dwell time of 24.8 ± 8.9 seconds (Figure 2.2C), similar to that of the non-canonical binding mode (Figure 2.1G). These results indicate that Cascade's interaction with target substrates through its non-canonical binding mode does not require an interfering PAM.

Given the results above, we hypothesize that the observed binding states represent two functional modes of Cascade. The first is the interference mode, in which Cascade binds a bona fide DNA target (i.e. interfering PAM and complementary protospacer) and triggers Cas3-mediated target degradation. The second is the priming mode (non-canonical mode), in which Cascade can associate with targets harboring a PAM mutation to initiate primed spacer acquisition.

2.3.3 Structural elements of two distinct binding modes

To investigate the structural elements of Cascade's two different binding modes, we employed a series of target DNA substrates bearing mutations in their PAM and/or protospacer sequence(s). Recent studies have reported that base pairing between Cascade's crRNA and the protospacer occurs over five segments of five-nucleotides (segments 1-5) and one segment of two nucleotides (segment 6) [26, 37–39]. We therefore chose to systematically mutate the protospacer in segments, starting from either the PAM-proximal or PAM-distal end of the protospacer (Figure 2.3, Figure S2.2, Figure S2.3 & Table S2.2).

Upon mutation of the first segment of the protospacer (Mut[S1], Figure 2.3A), which comprises the majority of the seed region, the non-canonical binding mode persisted as binding events exhibited nearly identical FRET values and dwell times to the Mut[PAM] targets (Figure 2.3B & Figure 2.3C). The same was observed for a DNA substrate containing both the PAM and seed mutations (Mut[PAM+S1], Figure

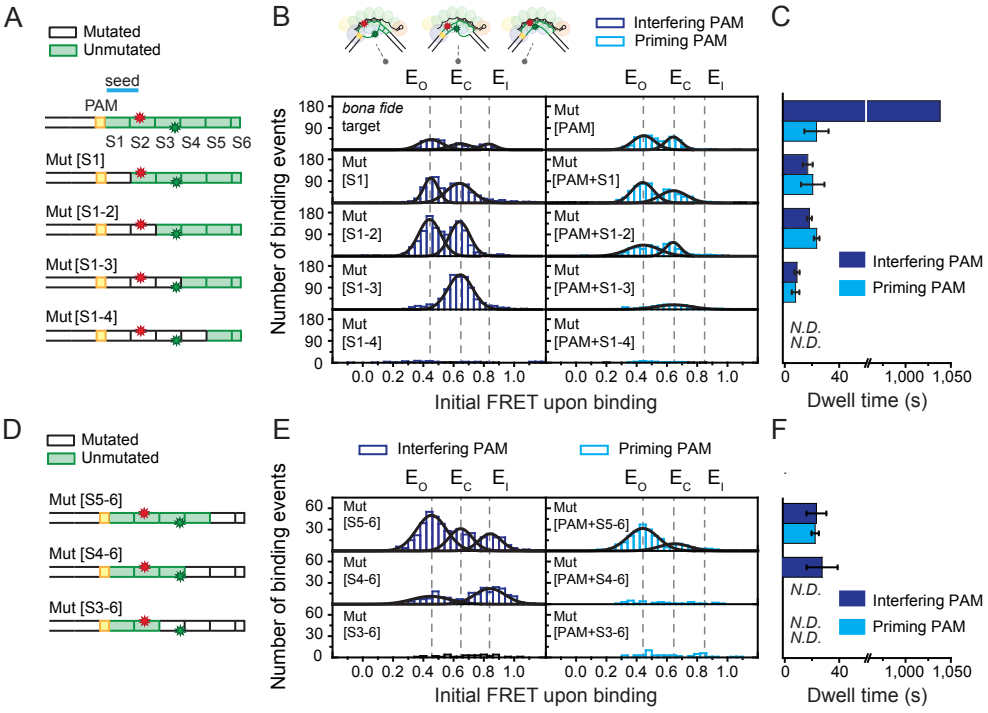


Figure 2.3: Cascade exhibits non-canonical binding to protospacers with PAM-proximal or PAM-distal segmented mutations

(A) Schematics of DNA targets in the PAM-proximal mutation series illustrating mutated (white) or unmutated (green) segments (S1-S6) of the protospacer. Mut[S1], Mut[S1-2], Mut[S1-3] and Mut[S1-4] have segments 1, 1-2, 1-3, and 1-4 mutated, respectively. **(B)** Histograms of the initial FRET upon binding (average of first 1.5 sec of each event) of each PAM-proximal mutant from [A] bearing either an interfering (purple bars, left column) or priming (light blue bars, right column) PAM exhibit peaks (Gaussian fits, black lines) positioned similar to that of the bona fide and Mut[PAM] targets (top row, same asFigure 2.1F and 2B) at E_O (0.44), E_C (0.65), or E_I (0.84) (dashed black lines). The recorded events are from one field-of-view of the detector. **(C)** Mean binding dwell time of each PAM-proximal mutant from [A] bearing either an interfering (purple bars) or a priming (light blue bars) PAM (derived from dwell time distributions, see Figure S2.2A). Error represents std (3 individual data sets). The dwell time of the bona fide target could not be measured accurately due to the photobleaching and thus arbitrarily set 1040 sec to represent the longer characteristic time scale in Figure 2.1G. **(D)** Schematics of DNA targets in the PAM-distal mutation series illustrated as in [A]. Mut[S5-6], Mut[S4-6], and Mut[S3-6] have segments 5-6, 4-6, and 3-6 mutated, respectively. **(E)** Histograms of the initial FRET upon binding of each PAM-distal mutant from [D] displayed in a similar fashion to [B]. **(F)** Mean binding dwell of each PAM-distal mutant from [D] bearing either an interfering (purple bars) or a priming (light blue bars) PAM (derived from dwell time distributions, see Figure S2.2A). Error represents std (3 individual data sets). N.D. is "Not Determined". See also Table S2.2.

2.3A, Figure 2.3B, & Figure 2.3C), indicating that the non-canonical binding mode is largely insensitive to the PAM and seed sequence. This observation is in stark contrast to the canonical binding mode, which requires both an intact seed sequence and an interfering PAM. Remarkably, when the first two PAM-proximal segments, including the entire seed, were mutated (Mut[S1-2]), the non-canonical binding mode was still evident with initial FRET values centered at E_O or E_C and an average dwell time of 19.6 ± 0.4 seconds (Figure 2.3B, Figure 2.3C, & Figure S2.2A).

Intriguingly, when the first three (Mut[S1-3], Figure 2.3A) PAM-proximal segments were mutated, the binding events exhibited only one major initial FRET population centered at E_C, with an average dwell time of 10.5 ± 1.9 seconds (Figure 2.3B, Figure 2.3C & Figure S2.2A), indicating that these events arise from sequence-specific interactions confined outside of the probed region of the protospacer (segments 4-6, Figure 2.3A). Removal of complementarity in the first four segments (Mut[S1-4], Figure 2.3A & Figure 2.3B) or all segments (Mut[S1-6], Figure S2.1E) disrupted binding to background levels. Taken together, the series of PAM-proximal mutations indicate that the non-canonical binding mode of Cascade comprises sequence-specific interactions with a minimum requirement of three full segments for target recognition.

The PAM-distal mutation series showed complementary behavior, consistent with the structural features of the non-canonical binding mode observed above (Figure 2.3D, Figure 2.3E & Figure 2.3F). Upon mutation of the last two segments of the protospacer (Mut[S5-6]), the non-canonical binding mode persisted with two peaks centered at E_O and E_C. When three segments (Mut[S4-6]) were mutated, the non-canonical binding mode exhibited only one peak centered at E_C, indicating that these interactions are confined within the probed region (segments 1-3). Further removal of complementarity disrupted binding to background levels, confirming that a minimum of three consecutive segments are required for non-canonical binding.

Besides the non-canonical mode, a fraction of binding events in the PAM-distal mutation series exhibited the signature initial FRET of the interference mode (E_I, left column, Figure 2.3E). Even though this initial FRET was identical to that of the canonical binding mode, binding events were transient and did not exhibit any FRET transitions until dissociation after 24.8 ± 7.3 seconds (Figure 2.3F and Figure S2.3). This state reports on the formation of an interference-like target-recognition complex that cannot be locked and is in line with a previous observation that the PAM-distal region is required for stable R-loop formation in the interference model [24].

Finally, to evaluate the role of the PAM in Cascade's non-canonical binding mode, we repeated both series of protospacer mutations in the presence of the escape-mutant PAM (named 'priming PAM', Figure 2.3 & Figure S2.2). Overall, mutation of the PAM substantially reduced the number of binding events for each mutant compared to its interfering PAM counterpart (compare columns, Figure 2.3B &Figure 2.3E), indicating that the PAM is not strictly required for, but facilitates, non-canonical binding. In addition, E_I state observed in Mut[S5-6] and Mut[S4-6] was completely abrogated upon PAM mutation, suggesting that this intermediate requires the coordinated ternary interaction of Cascade with the PAM and the seed.

In summary, our single-molecule results show that the non-canonical binding mode of Cascade is much more robust than its canonical mode, capable of binding a wide variety of mutated targets, yet still exhibiting sequence specificity. Such versatility could facilitate primed spacer acquisition, in which invading DNA variants that harbor mutations in their PAM or protospacer can still be detected by the CRISPR-Cas immune system.

2.3.4 Functional roles of two distinct binding modes

To investigate whether the canonical and non-canonical binding modes of Cascade lead to different functional outcomes, we reconstituted CRISPR interference in vitro. We cloned the segmented mutants that showed binding in our single-molecule experiments into plasmids (Table S2.3) and tested the plasmids for Cascade-directed degradation by Cas3. Our assay revealed that only the plasmid with a perfectly complementary protospacer accompanied by an interfering PAM led to target degradation, whereas target plasmids containing either an escape PAM mutation

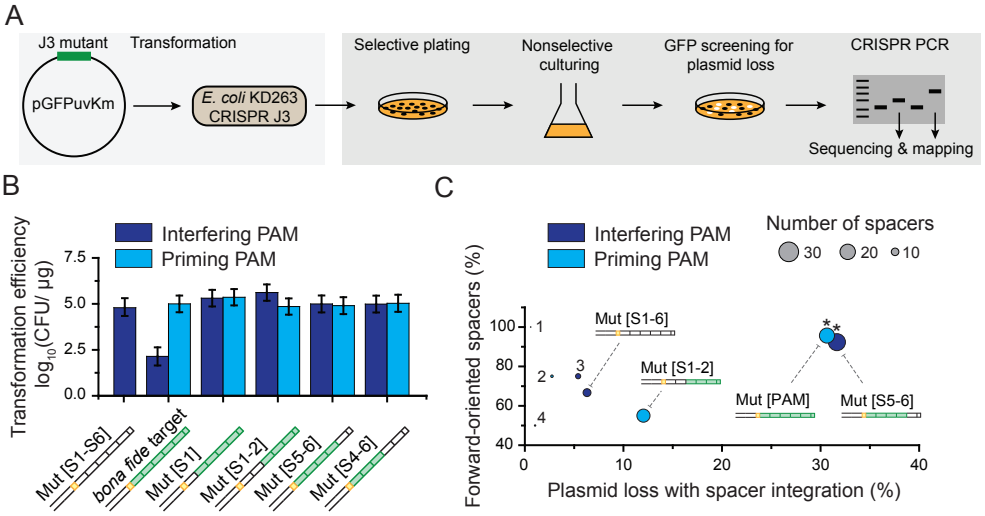


Figure 2.4: Non-canonical binding leads to primed spacer acquisition

(A) Cartoon representation of the in vivo assay used to determine primed plasmid loss and spacer acquisition. **(B)** Transformation efficiencies of plasmids harboring different target sequences (see schematics) with an interfering (purple bar) or a priming (light blue bar) PAM. CFU is "Colony-Forming Unit." Error is std of 3 individual measurements. **(C)** A two-dimensional bubble plot showing the fraction of forward-oriented spacers acquired versus the percentage of plasmid loss for those targets in [B] that exhibited spacer integration. Circle size represents the total number of spacers that were acquired and circle color represents an interfering (purple) or a priming (light blue) PAM. A star (*) indicates a forward directional bias (relative to random) with a P-value < 1×10^{-5} based on binomial statistics. The numbers of 1, 2, 3 and 4 indicate data points from constructs Mut[PAM+S5-6], Mut[PAM+S1], Mut[S4-6] and Mut[PAM+S4-6], respectively. See also Figure S2.4, Table S2.1 and Table S2.3.

and/or segmented mutations proximal or distal to the PAM were unaffected by Cas3 (Figure S2.4). These results suggest that only Cascade's canonical binding mode ($E_I \rightarrow E_O$) generates an R-loop structure that supports target degradation by Cas3.

Next, we sought to determine if the non-canonical binding mode of Cascade results in primed spacer acquisition in vivo. To assess primed spacer acquisition, we first transformed the target plasmids with segmented mutations into *E. coli* containing a targeting CRISPR array plasmid (Figure 2.4A). Notably, only the target with a perfectly complementary protospacer and interfering PAM led to a reduced transformation efficiency (Figure 2.4B), confirming that the CRISPR-Cas system exclusively targets the R-loops generated through the canonical binding mode of Cascade. Next, transformants were transferred to non-selective media, which allowed the CRISPR-Cas system to mount a primed response.

After two days of cell growth, three mutant constructs (Mut[PAM], Mut[S5-6], Mut[PAM+S1-2]) showed a higher degree of plasmid loss than the negative control construct Mut[S1-6] did (Figure 2.4C). To identify if these plasmids were lost through primed spacer acquisition, the genomic CRISPR-array was amplified by PCR and amplicons with increased size were sequenced (Figure 2.4A). In total, 23, 26, and 20 new spacers were obtained that originated from the target plasmids Mut[PAM], Mut[S5-6], Mut[PAM+S1-2], respectively. Sequencing of the genomic CRISPR-array also allowed us to determine whether the acquired spacers showed any strand bias that is typical of the priming process in Type I-E systems [25, 52]. Among the three constructs, Mut[PAM] and Mut[S5-6] exhibited bias in spacer acquisition toward the target strand (p-value $< 1 \times 10^{-5}$, Figure 2.4C), suggesting that these spacers were obtained by primed spacer acquisition. Taken together, the high frequency of plasmid loss and strand bias in the acquired spacers suggests that the non-canonical binding mode acts as a gateway to priming in vivo.

2.4 Discussion

Adaptive immune systems are found in both vertebrates and prokaryotes and provide specific defense against invading pathogens. The high specificity of this immunity is important for distinguishing self from non-self [40], yet it brings a downside that it can be readily overcome by rapidly evolving pathogens [41]. However, both vertebrates and prokaryotes have developed sophisticated fail-safe mechanisms to target these pathogens. For example, when vertebrates face invaders bearing mutated antigens, they may still be recognized by a pool of polyclonal antibodies [42]. The resulting secondary response proceeds more quickly and efficiently than the primary response, which allows vertebrate hosts to keep pace with their evolving pathogens [43].

The prokaryotic adaptive immune system faces similar challenges. Rapidly evolving pathogens readily overcome sequence-specific CRISPR-Cas-mediated host defense [29, 44], exposing a major limitation to prokaryotic adaptive immunity [45]. However, analogous to vertebrate adaptive immunity, once pre-exposed to an ancestral invader, CRISPR-Cas responds more rapidly and efficiently to future

variants then it can to a novel invader [5, 13, 19, 25–28]. Although Cascade was shown to be essential for this "primed" response [25], the underlying mechanism has remained enigmatic. Here we provide the first insights into this puzzle by showing that Cascade binds mutated targets through a distinct non-canonical mode with low-fidelity compared to the high-fidelity binding mode used for unmutated targets. We show that the canonical, high-fidelity binding mode is a stepwise process that locks, triggering recruitment of nuclease/helicase Cas3 only when all criteria are met, including: an interfering PAM, a matching seed, and pairing of all segments of the crRNA guide. In contrast, the non-canonical, low-fidelity binding mode initiates a downstream pathway that results in rapid spacer acquisition through the priming process (Figure 2.5).

2.4.1 Protein-mediated high fidelity target recognition

Our single-molecule data demonstrate in real time that high-fidelity target-DNA binding is a multi-step process and occurs in a directional manner from the PAM-proximal to PAM-distal end of the protospacer. Previous studies have shown that the recognition process is initiated when the Cse1 subunit recognizes the PAM [21] and the crRNA hybridizes with the seed sequence. After this initial recognition complex is formed, the R-loop propagates toward the PAM-distal region of the protospacer [24, 29, 36]. When the pairing of the crRNA reaches the PAM distal-end of the protospacer, Cascade senses the fully paired structure and stabilizes this complex into a lower energy state ("locking") [24]. This state acts as a flag for the destruction of the target DNA by Cas3 [17–19]. This stepwise mechanism involves both protein-nucleic acid interactions (Cse1-PAM) and progressive crRNA-DNA base pairing, ensuring efficient and high fidelity recognition, and degradation of targeted DNA.

Our study shows how Cascade maintains a strict regime to prevent non-specific cleavage by controlling the pathway toward the proposed locking process [24]. When Cascade encounters a target with mismatches (e.g. Mut[S5-6], Figure 2.3F), the initial recognition complex forms, but the R-loop does not propagate throughout the full protospacer (absence of a transition of $E_I \rightarrow E_O$) (Figure S2.3). As a result, Cascade will not lock the R-loop, and the initiation complex can disassemble using thermal energy. This process cannot be explained by the thermodynamic properties of base pairing alone since a target with mismatches often form a far larger number of consecutive base pairs than 7 (e.g. Mut[S5-6]), which has been shown to be the minimal number of base pairs required for stable binding [46]. Instead, the last step of stepwise recognition (locking) must involve protein-nucleic acid interactions that verify base pairing over the entire protospacer. This model is analogous to the stepwise conformational change observed with Argonaute proteins during its target search process [47, 48] and contrasts with the low fidelity RecA-mediated target search that does not seem to use protein-nucleic acid interaction in promoting specificity [49].

2.4.2 Structural view of the priming mode

The structure of Cascade supports our finding that low-fidelity target-DNA interactions can initiate from any segment of the crRNA (Figure 2.5B). Cascade is composed of five different Cas protein subunits assembled into a highly interlocked, crRNA-containing protein complex [37–39]. The backbone of the complex consists of six Cas7 subunits with a hand-like architecture. Each hand uses its thumb to hold and position the crRNA at 6 nt intervals. Consequently, every sixth base is flipped out of the plane and is unable to interact with the target DNA. This unusual configuration permits the crRNA to pair with a target in segments of five nucleotides in an underwound, ribbon-like structure [38]. Interestingly, individual segments of the crRNA in the apo-Cascade

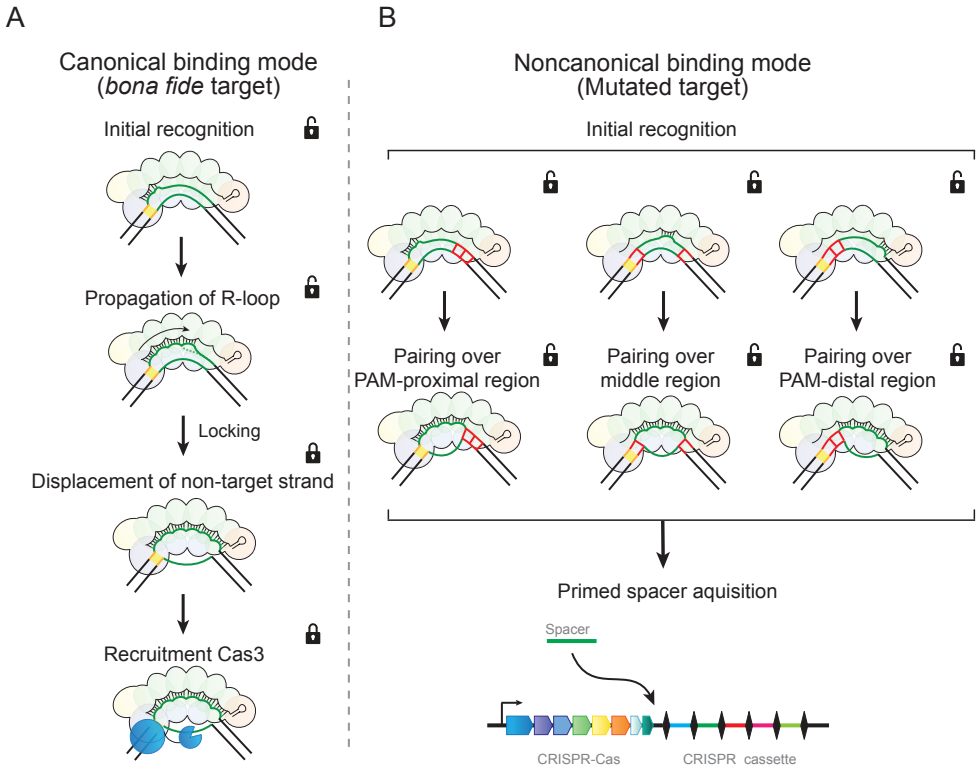


Figure 2.5: Two binding modes of Cascade lead to different functional outcomes

Cascade employs two distinct target-DNA binding modes that trigger **(A)** interference or **(B)** priming. **(A)** In the interference pathway, target recognition initiates from the PAM and PAM-proximal region. R-loop formation then propagates toward the PAM-distal region. When Cascade senses the fully paired structure, it brings this complex into a lower energy state ("locking") that displaces the non-target strand out of Cascade. This exposed strand is then cleaved by Cas3. **(B)** In the priming pathway, DNA is probed through brief interactions. PAM recognition facilitates this priming pathway but is not required. The brief interactions may initiate from the PAM-proximal (left), the PAM-distal region (right), or the middle of the protospacer (middle), which becomes stable when paired over 3 or more segments. This non-canonical ("unlocked") binding mode leads to a unique conformation of the R-loop and signals for primed spacer acquisition.

structure are already pre-ordered in a pseudo A-form helix with their nucleobases facing the solvent [37, 39]. Structural pre-ordering is a common strategy to facilitate target binding of nucleic acid guided complexes (e.g. Argonaute and RecA) [20, 47, 49], and thus the pre-ordered segments of crRNA of Cascade is in line with the idea that low-fidelity interactions can nucleate from any crRNA segment (Figure 2.5B). Although the low fidelity binding mode leads to relatively short-lived R-loops, it is their distinct conformation that likely signals for a primed spacer acquisition response in the cell.

The DNA recognition mechanism of Cascade contrasts that of Cas9, which has recently been shown to be strictly dependent on the PAM [23]. Furthermore, Cas9 does not base pair its crRNA in segments to the target DNA [50] but forms a contiguous double helix, making it more difficult to imagine that PAM-distal regions of Cas9's crRNA can initiate an interaction with the target DNA. Yet, off-target cleavage analysis of Cas9 during genome editing clearly indicates that Cas9 also tolerates mutations [51], but whether this leads to a priming response in bacteria with Type II CRISPR-Cas systems remains to be shown.

2.4.3 Mechanisms of the priming mode

Although the interference response of CRISPR immunity is a relatively well-characterized phenomenon, the molecular mechanism of priming remains poorly understood. First, our data shows that Cascade distinguishes mutated targets from bona fide targets using a low-fidelity binding mode that can initiate priming. A recent study showed that priming in *E. coli* is robust, tolerating up to 13 mutations throughout the 32 nt protospacer and 3 nt PAM [26]. Even when mutations were clustered in any of the crRNA defined segments, priming was not abolished. The low-fidelity binding mode of Cascade, in which individual segments may initiate pairing with a target, can explain the reported high tolerance for distributed and clustered mutations in a target during priming. In this mode, Cascade can probe DNA for complementarity to any of its crRNA segments, and extend such an interaction in either direction, thereby achieving sequence-specific detection of targets with limited base complementarity. However, the minimal number of base pairs required for priming [26] insures that detrimental self-priming of the bacterial genome at random sites is unlikely.

Second, we observed that the non-canonical binding mode occurs even for substrates containing an interfering PAM and an intact seed, suggesting that direct interference and priming may occur simultaneously. Indeed, we have previously observed that *E. coli* is cured from high copy number plasmids by using existing spacers to expand the CRISPR-array with a range of new spacers against the same target [3, 52]. For a host this is a highly advantageous strategy, by simultaneously using interference and priming, the CRISPR interference effect is amplified while the chance that invaders evade immunity through point mutations in their protospacers is reduced. Even though it remains to be seen how priming is coordinated in the presence of the remaining Cas protein machinery (Cas1, Cas2 and Cas3), the relatively short time that Cascade spends on a target in the priming mode suggests that other factors might stabilize this relatively weak interaction.

Finally, in CRISPR-Cas/I-E systems priming is a DNA strand dependent process in which approximately 90% of new spacers are integrated from the same strand as the spacer triggering priming [53]. Our results with Mut[PAM] and Mut[S5-6] in Figure 2.4 are consistent with this strand bias. In contrast, primed spacer acquisition in Type I-B and I-F systems does not exhibit such strand bias [27, 28]. Interestingly, for protospacers mutated in the PAM and segments 1 and 2 (Mut[PAM+S1-2]), including the seed, we observed a higher degree of spacer acquisition without the typical strand bias, suggesting that these types of targets lead to a priming behavior in which strand specificity is lost.

2.4.4 Conclusion

Faithful copying and decoding of genetic information is central to the most important processes in the cell, including DNA replication [8], RNA transcription [54], and protein translation [12]. But high fidelity always comes at the cost of reduced processing speed. Here we show how a crRNA guided complex solves this dilemma by employing both high and low fidelity target-DNA recognition modes. While the high fidelity mode ensures destruction of only perfectly matching targets, the low fidelity priming mode enables detection of a whole range of mutated invaders to initiate the priming process. The unique combination of these two properties in a single RNA-guided complex not only makes CRISPR immunity robust, but also reveals versatility of adaptive immunity against rapidly mutating pathogens.

2.5 Experimental Procedures

2.5.1 Preparation of Cascade, biotinylated Cascade, and Cas3

Cascade was expressed in *E. coli* BL21 (DE3) using plasmids listed in Table S2.3 and purified as described [16]. Elution buffer contained 20 mM HEPES pH 7.5, 75 mM NaCl, 1 mM DTT, 2 mM $MgCl_2$ (storage buffer) with 4 mM desthiobiotin. Primers for cloning are listed in Table S2.1. The nuclease-helicase Cas3 was produced and purified as described previously [17] with the following modifications. BL21-AI cells were used for over expression, and protein expression was induced with 0.5 mM IPTG and 0.2% L-Arabinose. The purification process was stopped after size exclusion chromatography and before the proteolytic removal of the Maltose Binding Protein (MBP) using the Tobacco Edge Virus protease [35]. MBP-Cas3 was flash frozen in liquid nitrogen and stored at -80 °C.

For site-specific Cascade labeling, plasmid pWUR706 (Cse1 with an N-terminal LCTPSR FGE recognition motif) was co-expressed with plasmid #16132 (fge gene, Addgene) [57], pWUR656 (CasBCDE) and pWUR630 (CRISPR J3). A solution of 45 µL purified Cascade (1.5 mg/ml) was mixed with 45 µL potassium acetate (0.5 M (pH 5.5)) and 40 µL Hydrazide-LC-Biotin (50 mM in DMSO, Thermo Scientific) and incubated overnight at room temperature. Labeled Cascade was purified by size exclusion chromatography (Superdex-200 HR 10/30 (GE Healthcare)). Fractions were concentrated using Vivaspin (50 kDa) spin columns and stored at 4 °C in storage buffer or at -20 °C in storage buffer containing 50% glycerol.

2.5.2 Preparation of DNA constructs

All the target dsDNA substrates that we used were 50 base pairs in length, bearing a protospacer, PAM, and a 15 bp sequence upstream of the PAM (Figure 2.1B, Table S2.2). These synthetic DNA targets (Integrated DNA Technologies) were internally labeled with a monoreactive acceptor dye (Cy7, GE Healthcare) at dT-C6 on the target strand (complementary to the crRNA) and a monoreactive donor dye (Cy3, GE Healthcare) at dT-C6 on the non-target strand (Figure 2.1B). After labeling, the ssDNA strands were annealed, which was followed by PAGE purification of the dsDNA constructs. To determine the initial FRET values of these constructs (Figure 2.1F), an additional 18 bp flank was added target strand downstream of the protospacer (Table S2.1). This flank allowed for hybridization with a 80 nt biotinylated poly(T)-linker that was used for immobilization of these constructs on a PEG-passivated surface.

2.5.3 Single-molecule FRET

Biotinylated Cascade complexes were anchored to polyethylene glycol-coated quartz microscope slides by biotin-streptavidin linkage Dye-labeled (Cy3 and Cy7) dsDNA targets were added to the immobilized Cascade complexes and detected by a prism-type TIRF microscope. In a typical field of view, 200-300 molecules were detected. dsDNA targets were excited with a 532 nm laser and fluorescence emissions from Cy3 and Cy7 were separated by dichroic mirrors and imaged onto two halves of a CCD camera after passing through various filters. Imaging buffer consisted of Cascade buffer (50mM HEPES (pH 7.5), 75mM NaCl, 2mM $MgCl_2$), an oxygen scavenging system (1% glucose (v/v), 0.1 mg/mL glucose oxidase (Sigma), 17 µg/µL Catalase (Roche)) to reduce photobleaching, and 1 mM Trolox (Sigma) to reduce photoblinking of the dyes [56]. Imaging was performed at room temperature (23± 2 °C). Fluorescence time traces of individual binding events were identified in recorded movies and subsequently analyzed using custom software developed in IDL and MATLAB, respectively. The FRET value was defined as $I_A/(I_D+I_A)$, where I_D and I_A represent the fluorescence signals detected in the Cy3 and Cy7 channels, respectively.

2.5.4 Single-molecule fluorescence

The fluorescent label Cy3 was imaged using prism-type total internal reflection microscopy, through excitation by a 532nm (Compass 215M-50, Coherent). Cy7 was detected via FRET with Cy3, but if necessary, Cy7 was directly excited using a 640nm solid-state laser (CUBE 640-100C, Coherent). Fluorescence signals from single molecules were collected through a 60x water immersion objective (UPlanSApo, Olympus) with an inverted microscope (IX71, Olympus). Scattering of the 532nm laser beam was blocked with a 550nm long-pass filter (LP03-532RU-25, SemRock). When the 640nm laser was used, 640nm laser scattering was blocked with a notch filter (633 ± 12.5nm, NF03-633E-25, SemRock). Subsequently, signals of Cy3 and Cy7 were spectrally split with a dichroic mirror (λ_{cutoff} = 645 nm, Chroma) and imaged

onto to halves of an electron multiplying CCD camera (iXon 897, Andor Technology). Given the reduced detection efficiency of the camera for Cy7 compared to Cy3 (~50%, Andor Technology), the measured Cy7 signal was multiplied by 2 prior to further analysis.

To eliminate non-specific surface adsorption of proteins and nucleic acids to a quartz surface (Finkenbeiner), piranha-etched slides were PEG-passivated over two rounds of PEGylation as described previously [58]. After assembly of a microfluidic flow chamber, slides were incubated for 1 minute with 20 µL streptavidin (0.1 mg/ml, S-888, Invitrogen) followed by a washing step with 100 µL of the Cascade buffer (50 mM HEPES (pH 7.5, AM9851, Ambion), 75 mM NaCl (AM9760G, Ambion), 2mM $MgCl_2$ (AM9530G, Ambion). Cascade molecules were end-specifically immobilized through biotin-streptavidin linkage by incubating the chamber with 100 µL of 1 nM biotinylated Cascade for 5 minutes. Remaining unbound Cascade molecule were flushed away with 100 µL Cascade buffer that was substituted with 60 nM J3-CasBCDE to reconstitute any Cse1 that lacks of CasBCDE subunits. After 5 minutes of incubation, unbound J3-CasBCDE were flushed away with 100 µL Cascade imaging buffer (50 mM HEPES (pH 7.5) ,75 mM NaCl, 2mM $MgCl_2$, 0.1 mg/mL glucose oxidase (G2133, Sigma), 4 µg/ml Catalase (10106810001, Roche) and 1 mM Trolox (((±)-6-Hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid, 238813, Sigma). Next, 3 nM labelled dsDNA substrate was introduced in the chamber while imaging at room temperature (23± 1 °C) to monitor strand opening in real time.

A series of CCD images were acquired with the AndorSolis software (Andor Technology) at a time resolution of 0.3 or 1.5 sec. Fluorescence time traces were extracted with an algorithm written in IDL (ITT Visual Information Solutions) that picked fluorescence spots above a threshold with a defined Gaussian profile. The extracted time traces were analysed using custom written MATLAB algorithms (MathWorks) that selectively picked anti-correlated traces above a defined threshold. These selected traces were further analyzed using a custom written MATLAB algorithm to extract dwell times and initial FRET values upon binding. To obtain histograms for initial FRET values upon binding (Figure 2.1F, Figure 2.2B, Figure S2.1E, Figure 2.3B & Figure 2.3E), the first five frames (1.5 s) of each binding event were averaged and plotted using MATLAB. Histograms were aligned by setting the donor-only signal to zero. Donor-only and low-FRET events (falling outside the most sensitive distance-range of FRET) were subsequently removed by discarding events with an acceptor intensity below 20% of the mean total-dye intensity (sum of the donor and acceptor) over the event.

To measure the initial FRET values upon binding, Cy3 molecules were excited an area of 50 x 50 μm^2 with a 16% of the full laser power (4 mW) green laser (532nm), while the time resolution was set to 0.3 seconds. Under these imaging conditions we obtained a high signal-to-noise ratio that allowed us to visualize kinetic intermediates while imaging over time periods of 8 min. In contrast, for dwell time measurements, Cy3 molecules were excited with 2% of the full laser power (1 mW) green laser (532nm) to minimize photobleaching of the donor and acceptor dye during our observation time. Meanwhile, the time resolution was set between 1.0 and 1.5 seconds to collect

a large enough number of photons per time bin despite the weak excitation. Under these imaging conditions we obtained a signal-to-noise ratio that allowed us to visualize kinetic intermediates while imaging over long periods of time (30 min).

2.5.5 Target degradation assays

All oligonucleotides are listed in Table S2.1. Target plasmids (pWUR738-pWUR747) were constructed using plasmid pGFPuv-Kan as a backbone [26]. PCR amplicons of the J3 target were cloned into BspHI and EcoRI sites of the pGFP-Kan plasmid, and confirmed sequencing (GATC-Biotech). Plasmids were prepared using GeneJET Plasmid Miniprep Kits (Thermo Scientific) and DNA from PCR and agarose gels was purified using the Thermo Scientific GeneJET PCR Purification and Gel Extraction Kits. Plasmid DNA (3.5 nM) was mixed with purified Cascade (35 nM or 70 nM) in a buffer containing 5 mM HEPES, pH 7.5, 60 mM KCl, 10 mM $MgCl_2$, 10 μM $CoCl_2$, and 2 mM ATP, and incubated at 37 ˚C for 15 min. After incubation purified Cas3 protein was added (70 nM) and incubated at 37 ˚C for 1, 10 or 30 minutes. Reactions were stopped by addition of 6x DNA Loading Dye (Thermo scientific). Samples were run on 0.8% TAE agarose gels containing SYBR Safe (Invitrogen) for 1h at 100 V. DNA band intensities were quantified using GeneTools Software (Syngene).

2.5.6 Direct interference and priming

The construction of *E. coli* strain KD263 was described previously [53]. The strain contains the cas3 gene under the control of the inducible lacUV5 promoter and the cse1 – cas2 operon under control of the inducible araBp8 promoter. The KD263 strains harbors a single CRISPR cassette containing the g8 spacer targeting bacteriophage M13. Plasmid pWUR564 containing the J3 spacer under control of the native CRISPR 2.1 promoter [59] was introduced by transformation. *E. coli* strains were grown at 37 °C in Luria Broth (LB; 5 g/L yeast extract, 10 g/L tryptone, 5 g/L NaCl) at 180 rpm or on LB-agar plates containing 1.5% (w/v) agar. When required, medium was supplemented with the following: ampicillin (Amp; 100 μg/L), chloramphenicol (Cam; 25 μg/mL), or kanamycin (Kan; 50 μg/mL). Bacterial growth was assessed spectrophotometrically at 600 nm (OD_{600}). To induce cas gene expression, IPTG (isopropyl β-D-1 thiogalactopyranoside) and L-arabinose were added to the final concentration of 1 mM each when an OD_{600} of approximately 0.4 was reached.

Direct interference was assessed by determining the transformation efficiency of target plasmid series pWUR738-pWUR747 to *E. coli* strain KD263 containing pWUR564. Cas gene expression was induced 30 minutes prior to making cells chemically competent. Priming was assessed using plasmid loss assays as described [26]. Briefly, *E. coli* transformants containing the target plasmid (pWUR738-747) were grown for 24 h in 5 mL LB in 15 mL tubes (Greiner) at 37 °C with shaking at 180 rpm. For further passaging, 100 μL of culture was subcultured into 5 mL LB in 15 mL tubes for a further 24 h at 37 °C at 180 rpm. Dilutions were plated on LBA and loss of fluorescence of individual colonies detected under mild UV light as described [26]. GFP-negative colonies were screened for spacer integration by colony PCR using

DreamTaq Green DNA polymerase (Fermentas). Briefly, acquisition of spacers in the former CRISPR 2.1 locus containing the g8 spacer was assessed by PCR using primers BG5301 and BG5302 for strains KD263. PCR products were visualized on 2% (wt•vol^{-1}) agarose gels and stained with SYBR-safe (Invitrogen). Newly acquired spacers were sequenced using primer BG5301 (Table S2.1). Spacer sequences were strand specifically mapped onto the target plasmid sequence to verify priming.

2.6 Supplementary information

2.6.1 Supplementary figures

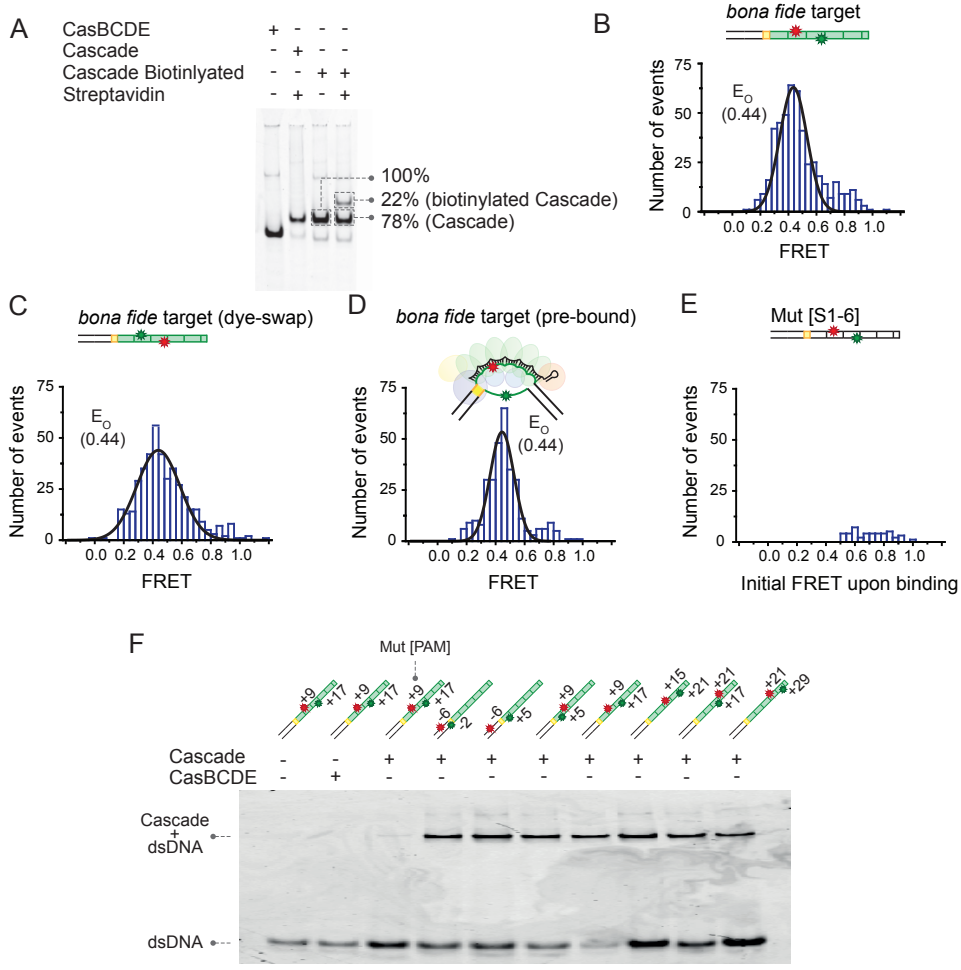


Figure S2.1: Cascade biotinylation efficiency; dye-labeling and surface-immobilization controls for single-molecule assay

(A) Efficiency of site-specific labeling of Cascade with a biotin on the Cse1 subunit. 50 nM of Cascade from the labeling reaction (see Extended Experimental Procedures) was incubated with 5 nM of a Cy5-labeled complementary oligo (TJ3_15 bona fide (+9), Table S2.2) and 500 nM Streptavidin in Cascade buffer. After 30 minutes of incubation at the room temperature, the protein/DNA solution was run on a 5% TBE PAGE gel at 100V for 1 hour, and imaged with Typhoon Trio scanner. Labeling efficiency was quantified with Imagequant software (GE Healthcare). Wild-type refers to unmodified Cascade. CasBCDE refers to wildtype Cascade lacking the Cse1 subunit. **(B)** A FRET histogram obtained after equilibration of a bona fide DNA target (target strand labeled with Cy7 (red star), non-target strand labeled with Cy3 (green star)) with immobilized Cascade. Peak centered at E_O (0.44) was derived from Gaussian fit (black line) (same data as Figure 2.1E, for reference). **(C)** Similar to [B] with Cy7 and Cy3 exchanged: target strand labeled with Cy3 (green star), non-target strand labeled with Cy7 (red star). Peak centered at E_O (0.44) was derived from Gaussian fit (black line). Data obtained from 5 fields of view (~200 molecules per field). **(D)** A FRET histogram obtained after immobilization of biotinylated Cascade and bona fide target DNA (Cy3 target strand, Cy7 non-target strand) which were pre-incubated in Cascade buffer for 30 min at room temperature (23 ± 1 °C) prior to immobilization. Peak centered at E_O (0.44) was derived from Gaussian fit (black line). Data obtained from 5 fields of view. **(E)** A histogram of initial FRET upon binding of a construct without complementarity (Mut [1-6]) to the crRNA. **(F)** EMSA of dsDNA constructs with different labeling positions (indicated by numbering, where the PAM sequence occupies positions -1, -2, and -3. Target strand was labeled with Cy7 (red star) and the non-target strand was labeled with Cy3 (green star). CasBCDE refers to wildtype Cascade lacking the Cse1 subunit. Cascade (50 nM) and dye-labeled dsDNA or ssDNA (5 nM) were incubated in Cascade buffer for 30 minutes at room temperature (23± 1 °C) and subsequently run on a 5% polyacrylamide TBE gel (Bio-Rad) at 100 V for 1 hour. This gel was imaged with a Typhoon Trio scanner (GE Healthcare).

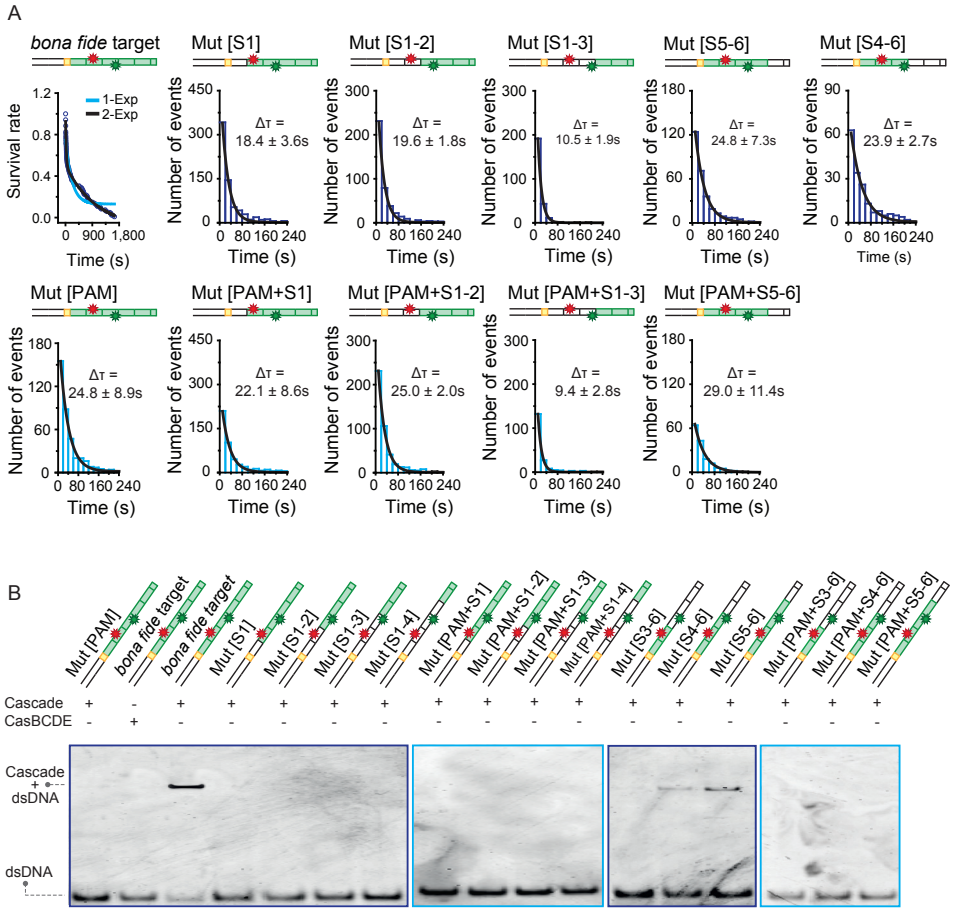


Figure S2.2: Binding dwell-time distributions and EMSAs of all DNA constructs

(A) Dwell-time (Δτ) histograms of bona fide and mutated target dsDNA binding to Cascade. The graph of the bona fide target shows the survival rate of events that start at E_i (0.84). Data was fitted using a single (light blue line) and a double (black line) exponential curve. The double exponential fit resulted in two characteristic times (25.9 and 1040 sec). The dwell time of the binding events of the mutant constructs was determined from a single exponential fit (black line). Error represents the standard deviation of 3 data sets from 3 different days. **(B)** EMSAs of DNA constructs. Mutants with an interference permissive PAM (5′-CAT-3′) or an escape mutant PAM (5′-CGT-3′) are indicated with a purple or light blue box, respectively. CasBCDE refers to wild-type Cascade lacking the Cse1 subunit. Cascade (50 nM) and dyelabeled dsDNA or ssDNA (5 nM) were incubated in Cascade buffer for 30 minutes at room temperature (23± 1 °C) and subsequently run on a 5% polyacrylamide TBE gel (Bio-Rad) at 100 V for 1 hour. This gel was imaged with a Typhoon Trio scanner (GE Healthcare).

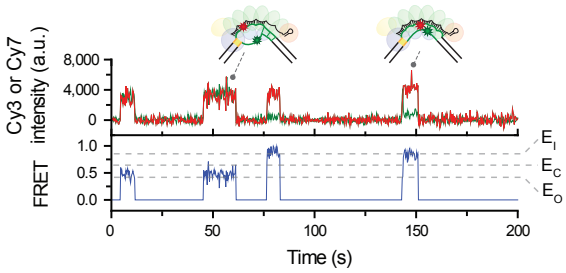


Figure S2.3: Short-lived binding of Mut[S5-6] exhibits two FRET states

A representative time trace exhibiting the short-lived binding of Mut[S5-6] exhibits two FRET states, E_O (0.44) and E_I (0.84). The duration of each state is measured as the dwell time ($\Delta\tau$).

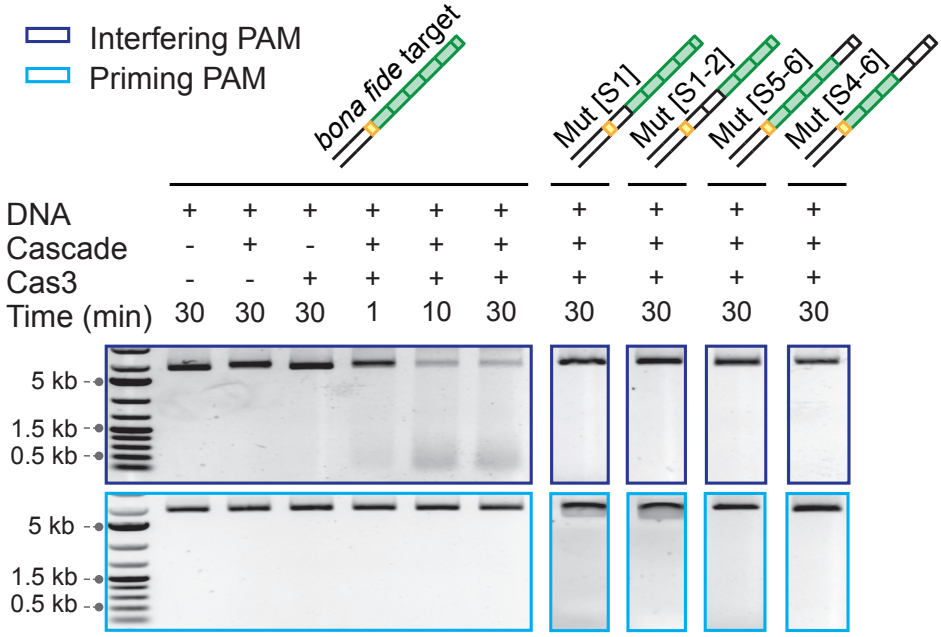


Figure S2.4: DNA degradation requires bona fide target

CRISPR interference reconstituted in vitro.Target plasmids (3.5 nM) harboring different target sequences (bona fide, Mut[S1], Mut[S1-2], Mut[S5-6], and Mut[S4-6], see schematics at top) with an interfering (purple outline, top row) or a priming (light blue outline, bottom row) PAM were incubated with Cascade (35 nM) for 15 min prior to addition of Cas3 (70 nM), which subsequently incubated for 1, 10, or 30 min. Reactions were run on a 0.8% agarose gel and stained for imaging. Degradation of the bona fide target can be seen a diffuse band at the bottom of the gel that increases in intensity with time.

2.6.2 Supplementary tables

Table S2.1: Oligos used for target DNA immobilization, and PCR

Construct (label pos.)	Sequence[a] (5′ → 3′)	Description[b]
TJ3_15-18 bona fide target (+9)	TTTGGTCTGCTCAATTTTGACAGCCCACATGGCATTCCACT/iAmMC6T/ATCACTGGCATCCTTCCACACTCCTG	Oligo for immobilisation
nTJ3_15 bona fide (+17)	CAGGAGTGTGGAAGGATGCCAGTGATAAGTGGAA/iAmMC6T/GCCATGTGGGCTGTC	Oligo for immobilisation
80nt biotin linker	AAAATTGAGCAGACCAAA$(PolyT)_{62}$- Biotin	Oligo for immobilisation
BG3903	CACCGGCCATGGCACTGTGCACACCATCGCGGAATTTGCTTATT-GATAACTGGATCCCTGTACGCC	Fw for N-term FGE tag CseI (NcoI)
BG3904	CCGGTGGGTACCTCAGCCATTTGATGGCCCTCCTTGCGGTTT-TAACTCCC	Rv CseI (KpnI)
BG4225	TTTGAATTCGCGCTGCATGCCTATTTG	Fw KmR in pRSF-1b (EcoRI)
BG5157	TTTTCCATGGGACAGCCCACATGGCATTCCACTTATCACTGGCAT-CCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer (NcoI)
BG5158	TTTTCCATGGACGGTGAGACATGGCATTCCACTTATCACTGGCAT-CCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer (Mut^{S5-6}) (NcoI)
BG5159	TTTTCCATGGACGGTGAGCACGCACATTCCACTTATCACTGGCAT-CCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer (Mut^{S4-6}) (NcoI)
BG5160	TTTTCCATGGGACAGCCCACATGGCATTCCACTTATCACTGGCGT-CCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer (Mut^{PAM}) (NcoI)
BG5161	TTTTCCATGGACGGTGAGACATGGCATTCCACTTATCACTGGCGT-CCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer ($Mut^{PAM\ S5-6}$) (NcoI)

[a] Underlined sequence indicates Nco1 restriction site.

[b] Fw stands for forward, Rv stands for reverse, FGE stands for formylglycine generating enzyme

Construct (label pos.)	Sequence[a] (5′ → 3′)	Description[b]
TJ3_15-18 bona fide target (+9)	TTTGGTCTGCTCAATTTTGACAGCCCACATGGCATTCCACT/iAmMC6T/ATCACTGGCATCCTTCCACACTCCTG	Oligo for immobilisation
BG5162	TTTTCCATGGACGGTGAGCACGCACATTCCACTTATCACTGGCGTCCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer ($Mut^{PAM\ S4\text{-}6}$) (NcoI)
BG5163	TTTTCCATGGGACAGCCCACATGGCATTCCACTTATACGGCCCATCCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer (Mut^{S1}) (NcoI)
BG5164	TTTTCCATGGGACAGCCCACATGGCATTCCCGGTGCACGGCCCATCCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer ($Mut^{S1\text{-}2}$) (NcoI)
BG5165	TTTTCCATGGGACAGCCCACATGGCATTCCACTTATACGGCCCGTCCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer ($Mut^{PAM\text{-}S1}$) (NcoI)
BG5166	TTTTCCATGGGACAGCCCACATGGCATTCCCGGTGCACGGCCCGTCCTTCCACACTCCTGTTAGAAAAACTCATCGAGCATC	Rv KmR in pRSF-1b with J3 protospacer ($Mut^{PAM\ S1\text{-}2}$) (NcoI)
BG5301	AAGGTTGGTGGGTTGTTTTTATGG	Fw Oligonucleotide annealing to the CRISPR 2.1 leader
BG5302	GGATCGTCACCCTCAGCAGCG	Rv Oligonucleotide annealing to spacer g8

Table S2.2: DNA oligos used in single-molecule studies of PAM-distal and PAM-proximal mutations

Construct (label pos.)	Sequence[a] (5′ → 3′)
TJ3_15 bona fide (+9)	GACAGCCCACATGGCATTCCACT/iAmMC6T/ATCACTGGCATCCTTCCACACTCCTG
nTJ3_15 bona fide (+17)	CAGGAGTGTGGAAGGATGCCAGTGATAAGTGGAA/iAmMC6T/GCCATGTGGGCTGTC
TJ3_15 Mut^{S5-6} (+9)	ACGGTGAGACATGGCATTCCACT/iAmMC6T/ATCACTGGCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S5-6} (+17)	CAGGAGTGTGGAAGGATGCCAGTGATAAGTGGAA/iAmMC6T/GCCATGTCTCACCGT
TJ3_15 Mut^{S4-6} (+9)	ACGGTGAGCACGCACATTCCACT/iAmMC6T/ATCACTGGCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S4-6} (+17)	CAGGAGTGTGGAAGGATGCCAGTGATAAGTGGAA/iAmMC6T/GTGCGTGCTCACCGT
TJ3_15 Mut^{S3-6} (+9)	ACGGTGAGCACGCAGACCGGACT/iAmMC6T/ATCACTGGCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S3-6} (+17)	CAGGAGTGTGGAAGGATGCCAGTGATAAGTCCGG/iAmMC6T/CTGCGTGCTCACCGT
TJ3_15 Mut^{S1-6} (+9)	ACGGTGAGCACGCAGACCGGCGG/iAmMC6T/GCACGGCCCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S1-6} (+17)	CAGGAGTGTGGAAGGATGGGCCGTGCACCGCCGG/iAmMC6T/CTGCGTGCTCACCGT
TJ3_15 Mut^{PAM} (+9)	GACAGCCCACATGGCATTCCACT/iAmMC6T/ATCACTGGCGTCCTTCCACACTCCTG
nTJ3_15 Mut^{PAM} (+17)	CAGGAGTGTGGAAGGACGCCAGTGATAAGTGGAA/iAmMC6T/GCCATGTGGGCTGTC
TJ3_15 Mut$^{PAM+S5-6}$ (+9)	ACGGTGAGACATGGCATTCCACT/iAmMC6T/ATCACTGGCGTCCTTCCACACTCCTG
nTJ3_15 Mut$^{PAM+S5-6}$ (+17)	CAGGAGTGTGGAAGGACGCCAGTGATAAGTGGAA/iAmMC6T/GCCATGTCTCACCGT
TJ3_15 Mut$^{PAM+S4-6}$ (+9)	ACGGTGAGCACGCACATTCCACT/iAmMC6T/ATCACTGGCGTCCTTCCACACTCCTG
nTJ3_15 Mut$^{PAM+S4-6}$ (+17)	CAGGAGTGTGGAAGGACGCCAGTGATAAGTGGAA/iAmMC6T/GTGCGTGCTCACCGT
TJ3_15 Mut$^{PAM+S3-6}$ (+9)	ACGGTGAGCACGCAGACCGGACT/iAmMC6T/ATCACTGGCGTCCTTCCACACTCCTG

[a] "/iAmMC6T/" refers to an amino-modified thymine base at the indicated position

Construct (label pos.)	Sequence[a] (5′ → 3′)
nTJ3_15 Mut$^{PAM+S3-6}$ (+17)	CAGGAGTGTGGAAGGACGCCAGTGATAAGTCCGG/iAmMC6T/CTGCGTGCTCACCGT
TJ3_15 Mut^{S1} (+9)	GACAGCCCACATGGCATTCCACT/iAmMC6T/ATACGGCCCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S1} (+17)	CAGGAGTGTGGAAGGATGGGCCGTATAAGTGGAA/iAmMC6T/GCCATGTGGGCTGTC
TJ3_15 Mut^{S1-2} (+9)	GACAGCCCACATGGCATTCCCGG/iAmMC6T/GCACGGCCCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S1-2} (+17)	CAGGAGTGTGGAAGGATGGGCCGTGCACCGGGAA/iAmMC6T/GCCATGTGGGCTGTC
TJ3_15 Mut^{S1-3} (+9)	GACAGCCCACATGGGACCGGCGG/iAmMC6T/GCACGGCCCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S1-3} (+17)	CAGGAGTGTGGAAGGATGGGCCGTGCACCGCCGG/iAmMC6T/CCCATGTGGGCTGTC
TJ3_15 Mut^{S1-4} (+9)	GACAGCCCCACGCAGACCGGCGG/iAmMC6T/GCACGGCCCATCCTTCCACACTCCTG
nTJ3_15 Mut^{S1-4} (+17)	CAGGAGTGTGGAAGGATGGGCCGTGCACCGCCGG/iAmMC6T/CTGCGTGGGGCTGTC
TJ3_15 Mut^{PAM+S1} (+9)	GACAGCCCACATGGCATTCCACT/iAmMC6T/ATACGGCCCGTCCTTCCACACTCCTG
nTJ3_15 Mut^{PAM+S1} (+17)	CAGGAGTGTGGAAGGACGGGCCGTATAAGTGGAA/iAmMC6T/GCCATGTGGGCTGTC
TJ3_15 Mut$^{PAM+S1-2}$ (+9)	GACAGCCCACATGGCATTCCCGG/iAmMC6T/GCACGGCCCGTCCTTCCACACTCCTG
nTJ3_15 Mut$^{PAM+S1-2}$ (+17)	CAGGAGTGTGGAAGGACGGGCCGTGCACCGGGAA/iAmMC6T/GCCATGTGGGCTGTC
TJ3_15 Mut$^{PAM+S1-3}$ (+9)	GACAGCCCACATGGGACCGGCGG/iAmMC6T/GCACGGCCCGTCCTTCCACACTCCTG
nTJ3_15 Mut$^{PAM+S1-3}$	CAGGAGTGTGGAAGGACGGGCCGTGCACCGCCGG/iAmMC6T/CCCATGTGGGCTGTC
TJ3_15 Mut$^{PAM-S1-4}$ (+9)	GACAGCCCCACGCAGACCGGCGG/iAmMC6T/GCACGGCCCGTCCTTCCACACTCCTG
nTJ3_15 Mut$^{PAM+S1-4}$ (+17)	CAGGAGTGTGGAAGGACGGGCCGTGCACCGCCGG/iAmMC6T/CTGCGTGGGGCTGTC

[a] "/iAmMC6T/" refers to an amino-modified thymine base at the indicated position

Table S2.3: Plasmid constructs

Plasmid	Description and order of genes (5'--> 3')	Restriction sites	Primers	Source
pWUR408	*cse1* in pRSF-1b, no tags			[60]
pWUR564	CRISPR containing J3 spacer in pACYCDuet-1. Derivative of pWUR477			[61]
pWUR610	pUC-3xJ3; pUC19 3 copies of the J3-protospacer, corresponding to a sequence derived from phage Lambda.			[59]
pWUR630	CRISPR poly J3, GA0936818 in pACYCDuet-1	NcoI/KpnI		[59]
pWUR656	*cse2* with Strep-tag II (N-term)-*cas7*-*cas5*-*cas6e* in pCDF-1b	NcoI/NotI		[16]
pWUR706	*cse1* containing an N-terminal LCTPSR FGE recognition sequence		BG3903/ BG3904	This study
pWUR738	J3 target plasmid. derivative of pGFPuv	NcoI(BspHI)/ EcoRI	BG4225/ BG5157	This study
pWUR739	Mutant J3 target plasmid (Mut^{S5-6}) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5158	This study
pWUR740	Mutant J3 target plasmid (Mut^{S4-6}) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5159	This study
pWUR741	Mutant J3 target plasmid (Mut^{PAM}) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5160	This study
pWUR742	Mutant J3 target plasmid ($Mut^{PAM\ S5-6}$) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5161	This study
pWUR743	Mutant J3 target plasmid ($Mut^{PAM\ S4-6}$) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5162	This study
pWUR744	Mutant J3 target plasmid (Mut^{S1}) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5163	This study

Plasmid	Description and order of genes (5'--> 3')	Restriction sites	Primers	Source
pWUR745	Mutant J3 target plasmid (Mut^{S1-2}) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5164	This study
pWUR746	Mutant J3 target plasmid ($Mut^{PAM\ S1}$) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5165	This study
pWUR747	Mutant J3 target plasmid ($Mut^{PAMS1-2}$) derivative of pGFPuv (Clontech)	NcoI(BspHI)/ EcoRI	BG4225/ BG5166	This study
pWUR748	pMAT11-MBP-Cas3	EcoRI/XhoI		[17]
#16132	pBAD/myc-his A Rv0712 (FGE)			Addgene

2.7 References

1 R. Barrangou, CRISPR-Cas systems and RNA-guided interference. *Wiley Interdiscip. Rev. RNA*. **4**, 267–278 (2013).

2 E. Charpentier, L. A. Marraffini, Harnessing CRISPR-Cas9 immunity for genetic engineering. *Curr. Opin. Microbiol.* **19**, 114–119 (2014).

3 P. C. Fineran, E. Charpentier, Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. *Virology*. **434**, 202–209 (2012).

4 J. Reeks, J. H. Naismith, M. F. White, CRISPR interference: a structural perspective. *Biochem. J.* **453**, 155–166 (2013).

5 J. E. Samson, A. H. Magadán, M. Sabri, S. Moineau, Revenge of the phages: defeating bacterial defences. *Nat. Rev. Microbiol.* **11**, 675–87 (2013).

6 R. Sorek, C. M. Lawrence, B. Wiedenheft, CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu Rev Biochem.* **82**, 237–266 (2013).

7 E. R. Westra *et al.*, The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339 (2012).

8 J. van der Oost, E. R. Westra, R. N. Jackson, B. Wiedenheft, Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **12**, 479–92 (2014).

9 G. W. Goldberg, W. Jiang, D. Bikard, L. a Marraffini, Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature*. **514**, 633–637 (2014).

10 C. R. Hale, A. Cocozaki, H. Li, R. M. Terns, M. P. Terns, Target RNA capture and cleavage by the Cmr type III-B CRISPR-cas effector complex. *Genes Dev.* **28**, 2432–2443 (2014).

11 C. Rouillon *et al.*, Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell.* **52**, 124–134 (2013).

12 R. H. J. Staals *et al.*, Structure and Activity of the RNA-Targeting Type III-B CRISPR-Cas Complex of Thermus thermophilus. *Mol. Cell.* ***52***, 135–145 (2013).

13 G. Tamulaitis *et al.*, Programmable RNA Shredding by the Type III-A CRISPR-Cas System of Streptococcus thermophilus. *Mol. Cell.* **56**, 506–517 (2014).

14 P. D. Hsu, E. S. Lander, F. Zhang, Development and applications of CRISPR-Cas9 for genome engineering. *Cell.* ***157***, 1262–1278 (2014).

15 R. M. Terns, M. P. Terns, CRISPR-based technologies: Prokaryotic defense weapons repurposed. *Trends Genet.* **30**, 111–118 (2014).

16 M. M. Jore *et al.*, Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18**, 529–536 (2011).

17 S. Mulepati, S. Bailey, In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA Target. *J. Biol. Chem.* **288**, 22184–22192 (2013).

18 T. Sinkunas *et al.*, In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *EMBO J.* **32**, 385–394 (2013).

19 E. R. Westra *et al.*, CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Mol. Cell.* **46**, 595–605 (2012).

20 T. Künne, D. C. Swarts, S. J. J. Brouns, Planting the seed: Target recognition of short guide RNAs. *Trends Microbiol.* **22**, 74–83 (2014).

21 D. G. Sashital, B. Wiedenheft, J. A. Doudna, Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. *Mol. Cell.* **46**, 606–615 (2012).

22 E. R. Westra *et al.*, Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. *PLoS Genet.* **9** (2013).

23 S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature.* **507**, 62–67 (2014).

24 M. D. Szczelkun *et al.*, Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9798–803 (2014).

25 K. A. Datsenko *et al.*, Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.* **3**, 945 (2012).

26 P. C. Fineran *et al.*, Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E1629-38 (2014).

27 M. Li, R. Wang, D. Zhao, H. Xiang, Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Res.* **42**, 2483–2492 (2014).

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

29 E. Semenova *et al.*, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10098–10103 (2011).

30 T. Ha, Single-molecule methods leap ahead. *Nat. Methods.* **11**, 1015–1018 (2014).

31 C. Joo, M. Fareh, V. Narry Kim, Bringing single-molecule spectroscopy to macromolecular protein complexes. *Trends Biochem. Sci.* **38**, 30–37 (2013).

32 M. F. Juette *et al.*, The bright future of single-molecule fluorescence imaging. *Curr. Opin. Chem. Biol.* **20**, 103–111 (2014).

33 Robinson, A. M. van Oijen, Bacterial replication, transcription and translation: mechanistic insights from single-molecule biochemical studies. *Nat. Rev. Microbiol.* **11**, 303–15 (2013).

34 B. Schuler, H. Hofmann, Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales. *Curr. Opin. Struct. Biol.* **23**, 36–47 (2013).

35 M. L. Hochstrasser *et al.*, CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6618–23 (2014).

36 B. Wiedenheft *et al.*, Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature.* **477**, 486–489 (2011).

37 R. N. Jackson *et al.*, Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science.* **345**, 1473–9 (2014).

38 S. Mulepati, A. Héroux, S. Bailey, Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science.* **345**, 1479–84 (2014).

39 H. Zhao *et al.*, Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. *Nature.* **515**, 147–50 (2014).

40 S. Gandon, P. F. Vale, The evolution of resistance against good and bad infections. *J. Evol. Biol.* **27**, 303–312 (2014).

41 B. F. Koel *et al.*, Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science.* **342**, 976–980 (2013).

42 W. E. Purtha, T. F. Tedder, S. Johnson, D. Bhattacharya, M. S. Diamond, Memory B cells, but not long-lived plasma cells, possess antigen specificities for viral escape mutants. *J. Exp. Med.* ***208***, 2599–606 (2011).

43 D. Tarlinton, K. Good-jacobson, Diversity Among Memory B Cells : *Science*. **341**, 1205–1211 (2013).

44 H. Deveau *et al.*, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1390–1400 (2008).

45 A. D. Weinberger, Y. I. Wolf, A. E. Lobkovsky, M. S. Gilmore, E. V. Koonin, Viral diversity threshold for adaptive immunity in prokaryotes. *MBio.* **3**, 1–10 (2012).

46 I. I. Cisse, H. Kim, T. Ha, A rule of seven in Watson-Crick base-pairing of mismatched sequences. *Nat. Struct. Mol. Biol.* **19**, 623–7 (2012).

47 Y. Wang *et al.*, Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature*. **461**, 754–761 (2009).

48 N. T. Schirle, J. Sheu-Gruttadauria, I. J. MacRae, Structural basis for microRNA targeting. *Science*. **346**, 608–613 (2014).

49 Z. Chen, H. Yang, N. P. Pavletich, Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. *Nature*. **453**, 489–4 (2008).

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

51 C. Kuscu, S. Arslan, R. Singh, J. Thorpe, M. Adli, Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol*. **32**, 677–683 (2014).

52 D. C. Swarts, C. Mosterd, M. W. J. van Passel, S. J. J. Brouns, CRISPR interference directs strand specific spacer acquisition. *PLoS One*. **7**, 1–7 (2012).

53 S. Shmakov *et al.*, Pervasive generation of oppositely oriented spacers during CRISPR adaptation. *Nucleic Acids Res.* **42**, 5907–5916 (2014).

54 X. Liang *et al.*, Molecular basis of transcriptional fidelity and DNA lesion- induced transcriptional mutagenesis. *DNA Repair (Amst).*, 71–83 (2014).

55 D. Rabuka, J. S. Rush, G. W. deHart, P. Wu, C. R. Bertozzi, Site-specific chemical protein conjugation using genetically encoded aldehyde tags. *Nat. Protoc.* **7**, 1052–1067 (2012).

56 I. Rasnik, S. a McKinney, T. Ha, Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods.* **3**, 891–893 (2006).

57 I. S. Carrico, B. L. Carlson, C. R. Bertozzi, Introducing genetically encoded aldehydes into proteins. *Nat. Chem. Biol.* **3**, 321–322 (2007).

58 S. D. Chandradoss *et al.*, Surface passivation for single-molecule protein studies. *J. Vis. Exp.* **50549**, 4–11 (2014).

59 E. R. Westra *et al.*, H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. *Mol. Microbiol.* **77**, 1380–1393 (2010).

60 S. J. J. Brouns *et al.*, Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science.* **340**, 216–219 (2008).

61 E. R. Westra, S. J. J. Brouns, The rise and fall of CRISPRs - dynamics of spacer acquisition and loss. *Mol. Microbiol.* **85**, 1021–1025 (2012).

3

The CRISPR associated Cas3 protein repetitively probes the target DNA with a 1-nt step size

Under Revision

Luuk Loeff [1], Stan J.J. Brouns** [1,2] & Chirlmin Joo** [1]

** Co-corresponding authors

[1] Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, 2628 CJ, Delft, The Netherlands

[2] Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University, 6703 HB, Wageningen, The Netherlands

3.1 Abstract

CRISPR-Cas loci provide an RNA-guided adaptive immune system that mediates defense against invading genetic elements. Interference in type I systems relies on the RNA-guided surveillance complex Cascade for target recognition and the trans-acting Cas3 helicase/nuclease protein for target degradation. Even though the biochemistry of CRISPR interference has been well understood, the biophysics of DNA unwinding and coupling of the helicase and nuclease domains of Cas3 remain elusive. Here we employed single-molecule FRET to probe the helicase activity with a high spatiotemporal resolution. We show that Cascade and Cas3 remain tightly associated, while Cas3 unwinds target DNA in distinctive steps of 3 basepairs with an underlying translocation step size of 1-nt. Unwinding is highly repetitive, allowing Cas3 to repeatedly present the intrinsically inefficient nuclease domain with unwound DNA. Our study reveals an unanticipated level of complexity, in which the discontinuous and burst-like helicase properties of Cas3 are the driving force behind CRISPR interference.

3.2 Introduction

Prokaryotes mediate defense against invading genetic elements using RNA guided adaptive immune systems that are encoded by CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated) loci [1, 2]. In the type I system, the most ubiquitous CRISPR-Cas system [3], foreign DNA targets (called protospacers) are recognized by the CRISPR RNA (crRNA)-guided surveillance complex Cascade [4]. Recognition of double stranded DNA targets results in the formation of an R-loop, in which the crRNA hybridizes with the complementary target strand and the non-complementary strand of the DNA is displaced (nontarget strand) [5–8]. This R-loop formation triggers a conformational change in the Cascade complex [6, 9, 10] and leads to the recruitment of the Cas3 protein for subsequent target degradation [11–13].

The *E. coli* Cas3 protein consists of two domains: a N-terminal metal-dependent histidine-aspartate (HD) nuclease domain and a C-terminal superfamily 2 helicase domain [3, 11, 14–17]. Cas3 is activated by the Cascade-marked R-loop, where it cleaves the displaced nontarget strand ~11 nucleotides into the R-loop region [14, 18]. Driven by ATP, Cas3 then moves along the nontarget strand in a 3′ to 5′ direction, while it catalyzes cobalt-dependent DNA degradation [12, 14, 18, 19]. Subsequently, Cas3 generates degradation products that are close to spacer length and enriched for NTT in their 3′ ends [20]. This makes a considerable fraction of the degradation products suitable substrates for integration by the Cas1-Cas2 integrases into the CRISPR locus [20]. Yet, the biophysics of DNA unwinding by Cas3 remains elusive In particular, it is not understood how the putative exonuclease HD domain can create degradation products of length suitable for spacer integration and how this process takes place in concert with Cascade.

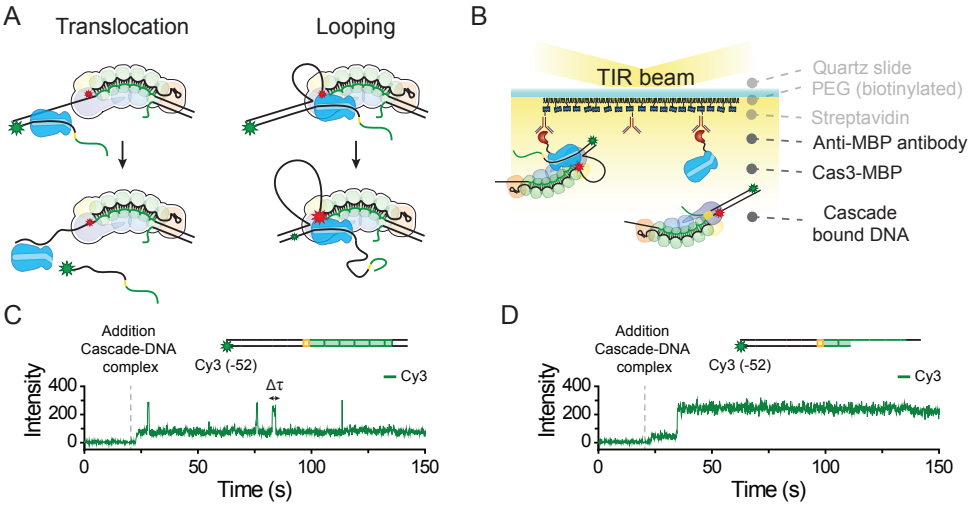


Figure 3.1: Single-molecule visualisation of the interaction between Cas3 and Cascade

(A) Schematic of two distinct model for DNA unwinding by Cas3. In the translocation model (left), Cas3 breaks its contacts with Cascade while it unwinds DNA. This results in translocation away from the Cascade target site. During DNA looping (right), Cascade and Cas3 remain tightly associated while Cas3 pulls on the DNA. The appearance of FRET during translocation or loop formation is indicated by the size of the star: low FRET, large green star) or high FRET, large red star. **(B)** Schematic of a single-molecule FRET assay used to probe the interaction between Cas3 and Cascade. **(C)** A representative time trace of the initial interaction between Cas3 and Cascade bound to a cognate target. **(D)** A representative time trace of the initial interaction between Cas3 and Cascade bound to a nicked target.

3.3 Results

3.3.1 Single-molecule observation of DNA unwinding by Cas3

We set out to understand how Cas3 unwinds dsDNA substrates. To date, two models prevail for DNA unwinding by the Cas3 helicase: a translocation model and a looping model. In the translocation model, Cas3 breaks its contacts with the Cascade complex while unwinding the DNA. Thereby Cas3 translocates away from the Cascade binding site and degrades single-stranded DNA fragments along the way (Figure 3.1A) [11–14, 19, 21]. In the looping model, Cas3 and Cascade remain in tight contact while Cas3 unwinds the DNA, which may result in loops in the target strand (Figure 3.1A) [19]. To distinguish between these two models, we sought to visualize the DNA unwinding activity of Cas3 with a high spatiotemporal resolution.

To visualize DNA unwinding by Cas3, we developed an assay based on single-molecule Förster resonance energy transfer (smFRET). In brief, anti-maltose binding protein (MBP) antibodies were anchored to the surface of a polyethylene glycol (PEG)-coated slide through biotin-streptavidin linkage followed by tethering of MBP-fused Cas3 monomers (Figure 3.1B & Figure S3.1A to Figure S3.1C). Notably, the immobilization of Cas3 did not appreciably affect its capability to degrade dsDNA substrates (Figure S3.1C to Figure S3.1G). Next, the antibody-tethered Cas3 molecules were presented to Cascade complexes bound to dye-labeled

dsDNA substrates and their interactions were probed in real time using total internal reflection fluorescence (TIRF) microscopy (Figure 3.1B). We first explored the interaction of Cas3 with Cascade complexes that were bound to a fully complementary dsDNA target. When the complexes were introduced in absence of cobalt, transient interactions were observed with a dwell-time ($\Delta\tau$) of 1.63 ± 0.236 s, which reflect the initial interaction between the Cse1 subunit of the Cascade complex and the Cas3 protein (Figure 3.1C & Figure S3.2A) [13]. This finding is consistent with DNA curtain experiments where no stable interaction between Cascade and Cas3 was observed when cobalt was omitted from the assay [19].

When the same experiment was repeated with a partial dsDNA construct that mimicked the nicked R-loop reaction intermediate formed by Cas3 (Figure 3.2A), a stable interaction was observed between Cascade and Cas3. This interaction lasted throughout the time course of the experiment and followed photo bleaching kinetics (Figure 3.1D & Figure S3.2B). This suggest that the initial nick made by Cas3 facilitates loading of the helicase domain, which stabilizes the interaction between Cas3 and the Cascade complex. Notably, the appearance of fluorescence signals was not observed when Cascade was omitted from the assay, confirming that Cas3 exclusively interacts with DNA in a Cascade-dependent manner (Figure S3.2C & Figure S3.2D) [8, 12, 13].

To focus on the mechanism by which Cas3 unwinds DNA, experiments were continued with the partial dsDNA construct that allowed for synchronized initiation of DNA unwinding. The DNA substrate was labelled with a donor (Cy3) and an acceptor (Cy5) dye that were positioned such that it could report on loop formation in the target strand via an increase in FRET (Figure 3.1A, Figure 3.2A & Table S3.1). The fluorescent probes where conjugated to the DNA using an amino-C6-linker (thymine-5-C6 amino linker), which has been shown not to interfere with the translocation and unwinding by helicases [22–25]. The target strand was labelled with Cy5 at nucleotide -7, which position is fixed near the Cascade complex [8]. The Cy3 dye was positioned further upstream of the PAM at position -52 such that high FRET would be observed upon loop formation in the target strand by Cas3 (Figure 3.1A). In absence of ATP, no FRET was observed between the donor and acceptor fluorophore, resulting in FRET values that were indistinguishable from background signals (E = 0.18) (Figure 3.1B & Figure S3.2E).

Upon introduction of ATP into the microfluidic chamber, a large fraction of the Cas3 molecules (201 out of 438 molecules) showed a gradual increase in FRET, which is consistent with loop formation (Figure 3.2B & Figure S3.3A). For remaining molecules, FRET stayed within background levels (E = 0.18). We hypothesize that these molecules either failed to initiate unwinding within our observation time (3.5 min) or form small loops outside the FRET range of approximately 20 base pairs (bp) (Figure S3.8A & Figure S3.8B). Consistent with the second hypothesis, the probability of unwinding scaled exponentially with the distance to the target-site (Figure 3.3D). This data shows that Cas3 remains anchored to Cascade while unwinding DNA. Notably, translocation of Cas3 away from the Cascade target site

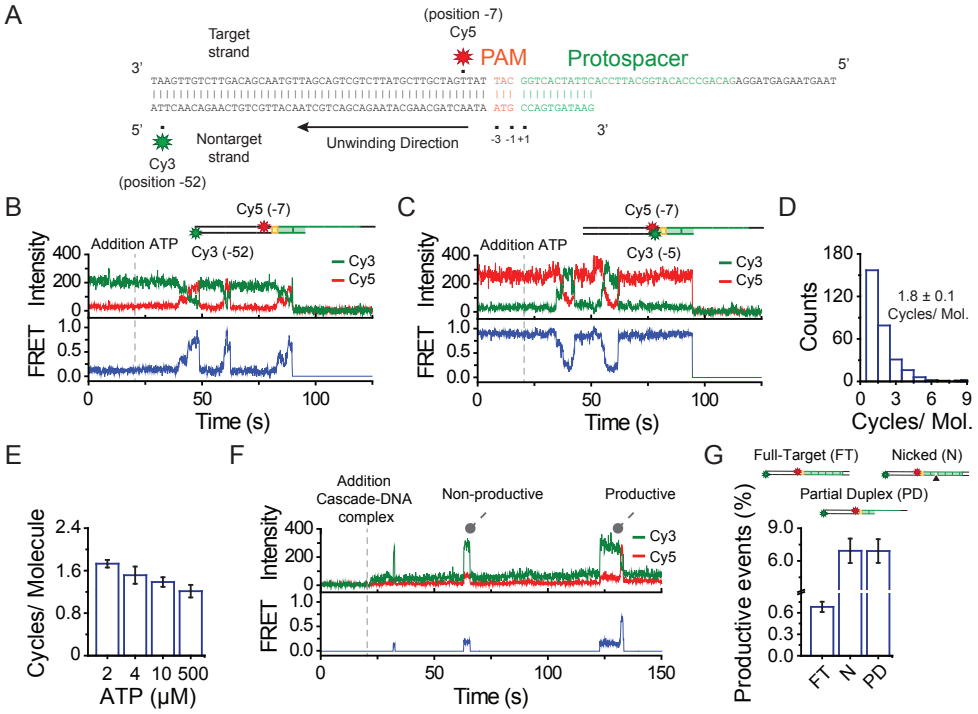


Figure 3.2: Real-time observation of DNA unwinding by Cas3

(A) Partial duplex DNA constructs consist of a PAM (orange), protospacer (green) and two flanks of 50 nt and 15 nt (black). Cy5 (red star) was attached to position -7 of the target strand and Cy3 (green star) to position -52 of the nontarget strand. **(B)** A representative time trace of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue) exhibiting multiple unwinding events. ATP (2 μM) was added at t = 20s (dashed gray line). **(C)** A representative time trace for a construct with Cy5 (red star) attached to position -7 of the target strand and Cy3 (green star) to position -5 of the nontarget strand. ATP (2 μM) was added at t = 20s (dashed gray line). **(D)** A histogram representing the number of unwinding cycles for each molecule. Error represents the standard error of the mean (SEM) from three independent measurements (N=3). **(E)** Quantification of the number of unwinding cycles per molecule at various ATP concentrations. Error bars represent the SEM (N=3). **(F)** Representative time traces of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue) obtained by tracking the interactions by Cas3 and Cascade in real time. Cascade bound DNA, ATP (500 μM) and Co^{2+} (10 μM) were added at t = 20s. **(G)** Quantification of the number of productive binding events for three distinct DNA constructs. Cy5 (red star) was attached to position -7 of the target strand and Cy3 (green star) to position -52 of the nontarget strand. Black triangle indicates the position of the nick (11 nt away from PAM).

[11–14, 19, 21] was not observed under our experimental conditions, which would have been manifested by a rapid loss of the total fluorescence signal (Figure 3.1A). Instead, loss of fluorescence was negligible and followed only the photo bleaching kinetics (Figure S3.2B).

To confirm loop formation, we tested various alternative immobilization and labelling schemes. When the DNA (Figure S3.3C & Figure S3.3D) or Cascade (Figure S3.3E & Figure S3.3F) was immobilized or when the donor and acceptor dyes were

swapped (Figure S3.4A & Figure S3.4D), identical behavior was observed. Next, an alternative labelling scheme with a donor and acceptor dye at position -5 of the nontarget strand and -7 of the target strand, respectively (Figure S3.2F), was tested. This construct initially yielded high FRET (E = 0.8, Figure S3.2F) and should lead to a decrease in FRET when unwinding is triggered. In agreement with our expectation, FRET decreased upon introduction of ATP (Figure 3.2C & Figure S3.3B). The same observation was made using PIFE (protein-induced fluorescence enhancement) (Figure S3.3G). In contrast, when a construct was used that was designed to detect loop formation on the nontarget strand (Cy3 position -52 target strand and Cy5 position -7 nontarget strand, Figure S3.4B) or when the PAM proximal and PAM distal flank were swapped (Figure S3.4C), a change in FRET was not observed (Figure S3.4E & Figure S3.4F). These control experiments support the model that Cas3 remains anchored to Cascade when pulling on the 3' end of nontarget strand, which results in DNA loops in the target strand during unwinding.

Our real-time analysis of DNA unwinding by Cas3 revealed, that Cas3 could go through multiple cycles of unwinding on a single substrate, by slipping back to its initial location (Figure 3.2B & Figure 3.2C). Analysis of this repetitive behavior showed that Cas3 undergoes an average of 1.8 ± 0.1 cycles per substrate (Figure 3.2D). Interestingly, the number of unwinding cycles per molecule decreased with an increase of ATP, reaching average unwinding frequency of 1.2 ± 0.1 cycles per molecule at saturating levels of ATP (Figure 3.2E). This data suggests that Cas3 is more effective in displacing the nontarget strand away from the Cas3-Cascade complex at higher levels of ATP, which is likely a result of using short DNA oligo's. Consistent with this hypothesis, the dwell time of the looping population that reached the end of the substrate was ~3 times shorter as compared to the seemingly stattic population (Figure S3.2B).

3.3.2 Cas3 exhibits sparse nuclease activity

Previous bulk experiments have shown that Cas3 degrades the nontarget strand while it moves along the DNA [14, 18]. Therefore, we hypothesized that activation of the nuclease domain, by the addition of cobalt, would result in a stark decrease in the number of unwinding cycles per molecule. However, no change in the number of cycles per molecule was observed when the nuclease domain was activated (Figure S3.3H), indicating that little nicking had occurred. Moreover, the addition of free Cas3 into the assay did not alter the behavior of Cas3 (data not shown).

To obtain a more quantitative estimate on the cleavage activity of Cas3, the initial interaction between Cas3 and Cascade was probed (Figure 3.2F). When Cascade bound to a full target substrate, without the initial nick, was introduced, only 0.7 ± 0.1% of the binding events resulted in DNA unwinding (Figure 3.2F & Figure 3.2G). However, when Cascade bound to a substrate mimicking the nicked intermediate was introduced (Figure 3.2A), the number of productive unwinding events increased with an order of magnitude (6.9 ± 1.1%, Figure 3.2F & Figure 3.2G). This suggest that the HD nuclease domain intrinsically exhibits a sparse nuclease

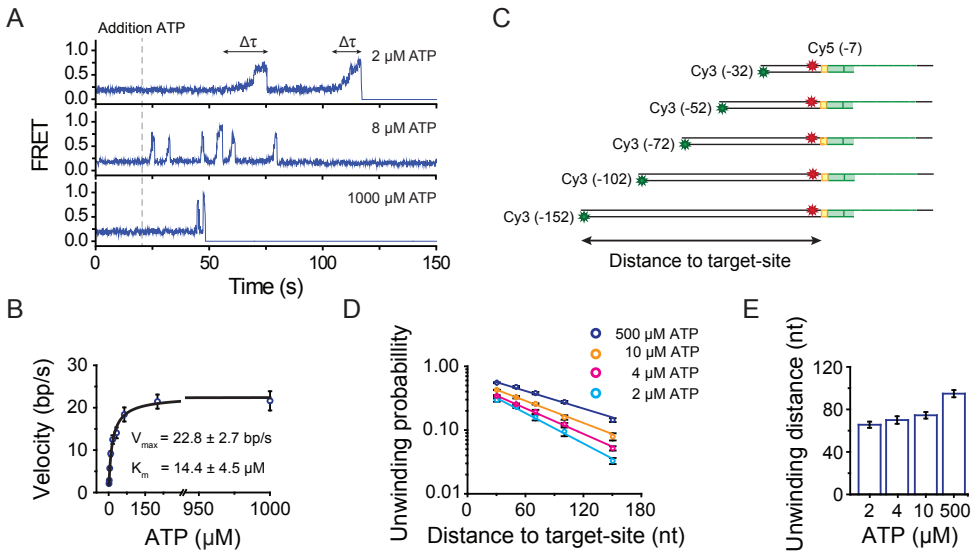


Figure 3.3: Velocity and processivity of the Cas3 helicase

(A) Representative FRET traces obtained at various ATP concentrations. **(B)** Michaelis-Menten fit (black line) of the velocity (bp/s) plotted against the ATP concentration (1, 2, 4, 8, 16, 32, 64, 200 and 1000 μM ATP). Error bars represent the 95% confidence interval obtained through bootstrap analysis. **(C)** Schematic overview of constructs used to determine the mean unwinding distance. Cy5 (red star) was attached to position -7 of the target strand and Cy3 (green star) was positioned at the end of the nontarget strand. The FRET range is indicated by the dashed gray lines filled with red gradient. **(D)** Unwinding probability over the distance to target-site at various ATP concentrations. Error bars represent the SEM (N=3). Solid lines represent a single-exponential fit used to determine the mean translocation distance. **(E)** Bar plot with average translocation distance (nt) at various ATP concentrations. Error bars represent the SEM (N=3).

activity, which contradicts previously published bulk data [12–14, 18, 20, 26]. Those bulk measurements were performed using a 10- to 500-fold excess of Cas3 [12–14, 18, 20, 26] that facilitated initial nicking and loading of the Cas3 helicase whereas here we used Cas3 in nano-molar concentrations. Our findings imply that the Cas3 protein compensates for the sparse nuclease activity by repeatedly feeding ssDNA in the HD nuclease domain, which ensures DNA cleavage.

3.3.3 Dynamics of DNA loop formation by Cas3

Next, we explored the molecular dynamics of DNA loop formation by Cas3. To determine the unwinding rate of Cas3, we performed DNA unwinding assays at various ATP concentrations (Figure S3.5D to Figure S3.5F). For every ATP concentration, the dwell time (Δτ) of each unwinding event was extracted (Figure 3.3A), followed by fitting of the histograms with a gamma distribution (Figure S3.5A to Figure S3.5C). Consistent with other helicases [22, 24, 27], the effective rate ($k_{effective}$, $1/\Delta\tau$) increased with increasing amounts of ATP, indicating that the unwinding velocity increases with ATP (Figure 3.3B). When ATP was replaced with a non hydrolyzable ATP analog ATP-γ-S, the unwinding activity of Cas3 was completely abrogated

(Figure S3.6A). To estimate the maximum velocity (V_{max}) of Cas3, the effective rate was converted to apparent velocity in base pairs per second (bp/s, see 3.5.5 on page 86). By plotting the velocity over the ATP concentration and fitting the data with a Michaelis-Menten fit (Figure 3.3B), a V_{max} = 22.8 ± 2.7 bp/s and K_m = 14.4 ± 4.5 μM was obtained. Notably, only a marginal change in velocity was observed when the nuclease domain was activated (Figure S3.5G), suggesting that the unwinding activity of the helicase domain dominates over the DNA degradation by the nuclease domain.

Recent DNA curtain experiments suggested that Cas3 is a highly processive molecular motor [19]. However, given that the Cas3 nuclease exhibits sparse activity, a highly processive motor would lead to single-stranded fragments that are much longer than the previously reported fragment size that is smaller than 200 nucleotides [14, 20]. Therefore, we sought to determine the average unwinding distance of Cas3 at saturating concentrations of ATP. To estimate the unwinding distance, a series of DNA substrates with an increasing length of the PAM proximal flank were used, while moving the donor dye towards the end of each substrate (Figure 3.3C). This set of constructs allowed for the determination of the probability that a Cas3 molecule reached the end of a DNA substrate within the observation time of 3.5 min.

Upon introduction of ATP, each construct yielded traces with identical behavior (Figure S3.7). However, we observed a decrease in the number of unwinding events with an increase in the flank length, suggesting that the unwinding probability decreased (Figure 3.3D). When the length of the flank was increased to 150 nt, the unwinding probability decreased to 0.13 ± 0.1 (Figure 3.3D), suggesting that the majority of molecules formed loops smaller than 150 nt. To estimate the average unwinding distance, the unwinding probability was plotted over the ATP concentration, followed by fitting each data series with a single-exponential decay. This yielded an average unwinding distance of 95 ± 3 nt at a saturating ATP concentration (Figure 3.3E). A decrease in the average unwinding distance was observed when the ATP concentration was lowered (Figure 3.3E). Notably, the addition of SSB did not alter the processivity of Cas3 (data not shown), implying that Cas3 may shelter the looped target strand. These observations are in good agreement previously reported bulk biochemical data, that showed Cas3 generates degradation products in the range of 30 to 150 nucleotides and become smaller at low ATP concentrations [14, 20, 28]. Taken together, these results suggest that the helicase domain of Cas3 limits the fragment size by repeatedly generating a distribution ssDNA fragments with an average size of ~90 nt.

3.3.4 Cas3 unwinds DNA in uniform steps

To understand what feature of the Cas3 helicase limits the unwinding distance, we sought to understand the molecular mechanism by which Cas3 unwinds the DNA. Close inspection of the FRET events revealed that FRET increased with a distinct pattern, marked by plateaus at specific FRET levels (Figure 3.4A). To elucidate this behavior, we employed an automated step-finder algorithm [29] (See Chapter 6 on page 177) that yielded the average FRET value for each plateau and the size of each step in between the plateaus (Figure 3.4A). Analysis of average the FRET value for each plateau, resulted in a histogram with four distinct peaks that were evenly separated (ΔE=0.15) (Figure 3.4B). To correlate these FRET values to distance in bp, we designed a series of DNA constructs, in which the distance between the dyes was systematically decreased (Figure S3.8A & Figure S3.8B). The calibration experiment yielded a conversion factor, in which 1-bp corresponds to a ΔE=0.05 FRET change. This conversion factor is in line with previously published work by [30]. Conversion of the FRET values suggests that Cas3 may move along the DNA with regular 3-bp steps.

To further characterize the stepping behavior of Cas3, a histogram was plotted with the distribution of step-sizes, the distance between each plateau. The distribution of the step-sizes exhibited a major peak centered at a step-size of approximately 3-bp (Figure 3.4C & Figure S3.8A to Figure S3.8C), which is consistent with the histogram of average the FRET value for each plateau (Figure 3.4B). Apart from the major peak at 3-bp, minor peaks that represented a multiplicity of this step-size (e.g. 6-bp) were observed (Figure 3.4C), which became more prominent when the ATP concentration was increased (Figure S3.8D). These larger steps are likely a result of a series of events that occur faster than our time resolution. Consistent with this hypothesis, a histogram of the average FRET levels at saturating concentration of ATP was skewed towards the high FRET states (Figure S3.8E). We confirmed the 3-bp step by designing a set of constructs in which the donor dye was shifted by 1, 2 or 3 nt (ΔN=1, ΔN=2 & ΔN=3) from the standard construct (ΔN=0) and observing 3-nt periodicity in FRET histograms (Figure 3.4D). These experiments provide strong evidence that Cas3 moves along the DNA in distinct steps of 3-bp at a time.

Apart from steps that led to an increase in FRET, we also observed slipping events where the FRET signal abruptly dropped to intermediate levels (Figure 3.4A). These events were represented as a negative value in our step-size analysis and showed a major peak centered at -3-bp (Figure 3.4C). Besides the slipping events to intermediate levels, we also observed slipping events that returned to their initial FRET state (Figure 3.2B & Figure 3.2C). We speculate that these slipping events occur through miscoordination of the RecA-like domains of the Cas3 helicase [16, 17], leaving the DNA to zip back over a short or long distances. The short and long-range slipping events result in discontinuous and burst-like unwinding behavior, that allows Cas3 to repeatedly feed ssDNA fragments into the nuclease domain for further processing.

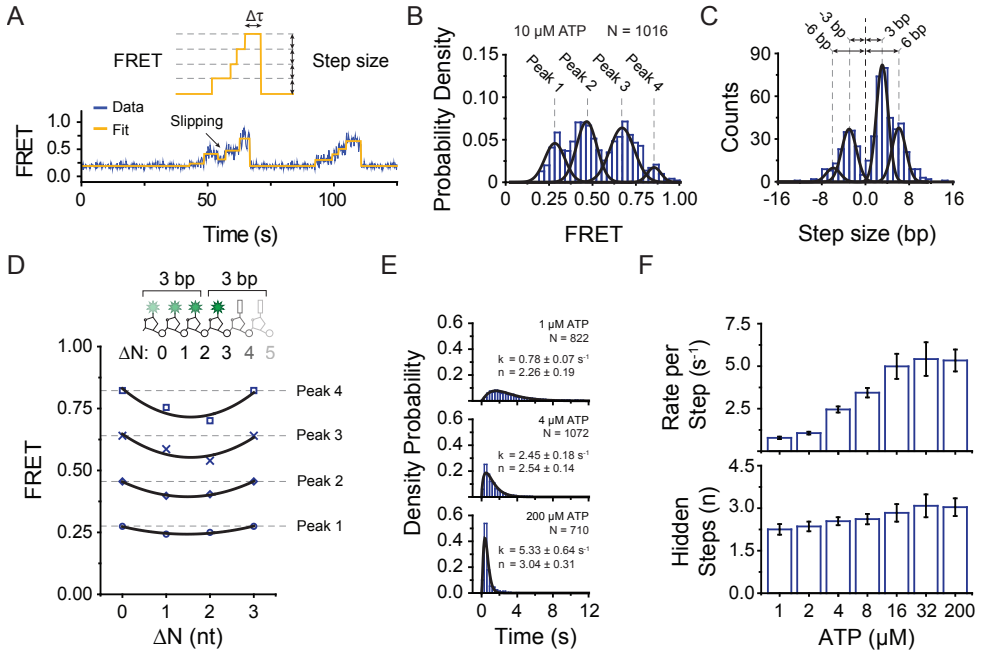


Figure 3.4: Cas3 unwinds DNA in uniform steps

(A) Representative FRET trace (dark blue) fitted with a step-finder algorithm (Orange). **(B)** Distribution of FRET levels obtained through the step-finder algorithm. Black lines represent a Gaussian fit. **(C)** Distribution of step-sizes obtained through a step-finder algorithm. Black lines represent a Gaussian fit. Dashed grey lines indicate the centre of each peak. Positive values represent processive unwinding whereas negative values represent slipping. **(D)** Location of the FRET levels for various positions of the donor dye. Given the remarkable regularity in the unwinding pattern of Cas3, we hypothesized that when moving the donor dye from its original position (ΔN=0) by one or two nucleotides (ΔN=1 & ΔN=2, respectively) would shift the position of the observed plateaus at specific FRET levels. In contrast, moving the donor dyes by three nucleotides (ΔN=3) locates the dye at a similar position as ΔN=0 (inset) and should yield identical FRET levels. Consistent with our hypothesis, the constructs with a donor dye at position ΔN=1 & ΔN=2 shifted the peak positions towards lower FRET values, whereas the construct with a dye at position ΔN=3 yielded identical FRET levels as ΔN=0 (Extended Data Fig. 8f). **(E)** Dwell-time distributions of the FRET levels at various ATP concentrations. Data was fitted with a gamma distribution (solid line) to obtain the number of hidden steps (n) and rate (k). Error represents the 95% confidence interval obtained through bootstrap analysis. **(F)** Bar plots representing the number of hidden steps (n) and rate (k) that was obtained through fitting dwell-time histograms with a gamma distribution. Error represents the 95% confidence interval obtained through bootstrap analysis.

Finally, we questioned if the observed 3-bp steps would correspond to the elementary step-size of the Cas3 helicase. If Cas3 would unwind 3-bp upon the hydrolysis of a single ATP molecule, the dwell time (Δτ, Figure 3.4A) histogram of the FRET levels would follow a single-exponential decay. However, a dwell time histogram of the FRET levels showed non-exponential behavior and followed a gamma distribution (Figure 3.4E). A fit of the histogram yielded a statistical description of the number of underlying hidden steps (n) and the rate per step (k) (Figure 3.4E). At various ATP concentrations, we obtained n values that remained

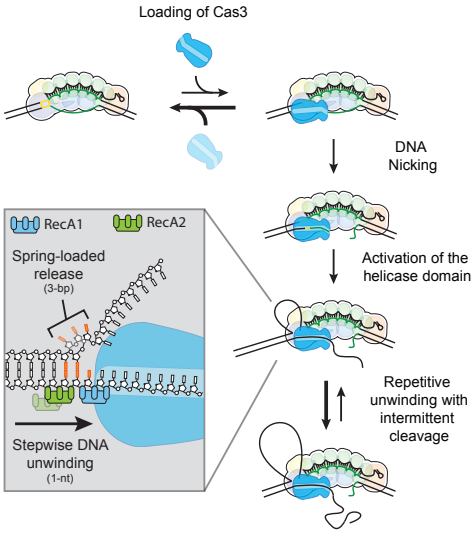


Figure 3.5: Model for CRISPR interference mediated DNA unwinding by Cas3

Model for CRISPR interference by Cas3. Interference starts with loading of the Cas3 protein onto the Cascade-DNA complex. Given the inherently sparse nuclease activity of Cas3 this may require multiple docking events. Once Cas3 nicks the R-loop, the loading of the helicase domain is facilitated that allows Cas3 and Cascade to form a stable complex. Upon hydrolysis of ATP, Cas3 initiates DNA unwinding that takes place in distinct spring-loaded steps and underlies an elementary step size of 1-bp. Cas3 repeatedly feeds ssDNA into its nuclease domain, which generates a distribution of degradation products with an average size of ~90 nt.

close to three hidden steps, whereas the rate increased with an increase of ATP (Figure 3.4E, Figure 3.4f & Figure S3.9A to Figure S3.9H). This analysis shows that each 3-bp step is composed of three hidden steps of 1-bp, suggesting that the elementary step-size of Cas3 is 1-nt. From this analysis, the model emerges that Cas3 successively unwinds three base pairs in 1-nt steps, using its RecA-like domains [16, 17]. During these successive 1-nt translocation events, the DNA is held in place by the Cas3 protein, resulting in an abrupt 3-bp burst upon release (Supplemental movie 1).

3.4 Discussion

CRISPR interference in the type I systems, relies on the interplay of multiple proteins to convey resistance against invading mobile genetic elements. Based on our results we propose a model for CRISPR interference, were the transacting Cas3 helicase/nuclease remains tightly anchored to the Cascade effector complex while reeling in the invader DNA (Figure 3.5). Together with the sparse nuclease activity of Cas3, this anchoring mechanism acts as a fail-safe to prevent the toxic effects of off-target DNA cleavage. Our data suggests that, Cas3 moves its RecA-like domains in an "inchworm" like fashion, breaking open the dsDNA helix 1-bp at a time (Figure 3.5, inset & Supplemental movie 1). While unwinding the DNA in 1-bp steps, the DNA is held in place by the Cas3 protein until three of such steps have taken place. Unwinding of the third nucleotide triggers the release of the DNA, resulting in a spring-loaded burst that moves the helicase by three base pairs (Supplemental movie 1). Such spring-loaded unwinding has been observed for both helicases with a RecA-like fold (e.g. NS3) [23] and nucleases [31, 32] and presumably reflects a general feature of Cas3 proteins [15]. Finally, our data suggests that Cas3 limits its translocation distance through slipping (Supplemental movie 2), which allows Cas3 to compensate for its low nuclease activity by repeatedly feeds ssDNA into the nuclease domain. We speculate that the joined action of repetitive unwinding by the helicase domain and the intrinsically inefficient DNA degradation by the nuclease domain generates ssDNA fragments of ~90 nt, suitable as new templates for integration by the Cas1-Cas2 spacer integration complex.

3.5 Experimental Procedures

3.5.1 Protein Purification

Cascade was expressed in *E. coli* BL21 (DE3) and purified using strep-tag affinity chromatography, as described previously [5]. Purified Cascade complexes were aliquoted and flash frozen in liquid nitrogen for long-term storage at -80 °C. The nuclease-helicase Cas3 was produced and purified as described previously [14] with the following modifications. BL21-AI cells were used for over-expression, and protein expression was induced with 0.5 mM IPTG and 0.2% L-Arabinose. The purification process was stopped after size exclusion chromatography and before the proteolytic removal of the Maltose Binding Protein (MBP) using the Tobacco Edge Virus protease [13]. MBP-Cas3 was aliquoted and flash frozen in liquid nitrogen before storage at -80 °C.

3.5.2 Cas3 degradation Assays

After purification, Cas3 nuclease activity was initially tested by a non-specific degradation assay on M13mp8 single-stranded circular DNA (Figure S3.1B). Non-specific nuclease activity was stimulated using Ni^{+2} ions as described previously [14]. To test specific degradation plasmid-based assays were performed in the presence of Cascade, ATP, Mg^{+2} and Co^{+2} ions (Figure S3.1C to Figure S3.1E), described previously by [14, 20]. Similar conditions were used for oligo based degradation assays (Figure S3.1H). In brief, 5 nM DNA was incubated with 50 nM Cascade and 100 nM Cas3 in buffer R (+10 μM $CoCl_2$ and 2 mM ATP) for 30 minutes at 37°C. Samples were immediately quenched by adding stop solution (20 mM Tris-HCl pH 8.0, 2% SDS, 50 mM EDTA), after which protein was removed by incubating the samples for 1 hour with 10 μg/ ml proteinase K (Sigma) at 50°C. Subsequently, DNA was precipitated with ethanol and and loaded on 10% denaturing PAGE gels (8M urea) with formamide. Gels were run for 2.5 hour at 350 V, followed by imaging with the Typhoon trio (GE healthcare).

3.5.3 DNA preparation

All the target dsDNA substrates that we used were bearing a protospacer, PAM, and two flanks of 50 and 15 nt (Figure 3.2A, Table S3.1). These synthetic DNA targets (Ella Biotech GmbH) were internally labelled with a monoreactive acceptor dye (Cy5, GE Healthcare) at dT-C6 on the target strand (complementary to the crRNA) and a monoreactive donor dye (Cy3, GE Healthcare) at dT-C6 on the nontarget strand (Figure 3.2A). After labelling, the ssDNA strands were annealed using a thermocycler (Biorad). To determine the initial FRET values of these constructs (Figure S3.2E to Figure S3.2F & Figure S3.4A to Figure S3.4C), Cascade bound DNA was docked on the surface immobilized Cas3 molecules in absence of ATP.

3.5.4 Single-molecule fluorescence data acquisition

The fluorescent label Cy3 and Cy5 were imaged using prism-type total internal reflection microscopy as described previously [6] with slight modifications. After assembly of a microfluidic flow chamber, slides were incubated for 10 minutes with 5% Tween20 to further improve slide quality [33]. Next, the chamber was incubated with 20 µL streptavidin (0.1 mg/ml, S-888, Invitrogen) for 5 minutes followed by a washing step with 100 µL of buffer R. Anti-Maltose Binding Protein (anti-MBP) antibodies (M2155-09P, US biological life sciences) were specifically immobilized through biotin-streptavidin linkage by incubating the chamber with 40 µL of 10 µg/ ml anti-MBP antibodies for 5 minutes. Remaining unbound anti-MBP antibodies were flushed away with 100 µL buffer R. Subsequently, 100 µL of 10 nM Cas3-MBP was incubated in the chamber, allowing the Cas3-MBP molecules to bind the surface immobilized anti-MBP antibodies. After 5 minutes of incubation, unbound Cas3-MBP molecules were flushed away with 100 µL buffer Rimaging (50 mM HEPES (pH 7.5), 60 mM KCl, 10 mM $MgCl_2$, 0.1 mg/mL glucose oxidase (G2133, Sigma), 4 µg/ml Catalase (10106810001, Roche) and 1 mM Trolox (((±)-6-Hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid, 238813, Sigma).

Cascade was incubated with 5 nM labelled dsDNA substrate with 50 nM Cascade for 5 minutes at 37°C. For docking experiments, pre-bound Cascade-DNA complexes were introduced in the chamber with 500 µM ATP and 10 µM Co^{2+} while imaging at room temperature (23 ± 1 °C) and binding events were monitored in real time. For DNA unwinding assays the Cascade-DNA complexes were incubated for 5 minutes, allowing the complexes to form a stable interaction with the surface immobilized Cas3 molecules. Unwinding was initiated by introducing ATP into the chamber while imaging at room temperature (23 ± 1 °C), allowing for visualisation of the dynamics of Cas3 in real time. To visualize the dynamics of Cas3, Cy3 molecules were excited an area of 50 x 50 μm^2 with a 28% of the full laser power (9 mW) green laser (532 nm), while the time resolution was set to 0.1 second. Under these imaging conditions we obtained a high signal-to-noise ratio that allowed us to visualize kinetic intermediates while imaging over time periods of 3.5 min. Under these conditions photobleaching of the donor and acceptor dye during our observation time was minimized.

3.5.5 Single-molecule fluorescence data analysis

A series of CCD images were acquired with laboratory-made software at a time resolution of 0.1 sec. Fluorescence time traces were extracted with an algorithm written in IDL (ITT Visual Information Solutions) that picked fluorescence spots above a threshold with a defined Gaussian profile. The extracted time traces were analysed using custom written MATLAB (MathWorks) and python algorithms. FRET efficiency was defined as the ratio between the acceptor intensity and the sum of the acceptor and donor intensities. The crosstalk between the two detection channels was not corrected to minimize any artefact in using the step-finder algorithm.

For dwell-time ($\Delta\tau$) analysis, the start and end of each unwinding event was determined (Figure 3.3A). The start of each event was marked by an abrupt decrease in the donor signal, whereas the end of each event was marked by an abrupt increase in the donor signal (Figure 3.2B & Figure 3.2C). Selecting the start and end of each event yielded the duration of each event, which was plotted in a histogram. These dwell-time distributions were fitted with a gamma distribution using maximum-likelihood estimations, which yielded an estimation of the number of hidden steps (N) and the rate per step (k). To obtain the global change in the velocity of Cas3 the number of hidden steps (N) and the rate per step (k) were converted to the effective rate ($k_{effective}$, $1/\Delta\tau$). The effective rate ($k_{effective}$, $1/\Delta\tau$) was obtained by dividing the rate per step (k) by the number of steps (N). Next, this effective rate was converted to velocity (bp/ s) by multiplying the effective rate by the FRET range of 22 base pairs (Figure S3.8A & Figure S3.8B). The 95% confidence intervals (errors) of the dwell-times were obtained by empirical bootstrap analysis as described by [34].

The step-size was characterized by adopting an automated step-finder algorithm, described previously by [23, 29]. The step-finder algorithm yielded the average FRET value for each plateau, the size of each step in between the plateaus and the duration/ dwell-time ($\Delta\tau$) of each plateau. To be able to correlate the size of each step in FRET to distance in base pairs, a set of constructs was generated where the distance between donor and acceptor was systematically increased (Figure S3.8A & Figure S3.8B). The slope of this calibration curve yielded a conversion factor, in which a change of $\Delta E=0.05$ corresponds to a distance of one base pair. This allowed direct conversion of the step-size in FRET to distance in base pairs.

The dwell-time distributions for each step were fitted with a gamma distribution using maximum-likelihood estimations (MLE), which yielded an estimation of the number of hidden steps (N) and the rate per step (k). During MLE, each data point is weighted with equal importance. As a consequence, the minor populations in the tail of the distribution are given a substantial amount of priority during minimization of the fit. This causes the fit to widen, which results in an under-estimation of the number of steps and thereby an over-estimation of the rate per step. To correctly interpret data, only the data in the peak of the distribution was fitted, through the use a threshold (Figure S3.9A to Figure S3.9G). Notably, the minor populations in the tail of the distribution may represent stalled helicases or enzymes that have a significantly slower velocity due to static disorder [24].

3.6 Supplementary information

3.6.1 Supplementary figures

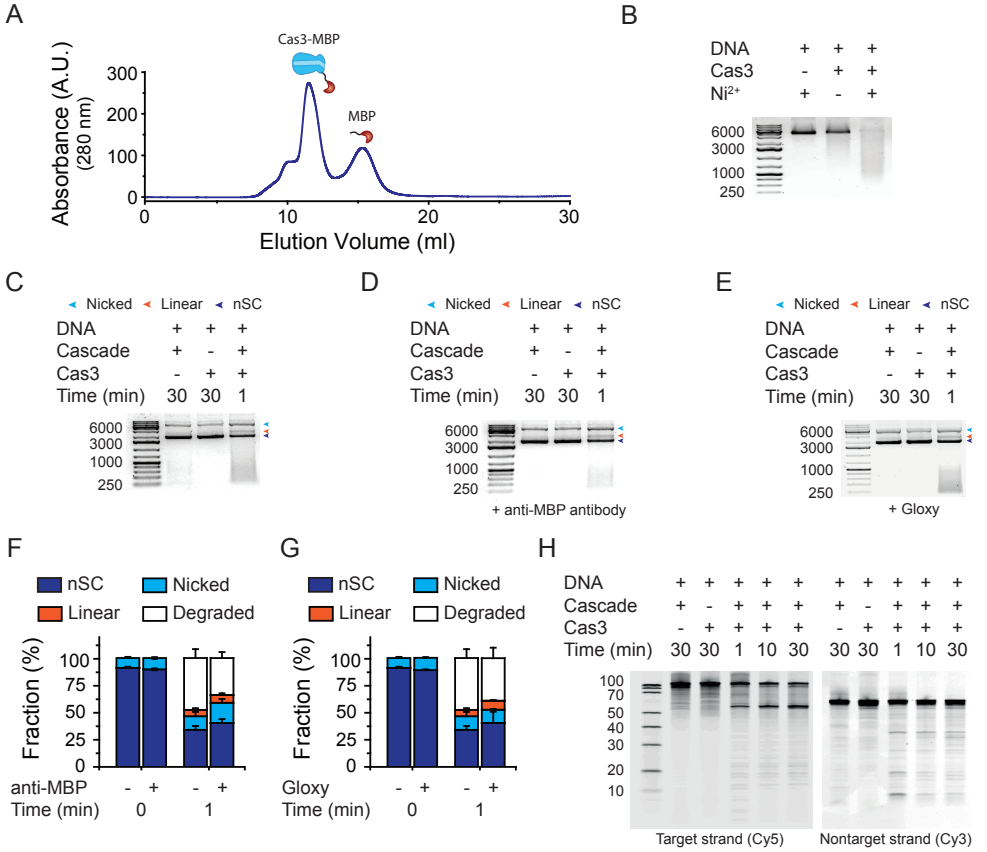


Figure S3.1: Purification and control experiments for Cas3 nuclease activity in bulk

(A) Gel filtration chromatogram of affinity purified Cas3-MBP. The two peaks correspond to Cas3-MBP and MBP. **(B)** Cas3 nuclease activity assay on single-stranded M13mp18 DNA. **(C)** Cas3 plasmid degradation assay. Negatively super-coiled (nSC), Nicked and linear DNA are indicated with the purple, cyan and orange arrows, respectively. **(D)** Cas3 plasmid degradation assay in presence of 600 nM anti-MBP antibody. Negatively super-coiled (nSC), Nicked and linear DNA are indicated with the purple, cyan and orange arrows, respectively. **(E)** Cas3 plasmid degradation assay in presence of 1x oxygen scavenging system. Negatively super-coiled (nSC), Nicked and linear DNA are indicated with the purple, cyan and orange arrows, respectively. **(F)** Comparison of the plasmid degradation efficiency in absence and presence of 600 nM anti-MBP antibody. Negatively super-coiled (nSC), Nicked and linear DNA are indicated with the purple, cyan and cyan bars, respectively. **(G)** Comparison of the plasmid degradation efficiency in absence and presence of 1x oxygen scavenging system. Negatively super-coiled (nSC), Nicked and linear DNA are indicated with the purple, cyan and cyan bars, respectively. **(H)** Cas3 degradation assay on dye-labelled oligonucleotides. The partial duplexed oligonucleotides (see Figure 3.2A) consist of a PAM, protospacer and two flanks of 50 nt and 15 nt. Cy5 was attached to position -7 of the target strand and Cy3 to position -52 of the nontarget strand.

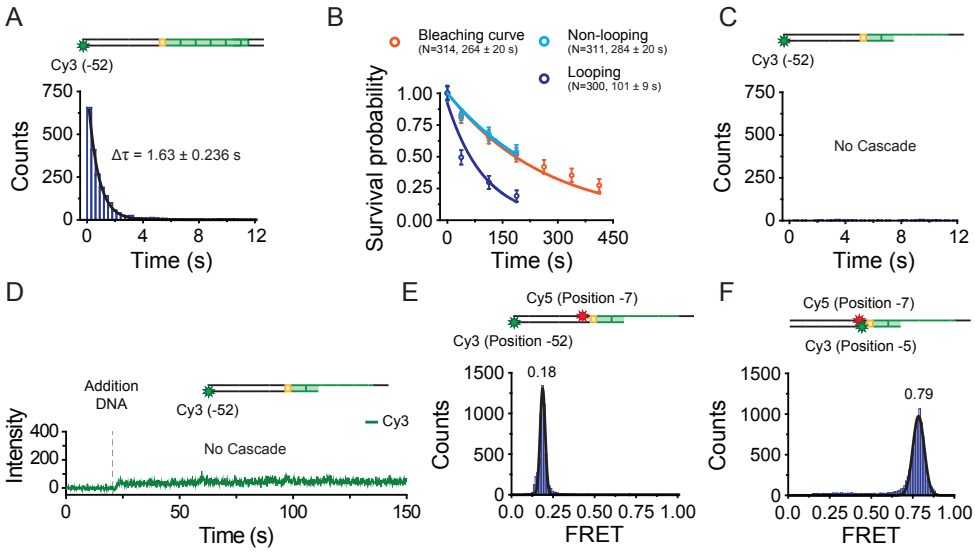


Figure S3.2: Dwell-times, representative traces and initial FRET values

(A) Dwell-time distribution of the interaction between Cas3 and Cascade bound to a cognate target, in absence of ATP and Co^{2+}. Black line indicates a single-exponential fit. Error represents the 95% confidence interval obtained by bootstrap analysis. **(B)** Survival probability of fluorescence signals in absence of ATP (bleaching curve, orange circles), Survival probability of the traces that displayed looping behaviour (purple circles) and the molecules that did not show looping behaviour (cyan circles). The lines represent a single-exponential fit. Error bars indicate the 95% confidence interval obtained by bootstrap analysis. **(C)** Dwell-time distribution of the interaction between Cas3 and a cognate target, in absence of Cascade. **(D)** A representative time trace of the initial interaction between Cas3 and a cognate target. **(E)** Initial FRET efficiency of a construct labelled at position -7 (Cy5, red star) of the target strand and at position -52 (Cy3, green star) of the nontarget strand. **(F)** Initial FRET efficiency of a construct labelled at position -7 (Cy5, red star) of the target strand and at position -5 (Cy3, green star) of the nontarget strand.

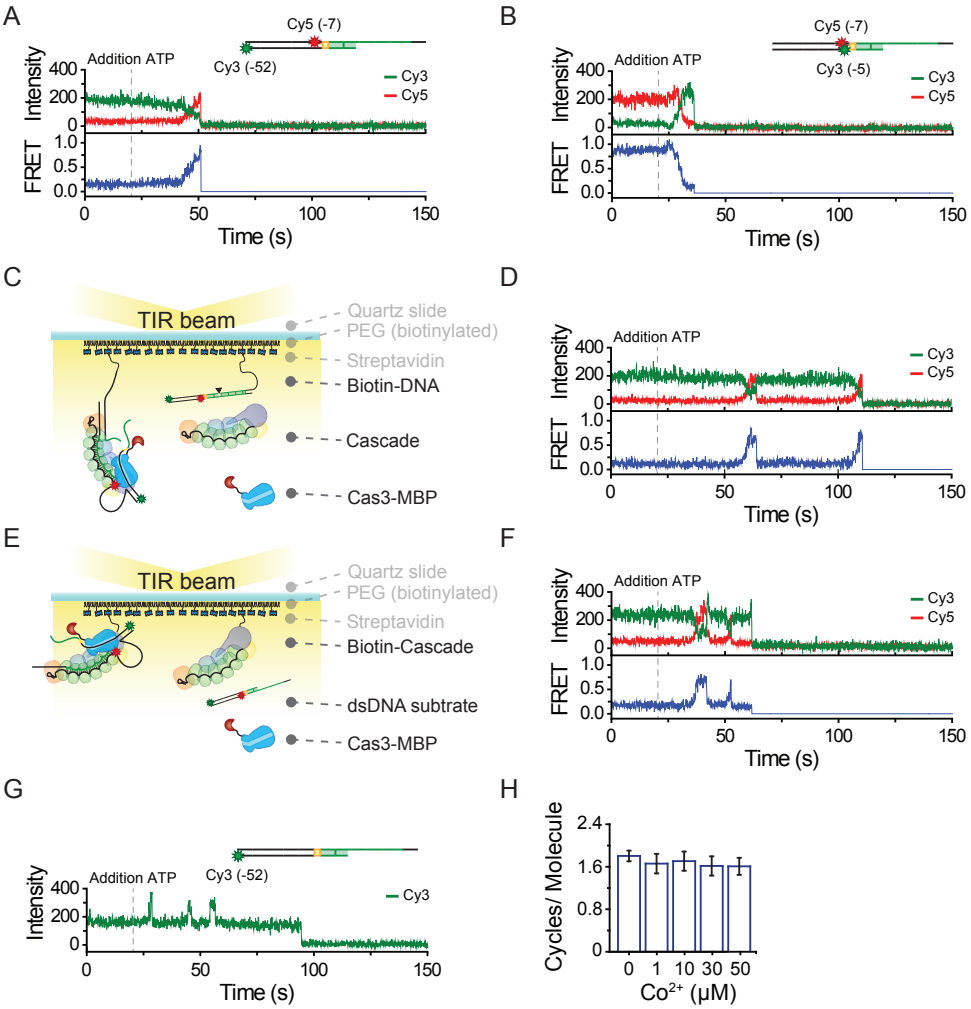


Figure S3.3: Representative traces and alternative immobilisation schemes

(A) A representative time trace of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue) exhibiting a single unwinding event. ATP (2 μM) was added at t = 20s (dashed gray line). **(B)** A representative time trace for a construct with Cy5 (red star) attached to position -7 of the target strand and Cy3 (green star) to position -5 of the nontarget strand. ATP (2 μM) was added at t = 20s (dashed gray line). **(C)** Schematic of an alternative single-molecule FRET assay used to probe the loop formation by Cas3. In brief, DNA was immobilized using biotin-streptavidin conjugation, followed by the addition of 10 nM Cascade. After incubation, Cas3 and ATP were introduced and the dynamics were followed in real-time. **(D)** A representative time trace for the immobilisation scheme depicted in [C]. ATP (2 μM) together with Cas3 (10 nM) were added at t = 20s (dashed gray line). **(E)** Schematic of an alternative single-molecule FRET assay used to probe the loop formation by Cas3. In brief, biotinylated Cascade complexes were immobilized using biotin-streptavidin conjugation, followed by the addition of the DNA substrate. After incubation, Cas3 and ATP were introduced and the dynamics were followed in real-time. **(F)** A representative time trace for the immobilisation scheme depicted in [e]. ATP (2 μM) together with Cas3 (10 nM) were added at t = 20s (dashed gray line). **(G)** A representative time trace of a construct labelled with a donor (Cy3, green) on the tracking strand, displaying PIFE. ATP (2 μM) was added at t = 20s (dashed gray line). **(H)** Quantification of the number of unwinding cycles per molecule at various Co^{2+} concentrations. Error bars represent the SEM (N=3).

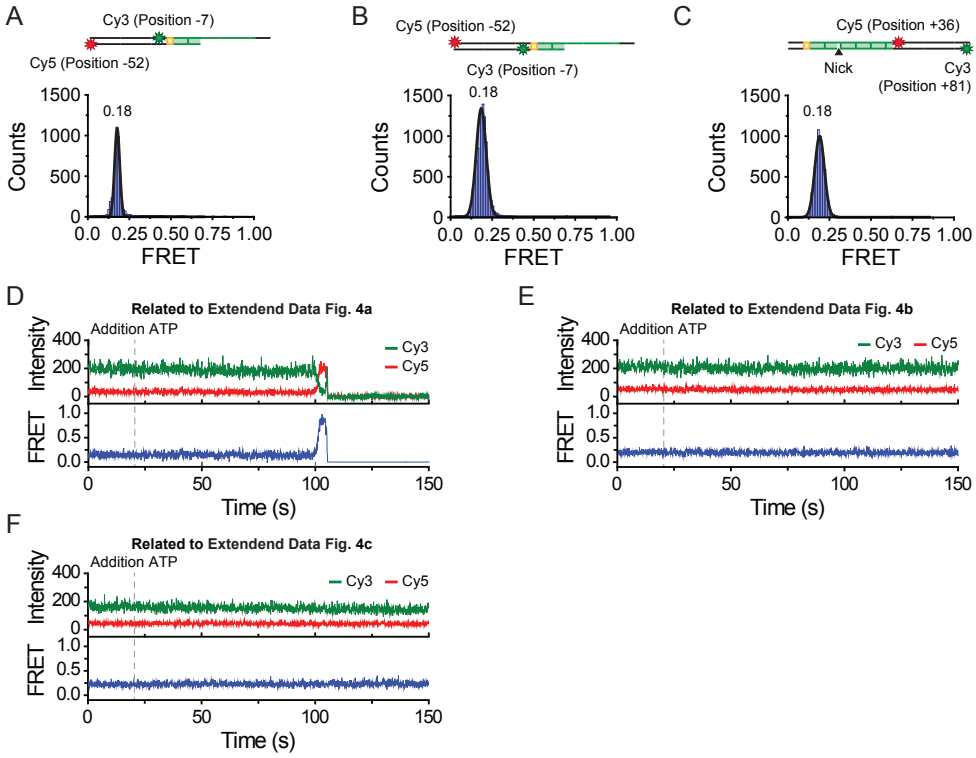


Figure S3.4: Controls for unwinding directionality by Cas3

(A) Initial FRET efficiency of a construct labelled at position -7 (Cy3, green star) of the target strand and at position -52 (Cy5, red star) of the nontarget strand. **(B)** Initial FRET efficiency of a construct labelled at position -52 (Cy5, red star) of the target strand and at position -7 (Cy3, green star) of the nontarget strand. **(C)** Initial FRET efficiency of a construct labelled at position +36 (Cy5, red star) of the target strand and at position +81 (Cy3, green star) of the nontarget strand. Black triangle indicates a nick at position +11. **(D)** A representative time trace for a construct labelled at position -7 (Cy3, green star) of the target strand and at position -52 (Cy5, red star) of the nontarget strand (see Figure S3.4A). ATP (2 μM) was added at t = 20s (dashed gray line). **(E)** A representative time trace for a construct labelled at position -52 (Cy5) of the target strand and at position -7 (Cy3) of the nontarget strand. (Figure S3.4B). ATP (2 μM) was added at t = 20s (dashed gray line). **(F)** A representative time trace for a construct labelled at position +36 (Cy5) of the target strand and at position +81 (Cy3) of the nontarget strand (Figure S3.4C). ATP (2 μM) was added at t = 20s (dashed gray line).

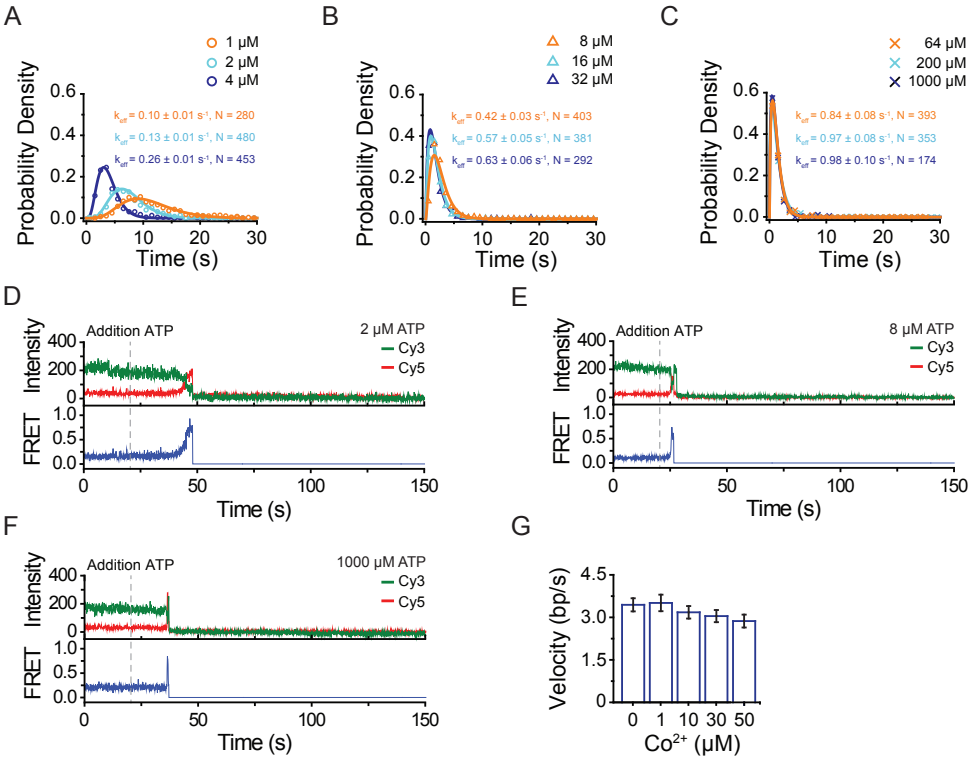


Figure S3.5: Dwell-time distributions and representative traces from ATP and cobalt titration

(A-C) Dwell-time (Δτ) distributions obtained by performing unwinding assays at various ATP concentrations. Data is fitted with a gamma distribution (solid line) to obtain the effective rate (k_{eff}, 1/ Δτ). Error represents the 95% confidence interval obtained through bootstrapping. **(D-E)** Representative time traces of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue) obtained at various ATP concentrations (2, 8, 1000 µM). **(G)** Quantification of the velocity at various Co^{2+} concentrations and 2 µM ATP. Error bars represent 95% confidence intervals obtained through bootstrap analysis.

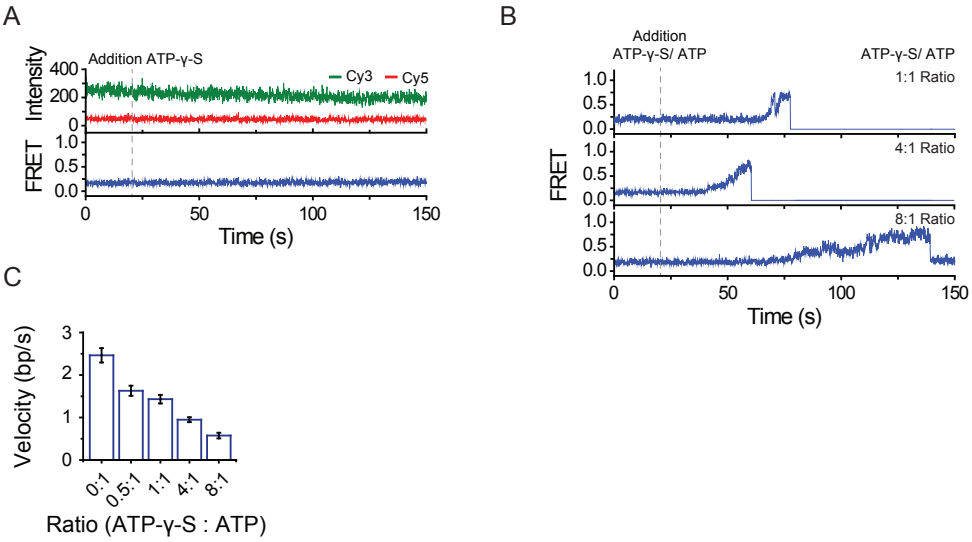


Figure S3.6: Characterisation of Cas3 unwinding activity in the presence of a non-hydrolysable ATP analogue

(A) A representative time trace of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue). ATP-γ-S (2 μM) was added at t = 20s (dashed gray line). **(B)** Representative FRET traces obtained at various ratios of ATP-γ-S and ATP. ATP was kept constant at 2 μM, whereas the concentration ATP-γ-S was increased. **(C)** Velocity of Cas3 at various ratios of ATP-γ-S and ATP.

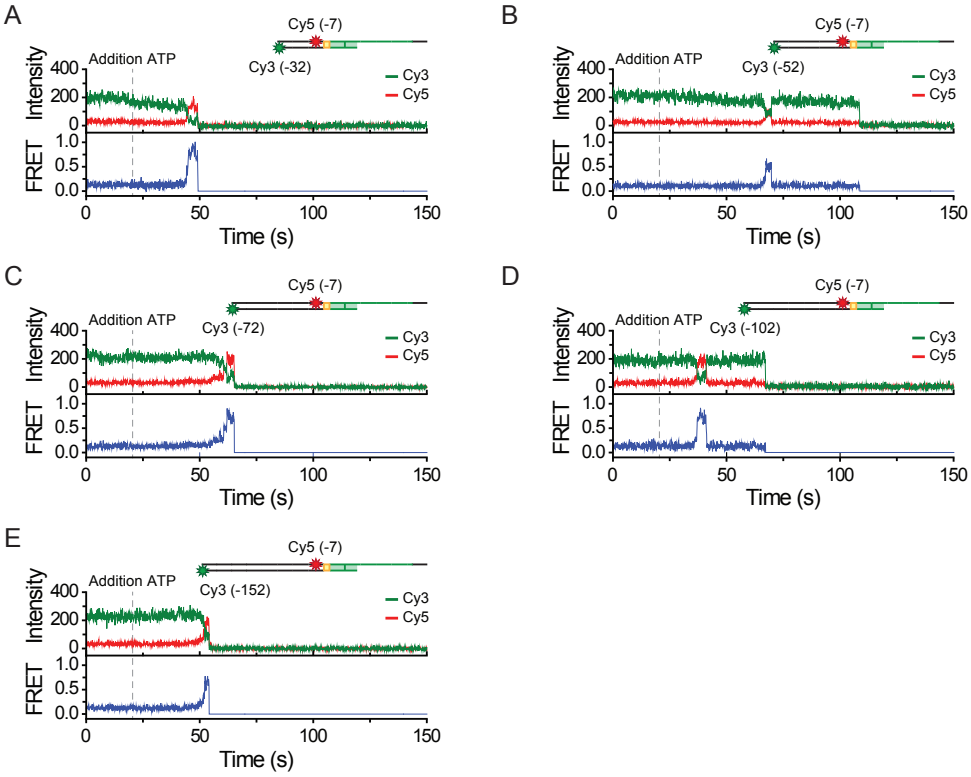


Figure S3.7: Representative traces from the length dependence experiment

(A-E) Representative time traces of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue) obtained from constructs with various flank lengths (30, 50, 70, 100 and 150 bp).

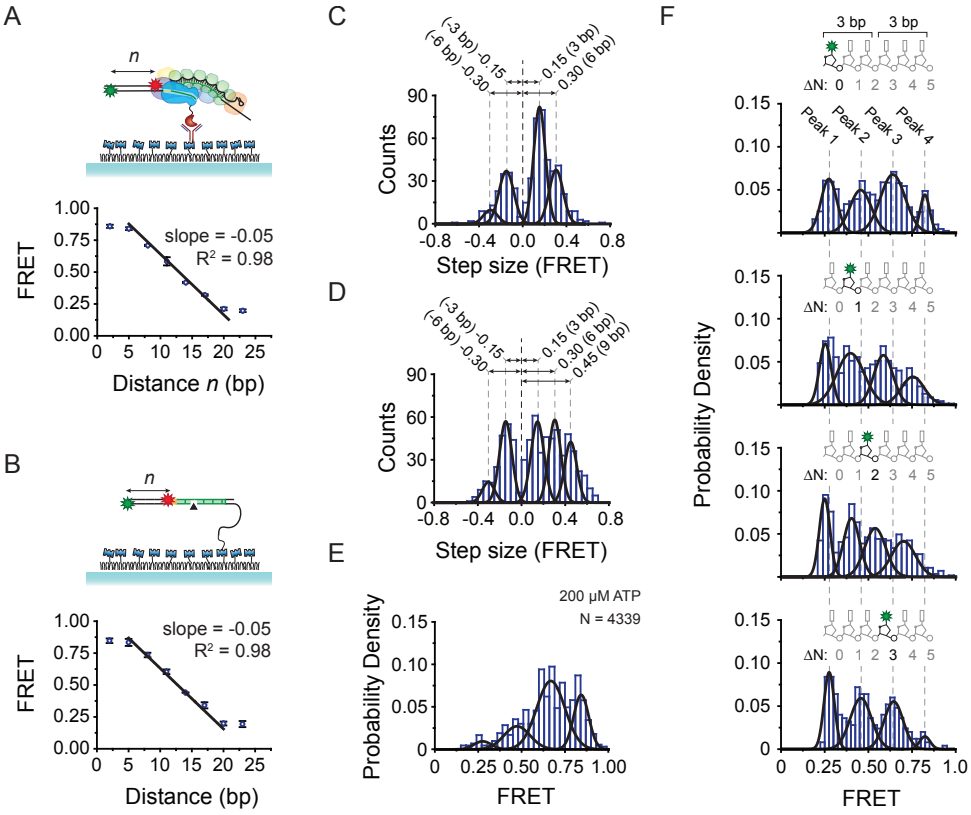


Figure S3.8: Calibration and distributions of the FRET levels and step-sizes obtained through a step-finder algorithm

(A-B) Distance dependence of FRET on double stranded DNA. For this control experiment, a series of DNA constructs were synthesized with different dye labelling positions. DNA was immobilized in either the presence of Cascade and Cas3 [a] or through direct immobilization of the DNA [B]. For both immobilization schemes, the FRET between donor and acceptor show a linear dependence on over a large FRET range. **(C)** Distribution of step-sizes in FRET obtained in the presence of 10 µM ATP through the use of a step-finder algorithm. Black lines represent a Gaussian fit. Dashed grey lines indicate the centre of each peak. Positive values represent processive unwinding whereas negative values represent slipping. **(D)** Distribution of step-sizes in FRET obtained in the presence of 200 µM ATP through the use of a step-finder algorithm. Black lines represent a Gaussian fit. Dashed grey lines indicate the centre of each peak. Positive values represent processive unwinding whereas negative values represent slipping. **(E)** Distribution of FRET levels obtained in the presence of 200 µM ATP through the use of a step-finder algorithm. Black lines represent a Gaussian fit. **(F)** Distribution of the FRET levels for various positions of the donor dye. ΔN=0 indicates the original dye position, whereas ΔN=1, 2 or 3 indicates by how many nucleotides the donor dyes has moved from its original position. Markedly, peak 1 shows a less prominent shift compared to the other peaks (Figure 3.4D), whereas peak 4 broadens for the constructs with a donor dye at position ΔN=1 & ΔN=2. Given that these two peaks are located on the lower and upper boundary of the FRET range, the subtle changes in the peak position and shape reflect the detection limit of FRET.

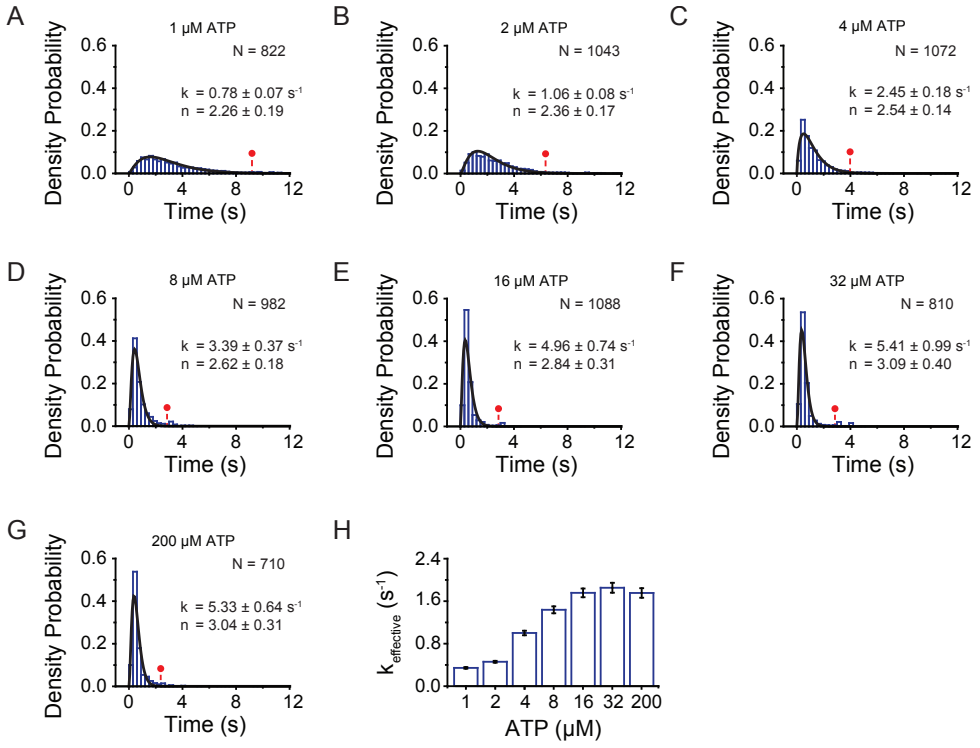


Figure S3.9: Calibration and distributions of the FRET levels and step-sizes obtained through a step-finder algorithm

(A-G) Dwell-time (Δτ) distributions per FRET level at various ATP concentrations. Data is fitted with a gamma distribution (black line) to obtain an estimate of the number of hidden steps (n) and the rate per step (k). The red dashed line indicates the threshold that was used to prevent fitting of the minor populations in the tail of the distribution. Error represents the 95% confidence interval obtained through bootstrap analysis. **(H)** The effective rate (K_{eff}) of individual plateaus at various ATP concentrations. Error represents the 95% confidence interval obtained through bootstrap analysis.

3.6.2 Supplementary tables

Table S3.1: List of used synthetic oligos for this study

Construct (label pos.)	Sequence[a] (5′ → 3′)	Description
TJ3_15_tar_30 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATAG	Substrate used for varying the length of flanking sequence
nTJ3_15_tar_30 (-32)	C/iAmMC6T/ATCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Substrate used for varying the length of flanking sequence
TJ3_15_tar_50 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATTGTAACGACAGTTCTGTTGAAT	Substrate to show loop formation
nTJ3_15_tar_50 (-52)	A/iAmMC6T/TCAACAGAACTGTCGTTACAATCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Substrate to show loop formation
nTJ3_15_tar_50 (-5)	ATTCAACAGAACTGTCGTTACAATCGTCAGCAGAATACGAACGATCAA/iAmMC6T/AATGCCAGTGATAAG	Control substrate to verify loop formation
TJ3_15_tar_70 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGGTCGTAACGACAGTTCTGTTGTTTTATCACTGGTACAATCCAAC	Control substrate to verify loop formation
nTJ3_15_tar_70 (-72)	G/iAmMC6T/TGGATTGTACCAGTGATAATTCAACAGAACTGTCGTTACGACCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Substrate used for varying the length of flanking sequence
TJ3_15_tar_100 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGGTCGTAACGACAGTTCTGTTGTTTTATCACTGGTACAATCCACTGCAACTGACACGATACTGTATCAATAATAG	Substrate used for varying the length of flanking sequence
nTJ3_15_tar_100 (-102)	C/iAmMC6T/ATTATTGATACAGTATCGTGTCAGTTGCAGTGGATTGTACCAGTGATAATTCAACAGAACTGTCGTTACGACCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Substrate used for varying the length of flanking sequence
TJ3_15_tar_150 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGGTCGTAACGACAGTTCTGTTGTTTTATCACTGGTACAATCCACTGCAACTGACACGATACTGTATCAATAATAGGTTCACCAACAGTCGATACTGAATGTCACACAGCACAGACAAATCACAAC	Substrate used for varying the length of flanking sequence

[a] /iAmMC6T/ refers to an amino-modified thymine base at the indicated position.

Construct (label pos.)	Sequence[a] (5′ → 3′)	Description
TJ3_15_tar_30 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATAG	Substrate used for varying the length of flanking sequence
nTJ3_15_tar_150 (-152)	G/iAmMC6T/TGTGATTTGTCTGTGCTGTGTGACATTCAGTATCGACTGTTGGTGAACCTATTATTGATACAGTATCGTGTCAGTTGCAGTGGATTGTACCAGTGATAATTCAACAGAACTGTCGTTACGACCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Substrate used for varying the length of flanking sequence
TJ3_15_tar_50 (-52)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTATAGATCGTTCGTATTCTGCTGACGATTGTAACGACAGTTCTGTTGA/iAmMC6T/T	Control substrate to verify loop formation
nTJ3_15_tar_50 (-7)	AATCAACAGAACTGTCGTTACAATCGTCAGCAGAATACGAACGATC/iAmMC6T/ATAATGCCAGTGATAAG	Control substrate to verify loop formation
TJ3_50_tar_15 (+36)	TAAGTTGTCTTGACAGCAATGTTAGCAGTCGTCTTATGCTTGCTAG/iAmMC6T/TATGACAGCCCACATGGCATTCCACTTATCACTGGCATAGGATGAGAATGAAT	Control substrate to verify directionality of Cas3
nTJ3_50_tar_15 (+81) (I)	TGTGGGCTGTCATAACTAGCAAGTATAAGACGACTGCTAACATTGCTGTCAAGACAACT/iAmMC6T/A	Control substrate to verify directionality of Cas3
nTJ3_50_tar_15 (II)	ATTCATTCTCATCCTATGCCAGTGATAAGTGGAATGCCA	Control substrate to verify directionality of Cas3
TJ3_15_tar_50 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATTGTAACGACAGTTCTGTGAATT	Control substrate to verify 3-bp periodicity
nTJ3_15_tar_50 (-51)	AA/iAmMC6T/TCACAGAACTGTCGTTACAATCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Control substrate to verify 3-bp periodicity
TJ3_15_tar_50 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATTGTAACGACAGTTCTGGAATTT	Control substrate to verify 3-bp periodicity
nTJ3_15_tar_50 (-50)	AAA/iAmMC6T/TCCAGAACTGTCGTTACAATCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Control substrate to verify 3-bp periodicity

[a] /iAmMC6T/ refers to an amino-modified thymine base at the indicated position.

Construct (label pos.)	Sequence[a] (5′ → 3′)	Description
TJ3_15_tar_30 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATAG	Substrate used for varying the length of flanking sequence
TJ3_15_tar_50 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTAT/iAmMC6T/GATCGTTCGTATTCTGCTGACGATTGTAACGACAGTTCTGAAGTTT	Control substrate to verify 3-bp periodicity
nTJ3_15_tar_50 (-49)	AAAC/iAmMC6T/TCAGAACTGTCGTTACAATCGTCAGCAGAATACGAACGATCAATAATGCCAGTGATAAG	Control substrate to verify 3-bp periodicity
TJ3_15_tar_33 (-7)	TAAGTAAGAGTAGGAGACAGCCCACATGGCATTCCACTTATCACTGGCATTA/iAmMC6T/TGAACGATCATAATCAGCAGCAGTATA	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-10)	TATACTGCTGCTGATTATGATCG/iAmMC6T/TCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-13)	TATACTGCTGCTGATTATGA/iAmMC6T/CGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-16)	TATACTGCTGCTGATTA/iAmMC6T/GATCGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-19)	TATACTGCTGCTGA/iAmMC6T/TATGATCGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-22)	TATACTGCTGC/iAmMC6T/GATTATGATCGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-25)	TATACTGC/iAmMC6T/GCTGATTATGATCGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-28)	TATAC/iAmMC6T/GCTGCTGATTATGATCGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
nTJ3_15_tar_33 (-31)	TA/iAmMC6T/ACTGCTGCTGATTATGATCGTTCAATAATGCCAGTGATAAG	Control substrate for FRET based ruler
Biotin_linker	AAAATTGAGCAGACCAAA(PolyT)$_{62}$- Biotin	Biotin linker used for immobilisation

[a] /iAmMC6T/ refers to an amino-modified thymine base at the indicated position.

3.7 References

1 R. Barrangou *et al.*, CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*. **315**, 1709–1712 (2007).

2 L. A. Marraffini, CRISPR-Cas immunity in prokaryotes. *Nature*. **526**, 55–61 (2015).

3 K. S. Makarova *et al.*, An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).

4 S. J. J. Brouns *et al.*, Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*. **340**, 216–219 (2008).

5 M. M. Jore *et al.*, Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* **18**, 529–536 (2011).

6 T. R. Blosser *et al.*, Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex. *Mol. Cell.* **58**, 60–70 (2015).

7 M. Rutkauskas *et al.*, Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep.* **10**, 1534–1543 (2015).

8 R. P. Hayes *et al.*, Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. *Nature*. **530**, 499–503 (2016).

9 B. Wiedenheft *et al.*, Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*. **477**, 486–489 (2011).

10 C. Xue *et al.*, Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity Short Article Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Mol. Cell.* **64**, 1–9 (2016).

11 T. Sinkunas *et al.*, Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).

12 E. R. Westra *et al.*, CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Mol. Cell.* **46**, 595–605 (2012).

13 M. L. Hochstrasser *et al.*, CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6618–23 (2014).

14 S. Mulepati, S. Bailey, In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA Target. *J. Biol. Chem.* **288**, 22184–22192 (2013).

15 R. N. Jackson, M. Lavin, J. Carter, B. Wiedenheft, Fitting CRISPR-associated Cas3 into the Helicase Family Tree. *Curr. Opin. Struct. Biol.* **24**, 106–114 (2014).

16 Y. Huo *et al.*, Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**, 771–7 (2014).

17 B. Gong *et al.*, Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16359–64 (2014).

18 T. Sinkunas *et al.*, In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. *EMBO J.* **32**, 385–394 (2013).

19 S. Redding *et al.*, Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. *Cell.* **163**, 854–865 (2015).

20 T. Künne *et al.*, Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Mol. Cell.* **63**, 1–13 (2016).

21 R. H. J. Staals *et al.*, Interference dominates and amplifies spacer acquisition in a native CRISPR-Cas system. *Nat. Commun.* **23**, 127–135 (2016).

22 S. Myong *et al.*, Cytosolic Viral Sensor RIG-I Is a 5′-Triphosphate–Dependent Translocase on Double-Stranded RNA. *Science.* **323**, 1070–1074 (2009).

23 S. Myong, M. M. Bruno, A. M. Pyle, T. Ha, Spring-Loaded Mechanism of DNA Unwinding by Hepatitis C Virus NS3 Helicase. *Science*, 513–517 (2007).

24 J. Park *et al.*, PcrA helicase dismantles RecA filaments by reeling in DNA in uniform steps. *Cell.* **142**, 544–555 (2010).

25 S. Myong, I. Rasnik, C. Joo, T. M. Lohman, T. Ha, Repetitive shuttling of a motor protein on DNA. *Nature.* **437**, 1321–1325 (2005).

26 M. F. Rollins *et al.*, Cas1 and the Csy complex are opposing regulators of Cas2 / 3 nuclease activity. *Proc. Natl. Acad. Sci.* **1** (2017)

27 S. Dumont *et al.*, RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature.* **439**, 105–108 (2006).

28 D. A. Dixon, S. C. Kowalczykowski, The recombination hotspot: is a regulatory sequence that acts by attenuating the nuclease activity of the E. coli RecBCD enzyme. *Cell.* **73**, 87–96 (1993).

29 J. W. J. Kerssemakers *et al.*, Assembly dynamics of microtubules at molecular resolution. *Nature.* **442**, 709–712 (2006).

30 T. R. Blosser, J. G. Yang, M. D. Stone, G. J. Narlikar, X. Zhuang, Dynamics of nucleosome remodelling by individual ACF complexes. *Nature.* **462**, 1022–1027 (2009).

31 G. Lee, J. Yoo, B. J. Leslie, T. Ha, Single-molecule analysis reveals three phases of DNA degradation by an exonuclease. *Nat. Chem. Biol.* **7**, 367–374 (2011).

32 G. Lee, M. A. Bratkowski, F. Ding, A. Ke, T. Ha, Elastic coupling between RNA degradation and unwinding by an exoribonuclease. *Science.* **336**, 1726–9 (2012).

33 H. Pan, Y. Xia, M. Qin, Y. Cao, W. Wang, A simple procedure to improve the surface passivation for single molecule fluorescence studies. *Phys. Biol.* **12**, 45006 (2015).

34 F. M. Dekking, A Modern Introduction to Probability and Statistics: Understanding why and how. (Springer Science & Business Media, 2005).

4

TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms

EMBO Journal
2015 Jul 2;34(13):1801-15. doi: 10.15252/embj.201590931.

Boseon Kim* [1, 2], Minju Ha* [1, 2], **Luuk Loeff*** [3], Hyeshik Chang [1, 2], Dhirendra K. Simanshu [4], Sisi Li [4], Mohamed Fareh [3], Dinshaw J. Patel [4], Chirlmin Joo** [3], & V. Narry Kim** [1,2]

* These authors have contributed equally to this work

** Co-corresponding authors

[1] Center for RNA Research, Institute for Basic Science, 08826, Seoul, Korea

[2] School of Biological Sciences, Seoul National University, 08826, Seoul, Korea

[3] Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, 2628 CJ, Delft, The Netherlands

[4] Structural Biology Program, Memorial Sloan-Kettering Cancer Center, NY 10065, New York, USA

4.1 Abstract

Terminal uridylyl transferases (TUTs) function as integral regulators of microRNA (miRNA) biogenesis. Using biochemistry, single-molecule, and deep sequencing techniques, we here investigate the mechanism by which human TUT7 (also known as ZCCHC6) recognizes and uridylates precursor miRNAs (pre-miRNAs) in the absence of Lin28. We find that the overhang of a pre-miRNA is the key structural element that is recognized by TUT7 and its paralogues, TUT4 (ZCCHC11) and TUT2 (GLD2/PAPD4). For group II pre-miRNAs, which have a 1 nt 3′ overhang, TUT7 restores the canonical end structure (2 nt 3′ overhang) through mono-uridylation, thereby promoting miRNA biogenesis. For pre-miRNAs where the 3′ end is further recessed into the stem (as in 3′ trimmed pre-miRNAs), TUT7 generates an oligo-U tail that leads to degradation. In contrast to Lin28-stimulated oligo-uridylation, which is processive, a distributive mode is employed by TUT7 for both mono- and oligo-uridylation in the absence of Lin28. The overhang length dictates the frequency (but not duration) of the TUT7-RNA interaction, thus explaining how TUT7 differentiates pre-miRNA species with different overhangs. Our study reveals dual roles and mechanisms of uridylation in repair and removal of defective pre-miRNAs.

4.2 Introduction

MicroRNAs (miRNAs) are generated by multiple maturation steps that consist of two endonucleolytic reactions [1]. First, the nuclear RNase III Drosha cleaves a primary miRNA transcript (pri-miRNA) and releases a ~70 nt hairpin-shaped RNA (pre-miRNA) with a 2 nt 3′ overhang [2]. The pre-miRNA is exported to the cytoplasm by exportin 5 [3–5] and is processed by another RNase III Dicer into a mature miRNA duplex [6–10]. The mature miRNA duplex is loaded onto an Argonaute (Ago) protein to form an effector complex called RNA-induced silencing complex (RISC) [11, 12].

In addition to the canonical miRNA biogenesis pathway, noncanonical cleavage of pre-miRNA has been reported [1, 13]. Pre-miRNAs are often heterogeneous at their 3′ ends, indicating that they are cleaved or trimmed after Drosha processing [14–17]. Ago2 contributes to the production of truncated species by cleaving pre-miRNAs in the middle of the 3′ strand. This results in truncated hairpins called Ago-cleaved pre-miRNAs (ac-pre-miRNAs) [18]. There is little evidence that ac-pre-miRNAs generate mature miRNAs with an exception of ac-pre-miR-451 that is shorter than others and trimmed further into mature miRNA [19–22]. Thus, it remains unclear whether ac-pre-miRNAs have a certain biological role in general or whether they are mostly degradation intermediates. Additional nucleases have been reported to cleave pre-miRNAs [23–25]. However, the molecular mechanism of how the truncated pre-miRNAs are removed is largely unknown.

Accumulating evidence indicates the importance of RNA tailing in the control of RNA stability and function [26–29]. Uridylation is one of the most frequent types of RNA tailing that occurs on diverse RNA species including miRNAs and mRNAs [29–33]. Uridylation is carried out by a group of noncanonical poly(A) polymerases (PAPs), also called terminal uridylyl transferases (TUTases or TUTs), which belong to DNA polymerase β superfamily [34]. TUTs are conserved throughout most eukaryotes [28, 29, 35–37]. Seven TUTs with distinct substrate specificity, localization and functions have been described in humans.

Recent studies have revealed that TUT4 (also known as ZCCHC11), TUT7 (ZCCHC6), and TUT2 (GLD2/PAPD4) play crucial roles in let-7 miRNA biogenesis in mammals [15]. In embryonic stem cells and cancer cells, TUT4 and TUT7 (TUT4/7) have been shown to oligo-uridylate precursors of let-7 family miRNAs in concert with the processivity factor Lin28 [38–42]. The oligo-U tail inhibits pre-miRNA processing by Dicer and promotes degradation by 3′ to 5′ exonuclease DIS3L2 [39, 43, 44]. In contrast, in somatic cells where Lin28 is not expressed, TUT7, TUT4, and TUT2 (TUT7/4/2) mono-uridylate group II pre-miRNAs redundantly to enhance Dicer processing [15]. Unlike prototypical group I pre-miRNAs which have an optimal 2 nt 3′ overhang for Dicer processing, group II pre-miRNAs have a shorter and defective overhang (1 nt 3′) due to a conserved bulge at the Drosha cleavage site. Mono-uridylation by TUT7/4/2 restores the optimal 2 nt 3′ overhang of group II pre-miRNAs resulting in efficient Dicer processing [15]. Between the two contrasting roles that TUTs play in miRNA biogenesis, Lin28-dependent oligo-uridylation by TUT4/7 has been intensively characterized via biochemical and structural studies [38–42, 45–47] whereas mechanism of mono-uridylation has been largely unknown.

In this study, we delineate the molecular mechanism of uridylation of pre-miRNAs with various structures. By mapping out the interactions between TUT7 and pre-miRNA, we show that the overhang of a pre-miRNA is the key structural element that TUT7 recognizes. Sensing the overhang structure, TUT7 preferentially uridylates 3′ truncated pre-miRNAs as well as group II pre-miRNA. Uridylation leads to two opposing consequences. Mono-uridylation of intact group II pre-let-7s (with a 1 nt 3′ overhang) restores functional pre-miRNAs (with a 2 nt 3′ overhang). On the contrary, recognition of pre-miRNAs with 5′ overhang (ac-pre-miRNA or trimmed decay intermediates) leads to oligo-uridylation and RNA degradation. Our single-molecule study further reveals that TUT7 employs a distributive mode for both uridylation pathways, and that TUT7 discriminates its substrates by interacting with them at different frequencies.

4.3 Results

4.3.1 TUT7 domains required for mono-uridylation

To map the interactions between TUTs and pre-miRNA, we first sought to identify the minimal domains used for pre-miRNA recognition and mono-uridylation. TUTs share a common catalytic motif consisting of a nucleotidyl transferase (Ntr) and the poly(A) polymerase-associated (PAP) domain [35, 48]. The Ntr domain contains three catalytic aspartates whereas the PAP domain provides nucleotide specificity through its contact with the base in the active site [35, 49–51]. While TUT2 has only one catalytic motif, TUT7 and TUT4 (TUT7/4) have a duplication of the catalytic motif at their N-terminus although it is inactive due to the lack of one of the catalytic aspartates (Figure S4.1A). Additionally, TUT7/4 possess a CCHH zinc finger motif at their N-terminus and three CCHC zinc finger motifs around the catalytic motif.

We focused on TUT7 as it is the major enzyme for pre-let-7 mono-uridylation (Heo et al, 2012). We generated three deletion mutants of TUT7 by deleting domains from the N-terminus of TUT7 (ΔZF1, ΔNtr1, and ΔPAP1) (Figure 4.1A). In addition, we produced Ntr2-PAP2 (NP) mutant that consists of only the active catalytic motif. The truncated proteins were immunopurified and incubated with unmodified pre-let-7a-1 (with a 1 nt 3′ overhang) or its mono-uridylated counterpart (+U, with a 2 nt 3′ overhang) (Figure 4.1B). The ΔZF1, ΔNtr1, and ΔPAP1 mutants mono-uridylated unmodified pre-let-7a-1 selectively and as efficiently as the full-length (FL) TUT7. However, NP mutant did not show any detectable activity in spite of its higher expression level than that of full-length TUT7 (Figure 4.1B & Figure S4.1B). It seems that NP mutant cannot uridylate RNA substrates, possibly because surrounding regions of catalytic motif are required to bind RNA and/or to maintain proper protein structure. These results indicate that the N-terminal half of TUT7 is dispensable while the C-terminal domains including the catalytic motif are required for pre-miRNA mono-uridylation. As the C-terminal half of TUT7 (ΔPAP1) is fully active, we generated recombinant TUT7 (rTUT7) encompassing 951-1495 a.a (Figure 4.1A) [33]. In vitro uridylation assays demonstrated that rTUT7 can mono-uridylate pre-let-7a-1 and that it has the same substrate preference as the immunopurified full-length TUT7 does (Figure 4.1C). This

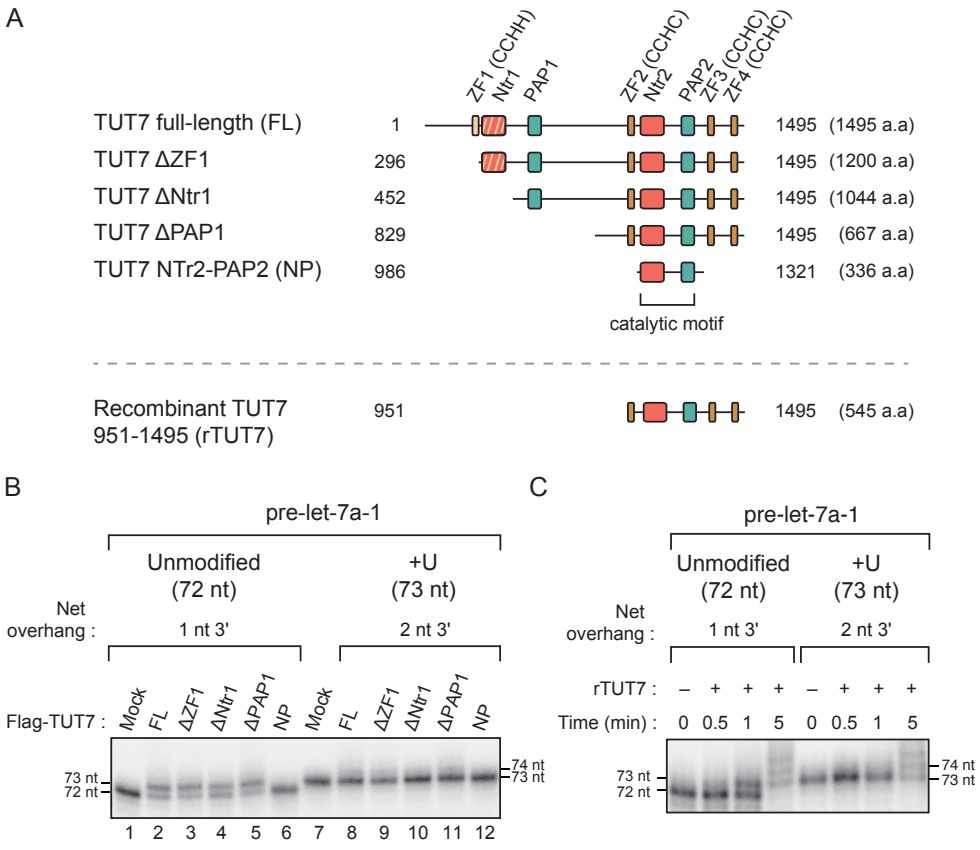


Figure 4.1: C terminal-half of TUT7 is sufficient to mono-uridylate pre-let-7a-1 specifically

(A) Domain organization of full-length (FL), recombinant TUT7 951-1495 (rTUT7) and deletion mutants (ΔZF1, ΔNtr1, ΔPAP1, and NP) of human TUT7. Yellow, CCHH-type zinc finger; hatched red, inactive nucleotidyl transferase domain due to a sequence variation; green, PAP-associated domain; orange, CCHC-type zinc finger; red, nucleotidyl transferase domain. **(B)** In vitro uridylation of unmodified pre-let-7a-1 and mono-uridylated pre-let-7a-1 (+U) by immunopurified full-length TUT7 and deletion mutants (15 min reaction). Deletion mutants except for NP showed mono-uridylation activity and the same substrate preference as that of full-length TUT7; they mono-uridylate unmodified pre-let-7a-1 with a 1 nt 3' overhang more efficiently than pre-let-7a-1 +U with a 2 nt 3' overhang. NP mutant lost its uridylation activity. **(C)** In vitro uridylation of unmodified pre-let-7a-1 and +U by rTUT7. rTUT7 exhibited mono-uridylation activity and the same substrate preference as full-length TUT7.

suggests that the residues 951-1495 are sufficient for mono-uridylation activity by TUT7. Note that prolonged incubation leads to "oligo"-uridylation (Figure 4.1C, lanes 4 and 8) due to multiple rounds of distributive "mono"-uridylation (see 4.4 on page 113 for further explanation). Given that the recombinant protein was produced in *E. coli* and purified to homogeneity, our result indicates that TUT7 does not require any additional cofactors for pre-miRNA mono-uridylation (Figure 4.1B & Figure 4.1C).

4.3.2 RNA motifs that are recognized by TUT7

To investigate which parts of pre-miRNA are recognized by TUT7, we generated mutants of pre-let-7a-1, a pre-miRNA that belongs to group II. Pre-let-7a-1 is divided up into three parts; a 27 nt terminal loop (green), a 21 bp base-paired stem (black), and a 1 nt 3′ overhang (red) (Figure 4.2A, left). First, we designed a terminal loop mutant (L4) by reducing the loop size from 27 nt to 4 nt (Figure 4.2A, center). Immunopurified full-length TUT7 failed to uridylate the L4 mutant efficiently, which suggested that TUT7 recognizes the terminal loop for mono-uridylation (Figure 4.2, lanes 1-4). Next, to test whether a stem of a certain length is necessary, we generated a stem mutant (S14) by shortening the stem from 21 bp to 14 bp (Figure 4.2A, right). The S14 mutant was uridylated as efficiently as the unmodified pre-let-7a-1, indicating that the overall length of the stem is not critical for pre-miRNA mono-uridylation by TUT7 (Figure 4.2A, lanes 1-2 and 5-6).

To find out how the overhang structure influences the mono-uridylation activity of TUT7, we designed six overhang variants of pre-let-7a-1 by shortening nucleotides from the 3′ end (Figure 4.2B). We included ac-pre-let-7a-1 (Ac-pre) with a 10 nt 5′ overhang, which is known to be uridylated in humans and mice [14, 16–18]. The substrates with a blunt end or a 5′ overhang were uridylated with comparable efficiency to (if not more efficiently than) unmodified pre-let-7a-1 by immunopurified full-length TUT7 (Figure 4.2C). To our surprise, RNAs containing a long 5′ overhang (ΔCUUUC and Ac-pre) were strongly oligo-uridylated. Similar results were obtained with rTUT7 951-1495 (Figure S4.2A & Figure S4.2B). These data indicate that TUT7 acts efficiently on 3′ truncated pre-miRNAs in the absence of any cofactor.

As TUT7/4/2 can act redundantly to mono-uridylate group II pre-let-7s (Heo et al, 2012), we also performed in vitro uridylation using immunopurified full-length TUT4 and TUT2 (TUT4/2) to compare their substrate preferences. By and large, TUT7/4/2 are highly similar to each other in specificity but they also displayed some distinct characteristics (Figure 4.2, Figure S4.2C & Figure S4.2D). For example, unlike TUT7, TUT4/2 uridylated the terminal loop mutant (L4) as efficiently as unmodified pre-let-7a-1, indicating that TUT4/2 do not interact with the terminal loop for mono-uridylation (Figure 4.2A & Figure S4.2C). Moreover, while TUT7 and TUT4 showed a strong oligo-uridylation activity on pre-miRNAs with a long 5′ overhang (pre-let-7a-1 ΔCUUUC and Ac-pre), TUT2 did not show such activity (Figure 4.2C & Figure S4.2D). Taken together, the primary cis-acting element recognized commonly by TUT7/4/2 is the overhang structure of pre-miRNA.

4.3.3 Differentiation of pre-miRNAs at the binding step

To further investigate the molecular mechanism by which TUT7 recognizes structural elements of its substrates, we employed single-molecule fluorescence spectroscopy. For long-term single-molecule observations, rTUT7 fused to a 6X-His tag was immobilized on a PEGylated quartz surface using anti-His antibodies (Figure 4.3A). Fluorescently labeled pre-let-7a-1 molecules were introduced to the microfluidic chamber, and the interactions between TUT7 and RNAs were monitored in real time.

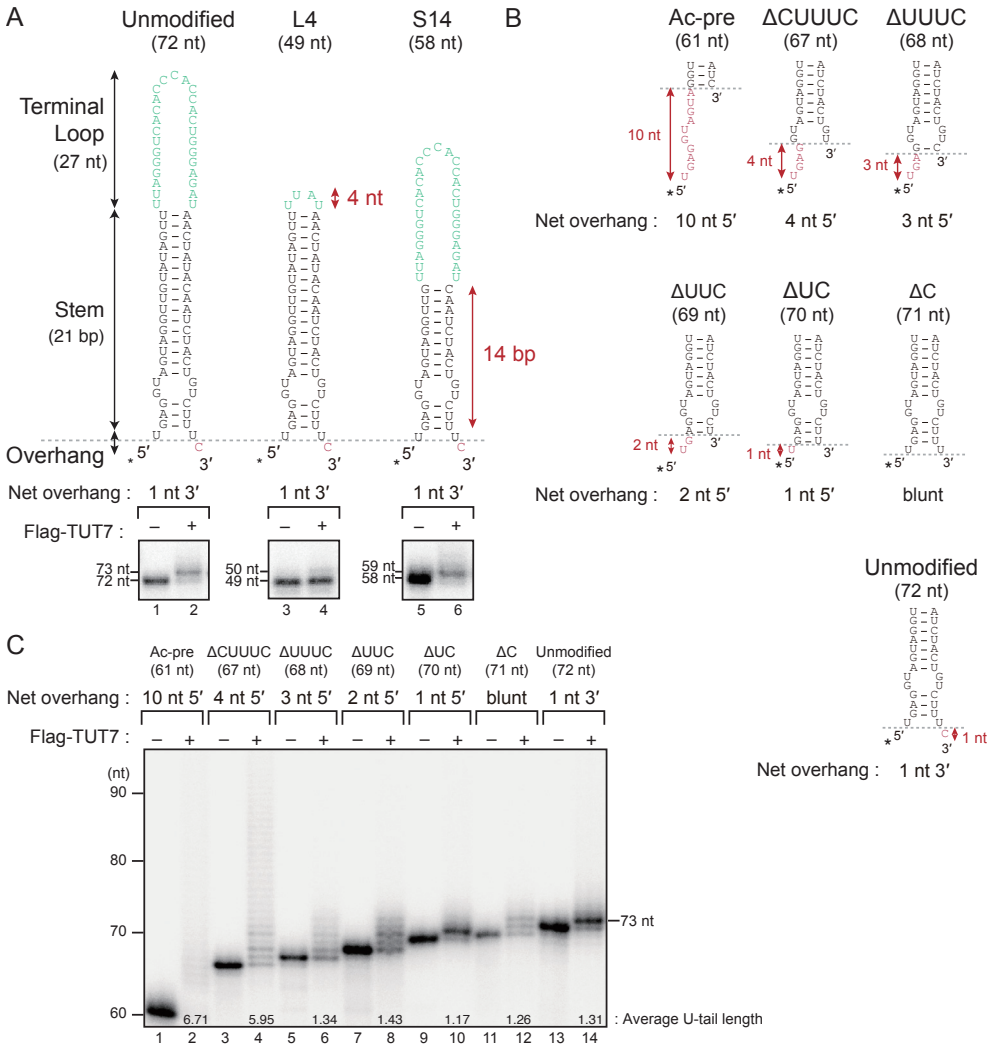


Figure 4.2: TUT7 recognizes overhang and terminal loop of pre-let-7a-1

(A) (top) Structure of unmodified pre-let-7a-1, terminal loop mutant (L4), and stem mutant (S14). Green, terminal loop; black, stem; red, overhang. Asterisks mark radio-labeled terminal phosphates. Red arrows indicate the length of the terminal loop or the stem of pre-let-7a-1 L4 or S14, respectively. (bottom) In vitro uridylation of unmodified pre-let-7a-1 and mutants by immunopurified full-length TUT7 (13 min reaction). Ablating the terminal loop (L4) dramatically reduced the mono-uridylation efficiency while shortening the stem (S14) did not affect the mono-uridylation efficiency. **(B)** Structure of unmodified pre-let-7a-1 and six overhang variants. Red, overhang. Asterisks mark radio-labeled terminal phosphates and red arrows indicate the net length of overhang. **(C)** In vitro uridylation of unmodified pre-let-7a-1 and overhang variants by immunopurified full-length TUT7 (15 min reaction). TUT7 uridylated blunt end or 5′ overhang variants efficiently and showed enhanced activity to the long 5′ overhang variants (ΔCUUUC and Ac-pre). The average length of U-tail is shown below each band. See "4.5 Experimental Procedures" on page 116 for quantification method.

As shown in representative time traces (Figure 4.3B & Figure 4.3C), the interaction of rTUT7 with RNA molecules was marked with a sudden increase and subsequent rapid decrease in the fluorescence intensity. This brief interaction suggests that uridylation by TUT7 is distributive. This is similar to Lin28-independent mono-uridylation by TUT4, which we previously reported to be distributive [42]. Control experiments showed that neither dye-labeling of pre-let-7a-1 nor immobilization of rTUT7 affected its uridylation efficiency (Unpublished observations).

Using this experimental system, we questioned at which kinetic step TUT7 discriminates between different RNA substrates. We first determined the dissociation rate (k_{off}) by analyzing the dwell time of interaction ($\Delta\tau$, the inverse of k_{off}) between TUT7 and unmodified pre-let-7a-1. The dwell time distribution from a total of 8,943 binding events followed a single-exponential decay with a time scale of 0.36 ± 0.03 s (Figure S4.3A). This indicates that dissociation of an RNA substrate from TUT7 is a single-step process. To gain more insights into the molecular mechanism of TUT7, we repeated this measurement for mono-uridylated pre-let-7a-1 (+U) and the terminal loop mutant (L4). Intriguingly, the dwell-time of the +U substrate (0.35 ± 0.05 s, 17,003 events; Figure 4.3D) was similar to that of unmodified substrate within an error whereas that of the terminal loop mutant showed a slight increase (0.54 ± 0.03 s, 8,012 events; Figure 4.3D). These results suggest that the binding strength between TUT7 and RNA (Figure S4.3B) is not a dominant factor in distinguishing between different RNA substrates, although the terminal loop might play a role in the release of the substrate.

Next, we asked whether the binding rate (k_{on}) might govern the substrate preference of TUT7. For this measurement, unmodified pre-let-7a-1 and a variant (e.g. +U) were labeled with spectrally separated fluorescent dyes (Cy3 and Cy5, respectively) and introduced together into a microfluidic chamber. Unmodified pre-let-7a-1 served as a reference. By monitoring the interactions between immobilized rTUT7 and two RNA substrates simultaneously, we were able to compare k_{on} of a variant to that of unmodified pre-let-7a-1 (Figure 4.3A & Figure 4.3C). This frequency measurement revealed that the mono-uridylated substrate (+U) binds less frequently than unmodified substrate does ($k^{+U}{}_{on}/k^{Unmodified}{}_{on}$ = 0.44 ± 0.07, Figure 4.3E), which indicates that the addition of a single uridine suppresses the TUT7-RNA interaction. Stronger suppression was observed with the terminal loop mutant (L4) ($k^{L4}{}_{on}/k^{Unmodified}{}_{on}$ = 0.24 ± 0.01). This is consistent with the decrease in uridylation efficiency observed in Figure 4.2A. Taken together, TUT7 may discriminate between the substrates during the binding step rather than after binding to RNA (Figure S4.3B).

Our biochemical study (Figure 4.2C) indicated that TUT7 oligo-uridylates truncated pre-let-7a-1 effectively under the condition where unmodified pre-let-7a-1 is mono-uridylated. We questioned whether uridylation is changed from a distributive to a processive mode in the presence of the 5′ overhang, or whether the 5′ overhang increases the frequency of the distributive interaction. We repeated our single-molecule kinetic measurements for two pre-let-7a-1 mutants with different 5′ overhang (ΔCUUUC with 4 nt 5′ overhang and Ac-pre with 10 nt 5′ overhang). Intriguingly, the dwell times ($\Delta\tau$ = $1/k_{off}$) of pre-let-7a-1 ΔCUUUC and Ac-pre were within the error comparable to that of

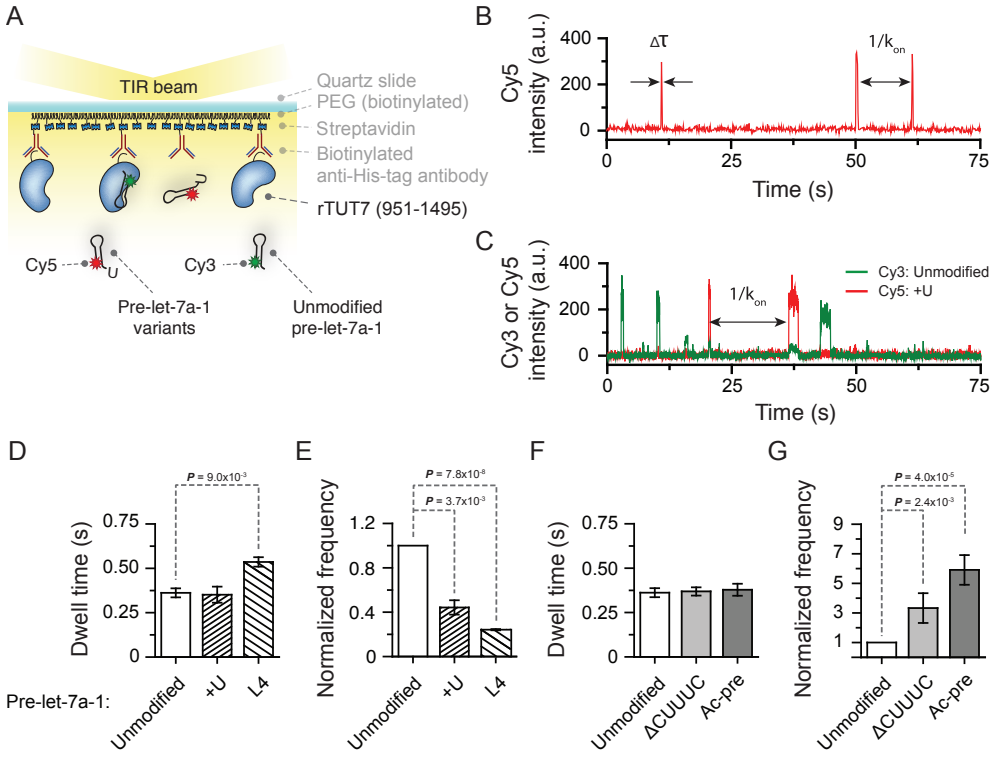


Figure 4.3: TUT7 distinguishes pre-miRNA substrates at binding step

(A) Schematic overview of the single-molecule assay. Recombinant TUT7 951-1495 (rTUT7) fused to a 6X-His were immobilized on a PEGylated surface using anti-His tag antibodies. Afterwards, fluorescently labeled RNA substrates were introduced into the chamber. **(B)** Representative time trajectory for the dwell time analysis. Δτ, dwell time of interaction; k_{on}, binding rate. **(C)** Representative time trajectory for binding frequency measurements. k_{on}, binding rate. **(D)** Average dwell time of pre-let-7a-1 +U and L4 mutants (n=3). Pre-let-7a-1 +U showed similar dwell time to unmodified pre-let-7a-1. Pre-let-7a-1 L4 yielded slightly increased dwell time (P-value $9.0x10^{-3}$, two-tailed t-test). Error bars represent standard error. **(E)** Binding frequency of pre-let-7 +U and L4 mutants relative to unmodified pre-let-7a-1 (n=3). Pre-let-7a-1 +U and L4 mutants showed much lower binding frequency compared to unmodified pre-let-7a-1 (P-values $3.7x10^{-3}$ and $7.8x10^{-8}$, respectively, two-tailed t-test). Error bars represent standard error. **(F)** Average dwell time of pre-let-7a-1 ΔCUUUC and Ac-pre mutants (n=3). Pre-let-7a-1 ΔCUUUC and Ac-pre had similar dwell times compared to unmodified pre-let-7a-1. Error bars represent standard error. **(G)** Binding frequency of pre-let-7 ΔCUUUC and Ac-pre mutants relative to unmodified pre-let-7a-1 (n=3). Pre-let-7a-1 ΔCUUUC and Ac-pre mutants displayed much higher binding frequency compared to unmodified pre-let-7a-1 (P-values $2.4x10^{-3}$ and $4.0x10^{-5}$, respectively, two-tailed t-test). Error bars represent standard error. **(D-G)** All data sets are normally distributed (Shapiro-Wilk test, P>0.1). All the datasets of the binding frequency measurements showed equality in variance (F-test). For the dwell-time measurements some data sets did not show equality in variance (+U and ΔCUUUC) in F-test, we have adjusted our two-tailed t-tests accordingly.

unmodified pre-let-7a-1 (ΔCUUUC, 0.37 ± 0.02 s, 3,495 events; Ac-pre, 0.38 ± 0.03 s, 10,308 events; Figure 4.3F). This suggests that the uridylation mode remains distributive (Figure 4.3F). Thus, in the absence of the processivity factor Lin28, TUT7 exclusively uses a distributive mode for both mono-uridylation and oligo-uridylation.

We next assessed k_{on} and compared the binding frequency of these RNAs with that of unmodified pre-let-7a-1. Indeed, TUT7 interacted with ΔCUUUC and Ac-pre substrates with a higher frequency than with the unmodified substrate ($k^{\Delta CUUUC}_{on}/k^{Unmodified}_{on}$ = 3.33 ± 1.01, $k^{Ac\text{-}pre}_{on}/k^{Unmodified}_{on}$ = 5.91 ± 1.00; Figure 4.3G). Moreover, binding frequency increased as the length of 5′ overhang got longer (from 4 nt to 10 nt). These results collectively hint that TUT7 distinguishes between its substrates at the binding step.

We further validated the distributive mode of TUT by an in vitro uridylation assay with dilution. The reaction mixture with recombinant TUT4 267-1312 (rTUT4, Figure S4.4A & Figure S4.4B) and RNA (unmodified pre-let-7a-1 or ac-pre-let-7a-1) was either not diluted or diluted four times with reaction buffer after 20 seconds, which should lower the uridylation efficiency in case of distributive uridylation. As a control for processive reaction, uridylation assay with rTUT4 and rLin28b was included. Of note, we used rTUT4 (267-1312) instead of rTUT7 (951-1495) because rTUT4 interacts with Lin28 while rTUT7 lacks the first zinc finger motif known to mediate the interaction with Lin28 [41]. We failed to produce soluble full-length rTUT7 protein. In the presence of Lin28b, rTUT4 oligo-uridylated unmodified pre-let-7a-1, which was not affected by dilution (Figure S4.4C, lanes 7-8). This is consistent with our previous single-molecule data that Lin28-mediated oligo-uridylation is a processive reaction [42]. In contrast, oligo-uridylation of ac-pre-let-7a-1 (Figure S4.4C, lanes 1-3) and mono-uridylation of unmodified pre-let-7a-1 (Figure S4.4C, lanes 4-6) was strongly reduced after dilution (Figure S4.4C, lanes 1-6). This result suggests that TUTases act distributively in the absence of Lin28, supporting our conclusion from single molecule measurements (Figure 4.3).

4.3.4 Uridylation of 3′ trimmed pre-miRNAs in cells

It is interesting that TUT7 is capable of oligo-uridylating pre-miRNAs with a blunt end or 5′ overhangs in vitro (Figure 4.2C). To investigate whether the 3′ truncated pre-miRNAs are indeed uridylated by TUT7 in cells, we carried out pre-miRNA deep sequencing in HeLa cells with or without TUT7/4/2 knockdown (Figure 4.4A & Figure S4.5). We depleted TUT7, TUT4, and TUT2 simultaneously due to their redundant activities [15]. Pre-miRNA library was constructed by size fractionation, 3′ adapter ligation, reverse transcription followed by PCR using primers specific to pre-miRNAs (Figure 4.4A). We selected 55 pre-miRNAs whose miRNAs are abundant in HeLa cells and/or those reported to produce ac-pre-miRNAs [14, 16, 18] (Table S4.1).

Knockdown of TUT7/4/2 resulted in a decrease of uridylation in the vast majority of pre-miRNAs, suggesting that TUT7/4/2 can uridylate most pre-miRNAs to some degrees (Figure 4.4B & Table S4.2). Adenylation was not affected significantly, confirming that TUT7/4/2 work mainly as uridylyl transferases on pre-miRNAs (Figure 4.4B & Table S4.2). Note that any terminal residue matching genomic sequence was considered as templated, so the modification rates are underestimated. To observe the effect of TUT7/4/2 knockdown, we first investigated mono-uridylation pattern of group I and group II pre-let-7s (Figure 4.4C). Consistent with the previous study [15], a significant portion (36.2%) of group II pre-let-7s is mono-uridylated in control cells while mono-uridylation decreased to 7.0% upon TUT7/4/2 knockdown. On the

other hand, group I pre-let-7s were rarely mono-uridylated (1.08%) in the control HeLa cells (Figure 4.4C). Notably, trimmed pre-miRNAs accumulated more than 2 fold in both group I and group II pre-let-7s when TUT7/4/2 are depleted, suggesting that TUT7/4/2 may act to facilitate removal of trimmed pre-miRNAs.

To observe trimming and uridylation pattern of pre-let-7s in detail, we drew dot plots which show the fractions of pre-miRNAs that were trimmed of specific length (x-axis) and gained a U-tail of certain size (y-axis) (Figure 4.4D). Although the majority of pre-let-7a-1 (group II) was mono-uridylated in control cells, uridylation reduced dramatically upon TUT7/4/2 knockdown (Figure 4.4D, right). We also observed that shorter pre-let-7a-1 species increased upon TUT7/4/2 knockdown. Pre-let-7e (group I) was not strongly affected by TUT7/4/2, yet we detected a modest decrease of uridylation and an accumulation of trimmed pre-let-7e (Figure 4.4D, left).

Next, we analyzed 54 pre-miRNAs that yielded sufficient reads for analysis (>400 total reads), including pre-let-7s. For most pre-miRNAs, a substantial fraction of reads corresponded to the 3′ truncated fragments (Figure 4.4E, left, 16.7% in control cells and 32.7% in TUT knockdown cells, median). A significant portion of the trimmed fragments was uridylated in control cells (Figure 4.4E, right, 16.3%, median). For some pre-miRNAs (7 of 54 pre-miRNAs), more than 40% of trimmed reads were uridylated (Table S4.3). When TUT7/4/2 were depleted, the uridylation frequency decreased to less than half (Figure 4.4E, right, 6.3%, median). These results indicate that uridylation is not restricted to the let-7 family.

Figure 4.4F presents pre-miR-26a-2 as an example. Nearly 33% of pre-miR-26a-2 reads were recessed from the 3′ end by 10-12 nt. U-tails are found mostly on the recessed pre-miR-26a-2; about 70% of 10 nt trimmed reads carried an oligo-U tail in control cells. In TUT7/4/2-depleted cells, uridylation was reduced (from 70% to 24%) and, at the same time, shorter reads (11 or 12 nt trimmed) accumulated more than twice (12 nt, from 9.2% to 18.7%; 13 nt, from 13.2% to 35.5%). Many other pre-miRNAs including pre-miR-191 showed similar patterns to that of pre-miR-26a-2 (Table S4.3). Thus, our results suggest that TUTases uridylate 3′ trimmed pre-miRNAs in general, which may lead to destabilization of defective pre-miRNAs. Of note, accumulation of trimmed pre-miRNAs upon TUT7/4/2 knockdown does not greatly affect the mature miRNA levels in general [15, 41, 52] because 3′ trimmed pre-miRNAs are defective for Dicer processing.

4.4 Discussion

Our work demonstrates that there are three distinct pathways of pre-miRNA uridylation (Figure 4.5A). (1) In embryonic cells and certain cancer cells, TUT4 (and TUT7, to a lesser extent) associates with Lin28 and oligo-uridylates pre-let-7 specifically. Lin28-mediated oligo-uridylation blocks pre-let-7 processing and promotes degradation by DIS3L2. (2) In the absence of Lin28, TUT7/4/2 mono-uridylate group II pre-miRNAs (with a 1 nt 3′ overhang) which include most of pre-let-7 members. Mono-uridylation of group II pre-miRNAs shapes an optimal 3′ end overhang for efficient processing. (3) In this study, we uncover another pathway in which oligo-U

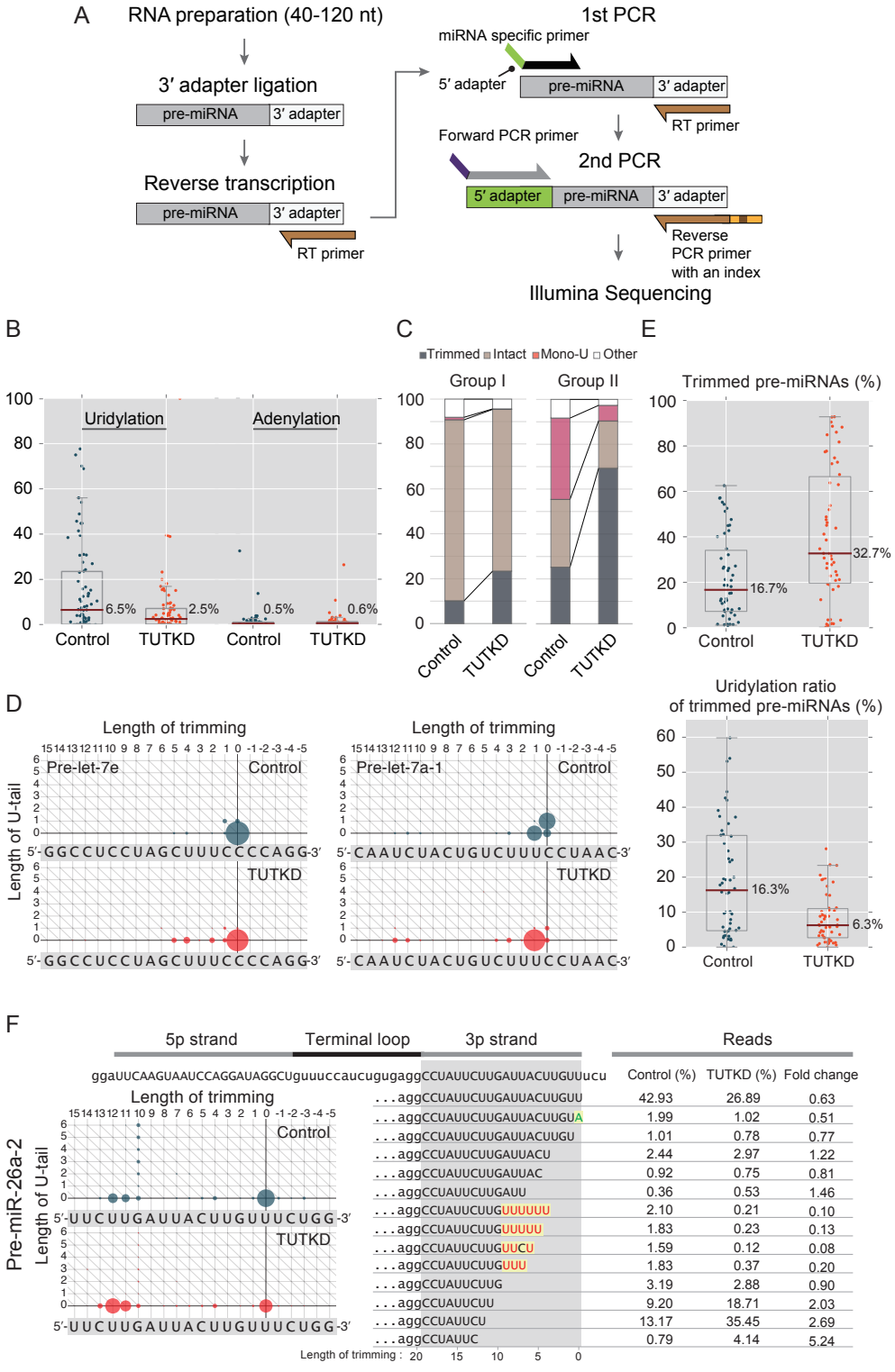

A
RNA preparation (40-120 nt)
3′ adapter ligation
pre-miRNA
3′ adapter
Reverse transcription
RT primer
1st PCR
miRNA specific primer
5′ adapter
Forward PCR primer
2nd PCR
Reverse PCR primer with an index
Illumina Sequencing
B
Uridylation
Adenylation
6.5%
2.5%
0.5%
0.6%
Control
TUTKD
C
Trimmed
Intact
Mono-U
Other
Group I
Group II
E
Trimmed pre-miRNAs (%)
16.7%
32.7%
D
Length of trimming
Pre-let-7e
Pre-let-7a-1
Length of U-tail
5′-GGCCUCCUAGCUUUCCCCAGG-3′
5′-CAAUCUACUGUCUUUCCUAAC-3′
Uridylation ratio of trimmed pre-miRNAs (%)
16.3%
6.3%
F
5p strand
Terminal loop
3p strand
Reads
Pre-miR-26a-2
5′-UUCUUGAUUACUUGUUUCUGG-3′
Control (%)
TUTKD (%)
Fold change
Length of trimming :

Figure 4.4: 3′ trimmed pre-miRNAs are uridylated by TUT7/4/2 *in vivo*

(A) Scheme of pre-miRNA deep sequencing. **(B)** Box plot of uridylation ratio and adenylation ratio of 78 pre-miRNAs which are sequenced in both control sample and TUT7/4/2 knockdown cells. Dark red line indicates the median. TUT7/4/2 knockdown suppressed uridylation specifically. **(C)** Mono-uridylation of group I (let-7a-2, let-7c, and let-7e) and group II (let-7a-1, let-7a-3, let-7b, let-7d, let-7f-1, let-7f-2, let-7g, let-7i, and miR-98) pre-let-7s in control and TUT7/4/2 knockdown (TUTKD) HeLa cells. A considerable portion of intact group II pre-let-7s was mono-uridylated ('Mono-U') in control. Upon TUT7/4/2 knockdown, mono-uridylation of intact group II pre-let-7s substantially decreased. The reads whose length of trimming are >0 were defined as 'Trimmed'. The reads whose length of trimming are = 0 with none non-templated addition were defined as 'Intact'. The reads whose length of trimming are = 0 with one non-templated uridine were defined as 'Mono-U'. The rest reads were defined as 'Other'. The percentage was calculated by normalizing with total read. **(D)** Dot plots of pre-let-7e (group I) and pre-let-7a-1 (group II). The status of 3′ trimming and 3′ U-tailing for each pre-miRNA read is represented by a circle on a two-dimensional matrix. The x-axis represents the length of 3′ trimming and the y-axis represents the length of U-tail. Area of a circle is proportional to the relative abundance of the pre-miRNA reads. Reference sequence of the hairpin is shown below the dot plot. Position 0 indicates the 3′ end of genomic sequence of most abundant read in control HeLa cells. Pre-let-7e was rarely mono-uridylated in control and was not affected by TUT7/4/2 knockdown. In contrast, about half of pre-let-7a-1 was mono-uridylated in control, and the mono-uridylated reads almost disappeared upon TUT7/4/2 knockdown. **(E)** Box plot of trimmed ratio and uridylation ratio of trimmed species of 54 pre-miRNAs which have more than 400 reads. Dark red line indicates the median. Uridylation of trimmed pre-miRNAs decreased while trimmed pre-miRNA reads increased upon TUT7/4/2 knockdown (TUTKD). **(F)** Dot plots and representative reads of pre-miR-26a-2. Representative reads are union of top 10 abundant reads in control or TUT7/4/2 knockdown (TUTKD) sample. Red letter, non-templated uridine; green letter, non-templated adenine; yellow box, non-templated tailing. Proportion (%) of each pre-miRNA species is indicated in each sample and fold change was calculated by dividing the proportion of TUTKD (%) by the proportion of Control (%). Trimmed pre-miR-26a-2 were substantially uridylated. In TUT7/4/2 knockdown cells, trimmed pre-miRNA reads increased.

tails are added by TUT7/4 to truncated pre-miRNAs with a 5′ overhang. TUT2 seems to be less active than TUT7/4 (Figure 4.2C & Figure S4.2D). The oligo-U tails on the trimmed pre-miRNAs may promote rapid degradation of nonfunctional pre-miRNA species.

This study reveals the molecular mechanism of uridylation of group II pre-miRNA (Figure 4.5). Our results indicate that TUT7 distinguishes pre-miRNA substrates at the binding step by recognizing the cis-acting elements. Pre-let-7a-1 with a 2 nt 3′ overhang is not efficiently uridylated due to infrequent binding of RNA to TUT7. Ablation of the terminal loop of pre-let-7a-1 also reduced the binding frequency, indicating that TUT7 recognizes both the overhang and the terminal loop. Given that TUT4/2 do not recognize the terminal loop (Figure S4.2C), TUT7 may be used preferentially to uridylate pre-let-7s while TUT4/2 may have a broader specificity.

We have drawn the energy landscape that explains kinetics of group II pre-miRNA mono-uridylation (Figure S4.6A & Figure S4.6B). RNA binding is represented as a transition from a 'free TUT7' state to 'RNA-bound TUT7' (termed RNA+TUT7). The structural motif of RNA is probed by TUT7 at the transient state, (RNA+TUT7)*. The energy barrier (ΔG, then $\Delta\Delta G = -RT \ln(k^{variant}_{on}/k^{Unmodified}_{on})$) between the 'free TUT7' state and the transient state becomes higher by 2.0 ± 0.5 or by 3.5 ± 0.1 (kJ/mol) if pre-let-7a-1 is mono-uridylated (pre-let-7a-1 +U) or if the terminal loop

is removed from the RNA substrate (pre-let-7a-1 L4), respectively (Figure S4.3B). In contrast, the energy barrier becomes lower when a pre-let-7a-1 variant contains a long 5′ overhang (ΔΔG(ΔCUUUC) = -2.7 ± 0.9 kJ/mol; and ΔΔG(Ac-pre) = -4.3 ± 0.6 kJ/mol) (Figure S4.3B). However, the energy barrier between (RNA+TUT7) and (RNA+TUT7)* does not appear to change as significantly as the energy barrier of binding does. In summary, in the energy landscape of the TUT7-RNA interaction, it is the transient state at which TUT7 probes two cis-acting elements (the terminal loop and the overhang) and discriminates its RNA substrates.

Noncanonical overhang structures of pre-miRNAs (e.g. 1 nt 3′ overhang or 5′ overhang) increase uridylation efficiency. TUT7 oligo-uridylates 3′ trimmed pre-miRNAs at a much higher rate than unmodified pre-let-7a-1 due to enhanced binding (Figure 4.2C, Figure 4.3G & Figure 4.5C). Intriguingly, unlike the processive reaction observed with Lin28-dependent oligo-uridylation of pre-let-7 [42], oligo-uridylation of 3′ trimmed pre-miRNAs results from successive uridylation in a distributive mode (Figure 4.5A and Figure 4.5C). This explains why apparent "oligo"-uridylation is observed even for unmodified pre-let-7a-1 when a large amount of TUT enzyme is used or when reaction time is extended (Figure 4.1C and Figure S4.2). The distributive activity of human TUT7 is consistent with a recent structural study which showed that Cid1, homolog of TUT7 in S. pombe, uridylates single-stranded RNA in a distributive manner [53].

Deep sequencing results suggest that TUT7 (and its paralogues) uridylates 3′ trimmed pre-miRNAs in vivo as well as in vitro, and that the uridylation is likely to induce degradation of the 3′ trimmed pre-miRNAs. This result is consistent with a recent finding by Mourelatos and colleagues that defective Ago-bound pre-miRNAs are uridylated by TUT7/4 and degraded by exosome (DIS3 and RRP6) in mouse embryonic fibroblasts [52]. They also reported that TUT7/4 associate with exosome and this interaction may facilitate degradation of pre-miRNA. Our study confirms and expands the role of uridylation in removal of defective pre-miRNAs, and further provides with mechanistic insights into the differential uridylation by TUTs. Moreover, our work explains their intrinsic preference for trimmed pre-miRNAs at the molecular level. Sensing the overhang structure, TUTs can employ multiple modes of action and thereby have versatile consequences of either repairing or removing pre-miRNAs depending on the molecular and cellular contexts.

4.5 Experimental Procedures

4.5.1 Cell culture and transfection

HeLa and HEK293T (mycoplasma-free) cells were maintained in DMEM (Welgene) supplemented with 9% fetal bovine serum (Welgene). For RNAi, HeLa cells were transfected with 42 nM of siRNA by using Lipofectamine 2000 (Life Technologies). For simultaneous knockdown of three TUTs, equal amounts (14 nM) of siTUT7, siTUT4, and siTUT2 were combined. Transfection was performed two times over 4 days. For ectopic expression of proteins, HEK293T cells were transfected with plasmids by calcium-phosphate method. The sequences of siRNA are listed in Table S4.4

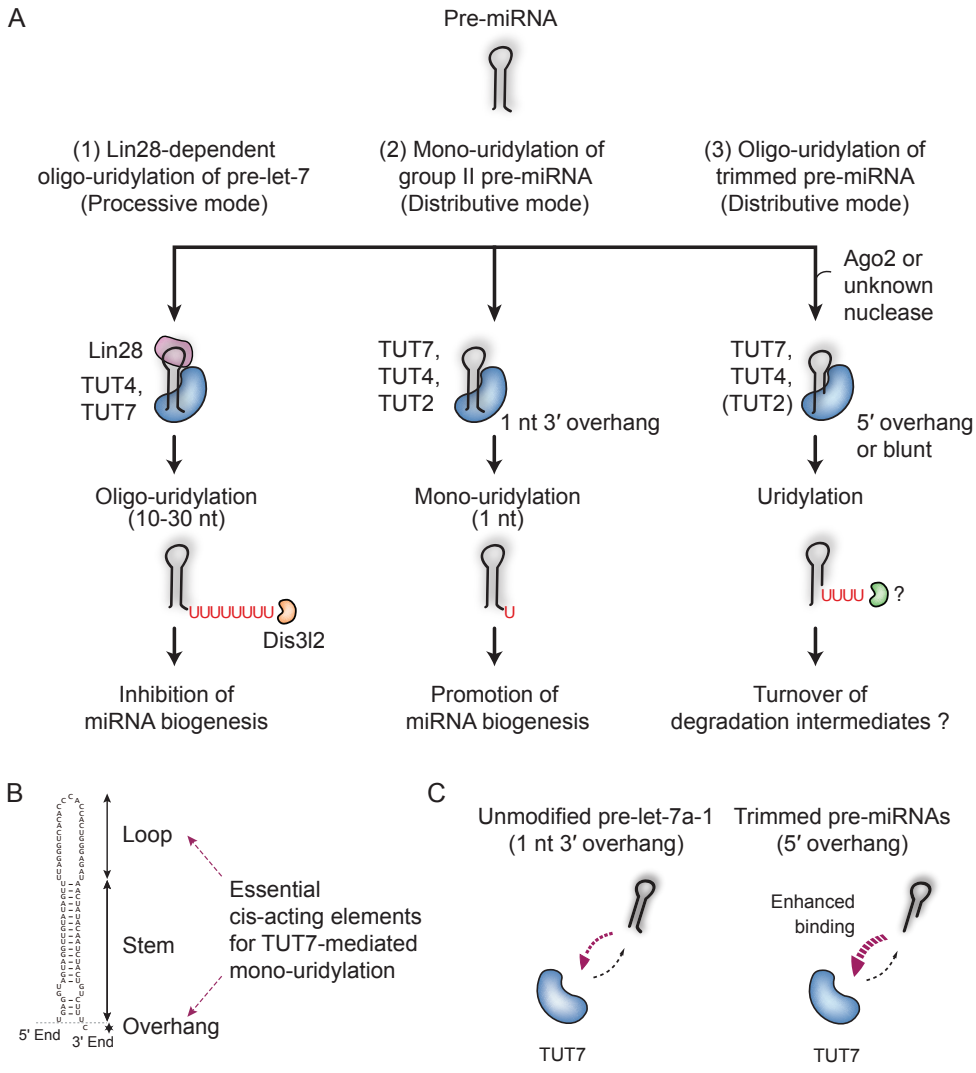


Figure 4.5: Models for pre-miRNA uridylation by TUT7

(A) Pre-miRNA uridylation by TUT7/4/2 in microRNA biogenesis pathway. (1) In embryonic cells and certain cancer cells, Lin28 recruits TUT4 (and TUT7, to a lesser extent) to pre-let-7 for oligo-uridylation. This reaction is processive. Oligo-uridylated pre-let-7 cannot be processed to mature miRNA and instead degraded by Dis3l2. (2) Group II pre-miRNA with a 1 nt 3′ overhang is mono-uridylated by TUT7/4/2, which creates an optimal end structure for Dicer-mediated processing. TUT7 plays a major role in HeLa cells but TUT4/2 also contribute to mono-uridylation. The reaction is distributive and the interaction between TUT and pre-let-7 is infrequent. (3) Trimmed pre-miRNAs are uridylated by TUT7 and TUT4 which promote their degradation. TUT2 seems to be less active than TUT7/4. As the interaction between enzyme and substrate is frequent, multiple cycles of distributive uridylation result in oligo-uridylation. **(B)** Cis-acting elements of pre-let-7a-1 for mono-uridylation by TUT7. The overhang and the terminal loop are recognized by the C-terminal half of TUT7. **(C)** Model of interaction between TUT7 and its RNA substrates. For 3′ trimmed pre-miRNAs, binding rate to TUT7 increased, resulting in enhanced uridylation activity.

4.5.2 Mutagenesis of TUT7

To prepare TUT7 deletion mutants, PCR products of TUT7 deletion mutants were subcloned into FLAG-pCK vector for expression in human cells, at the BamHI and NotI sites. Primer sequences used for PCR are as follows. For ΔZF1, 5′-GGCATTGC-CATTGACAAAGTGGTAC-3′ (forward) and 5′-TCATGATTCCTGCTGGGTCCTC-3′(reverse) were used. For ΔNTr1, 5′-CCTGAAGAAGGAGGTCTGCCACC-3′ (forward) and 5′-TCATGATTCCTGCTGGGTCCTC-3′ (reverse) were used. For ΔPAP1, 5′-CACTT-TACCCACTCAGTACAGGGCC-3′ (forward) and 5′-TCATGATTCCTGCTGGGTCCTC-3′ (reverse) were used. For NTr2-PAP2, 5′-CAGCTAGAACCTCTGCCACCATTAAC-3′ (forward) and 5′-GTCCTTTGGAAATCCCTTGACAGG-3′ (reverse) were used.

4.5.3 Immunoprecipitation and in vitro uridylation

For immunoprecipitation of FLAG-TUTases, HEK293T cells grown on 15 cm dishes were collected 48hr after transfection of FLAG-TUTase expression plasmids. The cells were incubated with buffer D (200 mM KCl, 20 mM Tris [pH 8.0], 0.2 mM EDTA) containing protease inhibitor for 20 min followed by sonication on ice and centrifugation for 30 min at 4 ˚C. The supernatant was incubated with 10 ul of anti-FLAG antibody-conjugated agarose beads (anti-FLAG M2 affinity gel, Sigma) with constant rotation for 2 hr at 4 ˚C. The beads were washed three times with buffer D. The reaction was performed in a total volume of 15 ul in 3.2 mM $MgCl_2$, 1 mM DTT, 0.25 mM UTP, 20 U RNasin® Ribonuclease Inhibitor (Promega), 5′ end labeled RNA of 0.2 nM, and 7.5 ul of immunopurified TUTases in buffer D. When uridylation assay was done with recombinant TUT7 (951-1495 a.a), 6.7 nM of rTUT7 was used in Figure 4.1C and Table S4.2A; 13.4nM of rTUT7 was used in Table S4.2B. The reaction mixture was incubated at 37 ˚C for 30 sec - 20 min. For in vitro uridylation assay with dilution experiment (Figure S4.4C), 26.7nM of recombinant TUT4 (rTUT4, 267-1312 a.a) and 53.6 nM of recombinant Lin28b (rLin28b) were used. The reaction mixture was diluted four times with prewarmed reaction buffer after 20 seconds. The RNA was purified from the reaction mixture by phenol extraction and analyzed on 6% urea polyacrylamide sequencing gel with 7M urea (20x40 cm, 0.4 mm thick). The gel was directly exposed to Phosphor Imaging Plate (Fujifilm) and was read with the Typhoon FLA 7000. Unmodified pre-let-7a-1 and variants were synthesized by ST pharm. The pre-miRNAs were labeled at the 5′ end with T4 polynucleotide kinase (Takara) and (γ-^{32}P) ATP. The sequences of pre-miRNAs are listed in Table S4.4.

4.5.4 Quantification of in vitro uridylation data

In vitro uridylation data are quantified as described in (Lim et al, 2014). The signal intensity profiles (20 pixels/mm) were calculated from the whole blot phosphorimages using Fujifilm MultiGauge v3.0. For each lane, background signal is estimated using the arithmetic mean of the 25th and the 50th percentiles of the signal intensities. The signal intensities were subtracted by the estimated background level, then clipped to zero so that all intensities have zero or positive values. For the alignment of size marker bands, the signals from a marker lane were transformed to the first

and second derivatives using Savitzky-Golay filter (window=31 pixels, order=3). The marker positions were detected by searching points where the sign of first derivative turns from positive to negative, and the second derivative is smaller than −100. The detected positions of marker bands were verified by visual inspection. The function between physical position in the gel and RNA size was defined using cubic spline interpolation. The density of RNA amount in size space was calculated using the first order discrete differences of equal-width samples (0.1 nt) from cumulative density of the original intensity values. For the average length of extensions, the position having maximum signal intensity in the 0 min sample is used as a reference position. The average length of extension was derived from an equation,

$$\bar{x} = \frac{\sum_p (8_p - r)\, I_p}{\sum_p I_p}$$

where x is the average length of extension, p is a position in the gel (by 0.1 nt-wide intervals), sp is the RNA size in nucleotides count for position p, r is the reference size of unextended RNA, and Ip is signal intensity for position p. We excluded signals from degraded products (shorter than the reference size by 3 nt) in the calculation of average extension.

4.5.5 Western blotting analysis

Proteins were resolved with 10% SDS-polyacrylamide gels and transferred to Immobilon-P transfer membrane (Millipore). Primary antibodies used were rabbit anti-FLAG (Sigma, F7425, 1:1000).

4.5.6 Purification of recombinant proteins

Recombinant TUT7 951-1495 (rTUT7) and His-rTUT7 were prepared as previously described [33]. For purification of recombinant TUT4 protein, human TUT4 267-1312 was inserted into a self-modified pMAL expression vector which fuses a hexa-His tag plus a Maltose-Binding Protein tag at the N-terminus to the target protein. The plasmid was transformed into *E. coli* BL21(DE3)-RIL strain (Stratagene). The cells were cultured at 37 °C until OD600 reached 1.0, and then the protein expression was induced with 1 mM IPTG at 16 °C overnight. The hexa-His-MBP tagged protein was purified using a HisTrap FF column (GE Healthcare). The tag was cleaved by TEV protease and further removed by a second step HisTrap FF column (GE Healthcare) purification. The target protein was further purified by a Heparin FF column (GE Healthcare) and a Hiload Superdex G200 16/60 column (GE Healthcare). Recombinant Lin28b was prepared as previously described [42].

4.5.7 Sample preparation and RNA labeling for single- molecule measurements

RNA samples were labeled and prepared as previously described [42].

4.5.8 Single-molecule fluorescence microscopy

The fluorescent label Cy3 was imaged using prism-type total internal reflection microscopy, through excitation by a 532nm (Compass 215M-50, Coherent). Cy5 was imaged using a 640 nm solid-state laser (CUBE 640-100C, Coherent). Fluorescence signals from single molecules were collected through a 60x water immersion objective (UPlanSApo, Olympus) with an inverted microscope (IX71, Olympus). Scattering of the 532 nm and 640 nm laser beams was blocked with a 488/532/635 nm notch filter (NF01-488/532/635, Chroma). Subsequently, signals of Cy3 and Cy5 were spatially split with a dichroic mirror (λ_{cutoff} = 645 nm, Chroma) and recorded.

4.5.9 Slide preparation and single-molecule assays

To eliminate non-specific surface adsorption of proteins and nucleic acids to a quartz surface (Finkenbeiner), piranha-etched slides were PEGylated over two rounds of PEGylation as described previously [54]. After assembly of a microfluidic chamber, slides were incubated for 1 minute with 20 µL streptavidin (0.1 mg/ml, S-888, Invitrogen) followed by a washing step with 100 µL of buffer A (12.5 mM Tris-HCl (pH 8.0, AM9855G, Ambion), 150 mM NaCl (AM9760G, Ambion), 1 mM DTT (D9779, Sigma). Anti-6X His tag antibodies were specifically immobilized through biotin-streptavidin linkage by incubating the chamber with 20 µL of 300 nM biotinylated Anti-6X His tag antibodies (ab27025, Abcam). After 5 minutes of incubation, remaining unbound anti-6X His tag antibodies were flushed away with 100 µL buffer A. Next, 30 µL of 200 nM recombinant TUT7 951-1495 fused to a 6X-His (His-rTUT7) was incubated on the slide, allowing the His-rTUT7 molecules to bind the surface immobilized antibodies. After 5 minutes of incubation, unbound His-rTUT7 molecules were flushed away with 100 µL imaging buffer A (0.5x buffer A substituted with, 0.1 mg/mL glucose oxidase (G2133, Sigma), 4 µg/ml Catalase (10106810001, Roche), 1 mM Trolox (((±)-6-Hydroxy-2,5,7,8-tetramethylchromane-2-carboxylic acid, 238813, Sigma) and 0.1 mM UTP (18333-013, Ambion). Next, 1 nM labeled RNA substrate(s) was/ were introduced in the chamber while imaging at room temperature (23 ± 1 °C) to monitor the interaction of TUT7 with RNA in real time.

The frequency measurement requires an accurate ratio between the concentrations of a sample of interest and a reference sample. To account for differences in concentration, we adsorbed RNA molecules to a positively charged surface as follows. KOH etched quartz slides were coated with a layer of positively charged Poly-L-lysine. After 5 minutes of incubation with 20 µL 0.01% Poly-L-lysine (P4707, Sigma), the chamber was washed with 100 µL of buffer A. After washing, two fluorescently labeled RNA substrates (Unmodified-Cy3, Variant-Cy5) were introduced into the microfluidic chamber. After 5 minutes of incubation, the remaining unbound substrate was washed away with 100 µL of imaging buffer A and data was obtained from 20 fields of view. For each construct this procedure was repeated with three individual dilutions on three different slides. Mean number of counts was used to correct the relative binding frequency for concentration.

4.5.10 Single-molecule data acquisition and analysis

A series of CCD images were acquired with lab-made imaging software at a time resolution of 0.03–0.1 sec. Fluorescence time traces were extracted with an algorithm written in IDL (ITT Visual Information Solutions) that picked fluorescence spots above a threshold with a defined gaussian profile. The extracted time traces were analysed using lab-made Matlab algorithms (MathWorks) that selectively picked anticorrelated traces above a defined threshold. These selected traces were further analyzed using a lab-made Matlab algorithm to extract dwell times and the number of binding events per trace. The relative binding frequency plots were generated by correcting the total number of binding events of each construct was divided with the correction factor obtained from the poly-L-lysine experiment, after which the variant was normalized against the unmodified construct.

To measure the binding frequency, Cy3 molecules were simultaneously excited over an area of 50 x 50 μm^2 with 16% of the full laser power of the (4 mW) green laser (532nm) and red laser (640nm), while the time resolution was set to 0.03 seconds. Under these imaging conditions we obtained a high signal-to-noise ratio that facilitated the automated analysis. For dwell time measurements, Cy5 molecules were excited with 8% of the full laser power (4 mW) green laser (640 nm) to minimize photobleaching of the Cy5 dye during our observation time. Meanwhile, the time resolution was set between at 0.1 seconds to collect a large enough number of photons per time bin.

4.5.11 Pre-miRNA library preparation

To prepare pre-miRNA cDNA library, total RNA was separated on 15% urea-PAGE and RNAs of 40-120 nt were gel-purified. Size fractionated RNAs were ligated to 3′ adaptor by using T4 RNA ligase 2, truncated (NEB). The 3′ adaptor-ligated RNA was separated on 12.5% urea-PAGE and RNAs of 60-140 nt were gel-purified. Size fractionated RNAs were reverse transcribed with a RT primer that is complementary to the 3′ adaptor by using SuperScript III (Life Technologies), followed by two-step PCR amplification. cDNA was firstly amplified with the RT primer and miRNA specific forward primers for 10 cycles and secondly amplified for 10 cycles (12 let-7 family primers) or 13 cycles (43 other miRNA primers) with Phusion DNA polymerase (NEB). The sequences of miRNA specific primers are shown in Table S4.1. The cDNA libraries were separated on 6% native polyacrylamide gel and DNAs of 150-225 bp were gel-purified. The library was sequenced on Illumina MiSeq (110 x 59 paired end run) with 50% of the PhiX control library (Illumina, FC-110-3001). All adapters and primers are synthesized by IDT. Oligonucleotide sequences except for miRNA specific forward primers are shown in Table S4.4.

4.5.12 Processing for Pre-miRNA Sequencing

We used miRBase release 21 [55] and the UCSC hg38 genome assembly for the reference sequences of human pre-miRNAs and flanking regions. To reduce misalignments near the ends of miRBase hairpin sequences, we extended the miRNA precursor spans by 10 bp to both ends. The sequences were extended and retrieved from the genome assembly using BEDTools [56] `slop` and `getfasta` commands. The cDNA library was sequenced for 110 cycles with the small RNA workflow in Illumina MiSeq. Sequences were processed using Cutadapt [17] to trim the 5′-most fifteen nucleotides and clip 3′ adapter sequences out. The sequence at the 5′-end was removed because they originate from PCR primers for specific enrichment of pre-miRNAs and often include significant number of mismatches to known pre-miRNA sequences. Short sequences (<15 nt) after trimming and clipping, and sequences without a 3′ adapter part were removed from the further analyses. The remaining sequences were aligned to 10bp-extended miRBase hairpins explained above using BLAT with options "`-noTrimA -tileSize=8 -stepSize=4 -minIdentity=70 -out=pslx`" [57]. From the output alignments, the best alignments among multi-mapped reads were chosen by following criteria in order: maximum matched bases, minimum mismatch, minimum number of gaps in query of alignment, minimum number of gaps in target of alignment, minimum number of gapped bases in query of alignment, minimum number of gapped bases in target of alignment (preferred first). For the selected alignments, all unaligned bases in 3′-ends of local alignments were regarded as non-templated additions. We additionally "rescued" the non-templated additions which were matched to the reference sequence by ambiguity in A/U-rich sites. First, the regions to be re-examined for the rescue were defined as all subsequent bases containing only A or U immediately starting from the 3′-end of a sequenced read (`/[AU]+$/` in the regular expression). Then, all subsequent bases including the first mismatched base in the re-examination region were rescued so as to be regarded as non-templated additions. The source codes, workflows implemented in Snakemake [57], interactive notebooks in IPython Notebook [58] used for the analyses in this study are freely available from https://github.com/hyeshik/bskim-2015-pre-miRNA.

4.5.13 Determination of length of trimming and length of U-tail

For each pre-miRNA, the most frequent 3′ end position of templated portions of reads in the control was considered as the "reference end position". Six hairpins (hsa-mir-16-2, hsa-mir-100, hsa-mir-222, hsa-mir-320a, hsa-mir-1248, and hsa-mir-1291) whose reference end positions were offset by more than 3 nt from the 3′ end of the mature miRNA from 3′ arm of the hairpin defined in the miRBase were removed from the subsequent analyses to exclude artifacts from the statistics. The length of trimming of each read was calculated by subtracting the position of last templated base in the read from the reference end position (positive for "trimmed", negative for "extended" reads). Length of U-tail was defined as number of U residues in the non-templated additions without any other kind of nucleotidyl additions.

4.5.14 Accession number

Sequenced reads have been deposited in the NCBI Gene Expression Omnibus (GEO) database (accession number GSE64482).

4.6 Supplementary information

4.6.1 Supplementary figures

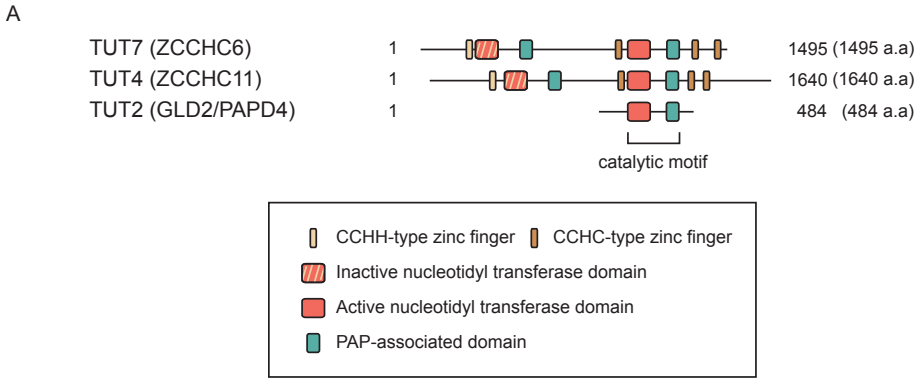


Figure S4.1: Domain organization of TUT7/4/2 and expression of TUT7 deletion mutants

(A) Domain organization of human TUT7, TUT4, and TUT2. Yellow, CCHH-type zinc finger; hatched red, inactive nucleotidyl transferase domain due to a sequence variation; green, PAP-associated domain; orange, CCHC-type zinc finger; red, nucleotidyl transferase domain. **(B)** Western blotting of immunoprecipitated TUT7 full-length (FL) and deletion mutants (ΔZF1, ΔNtr1, ΔPAP1, and NP). Each protein is indicated by red arrowheads. Dashed line indicates discontinuous lanes from the same gel.

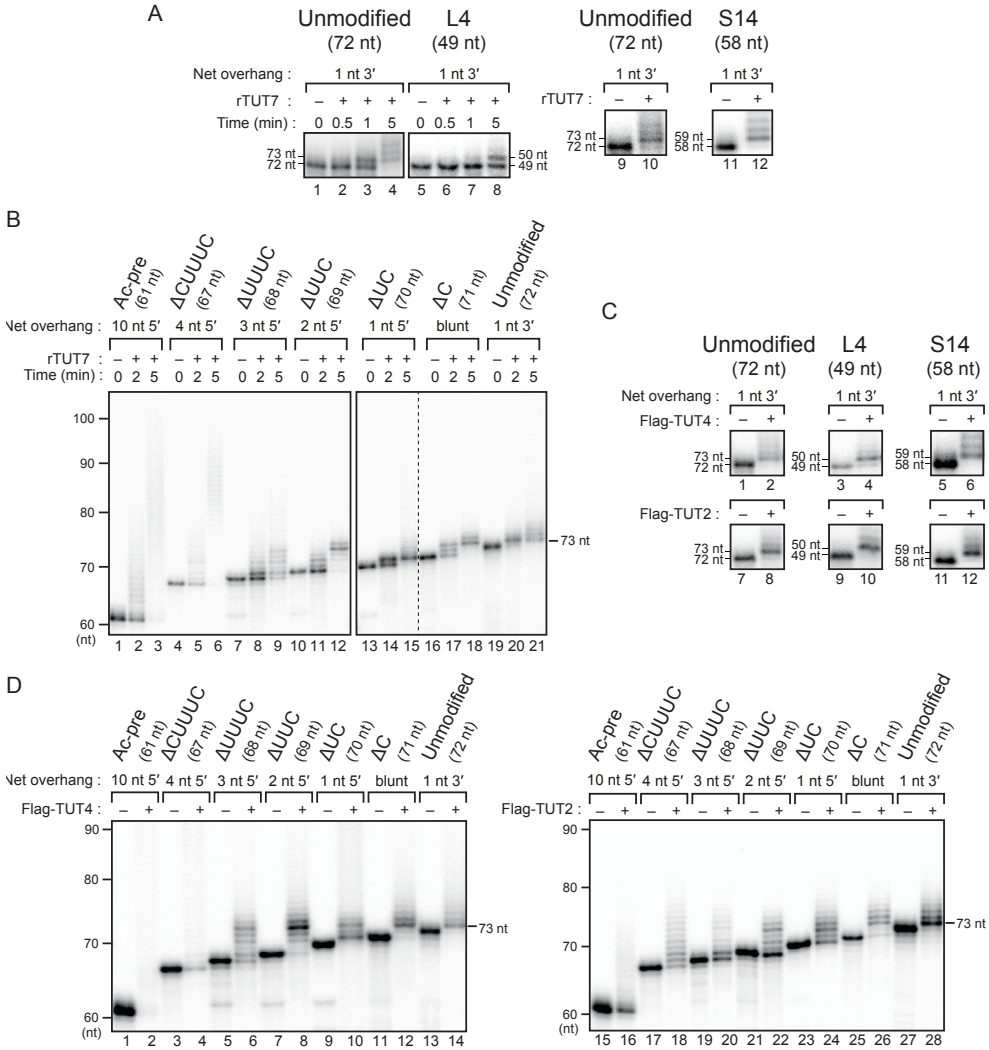
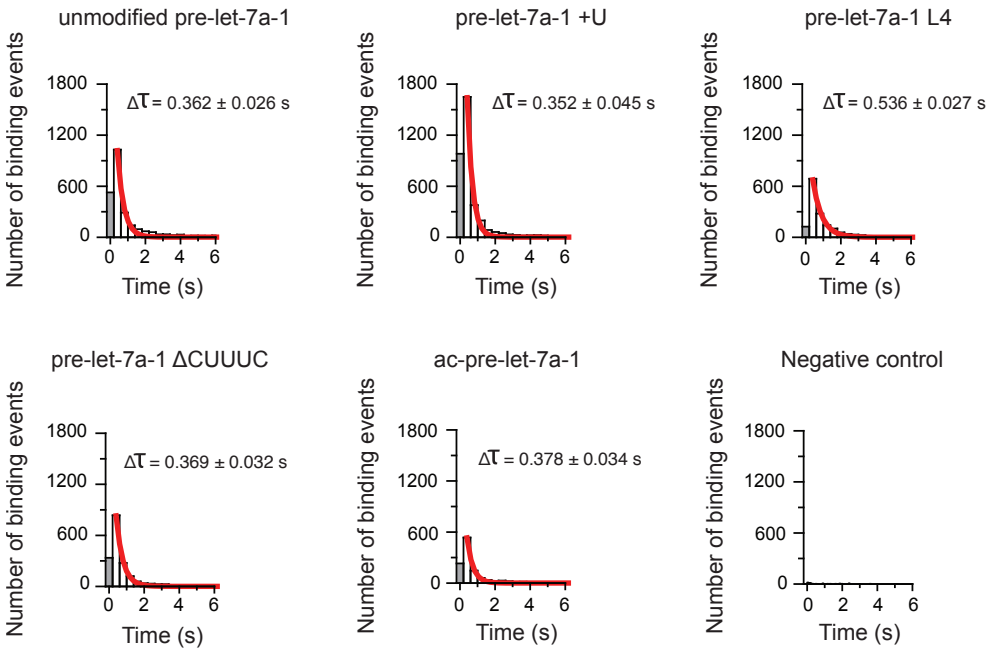


Figure S4.2: In vitro uridylation of pre-let-7a-1 mutants by recombinant TUT7 951-1495, TUT4, and TUT2

(A) In vitro uridylation of unmodified pre-let-7a-1, terminal loop mutant (L4), and stem mutant (S14) by recombinant TUT7 951-1495 (rTUT7) (4 min reaction for lanes 9 - 12). rTUT7 uridylated L4 mutant less efficiently than unmodified pre-let-7a-1 but uridylated S14 mutants as efficiently as unmodified pre-let-7a-1. **(B)** In vitro uridylation of pre-let-7a-1 overhang variants by rTUT7. 13.4 nM of rTUT7 was used. rTUT7 showed the same substrate preference as full-length TUT7. Dashed line indicates discontinuous lanes from the same gel. **(C)** In vitro uridylation of unmodified pre-let-7a-1, terminal loop mutant (L4), and stem mutant (S14) by immunopurified full-length TUT4 and immunopurified full-length TUT2 (20 min reaction for TUT4 and 15 min reaction for TUT2). TUT4 and TUT2, unlike TUT7, uridylated both L4 and S14 mutants efficiently. **(D)** In vitro uridylation of pre-let-7a-1 overhang variants by immunopurified full-length TUT4 and immunopurified full-length TUT2 (20 min reaction for TUT4 and 10 min reaction for TUT2). By and large, TUT7/4/2 showed similar substrate specificity to overhang variants. Unlike TUT7 and TUT4, TUT2 did not show enhanced uridylation to the pre-miRNAs with long 5′ overhangs (ΔCUUUC and Ac-pre).

A



B

	Unmodified	+U	L4	ΔCUUUC	Ac-pre
k_{off}	2.8 ± 0.2	2.9 ± 0.3	1.9 ± 0.1	2.7 ± 0.2	2.7 ± 0.2
$k_{on}^{Variant}/k_{on}^{Unmodified}$	1	0.4 ± 0.1	0.2 ± 0.0	3.3 ± 1.0	5.9 ± 1.0
ΔΔG (kJ/mol)	-	2.0 ± 0.5	3.5 ± 0.1	-2.7 ± 0.9	-4.3 ± 0.5

Figure S4.3: Single-molecule assay of unmodified pre-let-7a-1 and variants

(A) Representative dwell time distributions of unmodified pre-let-7a-1 and variants fitted with a single-exponential decay curve. As all RNA substrates followed single-exponential decay, dissociation of the RNA substrate from TUT7 is a single-step process. The first data point (grey) was not included in the fit due to limited time resolution. Δτ represents average dwell time (n=3) ± standard error. Negative control was performed with unmodified pre-let-7a-1 without recombinant TUT7 protein immobilized. **(B)** Table of dissociation rate (k_{off}), binding rate ($k^{Variant}_{on}/k^{Unmodified}_{on}$) and estimation of ΔΔG of unmodified pre-let-7a-1 and mutants.

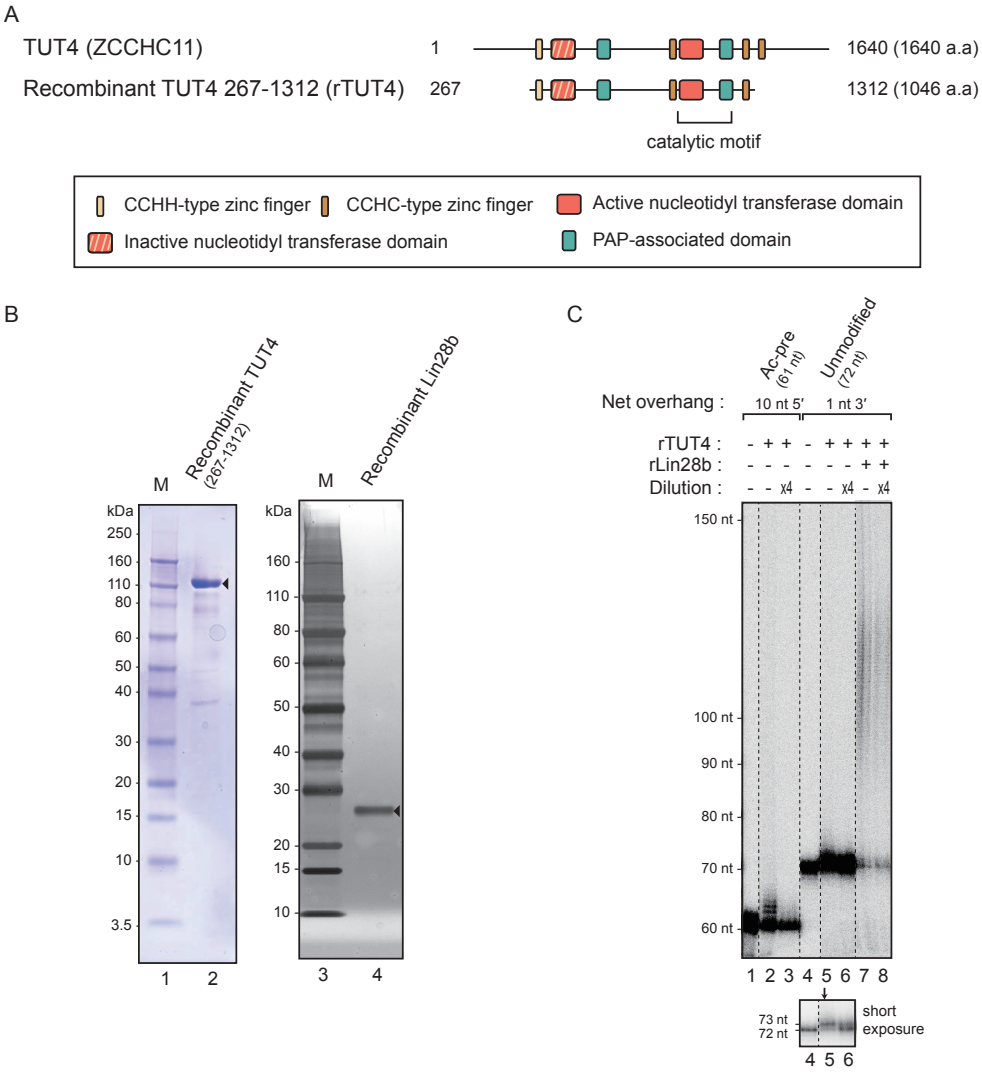


Figure S4.4: In vitro uridylation of pre-let-7a-1 unmodified and ac-pre-let-7a-1 by rTUT4

(A) Domain organization of recombinant protein of human TUT4 267-1312. Yellow, CCHH-type zinc finger; hatched red, inactive nucleotidyl transferase domain due to a sequence variation; green, PAP-associated domain; orange, CCHC-type zinc finger; red, nucleotidyl transferase domain. **(B)** Coomassie staining of recombinant TUT4 267-1312 resolved on NuPAGE® Bis-Tris gel and silver staining of recombinant Lin28b resolved on Bolt® 4-12% Bis-Tris Plus Gel. Each protein is indicated by arrowheads. M, size marker. **(C)** In vitro uridylation assay of ac-pre-let-7a-1 and unmodified pre-let-7a-1 by rTUT4 267-1312 with or without rLin28b. Reaction mixture was either not diluted or diluted after 20 seconds. While processive oligo-uridylation by rTUT4 and rLin28b was not inhibited by dilution, oligo-uridylation of ac-pre-let-7a-1 and mono-uridylation of unmodified pre-let-7a-1 were repressed by dilution, indicating that TUT4 is distributive enzyme. Dashed line indicates discontinuous lanes from the same gel. For over-exposed bands, image with short exposure is presented below.

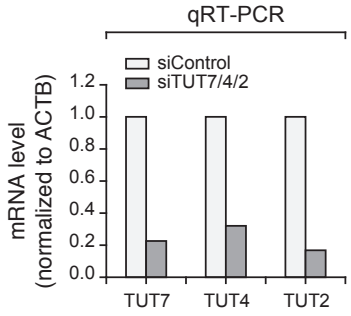


Figure S4.5: TUT7/4/2 knockdown in HeLa cells for pre-miRNA deep sequencing

The mRNA levels of TUT7, TUT4, and TUT2 were measured by qRT-PCR with sequencing samples. The mRNA of all the three TUTs decreased to about 20% upon TUT7/4/2 knockdown.

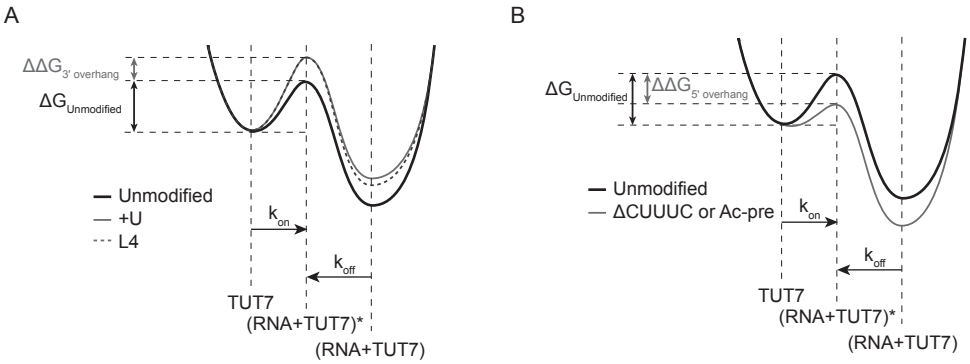


Figure S4.6: Energy landscape of unmodified pre-let-7a-1 and variants

(A, B) Energy landscape of pre-let-7a-1 +U, and L4 (A) and ΔCUUUC, and Ac-pre. **(B)** (RNA+TUT7)* indicates transient state of interaction between TUT7 and RNA substrate. When pre-let-7a-1 is mono-uridylated or when the terminal loop is diminished, the energy barrier increased. On the contrary, when the 3′ end of pre-let-7a-1 is shortened to have 5′ overhang, the energy barrier decreased. TUT7 distinguishes its RNA substrates at transient state.

4.6.2 Supplementary tables

Table S4.1: Pre-miRNA deep sequencing primers

Forward PCR primers used in the 1st PCR step of pre-miRNA deep sequencing. Primer consists of 5′ adapter and miRNA specific region.

Primer name	Sequence (5′ → 3′) 5′ Adapter + miRNA specific region
hsa-let-7a-1	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGGTTGTATAGTTTAGAATT
hsa-let-7a-3	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGGTTGTATAGTTTGG
hsa-let-7b	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGGTTGTGTGGTT
hsa-let-7c	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGGTTGTATGGTTTAGA
hsa-let-7d	GTTCAGAGTTCTACAGTCCGACGATCAGAGGTAGTAGGTTGCATAGTTTTAG
has-let-7e	GTTCAGAGTTCTACAGTCCGACGATCAGGAGGTTGTATAGTTGAGGAGGAC
hsa-let-7f-1	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGATTGTATAGTTGTGG
hsa-let-7f-2	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGATTGTATAGTTTTAGGG
hsa-let-7g	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGTTTGTACAGTTTGAG
hsa-let-7i	GTTCAGAGTTCTACAGTCCGACGATCTGAGGTAGTAGTTTGTACAGTTTGAG
hsa-miR-98	GTTCAGAGTTCTACAGTCCGACGATCGGTAGTAAGTTGTATTGTTGTGGGGTAG
hsa-miR-100	GTTCAGAGTTCTACAGTCCGACGATCAACCCGTAGATCCGAACTT
hsa-miR-103a-1	GTTCAGAGTTCTACAGTCCGACGATCGGCTTCTTTACAGTGCTGC
hsa-miR-103a-2	GTTCAGAGTTCTACAGTCCGACGATCAGCTTCTTTACAGTGCTGCC
hsa-miR-105-1	GTTCAGAGTTCTACAGTCCGACGATCTCAAATGCTCAGACTCCTGT
hsa-miR-106b	GTTCAGAGTTCTACAGTCCGACGATCTAAAGTGCTGACAGTGCAGATAG
hsa-miR-148b	GTTCAGAGTTCTACAGTCCGACGATCAAGTTCTGTTATACACTCAGGCTG

Primer name	Sequence (5′ → 3′) 5′ Adapter + miRNA specific region
hsa-miR-151a	GTTCAGAGTTCTACAGTCCGACGATCTCGAGGAGCTCACAGTCTAGTA
hsa-miR-182	GTTCAGAGTTCTACAGTCCGACGATCTTTGGCAATGGTAGAACTCA
hsa-mir-185	GTTCAGAGTTCTACAGTCCGACGATCTGGAGAGAAAGGCAGTTCC
hsa-miR-16-1	GTTCAGAGTTCTACAGTCCGACGATCTAGCAGCACGTAAATATTGGC
hsa-miR-191	GTTCAGAGTTCTACAGTCCGACGATCCAACGGAATCCCAAAAGC
hsa-miR-20a	GTTCAGAGTTCTACAGTCCGACGATCTAAAGTGCTTATAGTGCAGGTAGTG
hsa-miR-21	GTTCAGAGTTCTACAGTCCGACGATCTAGCTTATCAGACTGATGTTGACTG
hsa-miR-221	GTTCAGAGTTCTACAGTCCGACGATCACCTGGCATACAATGTAGATTTC
hsa-miR-222	GTTCAGAGTTCTACAGTCCGACGATCCTCAGTAGCCAGTGTAGATCCTG
hsa-miR-24-1	GTTCAGAGTTCTACAGTCCGACGATCTGCCTACTGAGCTGATATCAGT
hsa-miR-24-2	GTTCAGAGTTCTACAGTCCGACGATCTGCCTACTGAGCTGAAACAC
hsa-miR-26a-1	GTTCAGAGTTCTACAGTCCGACGATCTTCAAGTAATCCAGGATAGGCT
hsa-miR-27b	GTTCAGAGTTCTACAGTCCGACGATCAGAGCTTAGCTGATTGGTGAA
hsa-miR-30a	GTTCAGAGTTCTACAGTCCGACGATCTGTAAACATCCTCGACTGGA
hsa-miR-30c-1	GTTCAGAGTTCTACAGTCCGACGATCTGTAAACATCCTACACTCTCAGCT
hsa-miR-30d	GTTCAGAGTTCTACAGTCCGACGATCTGTAAACATCCCCGACTG
hsa-miR-320a	GTTCAGAGTTCTACAGTCCGACGATCGCCTTCTCTTCCCGGTT
hsa-miR-378a	GTTCAGAGTTCTACAGTCCGACGATCCTCCTGACTCCAGGTCCTG
hsa-miR-7-1	GTTCAGAGTTCTACAGTCCGACGATCTGGAAGACTAGTGATTTTGTTGTT
hsa-miR-93	GTTCAGAGTTCTACAGTCCGACGATCCAAAGTGCTGTTCGTGCA
hsa-miR-31	GTTCAGAGTTCTACAGTCCGACGATCAGGCAAGATGCTGGCATAGC

Primer name	Sequence (5′ → 3′) 5′ Adapter + miRNA specific region
hsa-miR-101-1	GTTCAGAGTTCTACAGTCCGACGATCCAGTTATCACAGTGCTGATGCT
hsa-miR-345	GTTCAGAGTTCTACAGTCCGACGATCGCTGACTCCTAGTCCAGGGC
hsa-miR-9-2	GTTCAGAGTTCTACAGTCCGACGATCTCTTTGGTTATCTAGCTGTATGAGTG
hsa-miR-18a	GTTCAGAGTTCTACAGTCCGACGATCTAAGGTGCATCTAGTGCAGATAGT
hsa-miR-30b	GTTCAGAGTTCTACAGTCCGACGATCTGTAAACATCCTACACTCAGCTGT
hsa-miR-10b	GTTCAGAGTTCTACAGTCCGACGATCTACCCTGTAGAACCGAATTTGTG
hsa-miR-15a	GTTCAGAGTTCTACAGTCCGACGATCTAGCAGCACATAATGGTTTGTGG
hsa-miR-423	GTTCAGAGTTCTACAGTCCGACGATCTGAGGGGCAGAGAGCGA
hsa-miR-183	GTTCAGAGTTCTACAGTCCGACGATCTATGGCACTGGTAGAATTCACTGT
hsa-miR-196a-2	GTTCAGAGTTCTACAGTCCGACGATCTAGGTAGTTTCATGTTGTTGGGATT
hsa-miR-1226	GTTCAGAGTTCTACAGTCCGACGATCGTGAGGGCATGCAGGCC
hsa-miR-1248	GTTCAGAGTTCTACAGTCCGACGATCACCTTCTTGTATAAGCACTGTGC
hsa-miR-1291	GTTCAGAGTTCTACAGTCCGACGATCTGGCCCTGACTGAAGACCA
hsa-miR-1307	GTTCAGAGTTCTACAGTCCGACGATCTCGACCGGACCTCGACC
hsa-miR-148b	GTTCAGAGTTCTACAGTCCGACGATCAAGTTCTGTTATACACTCAGGCTG
hsa-miR-449b	GTTCAGAGTTCTACAGTCCGACGATCAGGCAGTGTATTGTTAGCTGGT

Table S4.2: Uridylation ratio and adenylation ratio of pre-miRNAs

Control					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-21	867159	7681	0.89	7010	0.81
hsa-let-7f-1	834611	373627	44.77	7688	0.92
hsa-let-7b	238553	94091	39.44	2803	1.18
hsa-let-7a-3	233797	63023	26.96	1674	0.72
hsa-mir-423	230535	1807	0.78	328	0.14
hsa-mir-93	191255	25939	13.56	3523	1.84
hsa-let-7a-1	148857	68029	45.70	1162	0.78
hsa-let-7d	143835	1777	1.24	1444	1.00
hsa-mir-30a	106625	1533	1.44	136	0.13
hsa-mir-30c-2	89507	100	0.11	485	0.54
hsa-let-7e	74236	4853	6.54	314	0.42
hsa-mir-182	69259	86	0.12	649	0.94
hsa-let-7g	67155	52143	77.65	1146	1.71
hsa-mir-98	57305	32109	56.03	573	1.00
hsa-mir-106b	55067	4982	9.05	625	1.13
hsa-let-7f-2	52256	25562	48.92	439	0.84
hsa-mir-30c-1	43832	3477	7.93	495	1.13
hsa-mir-18a	39962	12405	31.04	201	0.50

Control					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-20a	39085	2468	6.31	68	0.17
hsa-mir-26a-2	36209	3787	10.46	883	2.44
hsa-mir-27b	34227	13172	38.48	44	0.13
hsa-mir-191	30486	4651	15.26	778	2.55
hsa-mir-30d	30172	1445	4.79	300	0.99
hsa-let-7a-2	28951	812	2.80	141	0.49
hsa-mir-1226	26450	14292	54.03	8623	32.60
hsa-let-7c	22036	1819	8.25	152	0.69
hsa-mir-24-2	21298	448	2.10	98	0.46
hsa-mir-15a	17914	12342	68.90	567	3.17
hsa-mir-183	15719	2180	13.87	26	0.17
hsa-mir-196b	15147	4661	30.77	229	1.51
hsa-mir-1307	14955	1536	10.27	197	1.32
hsa-mir-17	8292	1918	23.13	7	0.08
hsa-mir-148b	7756	5435	70.07	306	3.95
hsa-mir-26a-1	7629	511	6.70	48	0.63
hsa-mir-24-1	5635	98	1.74	33	0.59
hsa-let-7i	5118	2121	41.44	36	0.70
hsa-mir-103a-1	4918	548	11.14	19	0.39

Control					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-103a-2	4812	195	4.05	3	0.06
hsa-mir-7-1	3759	594	15.80	7	0.19
hsa-mir-16-1	3075	135	4.39	7	0.23
hsa-mir-196a-2	3062	744	24.30	17	0.56
hsa-mir-101-1	2778	22	0.79	7	0.25
hsa-mir-185	2144	271	12.64	28	1.31
hsa-mir-15b	2025	4	0.20	10	0.49
hsa-mir-105-1	1321	45	3.41	38	2.88
hsa-mir-105-2	1229	34	2.77	34	2.77
hsa-mir-3607	1141	0	0.00	7	0.61
hsa-mir-30b	860	64	7.44	1	0.12
hsa-mir-345	573	110	19.20	7	1.22
hsa-mir-221	527	236	44.78	2	0.38
hsa-mir-449a	506	7	1.38	0	0.00
hsa-mir-10b	496	15	3.02	0	0.00
hsa-mir-26b	459	18	3.92	18	3.92
hsa-mir-31	447	137	30.65	2	0.45
hsa-mir-10a	116	13	11.21	0	0.00
hsa-mir-34a	101	0	0.00	3	2.97

Control					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-449c	56	0	0.00	1	1.79
hsa-mir-107	55	13	23.64	0	0.00
hsa-mir-320e	46	3	6.52	1	2.17
hsa-mir-99a	41	0	0.00	0	0.00
hsa-mir-320c-1	29	8	27.59	4	13.79
hsa-mir-3653	26	0	0.00	0	0.00
hsa-mir-744	24	3	12.50	0	0.00
hsa-mir-6516	12	0	0.00	0	0.00
hsa-mir-30e	9	0	0.00	0	0.00
hsa-mir-4521	8	0	0.00	0	0.00
hsa-mir-18b	6	1	16.67	0	0.00
hsa-mir-4485	6	0	0.00	0	0.00
hsa-mir-1229	4	3	75.00	0	0.00
hsa-mir-664b	4	0	0.00	0	0.00
hsa-mir-195	3	0	0.00	0	0.00
hsa-mir-33b	3	0	0.00	0	0.00
hsa-mir-449b	3	0	0.00	0	0.00
hsa-mir-339	2	0	0.00	0	0.00
hsa-mir-6723	2	0	0.00	0	0.00

Control					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-9-1	2	0	0.00	0	0.00
hsa-mir-126	1	0	0.00	0	0.00
hsa-mir-6724-4	1	0	0.00	0	0.00

TUTKD (TUT7/4/2 KD)					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-21	666506	2783	0.42	4194	0.63
hsa-let-7f-1	1783537	301817	16.92	10938	0.61
hsa-let-7b	385241	69933	18.15	2923	0.76
hsa-let-7a-3	375958	16043	4.27	1823	0.48
hsa-mir-423	120155	292	0.24	144	0.12
hsa-mir-93	126946	11065	8.72	1625	1.28
hsa-let-7a-1	473040	23603	4.99	2201	0.47
hsa-let-7d	152561	1769	1.16	1662	1.09
hsa-mir-30a	93713	474	0.51	124	0.13
hsa-mir-30c-2	74924	136	0.18	487	0.65
hsa-let-7e	39590	935	2.36	302	0.76
hsa-mir-182	116461	77	0.07	1238	1.06
hsa-let-7g	35234	13723	38.95	678	1.92

TUTKD (TUT7/4/2 KD)					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-98	79126	18401	23.26	955	1.21
hsa-mir-106b	115677	6217	5.37	2041	1.76
hsa-let-7f-2	164831	7639	4.63	868	0.53
hsa-mir-30c-1	37859	1590	4.20	303	0.80
hsa-mir-18a	97841	4873	4.98	1432	1.46
hsa-mir-20a	26602	584	2.20	57	0.21
hsa-mir-26a-2	62552	1199	1.92	1041	1.66
hsa-mir-27b	16735	1111	6.64	24	0.14
hsa-mir-191	32783	2159	6.59	688	2.10
hsa-mir-30d	15414	152	0.99	161	1.04
hsa-let-7a-2	14769	234	1.58	107	0.72
hsa-mir-1226	16448	6510	39.58	4353	26.47
hsa-let-7c	11607	375	3.23	95	0.82
hsa-mir-24-2	15486	96	0.62	89	0.57
hsa-mir-15a	24921	3973	15.94	2696	10.82
hsa-mir-183	21571	1386	6.43	156	0.72
hsa-mir-196b	8656	640	7.39	281	3.25
hsa-mir-1307	10234	299	2.92	105	1.03
hsa-mir-17	6020	432	7.18	18	0.30

TUTKD (TUT7/4/2 KD)					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-148b	3639	1431	39.32	98	2.69
hsa-mir-26a-1	8549	357	4.18	119	1.39
hsa-mir-24-1	5284	21	0.40	36	0.68
hsa-let-7i	5933	557	9.39	78	1.31
hsa-mir-103a-1	3158	93	2.94	26	0.82
hsa-mir-103a-2	4472	25	0.56	6	0.13
hsa-mir-7-1	7590	538	7.09	55	0.72
hsa-mir-16-1	2194	36	1.64	7	0.32
hsa-mir-196a-2	2111	317	15.02	17	0.81
hsa-mir-101-1	2544	12	0.47	4	0.16
hsa-mir-185	987	21	2.13	3	0.30
hsa-mir-15b	2149	1	0.05	17	0.79
hsa-mir-105-1	1116	15	1.34	35	3.14
hsa-mir-105-2	867	8	0.92	46	5.31
hsa-mir-3607	625	0	0.00	3	0.48
hsa-mir-30b	1263	41	3.25	7	0.55
hsa-mir-345	587	50	8.52	8	1.36
hsa-mir-221	731	113	15.46	3	0.41
hsa-mir-449a	425	1	0.24	0	0.00

TUTKD (TUT7/4/2 KD)					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-10b	332	9	2.71	4	1.20
hsa-mir-26b	351	4	1.14	9	2.56
hsa-mir-31	344	52	15.12	1	0.29
hsa-mir-10a	64	6	9.38	1	1.56
hsa-mir-34a	74	3	4.05	3	4.05
hsa-mir-449c	35	1	2.86	1	2.86
hsa-mir-107	50	0	0.00	0	0.00
hsa-mir-320e	29	2	6.90	0	0.00
hsa-mir-99a	42	0	0.00	1	2.38
hsa-mir-320c-1	31	3	9.68	1	3.23
hsa-mir-3653	2	0	0.00	0	0.00
hsa-mir-744	23	3	13.04	0	0.00
hsa-mir-6516	7	0	0.00	0	0.00
hsa-mir-30e	6	0	0.00	0	0.00
hsa-mir-4521	1	0	0.00	0	0.00
hsa-mir-18b	17	0	0.00	0	0.00
hsa-mir-4485	2	0	0.00	0	0.00
hsa-mir-1229	3	3	100.00	0	0.00
hsa-mir-664b	2	0	0.00	0	0.00

TUTKD (TUT7/4/2 KD)					
hairpin	Total reads	Uridylated reads	Uridylation ratio (%)	Adenylated reads	Adenylation ratio (%)
hsa-mir-195	4	0	0.00	0	0.00
hsa-mir-33b	4	0	0.00	0	0.00
hsa-mir-449b	2	0	0.00	0	0.00
hsa-mir-339	5	1	20.00	0	0.00
hsa-mir-6723	3	0	0.00	0	0.00
hsa-mir-9-1	4	0	0.00	0	0.00
hsa-mir-126	1	0	0.00	0	0.00
hsa-mir-6724-4	3	0	0.00	0	0.00

Table S4.3: Trimming and uridylation of pre-miRNAs

Control					
Hairpin	Total reads	Trimmed reads	Trimmed (%)	Uridylated trimmed reads	Uridylated trimmed (%)
hsa-mir-21	867159	9638	1.11	3766	39.07
hsa-let-7f-1	834611	226588	27.15	67252	29.68
hsa-let-7b	238553	99129	41.55	31917	32.20
hsa-let-7a-3	233797	81163	34.72	2681	3.30
hsa-mir-423	230535	2963	1.29	300	10.12
hsa-mir-93	191255	76663	40.08	24307	31.71

Control					
Hairpin	**Total reads**	**Trimmed reads**	**Trimmed (%)**	**Uridylated trimmed reads**	**Uridylated trimmed (%)**
hsa-let-7a-1	148857	66890	44.94	1805	2.70
hsa-let-7d	143835	12074	8.39	577	4.78
hsa-mir-30a	106625	2517	2.36	1337	53.12
hsa-mir-30c-2	89507	4654	5.20	93	2.00
hsa-let-7e	74236	5509	7.42	2447	44.42
hsa-mir-182	69259	1030	1.49	54	5.24
hsa-let-7g	67155	2565	3.82	243	9.47
hsa-mir-98	57305	8678	15.14	2116	24.38
hsa-mir-106b	55067	34440	62.54	1079	3.13
hsa-let-7f-2	52256	22328	42.73	504	2.26
hsa-mir-30c-1	43832	9302	21.22	3477	37.38
hsa-mir-18a	39962	12830	32.11	735	5.73
hsa-mir-20a	39085	11917	30.49	1996	16.75
hsa-mir-26a-2	36209	19666	54.31	3787	19.26
hsa-mir-27b	34227	381	1.11	73	19.16
hsa-mir-191	30486	14521	47.63	4648	32.01
hsa-mir-30d	30172	2359	7.82	228	9.67
hsa-let-7a-2	28951	2291	7.91	809	35.31
hsa-mir-1226	26450	1883	7.12	700	37.17

Control					
Hairpin	**Total reads**	**Trimmed reads**	**Trimmed (%)**	**Uridylated trimmed reads**	**Uridylated trimmed (%)**
hsa-let-7c	22036	3356	15.23	1811	53.96
hsa-mir-24-2	21298	2691	12.63	63	2.34
hsa-mir-15a	17914	4504	25.14	306	6.79
hsa-mir-183	15719	1055	6.71	443	41.99
hsa-mir-196b	15147	1824	12.04	334	18.31
hsa-mir-1307	14955	2710	18.12	423	15.61
hsa-mir-17	8292	886	10.68	177	19.98
hsa-mir-148b	7756	428	5.52	1	0.23
hsa-mir-26a-1	7629	1995	26.15	497	24.91
hsa-mir-24-1	5635	1302	23.11	59	4.53
hsa-let-7i	5118	424	8.28	19	4.48
hsa-mir-103a-1	4918	1313	26.70	233	17.75
hsa-mir-103a-2	4812	901	18.72	71	7.88
hsa-mir-7-1	3759	2149	57.17	591	27.50
hsa-mir-16-1	3075	394	12.81	77	19.54
hsa-mir-196a-2	3062	1745	56.99	744	42.64
hsa-mir-101-1	2778	579	20.84	18	3.11
hsa-mir-185	2144	33	1.54	11	33.33
hsa-mir-15b	2025	30	1.48	3	10.00

Control					
Hairpin	**Total reads**	**Trimmed reads**	**Trimmed (%)**	**Uridylated trimmed reads**	**Uridylated trimmed (%)**
hsa-mir-105-1	1321	150	11.36	23	15.33
hsa-mir-105-2	1229	146	11.88	23	15.75
hsa-mir-3607	1141	16	1.40	0	0.00
hsa-mir-30b	860	388	45.12	8	2.06
hsa-mir-345	573	107	18.67	5	4.67
hsa-mir-221	527	277	52.56	70	25.27
hsa-mir-449a	506	7	1.38	7	100.00
hsa-mir-10b	496	274	55.24	15	5.47
hsa-mir-26b	459	120	26.14	1	0.83

TUTKD (TUT7/4/2 KD)					
Hairpin	**Total reads**	**Trimmed reads**	**Trimmed (%)**	**Uridylated trimmed reads**	**Uridylated trimmed (%)**
hsa-mir-21	666506	11043	1.66	2179	19.73
hsa-let-7f-1	1783537	1385900	77.71	241235	17.41
hsa-let-7b	385241	298189	77.40	61390	20.59
hsa-let-7a-3	375958	307254	81.73	4592	1.49
hsa-mir-423	120155	1897	1.58	109	5.75
hsa-mir-93	126946	58714	46.25	10929	18.61

TUTKD (TUT7/4/2 KD)					
Hairpin	Total reads	Trimmed reads	Trimmed (%)	Uridylated trimmed reads	Uridylated trimmed (%)
hsa-let-7a-1	473040	439060	92.82	7856	1.79
hsa-let-7d	152561	74402	48.77	1691	2.27
hsa-mir-30a	93713	3129	3.34	470	15.02
hsa-mir-30c-2	74924	9239	12.33	136	1.47
hsa-let-7e	39590	9182	23.19	721	7.85
hsa-mir-182	116461	2053	1.76	74	3.60
hsa-let-7g	35234	15450	43.85	1719	11.13
hsa-mir-98	79126	53334	67.40	10266	19.25
hsa-mir-106b	115677	104689	90.50	4693	4.48
hsa-let-7f-2	164831	152765	92.68	1821	1.19
hsa-mir-30c-1	37859	17996	47.53	1590	8.84
hsa-mir-18a	97841	84185	86.04	2334	2.77
hsa-mir-20a	26602	9240	34.73	542	5.87
hsa-mir-26a-2	62552	45143	72.17	1199	2.66
hsa-mir-27b	16735	311	1.86	33	10.61
hsa-mir-191	32783	24502	74.74	2159	8.81
hsa-mir-30d	15414	3807	24.70	54	1.42
hsa-let-7a-2	14769	2690	18.21	234	8.70
hsa-mir-1226	16448	3752	22.81	729	19.43

TUTKD (TUT7/4/2 KD)					
Hairpin	Total reads	Trimmed reads	Trimmed (%)	Uridylated trimmed reads	Uridylated trimmed (%)
hsa-let-7c	11607	3337	28.75	375	11.24
hsa-mir-24-2	15486	2599	16.78	30	1.15
hsa-mir-15a	24921	21157	84.90	1351	6.39
hsa-mir-183	21571	4565	21.16	1067	23.37
hsa-mir-196b	8656	2830	32.69	134	4.73
hsa-mir-1307	10234	3139	30.67	198	6.31
hsa-mir-17	6020	1903	31.61	198	10.40
hsa-mir-148b	3639	1029	28.28	10	0.97
hsa-mir-26a-1	8549	3853	45.07	355	9.21
hsa-mir-24-1	5284	1466	27.74	8	0.55
hsa-let-7i	5933	2444	41.19	218	8.92
hsa-mir-103a-1	3158	1033	32.71	64	6.20
hsa-mir-103a-2	4472	1351	30.21	8	0.59
hsa-mir-7-1	7590	6698	88.25	538	8.03
hsa-mir-16-1	2194	417	19.01	30	7.19
hsa-mir-196a-2	2111	1346	63.76	317	23.55
hsa-mir-101-1	2544	982	38.60	10	1.02
hsa-mir-185	987	34	3.44	2	5.88
hsa-mir-15b	2149	77	3.58	1	1.30

TUTKD (TUT7/4/2 KD)					
Hairpin	Total reads	Trimmed reads	Trimmed (%)	Uridylated trimmed reads	Uridylated trimmed (%)
hsa-mir-105-1	1116	232	20.79	13	5.60
hsa-mir-105-2	867	167	19.26	7	4.19
hsa-mir-3607	625	6	0.96	0	0.00
hsa-mir-30b	1263	920	72.84	31	3.37
hsa-mir-345	587	305	51.96	31	10.16
hsa-mir-221	731	648	88.65	96	14.81
hsa-mir-449a	425	1	0.24	1	100.00
hsa-mir-10b	332	209	62.95	9	4.31
hsa-mir-26b	351	92	26.21	3	3.26

Table S4.4: List of oligonucleotides used in this study

Oligo name	Sequence (5′ → 3′)
siCont	ACGAAAUUGGUGGCGUAGGTT
siTUT2_1	UUAAUCACCAGCACUAACGTT
siTUT2_2	AUUACAUGGAGCUUGAUGUTT
siTUT2_3	UAAAUCACCAUCACUGCUCTT
siTUT2_4	UUGAUCUCAGUUUCUGUUGTT

Oligo name	Sequence (5′ → 3′)
siTUT4_1	UAUAAAGUCUGAAGCAACCTT
siTUT4_2	UCUUUCUCUUCUUCAUUCCTT
siTUT4_3	UUUCUUAUGUCGUUUCUCCTT
siTUT4_4	AAUUUAAGCAGCUCUAACCTT
siTUT7_1	UUUUCUUGUGCCUCUUUUCTT
siTUT7_2	AUUUCUUUGUCCUCUUUGCTT
siTUT7_3	UUUGACACGAAUACUUAUCTT
siTUT7_4	UAAAUAGGUACUCAUGUUCTT
pre-let-7a-1 unmodified	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUUC
pre-let-7a-1 +U	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUUCU
pre-let-7a-1 L4	UGAGGUAGUAGGUUGUAUAGUUUUAUAACUAUACAAUCUACUGUCUUUC
pre-let-7a-1 S14	UGAGGUAGUAGGUUGUUAGGGUCACACCCACCACUGGGAGAUCAAUCUACUGUCUUUC
pre-let-7a-1 ΔC	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUUU
pre-let-7a-1 ΔUC	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCUU
pre-let-7a-1 ΔUUC	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUCU
pre-let-7a-1 ΔUUUC	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGUC
pre-let-7a-1 ΔCUUUC	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUCUACUGU

Oligo name	Sequence (5′ → 3′)
Ac-pre-let-7a-1	UGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAUC
3′ adapter	TGGAATTCTCGGGTGCCAAGG
RT primer	CAAGCAGAAGACGGCATACGA
2nd PCR Forward primer	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA
2nd PCR Reverse primer	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

4.7 References

1 M. Ha, V. N. Kim, Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014).

2 Y. Lee *et al.*, The nuclear RNase III Drosha initiates microRNA processing. *Nature*. **425**, 1–5 (2003).

3 M. T. Bohnsack, K. Czaplinski, D. Gorlich, Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*. **10**, 185–91 (2004).

4 R. Yi, Y. Qin, I. G. Macara, B. R. Cullen, Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.*, 3011–3016 (2003).

5 E. Lund, S. Guttinger, A. Calado, J. E. Dahlberg, U. Kutay, Nuclear Export of MicroRNA Precursors. *Science*. **303**, 95–98 (2004).

6 E. Bernstein, A. A. Caudy, S. M. Hammond, G. J. Hannon, Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*. **409**, 363–366 (2001).

7 A. Grishok *et al.*, Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*. **106**, 23–34 (2001).

8 G. Hutvágner *et al.*, A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*. **293**, 834–838 (2001).

9 R. F. Ketting *et al.*, Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C . elegans. *Genes Dev.* **15**, 2654–2659 (2001).

10 S. W. Knight, B. L. Bass, A Role for the RNase III Enzyme DCR-1 in RNA Interference and Germ Line Development in Caenorhabditis elegans. *Science*. **151**, 2269–2271 (2001).

11 T. Kawamata, Y. Tomari, Making RISC. *Trends Biochem. Sci.* **35**, 368–376 (2010).

12 Z. Mourelatos *et al.*, miRNPs : a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.*, 720–728 (2002).

13 J. Yang, E. C. Lai, Review Alternative miRNA Biogenesis Pathways and the Interpretation of Core miRNA Pathway Mutants. *Mol. Cell.* **43**, 892–903 (2011).

14 A. M. Burroughs, M. Kawano, Y. Ando, C. O. Daub, Y. Hayashizaki, Pre-miRNA profiles obtained through application of locked nucleic acids and deep sequencing reveals

complex 5′/3′ arm variation including concomitant cleavage and polyuridylation patterns. *Nucleic Acids Res.* **40**, 1424–1437 (2012).

15 I. Heo *et al.*, Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell.* **151**, 521–532 (2012).

16 N. Li *et al.*, Global profiling of miRNAs and the hairpin precursors : insights into miRNA processing and novel miRNA discovery. *Nucleic Acids Res.* **41**, 3619–3634 (2013).

17 M. A. Newman, V. Mani, S. M. Hammond, Deep sequencing of microRNA precursors reveals extensive 3′ end modification. *RNA*. **17**, 1795–1803 (2011).

18 S. Diederichs, D. A. Haber, Dual Role for Argonautes in MicroRNA Processing and Posttranscriptional Regulation of MicroRNA Expression. *Cell.* **131**, 1097–1108 (2007).

19 S. Cheloufi, C. O. Dos Santos, M. M. Chong, G. J. Hannon, A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature.* **465**, 584–589 (2010).

20 D. Cifuentes *et al.*, A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science.* **328**, 1694–8 (2010).

21 M. Yoda *et al.*, Poly (A)-Specific Ribonuclease Mediates Precursor MicroRNAs. *Cell Rep.* **5**, 715–726 (2013).

22 J. Yang *et al.*, Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. *Proc. Natl. Acad. Sci.* **107**, 15163–15168 (2010).

23 K. Asada *et al.*, Rescuing dicer Defects via Inhibition of an Anti-Dicing Nuclease. *Cell Rep.* **9**, 1471–1481 (2014).

24 H. I. Suzuki *et al.*, MCPIP1 Ribonuclease Antagonizes Dicer and Terminates MicroRNA Biogenesis through Precursor MicroRNA Degradation. *Mol. Cell.* **44**, 424–436 (2011).

25 J.-P. Upton *et al.*, IRE1a Cleaves Select microRNAs During ER Stress to Derepress Translation of Proapoptotic Caspase-2. *Science.* **338**, 818–822 (2012).

26 S. L. Ameres, P. D. Zamore, Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* **14**, 475–88 (2013).

27 L. Ji, X. Chen, Regulation of small RNA stability : methylation and beyond. *Cell Res.* **22**, 624–636 (2012).

28 C. J. Norbury, Cytoplasmic RNA : a case of the tail wagging the dog. *Nat. Rev. Mol. Cell Biol.* **13**, 643–653 (2013).

29 D. D. Scott, C. J. Norbury, RNA decay via 3′ uridylation. *Biochim. Biophys. Acta J.* **1829**, 654–665 (2013).

30 H. Chang, J. Lim, M. Ha, V. N. Kim, TAIL-seq: Genome-wide determination of poly(A) tail length and 3′ end modifications. *Mol. Cell.* **53**, 1044–1052 (2014).

31 M. Lee, B. Kim, V. N. Kim, Emerging Roles of RNA Modification : m6A and U-Tail. *Cell.* **158**, 980–987 (2014).

32 J. Li, Z. Yang, B. Yu, J. Liu, X. Chen, Methylation Protects miRNAs and siRNAs from a 3′-End Uridylation Activity in Arabidopsis. *Curr. Biol.* **15**, 1501–1507 (2005).

33 J. Lim *et al.*, Article Uridylation by TUT4 and TUT7 Marks mRNA for Degradation. *Cell.* **159**, 1365–1376 (2014).

34 L. Aravind, E. V Koonin, DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucleic Acids Res.* **27**, 1609–18 (1999).

35 G. Martin, W. Keller, RNA-specific ribonucleotidyl transferases. *RNA*. **13**, 1834–1849 (2007).

36 A. L. Stevenson, C. J. Norbury, The Cid1 family of non-canonical poly (A) polymerases. *Yeast.* **23**, 991–1000 (2006).

37 C. J. Wilusz, J. Wilusz, New ways to meet your (3′) end — oligouridylation as a step on the path to destruction. *Genes Dev.* **22**, 1–7 (2008).

38 J. P. Hagan, E. Piskounova, R. I. Gregory, Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* **16**, 1021–5 (2009).

39 I. Heo *et al.*, Lin28 Mediates the Terminal Uridylation of let-7 Precursor MicroRNA. *Mol. Cell.* **32**, 276–284 (2008).

40 I. Heo *et al.*, TUT4 in Concert with Lin28 Suppresses MicroRNA Biogenesis through Pre-MicroRNA Uridylation. *Cell.* **138**, 696–708 (2009).

41 J. E. Thornton, H. Chang, E. Piskounova, R. I. Gregory, Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7). *RNA*. **18**, 1875–1885 (2012).

42 K. Yeom *et al.*, Single-molecule approach to immunoprecipitated protein complexes: insights intomiRNA uridylation. *EMBO Rep.* **12**, 690–696 (2011).

43 H. Chang, R. Triboulet, J. E. Thornton, R. I. Gregory, A role for the Perlman syndrome exonuclease Dis3l2 in the Lin28-let-7 pathway. *Nature.* **497**, 244–8 (2013).

44 D. Ustianenko *et al.*, Mammalian DIS3L2 exoribonuclease targets the uridylated precursors of let-7 miRNAs. *RNA*. **19**, 1632–1638 (2013).

45 F. E. Loughlin *et al.*, Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28. *Nat. Struct. Mol. Biol.* **19**, 84–89 (2012).

46 F. Mayr, A. Schu, N. Doge, U. Heinemann, The Lin28 cold-shock domain remodels pre- let-7 microRNA. *Nucleic Acids Res.* **40**, 7492–7506 (2012).

47 Y. Nam, C. Chen, R. I. Gregory, J. J. Chou, P. Sliz, Molecular Basis for Interaction of let-7 MicroRNAs with Lin28. *Cell.* **147**, 1080–1091 (2011).

48 J. E. Kwak, M. Wickens, A family of poly (U) polymerases. *RNA*. **13**, 860–867 (2007).

49 B. M. Lunde, I. Magler, A. Meinhart, Crystal structures of the Cid1 poly (U) polymerase reveal the mechanism for UTP selectivity. *Nucleic Acids Res.* **40**, 13–16 (2012).

50 P. Munoz-tello, C. Gabus, Functional Implications from the Cid1 Poly (U) Polymerase Crystal Structure. *Structure.* **20**, 977–986 (2012).

51 L. A. Yates *et al.*, Structural basis for the activity of a cytoplasmic RNA terminal uridylyl transferase. *Nat. Struct. Mol. Biol.* **19**, 782–787 (2012).

52 X. Liu *et al.*, Article A MicroRNA Precursor Surveillance System in Quality Control of MicroRNA Synthesis. *Mol. Cell.* **55**, 868–879 (2014).

53 L. A. Yates *et al.*, Structural plasticity of Cid1 provides a basis for its distributive RNA terminal uridylyl transferase activity. *Nucleic Acids Res.* **43**, 2968–2979 (2015).

54 S. D. Chandradoss *et al.*, Surface passivation for single-molecule protein studies. *J. Vis. Exp.* **50549**, 4–11 (2014).

55 A. Kozomara, S. Griffiths-jones, miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, 68–73 (2014).

56 A. R. Quinlan, I. M. Hall, BEDTools : a flexible suite of utilities for comparing genomic features. Bioinformatics. **26**, 841–842 (2010).

57 W. J. Kent, BLAT — The BLAST -Like Alignment Tool, 656–664 (2002).

58 J. Koster, S. Rahmann, Snakemake — a scalable bioinformatics workflow engine. *Bioinformatics*. **28**, 2520–2522 (2012).

59 F. Perez, B. E. Granger, IPython: A System for Interactive Scientific Computing Python. *Comput. Sci. Eng.*, 21–29 (2007).

5

Single-molecule pull-down for investigating protein–nucleic acid interactions

Methods
2016 Aug 1;105:99-108. doi: 10.1016/j.ymeth.2016.03.022.

Mohamed Fareh [1], **Luuk Loeff** [1], Malwina Szczepaniak [1], Anna C. Haagsma [1], Kyu-Hyeon Yeom [1, 2] & Chirlmin Joo** [1]

** Corresponding author

[1] Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, 2628 CJ, Delft, The Netherlands.

[2] Present address: MacDonald Research Laboratories, University of California at Los Angeles, Los Angeles, CA 90095-1662, USA.

5.1 Abstract

The genome and transcriptome are constantly modified by proteins in the cell. Recent advances in single-molecule techniques allow for high spatial and temporal observations of these interactions between proteins and nucleic acids. However, due to the difficulty of obtaining functional protein complexes, it remains challenging to study the interactions between macromolecular protein complexes and nucleic acids. Here, we combined single-molecule fluorescence with various protein complex pull-down techniques and determined the function and stoichiometry of ribonucleoprotein complexes. Through the use of three examples from eukaryotic cells (Drosha, Dicer, and TUT4 protein complexes), we provide step-by-step guidance for using novel single-molecule techniques. Our single-molecule methods provide sub-second and nanometer resolution and can be applied to other nucleoprotein complexes that are essential for cellular processes.

5.2 Introduction

Interactions between protein assemblies and nucleic acids are essential elements of cellular processes, such as transcription, translation, and chromatin remodeling. A well-known example of such a protein assembly is the spliceosome, a multi-megadalton ribonucleoprotein complex that uses numerous cofactors to catalyze the splicing of precursor messenger RNA [1, 2]. The ribonucleoprotein complex called RISC (RNA-induced silencing complex) is a key player in RNA interference-a cellular process of translational repression [3]. The biogenesis and regulation of microRNA (non-coding RNA that mediates RNA interference) involves several protein complexes such as human Drosha-DGCR8 [4, 5], human Dicer-TRBP [6, 7], Drosophila Dicer-Loqs [8, 9] and human TUTase-Trim25 [10].

A comprehensive analysis of nucleoprotein complexes is a stepping stone to understanding cellular processes. Recent advances in analytical and biochemical methods have led to numerous breakthroughs in the characterization of multi-component protein assemblies in complexes with nucleic acids. High-throughput approaches, including large-scale tandem affinity purification, the yeast two-hybrid system, and mass spectrometry analysis, have been used to identify thousands of new protein complexes in yeast [11–15], *Drosophila melanogaster* [16, 17] and *Caenorhabditis elegans* [18]. In parallel, advanced computational methods have emerged during the past decade, which made it possible to predict the formation of protein complexes [19]. Major advances in sample preparation and detection techniques have also enabled crystallographers and electron microscopists to determine the structure of large protein complexes interacting with nucleic acid substrates at an atomic resolution [20, 21].

Despite the wealth of information acquired from these analytical and biochemical methods, there is a need for complementary techniques that allow for real-time observations of the assembly and function of nucleoprotein complexes. Recently, we and other groups developed such single-molecule fluorescence methods. Hoskins et al. revealed the order of spliceosome assembly during pre-mRNA maturation in cell extract via single-molecule multi-color fluorescence [22, 23]. Single-molecule pull-down FRET allowed Nils et al. to visualize in real time the splicing of pre-mRNA by the spliceosome [24, 25]. Lee et al. used a single-molecule co-immunoprecipitation approach to investigate weak interactions between different proteins [26, 27]. Jain et al. developed single-molecule pull-down techniques to determine the stoichiometry of protein complexes [28–33]. We developed a single-molecule pull-down method to gain insight into the molecular mechanism of large nucleoprotein complexes involved in microRNA uridylation [34].

Here, we describe various single-molecule pull-down approaches and provide protocols for the purification and immobilization of ribonucleoprotein complexes associated with their native cofactors. Our pull-down methods in combination with single-molecule fluorescence allow for real-time visualization of protein complexes and RNA interactions. We describe several different strategies used in our laboratory and list the challenges that we encountered during the development of these

techniques. As a proof-of-concept, we show three examples of protein complexes involved in small RNA biogenesis (Drosha-DGCR8, human Dicer-TRBP, Drosophila Dicer 2-Loqs-PD, and a TUT4 complex) and illustrate how we elucidate the molecular bases of their functions. With this protocol, single-molecule fluorescence can be widely used to study nucleoprotein complexes.

5.3 Results & Discussion

5.3.1 Stoichiometry determination: Drosha-DGCR8 protein complex

The microprocessor complex, composed of Drosha and its cofactor DGCR8, plays an essential role in the initial stage of microRNA (miRNA) biogenesis. In the nucleus, the microprocessor binds to and subsequently cleaves pri-miRNA transcripts, resulting in the production of hairpin-structured pre-miRNAs [35]. Drosha hosts catalytic sites that are required for cleavage, while its cofactor DGCR8 enhances binding to the substrate pri-miRNA [35]. Using a single-molecule pull-down method, we determined the stoichiometry of Drosha and its cofactor DGCR8 in naturally formed protein complexes.

We advanced a single-molecule pull-down method from Jain et al. [28] by introducing a biotin-streptavidin conjugation scheme (Figure 5.1). We fused the N-terminus of Drosha with the acceptor peptide (AP) (Figure 5.2A), which is covalently coupled with biotin by the *E. coli* enzyme biotin ligase (BirA) [36, 37]. We fused DGCR8 with a fluorescent protein (eGFP) to be able to observe interactions between Drosha and DGCR8.

We immobilized the Drosha-DGCR8 complexes from the crude cell extract on the surface of the microfluidic chamber that was passivated with PEG with two rounds of PEGylation, as described by Chandradoss et al. in [38], and afterwards, coated with streptavidin (Figure 5.3A). However, we observed a prominent non-specific adsorption of cellular proteins to the glass surface. Comparison with a control chamber that was not treated with streptavidin, which thus should not show immobilized proteins, revealed little difference in the number of detected molecules (data not shown). Tween-20 was recently reported to improve the surface passivation [38]. We observed that the additional treatment of the PEGylated surface with 5 % Tween-20 reduced the number of non-specific adsorption of cellular proteins by factor of two (data not shown). We note that it is not recommended to use BSA (bovine serum album) for surface passivation since BSA increases the degree of non-specific binding of proteins (unpublished observation).

To determine the stoichiometry of the Drosha-DGCR8 complexes immobilized on the surface, we excited eGFP with a laser beam and recorded its fluorescence signal until all of the eGFP molecules were photobleached. The number of photobleaching steps, defined as a sudden decrease in eGFP fluorescence intensity (Figure 5.3B), reflects the number of DGCR8 molecules associated with a single Drosha protein. Our photobleaching data Figure 5.3C) show that ~46% (236 among 513 analyzed molecules) of the microprocessor complexes are composed of one Drosha and two DGCR8 proteins, in agreement with recently published work by Nguyen et al. [39].

To determine whether the immobilized Drosha-DGCR8 complexes were capable of binding RNA substrates, we introduced Cy5-labeled pri-miRNA, a known substrate of Microprocessor [40], into the microfluidic chamber (Figure 5.3D). Simultaneous illumination of eGFP and Cy5 dyes allowed us to co-localize RNA bound to Drosha-DGCR8, suggesting that the immobilized Microprocessor complexes retained its RNA-binding activity.

We emphasize that when attempting to immobilize a protein of interest directly from the cell extract, it is important to pay attention to the high amount of proteins in the cell extract. It is crucial to maximize the quality of the glass surface and minimize the incubation time of the cell extract (we recommend an incubation for 30 seconds or less). As an alternative, one could employ an additional purification step to reduce the content of unwanted proteins before applying the protein sample to a microfluidic chamber. Examples of a tandem purification scheme are described in the following sections.

5.3.2 Drosophila Dicer-2 associated with Loquacious-PD

Drosophila melanogaster Dicer-2 (dmDicer-2) is an endoribonuclease that processes long double stranded RNA (dsRNA) molecules into 21-nt small interfering RNAs (siRNAs). For efficient cleavage of RNA substrates, dmDicer-2 requires a cofactor, Loquacious-PD [41]. Loquacious-PD also facilitates the loading of dsRNA substrates onto dmDicer-2 [42]. To visualize the binding of a dsRNA substrate by dmDicer-2 at the single-molecule level, we sought to develop a single-molecule pull-down assay advancing our previously reported SIMPlex technique (Single-molecule approach to Immunoprecipitated Protein complexes) [34]. For this purpose, we tested various immobilization schemes.

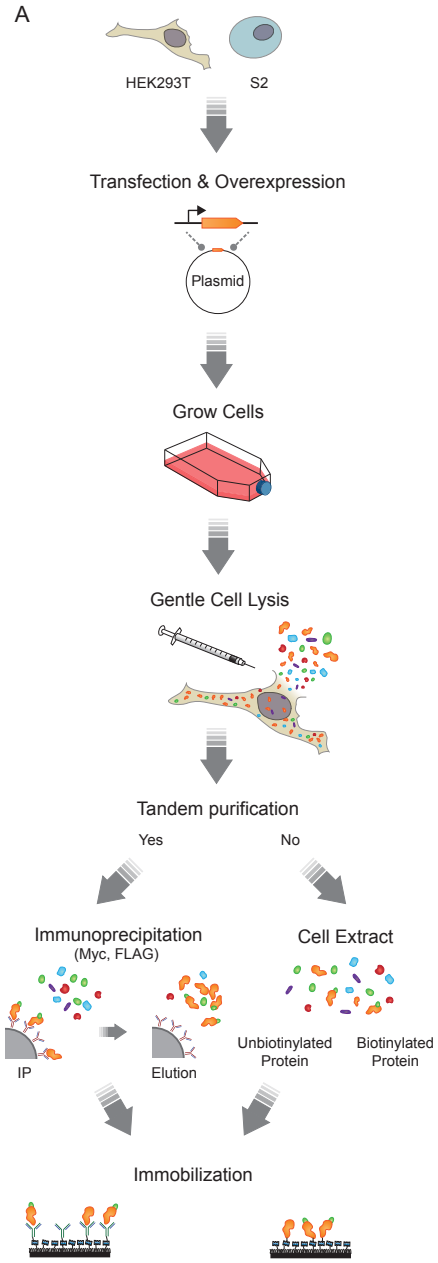


Figure 5.1: Overview of the single molecule pull-down techniques

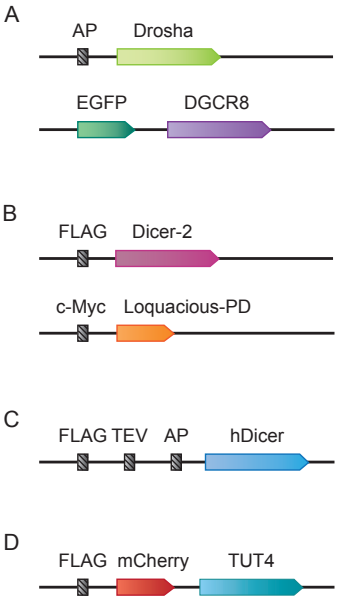


Figure 5.2: Plasmid constructs

(A) Illustrated are the constructs used for Figure 5.3. AP stands for the acceptor peptide that is recognized by BirA. DGCR8 is tagged with a fluorescent protein, eGFP. **(B)** Illustrated are the constructs used for Figure 5.4. Dicer-2 is tagged with the FLAG epitope. Loquacious-PD is tagged with the 1xc-Myc epitope. (**C**) Illustrated is a construct used for Figure 5.5. hDicer is tagged with 1xFLAG, TEV (Tobacco Etch Virus), and AP. (**D**) Illustrated is a construct used for Figure 5.6. TUT4 is tagged with FLAG epitope and a fluorescent protein (mCherry).

In the first attempt, we immobilized the dmDicer-2/Loquacious-PD complex from the crude cell extract using antibodies. We immobilized a primary biotinylated anti-rabbit IgG antibody, which allowed us to immobilize the secondary anti-c-Myc antibody that targets the 1xc-Myc tag fused to Loquacious-PD (Figure 5.2B and Figure 5.4A). Upon introduction of a 70-nt long Cy3-labeled dsRNA substrate, we observed a large number of binding events (dark spots on the CCD image, Figure 5.4B, left panel). However, a control without antibodies also showed a substantial number of binding events (CCD image in the Figure 5.4B, right panel). These results suggested that other RNA-binding proteins in the cell extract were non-specifically adsorbed on the surface and mediated RNA binding, which is consistent with our previous observation [34]. We note that the high concentration of proteins in the cell extract lead to a non-specific adsorption of many proteins on the surface of the imaging chamber, among which RNA binding proteins can interact with dye-labeled RNA molecules.

To overcome non-specific adsorption, we prepared higher purity immunoprecipitates via two rounds of immunoprecipitation. In the first round of immunoprecipitation, the dmDicer-2/Loquacious-PD complex was pulled down using anti-c-Myc coated beads that target Loquacious-PD. After elution of the dmDicer-2/Loquacious-PD complex from the beads, a second round of immunoprecipitation was conducted on a single-molecule surface coated with a primary biotinylated anti-rabbit IgG antibody and a secondary anti-FLAG antibody that targeted FLAG-tagged dmDicer-2. Upon introduction of the Cy3-labeled dsRNA, we observed that this immobilization scheme generated little fluorescence (34.3 ± 9.0 binding events per field of view; CCD image in Figure 5.4C, left panel). This observation suggests that two rounds of immunoprecipitation improved the purity of the IP, resulting in a reduced background of non-specific interactions. However, the lack of Cy3 fluorescence signals indicated that there were few dmDicer-2/Loquacious-PD immobilized on the surface. This could be due to the overrepresentation of Loquacious-PD compared to dmDicer-2 in the crude cell extract. Pull-down with anti-c-Myc coated beads might have resulted in a large quantity of Loquacious-PD that was not associated with dmDicer-2. Alternatively, even if dmDicer-2/Loquacious-PD complex was immobilized, the FLAG antibody might have affected the ability of dmDicer-2 to bind to RNA substrates.

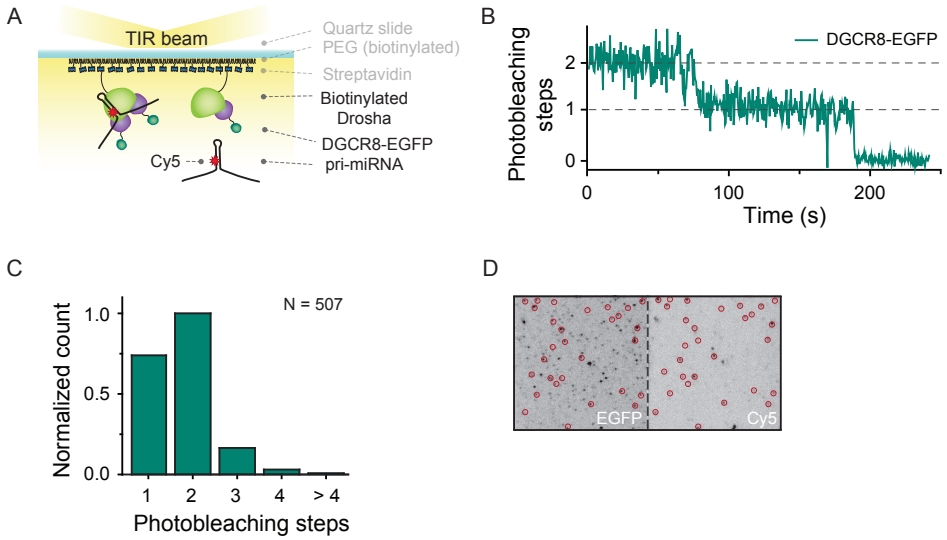


Figure 5.3: Single-molecule stoichiometry measurement

(A) Schematic overview of a single-molecule stoichiometry measurement. Biotinylated Drosha-DGCR8/eGFP complexes were immobilized on a PEGylated surface using biotin-streptavidin conjugation. Cy5-labeled pri-miRNA molecules were introduced into the imaging chamber. **(B)** Representative time trajectory of eGFP fluorescence. The stoichiometry of DGCR8 molecules associated with Drosha molecules was determined by counting the number of eGFP photobleaching steps (indicated with dashed lines). **(C)** Bar plot showing the distribution of the Drosha/DGCR8 stoichiometry. **(D)** Ability of Drosha-DGCR8 complexes to bind RNA molecules was confirmed using the co-localization of pri-miRNA and DGCR8. The figure shows a camera screenshot with eGFP molecules in the left channel and Cy5 molecules in the right channel. The red circles indicate co-localized DGCR8 (eGFP) and pri-miRNA (Cy5).

To enrich the immunoprecipitate with dmDicer-2, we changed the order of the purification scheme. We pulled down dmDicer-2 in the first round of immunoprecipitation using an anti-FLAG antibody. For the second round of immunoprecipitation on the single-molecule surface, we used the anti-c-Myc antibody that targets Loquacious-PD. With this scheme, we observed a large amount of fluorescence signal upon introduction of Cy3-labeled dsRNA, suggesting that the immobilized complexes are potent for RNA binding (546.0 ± 44.7 binding events per field of view; CCD image in Figure 5.4C, right panel). In addition, a control surface without antibodies (data not shown) showed hardly any fluorescence signal. These results suggest that a tandem purification scheme can be used to purify protein complexes with high purity. In summary, it is essential to determine how many rounds of pull-down are required to reach purity that is suitable for single-molecule observation, to optimize the order of pull-down to obtain protein complexes in a high yield, and to empirically select the position of tags that allows for reliable immobilization.

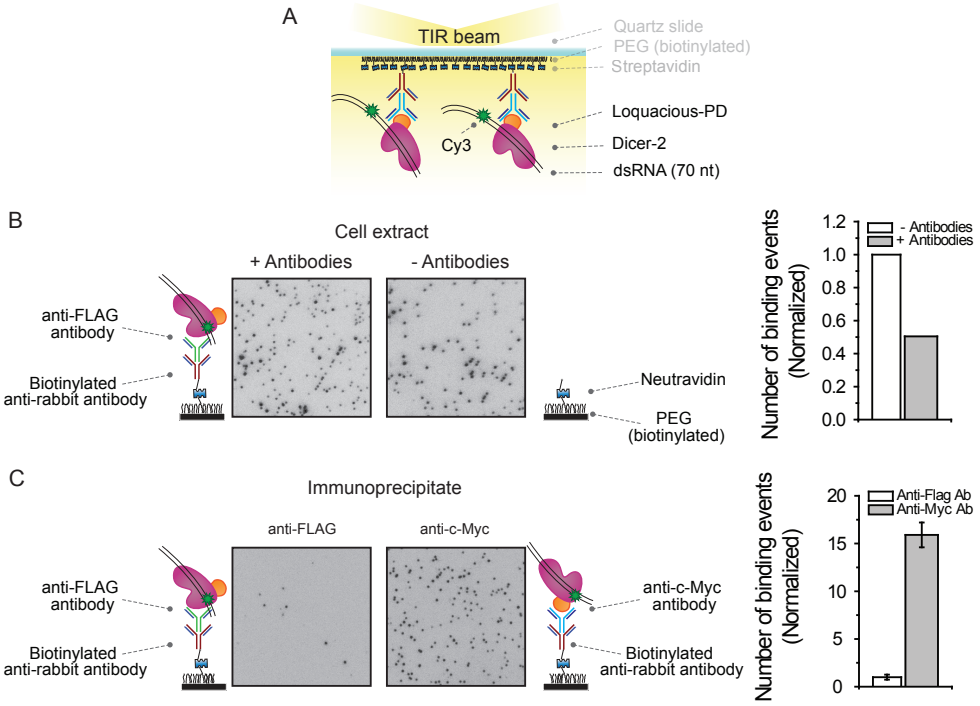


Figure 5.4: Single-molecule binding measurement

(A) Schematic overview of a single-molecule pull-down assay for Dicer-RNA interactions. Immunoprecipitated dmDicer-2/Loquacious-PD is immobilized on a PEGylated surface using various antibodies. Binding of RNA was observed after injecting a 70-nt Cy3 labeled dsRNA. **(B)** Crude cell extract is added on a PEGylated surface coated with a primary biotinylated anti-rabbit IgG antibody bound to a secondary anti-c-Myc antibody (left panel) or without a primary biotinylated anti-rabbit IgG antibody bound to a secondary anti-c-Myc antibody (right panel). Binding of RNA was observed after introducing a 70-nt Cy3 labeled dsRNA. The histogram on the right displays a normalized number of dsRNA molecules docked to the surface. **(C)** Immunoprecipitated dmDicer-2/Loquacious-PD is immobilized on a PEGylated surface using biotinylated anti-rabbit IgG antibody bound to anti-FLAG antibody (left panel) or using biotinylated anti-rabbit IgG antibody bound to anti-c-Myc antibody (right panel). Binding of RNA was observed after introducing a 70-nt Cy3 labeled dsRNA. The histogram on the right displays a normalized number of dsRNA molecules docked to the surface.

5.3.3 Human Dicer associated with TRBP

In humans, the endoribonuclease Dicer processes pre-miRNAs into mature miRNAs [3]. Dicer is associated with dsRNA-binding protein TRBP. A static picture of pre-miRNA maturation by the Dicer complex was provided by structural and biochemical studies, but a more dynamic view of this process remains to be established.

To visualize miRNA processing by the Dicer-TRBP complex at the single-molecule level, we employed a tandem purification method that allows the pull-down and surface immobilization of protein complexes. To pull down and immobilize TRBP-associated Dicer on a surface, we cloned FLAG, TEV (tobacco etch virus) and AP tags upstream of the human Dicer coding sequence [43]. The three tags were used for

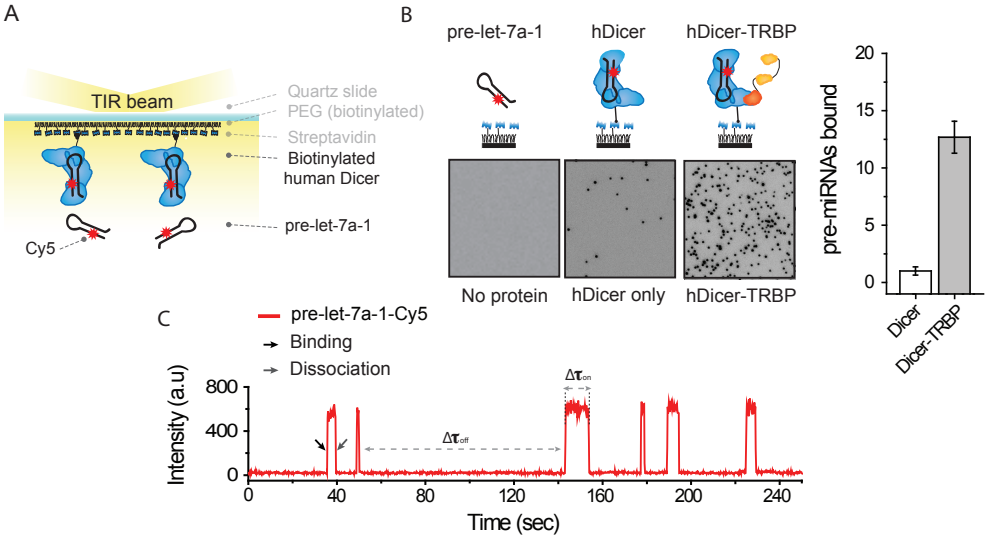


Figure 5.5: Single-molecule kinetics measurement

(A) Schematic overview of a single-molecule pull-down assay for Dicer-RNA interactions. Immunoprecipitated human Dicer-TRBP complexes were immobilized on a PEGylated surface using biotin-streptavidin interaction. Cy5-labeled pre-let-7a-1 was introduced into the imaging chamber by flow. Interactions between surface-immobilized Dicer-TRBP and Cy5-labeled pre-let-7a-1 were recorded in real time. **(B)** CCD images display the RNA binding activity of Dicer (middle) and the Dicer-TRBP complex (right). Passivated surface without Dicer immobilized was used as negative control (left). The histogram on the right displays a normalized number of pre-let-7a-1 stably bound to Dicer (white) or to Dicer-TRBP (grey). **(C)** Representative time trajectory (obtained with a time resolution 300 ms) displays six binding events of Cy5-labeled pre-let-7a-1 to a single Dicer-TRBP complex. The black arrow indicates the binding and the grey arrow indicates the dissociation of pre-let-7a-1.

immunoprecipitation, elution and in vivo biotinylation, respectively (Figure 5.1 & Figure 5.2C). BirA enzyme was co-expressed for in vivo biotinylation of the AP tag. After the purification process (in vivo biotinylation, FLAG immunoprecipitation and TEV elution), the IPs were immobilized on the surface of the imaging chamber via biotin-streptavidin conjugation (Figure 5.1 & Figure 5.5A).

We introduced a dye-labeled pre-miRNA into the imaging chamber. After 5 minutes of incubation, we washed away the unbound pre-miRNA and quantified the RNA binding activity of each protein complex by taking snapshots of different fields of view (Figure 5.5B). The dark spots on the CCD image represent single Cy5-labeled pre-miRNAs that are stably bound to single Dicer proteins. Compared to Dicer alone, Dicer-TRBP showed an increase in the RNA binding activity of one order of magnitude (Figure 5.5B). A passivated surface without immobilized Dicer did not show any adsorption of RNA, ruling out any nonspecific interactions between RNA and the glass surface (Figure 5.5B). This single-molecule approach can be used to quantify the previously reported enhancement of RNA binding activity mediated by TRBP [44].

To visualize the interaction between the Dicer-TRBP complex and pre-miRNA in real time, we introduced dye-labeled RNA into the imaging chamber while recording the binding events. The representative time trace (Figure 5.5C) shows a sudden increase of the fluorescence, which reflects the interaction of surface-immobilized Dicer-TRBP with the RNA molecule, followed by RNA dissociation that is reflected by the loss of the fluorescence signal. The time trace can be further analyzed to determine kinetic parameters, including the binding and dissociation rates, which allows to draw the energy landscape of substrate recognition and processing by Dicer complexes. The average binding dwell-time ($<\Delta\tau_{on}>$) is obtained from a distribution of the interaction time between Dicer complexes and pre-miRNA ($\Delta\tau_{on}$). The dissociation rate (k_{off}) is the inverse of $<\Delta\tau_{on}>$. The association rate (k_{on}) is the inverse of the average time interval between two successive binding events ($\Delta\tau_{off}$).

5.3.4 Single-molecule FRET measurements on TUT4 protein complexes

Terminal uridylyl transferases (TUTs) function as integral regulators of miRNA biogenesis. Recent studies have shown that TUT4 (ZCCHC11), TUT7 (ZCCHC6) and TUT2 (GLD2/PAPD4) enhance the maturation of pre-miRNAs through distributive mono-uridylation [45, 46]. In contrast, in embryonic stem cells and cancer cells, where Lin28 is enriched, TUT4 and TUT7 inhibit pre-miRNA maturation through oligo-uridylation [46–49]. The oligo U-tail promotes degradation by the exonuclease DIS3L2 [34, 47, 50, 51]. Although the general mechanism of oligo-uridylation has been well established, the underlying molecular mechanism remains poorly understood.

This limited understanding is mainly due to the lack of full-length recombinant TUT4. Therefore, we employed our single-molecule pull-down method that makes use of tandem purification to obtain high purity full-length TUT4 [34]. In brief, TUT4-FLAG-mCherry proteins were pulled down from a crude cell extract using beads coated with FLAG antibodies, followed by a second round of immunoprecipitation directly in the single-molecule chamber (Figure 5.1). This scheme resulted in the immobilization of TUT4 through mCherry anti-RFP conjugation (Figure 5.6A). Next, we reconstituted the ternary complex required for oligo-uridylation by introducing a Cy5-labeled pre-miRNA substrate that was pre-incubated with the processivity factor Lin28b. After equilibration, the Lin28b-bound pre-miRNA complex docked to the immobilized TUT4 proteins (Figure 5.6A) [34].

To initiate oligo-uridylation by the immobilized ternary TUT4/Lin28b/pre-miRNA complex, we injected a solution containing 100 μM UTP into the microfluidic chamber. To track the molecular dynamics of oligo-uridylation in real time, we included 10 nM Cy3-labeled oligonucleotide dA_{15} (oligo-dA_{15}). Upon elongation of the U-tail, oligo-dA_{15} hybridized with the U-tail, resulting in a stepwise increase of the total fluorescence intensity (Figure 5.6B, black line). Intriguingly, we obtained a signal from the Cy5-labeled pre-miRNA while exciting the Cy3-labeled oligo-dA_{15}, suggesting that Förster resonance energy transfer (FRET) occurred between these two dyes (Figure 5.6B, green and red line). Apparent FRET efficiency is the ratio between I_A (acceptor signal) and ($I_D + I_A$) (total signals summing donor and acceptor signals). Upon hybridization of the first oligo-dA_{15},

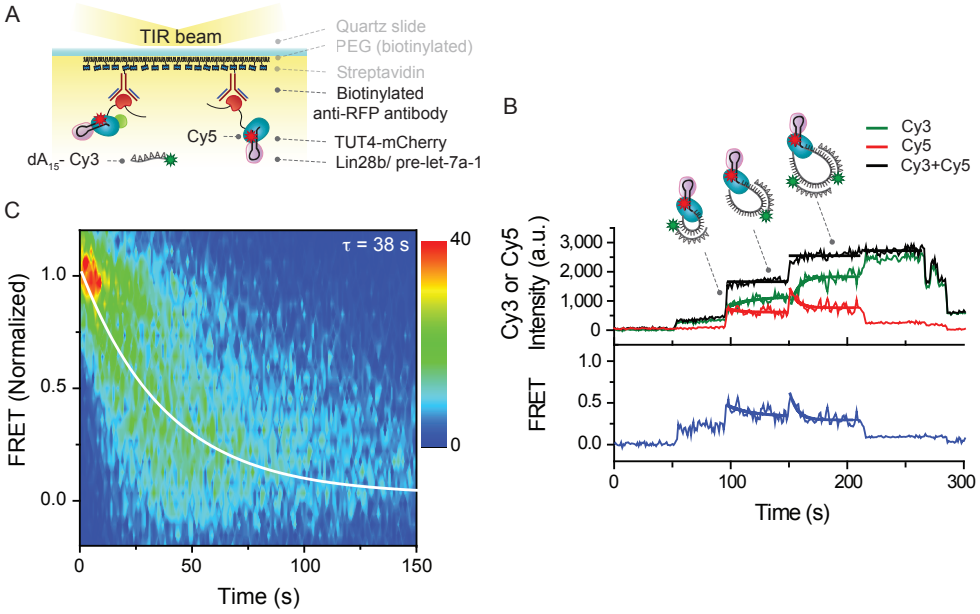


Figure 5.6: Single-molecule FRET measurement

(A) Schematic overview of a single-molecule FRET assay. Immunoprecipitated TUT4-Flag-mCherry proteins were immobilized on a PEGylated surface using anti-RFP antibodies (RFP, red fluorescent protein). Next, Lin28b/pre-let-7a-1 complexes were immobilized to the surface in a TUT4 specific manner. Oligo uridylation was tracked in real-time by simultaneously injecting UTP with Cy3-labeled oligo dA_{15}. **(B)** Representative time trajectory of the donor (Cy3, green), acceptor (Cy5, red), total fluorescence intensity (black), and the corresponding FRET values are in blue. Thick lines in green (Cy3), red (Cy5), and blue represent single-exponential fits of each oligo dA_{15} hybridization event. Thick lines in black represent the mean value of the total intensity during the hybridization of an oligo dA_{15}. **(C)** Contour plot of the evolution of FRET over time, measured with 749 single TUT4 complexes.

we observed an increase in FRET (Figure 5.6B, blue line) that gradually decreased during the elongation of the U-tail. When the U-tail had reached a sufficient length, a second oligo-dA_{15} hybridized, leading to another increase in FRET that again gradually decreased over time (Figure 5.6B and 6C). This suggests that TUT4 and Lin28 maintain a tightly associated complex, which captures the 3' end of pre-miRNA and brings it to its own catalytic domain. This mechanism hints at the formation of a unique closed loop of the U-tail during oligo-uridylation by TUT4 (Figure 5.6B).

To obtain a kinetic understanding of the oligo-uridylation process, we analyzed the gradually decreasing FRET events (Figure 5.6C). Each oligo-dA_{15} hybridization event was selected using home-written Matlab software. The starting time and the FRET values of selected events were normalized such that all the events start with a FRET efficiency value 1. To visualize the distribution among the decaying FRET traces, a contour plot was made of the normalized FRET data. The kinetic rate was extracted by fitting the data from the contour plot with a single-exponential decay, which resulted in an average dwell time of 38 sec per oligo-dA_{15} hybridization event.

Our oligo-uridylation data suggest that we were able to obtain functional full-length TUT4 molecules with high purity through the single-molecule pull-down method. In addition, by using FRET, we uncovered that TUT4 and Lin28 remain in tight contact while making use of a unique closed loop formation during oligo-uridylation. Recent crystal structures of the yeast homolog Cid1 showed that the surface of the C-terminal domain of Cid1 is mostly positively charged [52], which might facilitate loop formation by wrapping the U-tail around the protein. This suggests that the loop formation may be a general feature of TUTs.

5.4 Conclusion

We have shown that when integrated with protein complex pull-down methods, single-molecule fluorescence techniques become pertinent tools to obtain mechanistic insights into ribonucleoprotein complexes. These techniques can be applied to study the function and stoichiometry of any nucleoprotein that is difficult to obtain using traditional biochemical methods. However, special consideration must be given to protein complex purification and surface immobilization to attain single-molecule observation of functionally active proteins that are free from surface artifacts.

5.5 Experimental procedures

5.5.1 Cell culture: HEK-293T cells

Human embryonic kidney cells (HEK-293T) were maintained in Dulbecco's Modified Eagle's Medium (DMEM, 31885023, Gibco®) supplemented with 10 % fetal bovine serum (FBS, heat-inactivated, Greiner Bio-One) at 37°C and 5 % CO_2. Before transfection, cells were split into 10 cm cell culture dishes to a confluence of 25 %. After 24 hours of growth, plasmids of interest were transfected using a $CaPO_4$ transfection method [53]. For the in vivo biotinylation of human Dicer and Drosha proteins, an additional plasmid coding for the BirA enzyme was co-transfected. After 5 hours, the medium was exchanged with fresh DMEM containing 1 µg/ml biotin (B4639, Sigma), and the transfected cells were incubated for another 48 hours to enable protein expression and in vivo biotinylation.

5.5.2 Cell culture: SL2 cells

Schneider's Drosophila Line 2 (SL2, CRL-1963™, ATCC®) was maintained in HyClone SFX-Insect Cell Culture medium (SH30278.LS, GE Healthcare HYCLONE) supplemented with 10% FBS (heat-inactivated, Greiner Bio-One) at 25°C. When the culture reached a density of 0.5 x 10^6 cells/ mL, the cells were transfected using the FuGENE® HD transfection method (E2311, Promega). After 24 hours of incubation, 1 mM $CuSO_4$ was added to the medium, and the cells were incubated for an additional 48 hours.

5.5.3 Cell harvest and lysis

Before the transfected cells were harvested with scrapers, DMEM was removed and the cells were washed with ice-cold Dulbecco's Phosphate-Buffered Saline (DPBS, 14200 Gibco®). Subsequently, the cells were transferred to 15 mL tubes

and centrifuged at 276 x *g* and 4°C for 5 min to form cell pellets. After the removal of the supernatant, the cell pellets were frozen and stored at -80 °C until further processing. Before lysis, the cells were thawed on ice for over 30 min. Subsequently, HEK-293T and SL2 cells were resuspended in buffer D (20 mM Tris-HCl [pH 8.0], 200 mM KCl and 0.2 mM EDTA) and lysis buffer SL2 (30 mM HEPES-KOH [pH 7.4], 100 mM KOAc, 10 % glycerol, 0.1 % Triton X-100), respectively. Lysis was carried out by carefully passing the cells 10 times through a needle (30 ½ gauge, BD), while avoiding the formation of air bubbles. Afterwards, the lysate was centrifuged twice (16,100 x *g* at 4 °C, for 20 min) to remove cell debris (pellet). The recovered cell extract (supernatant) was either directly used for single-molecule experiments (Drosha-DGCR8), or alternatively, tandem purification steps were carried out to obtain higher purity samples (dmDicer-2, hDicer and TUT4). To prevent disturbing the protein complexes, it is important to perform the cell lysis and immunoprecipitation in a gentle manner and in a physiologically relevant buffer. We do not recommend the use of sonication as a cell lysis method because this may cause protein complexes to disassemble and form aggregates [54].

5.5.4 Immunoprecipitation and elution

For immunoprecipitation of 1xFLAG-tagged proteins (dmDicer-2, hDicer, and TUT4), 1 mg of total protein in the cell extract was incubated with 2.5 µL of anti-FLAG antibody-conjugated agarose beads (50 % slurry, anti-FLAG® M2 affinity gel, A2220, Sigma) under gentle agitation at 4°C for 30 to 60 min. It is noted that a longer incubation time may increase the number of non-specific interactions and result in the pull-down of contaminant proteins. After incubation, the beads were gently washed five times with buffer D or buffer SL2 and resuspended in 10 µL of buffer D or buffer SL2, resulting in 100 µg/µL of total protein concentration. hDicer was eluted from the beads by site-specific cleavage using Tobacco Etch Virus TEV protease (0.05 U/µL) (V6101, Promega) at 30°C for 90 min. Alternatively, the proteins of interest (dmDicer-2 and TUT4) were eluted from the beads using 2 mM 3xFLAG® peptide (F4799, Sigma). The eluted proteins were supplemented with glycerol to a final concentration of 10 %, aliquoted and snap-frozen with liquid nitrogen for long-term storage at -80°C. The immunoprecipitates (IPs) were tested for the enrichment of the proteins of interest using western blot analysis, while the catalytic activities of the IPs were tested with bulk assays (data not shown).

5.5.5 Single-molecule pull-down

To increase purity of the IPs, an additional purification step was carried out directly on the surface of the imaging chamber using streptavidin or specific antibodies targeting the proteins of interest with nanomolar affinity range. This allowed for an efficient immobilization of the protein of interest, while discarding unwanted contaminant proteins (Figure 5.1). Single-molecule pull-down procedures are described case by case in the Results and Discussion sections.

5.5.6 Nucleic acids preparation: Stem-loop RNA

All of the RNA constructs used in this study were synthesized by ST Pharm Co., Ltd., South Korea. Precursor-microRNA (pre-miRNA) molecules were constructed by ligating two synthetic RNAs. First, a single-stranded RNA containing a 5′ phosphate and a half of the terminal loop of pre-miRNA (100 pmol, strand J in Table S5.1) was mixed with the 5′ strand that contained the other half of the terminal loop (200 pmol, strand K in Table S5.1). The mixture (20 µL) in TE buffer supplemented with 100 mM NaCl was annealed by heating it to 80°C, followed by a slow cooling down to 4 o C (-1°C/ 4 min in a thermal cycler). The annealed substrate was ligated using 3 µL of T4 RNA ligase (5 U/µL, AM2140, Invitrogen), 3 µL of 0.1 % BSA (AM2616, Ambion), 5 µL of the 10x ligation buffer provided, and 19 µL of H_2O at 16°C for 24 hrs. After acid phenol-chloroform extraction and ethanol precipitation, the RNA was purified with 12.5 % urea polyacrylamide gel.

The primary miRNA (pri-miRNA) substrate was constructed using the method described above. However, due to its length of 116 nucleotides (nt), pri-miRNA had to be ligated in two ligation steps. In the first ligation, a stem-loop structure was constructed (strands A and B in Table S5.1), followed by an additional ligation with a supplementary single-stranded RNA tail (strand C in Table S5.1) to obtain the full-length construct.

5.5.7 Nucleic acids preparation: Double-stranded RNA

First, two 70-nt ssRNA strands were constructed by ligating two synthetic RNAs with a DNA splint following the method described in section 2.5.1. The DNA splint was used to facilitate the ligation of the two RNA strands by T4 RNA ligase 2 (10 U/µL, M0239L, NEB). For the first strand, the ligation mixture contained a 34-nt Cy3-labeled RNA (200 pmol, strand D in Table S5.1), a 36-nt RNA containing a 5′ phosphate (200 pmol, strand E in Table S5.1) and a DNA splint (300 pmol, strand F in Table S5.1). For the second strand, the ligation mixture contained a 45-nt RNA with a 5′ phosphate (200 pmol, strand H in Table S5.1), a 25-nt RNA with a 5′ phosphate (200 pmol, strand H in Table S5.1) and a DNA splint (300 pmol, strand I in Table S5.1). After acid phenol-chloroform extraction and ethanol precipitation, both RNA strands were purified using a 10 % urea polyacrylamide gel. Both RNA strands were annealed following the method described in section 5.5.6 on page 166.

5.5.8 Nucleic acids preparation: DNA

The fluorescently labeled ssDNA (dA_{15}) was purchased from IDT DNA, USA.

5.5.9 Nucleic acids preparation: RNA labeling

All RNA strands were labeled with the NHS-ester form of cyanine dyes, Cy3 or Cy5, (GE Healthcare), with an almost 100% efficiency, as described elsewhere (Selvin & Ha, 2007). The positions of the labeled bases are indicated in Table S5.1.

5.5.10 Single-molecule fluorescence microscopy

A prism-type total internal reflection microscope was used for the single-molecule experiments. eGFP molecules were excited with a 473nm solid-state laser (OBIS LX 75 mW, Coherent), Cy3 molecules were excited with a 532nm solid-state laser (Compass 215M-50, Coherent), and Cy5 molecules were excited with a 632nm solid-state laser (25 LHP 928, CVI Melles Griot). To obtain the time traces, we excited eGFP, Cy3 and Cy5 molecules as weakly as possible to minimize their rapid photobleaching during the observation time. The fluorescence signals from single molecules were collected through a 60x water immersion objective (UPlanSApo, Olympus) with an inverted microscope (IX71, Olympus). To block 473nm, 532nm and 632nm laser scattering, we used a 473nm long-pass filter (Chroma), a 550nm long-pass filter (Chroma) and a 633nm notch filter (SemRock), respectively. Data were obtained in either single color or dual color mode. For dual color measurements, fluorescence signals were spatially split with a dichroic mirror (λ_{cutoff} = 645 nm, Chroma) and imaged onto two halves of an EMCCD camera (iXon 897, Andor Technology).

5.5.11 Microfluidic chamber preparation and immobilization schemes

To eliminate the nonspecific surface adsorption of proteins and nucleic acids to a quartz surface, piranha-etched slides (Finkenbeiner) were passivated with polyethylene glycol (PEG) over two rounds of PEGylation as described previously [38]. To further improve the surface quality for the experiments where crude cell extracts were used (Section 3.1), the assembled microfluidic flow chambers were incubated with 5 % Tween-20 (v/v in T50 buffer: 10 mM Tris [pH 8.0], 50 mM NaCl) for 10 minutes, followed by a washing step with 100 µL of T50 buffer [55]. Afterwards, slides were incubated with 50 µL of streptavidin (0.1 mg/mL, S888, Invitrogen) for 2 minutes followed by a washing step with 100 µL of buffer of interest. Biotinylated proteins were specifically immobilized by incubating the chamber with 50 µL of immunoprecipitated protein or crude cell extract (500x diluted in buffer D) for 5 or 0.5 minutes, respectively. The remaining unbound proteins were washed away with 100 µL of the buffer of interest supplemented with an oxygen scavenging system (0.8 % glucose (v/v), 0.1 mg/mL glucose oxidase (G2133, Sigma), 17 µg/µL catalase (Roche)). Oxygen scavenging system was used to reduce photobleaching and 1 mM Trolox (238813, Aldrich) was used to reduce photoblinking of the dyes [56]. The interaction between the biotinylated proteins and streptavidin was stable for several hours without any noticeable dissociation (data not shown).

Alternatively, when the proteins of interest were not biotinylated, commercially available biotinylated antibodies were used for specific immobilization. In brief, after flushing the unbound streptavidin away, the chamber was incubated with 50 µL of biotinylated-antibody (66 nM) for 5 minutes. The remaining unbound antibodies were washed away with 100 µL of buffer of interest, and 20-50 µL of diluted immunoprecipitated protein or crude cell extract was introduced to the microfluidic chamber. After 5 minutes of incubation, the unbound proteins were washed away with the buffer of interest supplemented with an oxygen scavenging system and Trolox.

5.5.12 Single-molecule data acquisition and analysis

A series of CCD images were acquired with lab-made software written in Visual C++ with a time resolution of 0.03 – 1 sec. Fluorescence images and time traces were extracted with programs written in IDL (ITT Visual Information Solutions) and analyzed with Matlab (MathWorks) and Origin (OriginLab Corporation). To systematically select single-molecule fluorescence signals of eGFP, Cy3 or Cy5 from the acquired images, we employed an algorithm written in IDL that searched for fluorescence spots with a defined Gaussian profile and with signals above a threshold. Apparent FRET (Förster resonance energy transfer) efficiency was defined as $I_A/(I_D+I_A)$, where I_D and I_A represent the donor (Cy3) and acceptor (Cy5) signals from two fluorescence spots from an identical same molecule, respectively.

5.6 Supplementary information

5.6.1 Supplementary tables

Table S5.1: DNA and RNA sequences

Name	Sequence (5′ → 3′)[a]
Pri-miRNA Drosha-DGCR8 Strand C was Cy5-labeled. (5.3.1 on page 156)	**A:** GAU ACU AUA CUG AGA GCA UUC CGU UAU GUA GCA UUU CUU GGU UGU GAG GGG UUG UGC
	B: AAG AAG AAU CUC ACG AUC AAG GAA UGC UAC AU
	C: AAC GGA G***u***G UUU GAG CAG ACC CGC GAC U
dsRNA Drosophila Dicer-2 Strand D was Cy3-labeled (5.3.2 on page 157)	**D:** AAG AAG AAU CUC ACG AUC AAG GAA UGC ***u***AC AUA A
	E: pCGG AGU GUU UGA GCA GAC CCG CGA UCU UUC AUU GCC
	F: CTC AAA CAC TCC GTT ATG TAG CAT TC
	G: pGGC AAU GAA AGA UCG CGG GUC UGC UCA AAC ACU CCG UUA UGU AGC
	H: pAUU CCU UGA UCG UGA GAU UCU UCU U
	I: ACG ATC AAG GAA TGC TAC ATA ACG GA
pre-let-7a-1 Human Dicer Strand K was Cy5-labeled. (5.3.3 on page 160)	**J: (5P strand):** UGA GGU AGU AGG UUG UAU AGU UUU AGG GUC ACA CC
	K: (3P strand): pCAC CAC UGG GAG AUA ACU AUA CAA UCU ACU GUC ***u***UU CU
Cy3-dA_{15} (5.3.4 on page 162)	**X:** Cy3–AAA AAA AAA AAA AAA

[a] p indicates phosphate; **u** represents dye-labeled nucleotide

Table S5.2: Antibodies used for immobilizing the proteins of interest

Antibody	Final concentration	Reference
anti-FLAG antibody	66 nM	F7425, Sigma
c-Myc antibody (A-14)	66 nM	sc-789, Santa Cruz Biotechnology
Biotin-SP (long spacer) AffinePure Goat Anti-Rabbit IgG (H+L)	66 nM	111-065-003, Jackson ImmunoResearch

5.7 References

1 J. Hang, R. Wan, C. Yan, Y. Shi, Structural basis of pre-mRNA splicing. *Science*. **349**, 1191–1198 (2015).

2 C. Yan *et al.*, Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science*. **349**, 1182–1190 (2015).

3 M. Ha, V. N. Kim, Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.* **15**, 509–524 (2014).

4 R. I. Gregory *et al.*, The Microprocessor complex mediates the genesis of microRNAs. *Nature*. **432**, 235–240 (2004).

5 J. Han *et al.*, The Drosha – DGCR8 complex in primary microRNA processing. *Genes Dev.*, 3016–3027 (2004).

6 T. P. Chendrimada *et al.*, TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*. **436**, 740–744 (2005).

7 A. D. Haase *et al.*, TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Rep.* **6**, 961–7 (2005).

8 K. Saito, A. Ishizuka, H. Siomi, M. C. Siomi, Processing of pre-microRNAs by the Dicer-1-Loquacious complex in drosophila cells. *PLoS Biol.* **3**, 1202–1212 (2005).

9 K. Förstemann *et al.*, Normal microRNA maturation and germ-line stem cell maintenance requires loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol.* **3**, 1187–1201 (2005).

10 N. R. Choudhury *et al.*, Trim25 Is an RNA-Specific Activator of Lin28a/TuT4-Mediated Uridylation. *Cell Rep.* **9**, 1265–1272 (2014).

11 P. Uetz *et al.*, A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. **403**, 623–627 (2000).

12 Y. Ho *et al.*, Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*. **415**, 180–3 (2002).

13 A.C. Gavin *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. **415**, 141–147 (2002).

14 N. J. Krogan *et al.*, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*. **440**, 637–643 (2006).

15 A.C. Gavin *et al.*, Proteome survey reveals modularity of the yeast cell machinery. *Nature*. **440**, 631–6 (2006).

16 L. Giot *et al.*, A Protein Interaction Map of Drosophila melanogaster. **302**, 1727–1737 (2003).

17 K. G. Guruharsha *et al.*, A protein complex network of Drosophila melanogaster. *Cell*. **147**, 690–703 (2011).

18 S. Li *et al.*, A map of the interactome network of the metazoan C. elegans. *Science*. **303**, 540–543 (2004).

19 S. Srihari, C. H. Yong, A. Patil, L. Wong, Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes. *FEBS Lett.* **589**, 2590–2602 (2015).

20 D. P. Maskell *et al.*, Structural basis for retroviral integration into nucleosomes. *Nature*. **523**, 366–369 (2015).

21 R. K. McGinty, R. C. Henrici, S. Tan, Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome. *Nature*. **514**, 591–6 (2014).

22 A. A. Hoskins *et al.*, Ordered and Dynamic Assembly of Single Spliceosomes Aaron. *Science*. **331**, 1289–1295 (2011).

23 M. L. Rodgers, J. Paulson, A. a Hoskins, Rapid isolation and single-molecule analysis of ribonucleoproteins from cell lysate by SNAP-SiMPull. *RNA*. **21**, 1031–1041 (2015).

24 R. Krishnan *et al.*, Biased Brownian ratcheting leads to pre-mRNA remodeling and capture prior to first-step splicing. *Nat. Struct. Mol. Biol.* **20**, 1450–7 (2013).

25 M. L. Kahlscheuer, J. Widom, N. G. Walter, Single-molecule pull-down FRET to dissect the mechanisms of biomolecular machines, *Methods in Enzymology*. **558**, (2015).

26 H. W. Lee *et al.*, Real-time single-molecule co-immunoprecipitation analyses reveal cancer-specific Ras signalling dynamics. *Nat. Commun.* **4** (2013)

27 H. W. Lee *et al.*, Real-time single-molecule coimmunoprecipitation of weak protein-protein interactions. *Nat. Protoc.* **8**, 2045–2060 (2013).

28 A. Jain *et al.*, Probing cellular protein complexes using single-molecule pull-down. *Nature*. **473**, 484–488 (2011).

29 A. Jain *et al.*, Stoichiometry and assembly of mTOR complexes revealed by single-molecule pulldown. **111** (2014), doi:10.1073/pnas.1419425111.

30 M. S. Panter, A. Jain, R. M. Leonhardt, T. Ha, P. Cresswell, Dynamics of major histocompatibility complex class I association with the human peptide-loading complex. *J. Biol. Chem.* **287**, 31172–31184 (2012).

31 Z. Shen *et al.*, Dynamic Association of ORCA with Prereplicative Complex Components Regulates DNA Replication Initiation. *Mol Cell Biol.* **32**, 3107–3120 (2012).

32 A. Jain, R. Liu, Y. K. Xiang, T. Ha, Single-molecule pull-down for studying protein interactions. *Nat. Protoc.* **7**, 445–452 (2012).

33 C. K. Means *et al.*, An entirely specific type I A-kinase anchoring protein that can sequester two molecules of protein kinase A at mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1227-35 (2011).

34 K. Yeom *et al.*, Single-molecule approach to immunoprecipitated protein complexes: insights intomiRNA uridylation. *EMBO Rep.* **12**, 690–696 (2011).

35 A. M. Denli, B. B. J. Tops, R. H. a Plasterk, R. F. Ketting, G. J. Hannon, Processing of primary microRNAs by the Microprocessor complex. *Nature.* **432**, 231–5 (2004).

36 D. Beckett, E. Kovaleva, P. J. Schatz, A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci.* **8**, 921–929 (2008).

37 E. de Boer *et al.*, Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7480–5 (2003).

38 S. D. Chandradoss *et al.*, Surface passivation for single-molecule protein studies. *J. Vis. Exp.* **50549**, 4–11 (2014).

39 T. A. Nguyen *et al.*, Functional Anatomy of the Human Microprocessor. *Cell.* **161**, 1374–87 (2015).

40 J. Han *et al.*, Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell.* **125**, 887–901 (2006).

41 R. Fukunaga *et al.*, Dicer partner proteins tune the length of mature miRNAs in flies and mammals. *Cell.* **151**, 533–546 (2012).

42 N. K. Sinha, K. D. Trettin, P. J. Aruscavage, B. L. Bass, Drosophila dicer-2 cleavage is mediated by helicase- and dsrna termini-dependent states that are modulated by loquacious-PD. *Mol. Cell.* **58**, 406–417 (2015).

43 M. Fareh *et al.*, TRBP ensures efficient Dicer processing of precursor microRNA in RNA-crowded environments. *Nat. Commun.* **7**, 13694 (2016).

44 S. Chakravarthy, S. H. Sternberg, C. A. Kellenberger, J. A. Doudna, Substrate-specific kinetics of dicer-catalyzed RNA processing. *J. Mol. Biol.* **404**, 392–402 (2010).

45 I. Heo *et al.*, Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell.* **151**, 521–532 (2012).

46 B. Kim *et al.*, TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms. *EMBO J.* **34**, 1801–15 (2015).

47 I. Heo *et al.*, Lin28 Mediates the Terminal Uridylation of let-7 Precursor MicroRNA. *Mol. Cell.* **32**, 276–284 (2008).

48 4I. Heo *et al.*, TUT4 in Concert with Lin28 Suppresses MicroRNA Biogenesis through Pre-MicroRNA Uridylation. *Cell.* **138**, 696–708 (2009).

49 J. P. Hagan, E. Piskounova, R. I. Gregory, Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* **16**, 1021–5 (2009).

50 H. Chang, R. Triboulet, J. E. Thornton, R. I. Gregory, A role for the Perlman syndrome exonuclease Dis3l2 in the Lin28-let-7 pathway. *Nature.* **497**, 244–8 (2013).

51 D. Ustianenko *et al.*, Mammalian DIS3L2 exoribonuclease targets the uridylated precursors of let-7 miRNAs. *RNA.* **19**, 1632–1638 (2013).

52 P. Munoz-tello, C. Gabus, Functional Implications from the Cid1 Poly (U) Polymerase Crystal Structure. *Structure.* **20**, 977–986 (2012).

53 J. F. Sambrook, D. W. Russell, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, ed. 3, 2001).

54 P. B. Stathopulos *et al.*, Sonication of proteins causes formation of aggregates that resemble amyloid. *Protein Sci.* **13**, 3017–27 (2004).

55 H. Pan, Y. Xia, M. Qin, Y. Cao, W. Wang, A simple procedure to improve the surface passivation for single molecule fluorescence studies. *Phys. Biol.* **12**, 45006 (2015).

56 I. Rasnik, S. a McKinney, T. Ha, Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat. Methods.* **3**, 891–893 (2006).

6

A fast and automated step detection method for analysing single-molecule trajectories

In preparation .

Luuk Loeff*, Jacob Kerssemakers*, Chirlmin Joo** & Cees Dekker**

* These authors have contributed equally to this work

** Co-corresponding authors

Kavli Institute of NanoScience and Department of BioNanoScience, Delft University of Technology, 2628 CJ, Delft, The Netherlands.

6.1 Abstract

Single-molecule techniques have made it possible to study the molecular dynamics of nucleic acids and proteins with a high spatial and temporal resolution. Accurate determination of the states in single-molecule events provides value information about the inherent kinetic properties of biomolecules. Here we present a fast and automated step detection method that is capable of detecting steps in large datasets without any prior knowledge on the distribution of step sizes or their location. Step detection is based on a series of partition events that minimize the variance between fit and the data (chi-squared). After each step fit, the quality of the fit is assessed by performing a secondary fit on the data. A multi-pass strategy that determines the optimal fit for the data over two rounds allows *Stepfinder* to automatically detect steps. The user-friendly interface and the automated step detection of the enhanced *Stepfinder* algorithm provides a robust "hands-off" fitting procedure that can be executed by anyone without programming knowledge in less than 10 minutes.

6.2 Introduction

Over the last two decades, fluorescence based single-molecule techniques have greatly enhanced our understanding of complex biological processes [1–5]. These techniques have made it possible to track the molecular dynamics of single proteins and protein complexes with a (sub)nanometer spatial resolution and a (sub)millisecond timescale [1, 2, 4]. Single-molecule fluorescence techniques have been used to determine the stoichiometry, binding kinetics and conformational dynamics of both nucleic acids and proteins [6–11]. Force spectroscopy (e.g. optical or magnetic tweezers) has been exploited as a versatile tool for probing the forces and motions that are associated with biological molecules [4, 12–14]. Lastly, nanopores have provided a powerful tool for label-free detection of nucleic acids and proteins [15–18].

Accurate determination of the states in single-molecule events provides valuable information about the kinetic properties that drive the biological function of proteins. The states in single-molecule fluorescence events often have to be analysed by a trained person, picking out each state manually. This manual analysis poses several drawbacks: (i) manual analysis relies on one's experience in distinguishing background noise from legitimate state-to-state transitions; (ii) it may be subject to a user bias; (iii) short events are likely to be missed; (iv) manually picking out events is a time-consuming process. These shortcomings affect both the reliability and reproducibility of data analysis and thereby manual analysis becomes almost impossible when the data exhibits a large number of states (e.g. more than three).

To reliably and reproducibly analyse data, several automated step detection algorithms have been developed over the past two decades [19–22]. Commonly used approaches for automated step detection rely on thresholding [21] or pairwise distribution analysis [23, 24], which do not suffice when the data exhibits more than three distinct steps of variable size. Another commonly used approach is based on Hidden Markov Modelling (HMM) [19], that requires specification of the number of different states that are visited during the time course of an experiment, which is commonly a unknown variable when doing experiments [25]. To work around this limitation, HMM can be combined with Bayesian nonparametrics [26], allowing one to use HMM without any knowledge on the number of visited states a priori [22, 27, 28]. However, HMM assumes that each state is visited successively, making the algorithm only suitable for events that occur frequently.

We previously reported on a step finding algorithm (*Stepfinder*) that is based on chi-squared minimization, capable of detecting steps in step trains without any prior knowledge on their size or location [20]. After its release, the chi-squared minimization algorithm gained a great interest in the field of biophysics [29] and has been applied on the analysis of trajectories of a wide variety of techniques. These techniques include: optical and magnetic tweezers, single-molecule FRET and nanopores (Figure 6.1) [20, 30–36]. Despite its popularity, the algorithm faced several caveats: (i) the algorithm was subject to user bias, requiring the user to determine final the number of steps (ii) the algorithm was computationally demanding when presented with large datasets (iii) step evaluation failed when presented with data that exhibited a large variety of step-sizes, which especially holds true for baseline type trajectories (iv) the algorithm lacked a user-friendly interface.

Here, we present an enhanced version of *Stepfinder* that allows for high-throughput and automated step detection (Chapter 3 on page 73). We created a user-friendly interface that simplifies step detection in single-molecule trajectories. First, we recapitulate how the step finding procedure is performed and how the optimal number of fitted steps is determined. Next, we elaborate on how the selection criteria for the optimal number of steps change when data exhibits a wide variety of step-sizes and plateau lengths. We show that these considerations lead to a robust "hands-off" fitting procedure that is suitable for single-molecule trajectories.

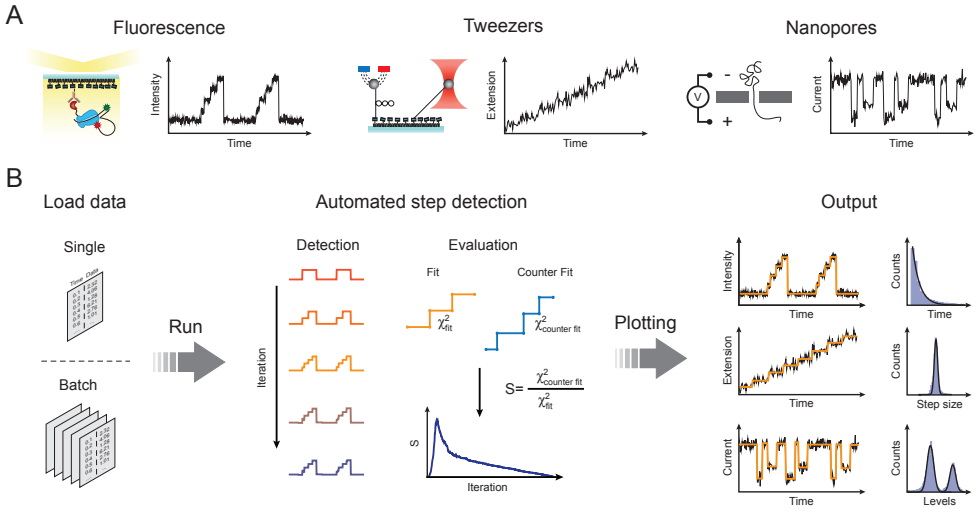


Figure 6.1: Workflow of the automated *Stepfinder*

(A) The stepfinder algorithm can be applied on a wide variety single-molecule of trajectories, including single-molecule fluorescence, magnetic & optical tweezers and nanopore data. **(B)** The algorithm requires input in the form of one or multiple .txt files with two columns (time and data). After pressing run, the algorithm iteratively adds single steps to the data that minimize χ^2. For each iteration, the quality is assessed by means of a secondary counter fit. Lastly, the best fit is selected and the algorithm outputs the corresponding fit, dwell-times, step sizes and levels. Fitting large data sets (>10^6 datapoints) can be done in less than 1 minute with a desktop computer.

6.3 Results

6.3.1 Overview of the procedure

The workflow for the *Stepfinder* algorithm is outlined in Figure 6.1. After single-molecule trajectories have been obtained, the step finding procedure can be divided in three basic steps: loading of the data, step detection and output of the result. The *Stepfinder* algorithm can be run on multiple files (batch mode) or on a single data file (Figure 6.1A). After loading the data, *Stepfinder* iteratively executes a series of partition events that allows the algorithm to determine the optimal fit for the data (Figure 6.1B). During each partition event the algorithm calculates chi-squared (χ^2), which provides a means to determine the variance between fit and the data. For each iteration, the next step is fitted at a location that yields the biggest reduction in χ^2. Subsequently, *Stepfinder* evaluates the quality of the fit by performing a secondary fit (called a counter fit) (Figure 6.1B) [20]. Once the optimal fit is determined the algorithm outputs several files that allow post-processing of the results (Figure 6.1B).

6.3.2 Step fitting

The *Stepfinder* algorithm fits data through a series of partition events that minimize chi-squared (χ^2). To fit data, the algorithm makes the sole assumption that the data contains steps with variable size (Δ) and plateau length (N) that are subject to noise (σ^2) (Figure 6.2A). The algorithm initiates the fitting procedure by splitting the data at a location that gives the lowest value of χ^2. This initial partition event generates a fit with two plateaus at a position that represents the average of the data points within the plateau (Figure 6.2A) [20]. After the first fit, the plateau that exhibits a step yielding the largest reduction χ^2 is selected for the next partition event, resulting in a fit with three plateaus (Figure 6.2A, dashed red line). The algorithm continues this process of adding a single-step to one of the plateaus for each iteration (Figure 6.2B, cyan arrow heads), until *Stepfinder* finds the user defined maximum number of steps.

Stepfinder successively selects a previously fitted step for the next partition event based on the biggest reduction in χ^2 (Figure 6.2A). By iteratively prioritizing the next fit that gives the biggest reduction in χ^2, the most prominent features of the data are fitted first followed by fits for the more refined features. As this process continues until the user defined number of steps are found, the number of step fits is likely to go beyond the 'optimal fit' (Figure 6.2B, middle). This result in 'over fitting', where new steps are fitted within the noise of the data (Figure 6.2B, bottom). To determine to optimal fit for a given dataset, it is important to evaluate the quality of the fit for every step that is added to the fit (Figure 6.1).

The quality of the existing fit is evaluated by performing a secondary fit for each iteration, hereafter called a counter fit [20]. *Stepfinder* generates counter fits by means of three steps: (i) *Stepfinder* first determines the next partition location (i_{next}) within each plateau (Figure 6.3A); (ii) next the algorithm rejects the existing step locations; (iii) *Stepfinder* builds a new fit based on the i_{next} locations, generating new plateaus with a position that represents the average of the data points within each plateau (Figure 6.3A). These three steps result in a counter fit with steps that are all

located in between the existing best-fit locations (Figure 6.3A). If the analyzed data does not display step-like behavior, both the existing fit and counter fit will have similar values of χ^2 [20]. However, when the data does display step-like behavior, counter fitting results in a fit that is much worse than the existing fit (Figure 6.3A) and thereby yields a larger value of χ^2 [20].

To evaluate the quality of the a fit, the *Stepfinder* algorithm takes advantage of the changing χ^2 landscape upon counter fitting. The quality of a fit (S), can be quantified by taking the ratio of the χ^2 from the existing fit and the counter fit, which is defined as:

$$S = \frac{\chi^2_{counter\ fit}}{\chi^2_{existing\ fit}}$$

If the existing fit is at the optimal number of iterations, the χ^2 of the existing fit approximates the noise in the data, whereas the χ^2 of the counter fit reaches its maximum value ($\sim\Delta^2/4\sigma^2$). Thereby, the maximum S-value (S^{max}) can be described by: $S^{max} = 1+P$, where P equals the maximum value of the counter fit ($\Delta^2/4\sigma^2$). The strong difference of χ^2 between the fit and the counter fit when an optimal number of iterations is reached, results in a S-value that is much larger than one (Figure 6.3). In contrast, when the data is under fitted, the χ^2 of the counter fit and existing fit approximate each other, resulting in a S-value that is close to one (Figure 6.3B).

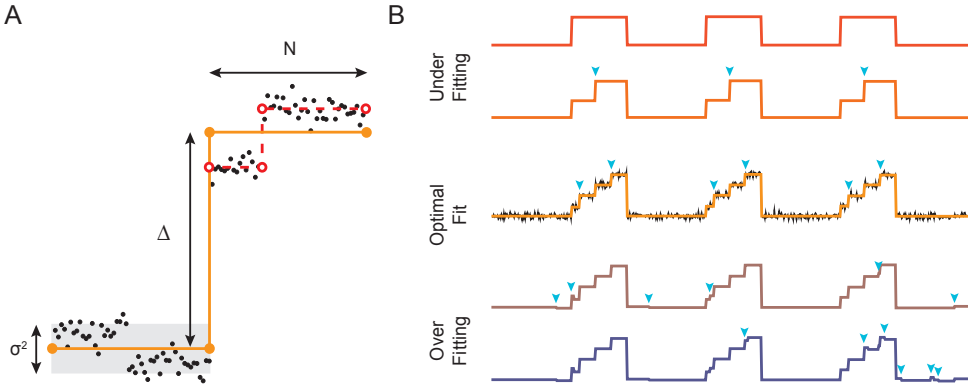


Figure 6.2: Global arrangement of the *Stepfinder* algorithm

(A) An example of an iterative step fit (orange line) on a single-molecule trajectory (black dots). Single-molecule trajectories are fitted by the *Stepfinder* algorithm by iteratively minimizing chi-squared (χ^2). To perform a step fit the program makes the assumption that the data contains steps (Δ), bounded by a plateau (N) that is subject to noise (σ^2, grey box). After the first step fit, *Stepfinder* selects the plateau with the largest value of χ^2, for the next partition event (red dotted lines). This process continues until the user defined number of steps is reached. **(B)** An example of the iterative process of step fitting by the *Stepfinder* algorithm. The algorithm successively adds a single step to the data (cyan triangles) and thereby minimizes χ^2. Step fitting below the optimal number of steps is considered under fitting, whereas step fitting beyond the optimal number of steps is considered over fitting.

Similarly, over fitting a dataset with steps that follow the noise, only results in a marginal change in the χ^2 of the counter fit (Figure 6.2B & Figure 6.3B). Therefore, the S-curve is a powerful indicator for the quality of the fit, displaying a sharp peak when the optimal fit is reached (Figure 6.3B).

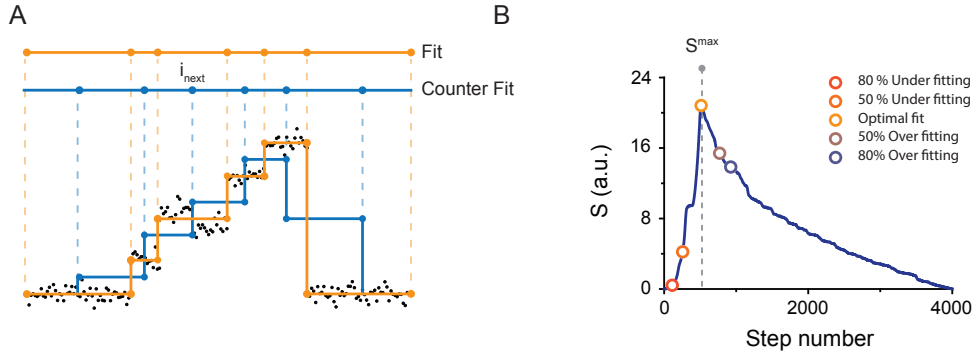


Figure 6.3: Determining the quality of a step fit

(A) For every step fit the algorithm performs, the quality of the fit (orange line) is evaluated by means of a secondary fit (blue line, called a counter fit). The counter fit is built by determining the next partition point (i_{next}), after which the current is rejected. Subsequently, the algorithm places the counter fit (blue) plateaus at a located within the existing fit (orange). **(B)** A representative example of an S-curve. The S value can be calculated by taking the χ^2 of the fit and dividing it by the χ^2 of the counter fit. When the existing fit is close to the optimal fit, the counter fit is at its worst, yielding a large value of S. However, when the data is over- or under fitted, little change in χ^2 is observed, resulting in S values close to 1. Thereby, the S-curve is a powerful predictor of the quality of the fit, exhibiting a sharp peak when the optimal number of step fits is performed.

6.3.3 A multi-pass strategy for automated step fitting

The S-curve is a robust measure to determine the quality of a fit, showing a distinct peak when the optimal number of iterations is reached. When the data exhibits steps that are in the same order of size and duration (e.g. Δ_1 or Δ_2, Figure 6.4A & Figure 6.4B), the optimal fit could be determined by finding the global maximum of the S-curve (S^{max}) (Figure 6.4D). However, this assumption cannot be made when the data exhibits steps that vary widely in size and duration (e.g. Δ_1 and Δ_2, Figure 6.4C). In this case, the S-curve exhibits a secondary peak (S_{P2}) that has a lower S_{P2}^{max} than the first peak (S_{P1}) (Figure 6.4D). Notably, the position of these peaks is identical to the peaks observed for a dataset with either Δ_1 or Δ_2 (Figure 6.4D). In this case, a significant portion of the refined steps would not be fitted, when the optimal fit could be determined by finding S^{max}.

To automate step detection, we developed a multi-pass strategy that determines the optimal fit for the data over two rounds. The *Stepfinder* algorithm first performs a step-fit, which yields a S-curve with a global maximum that corresponds to the most prominent features in the data. This step fit is then subtracted from the data and a secondary step-fit is performed on the 'residual data'. Only if the global maximum of the secondary step-fit is above the user defined threshold, coined acceptance

threshold, the fit will be accepted (Figure 6.6D). In summary, the multi-pass approach combined with the acceptance threshold on the second round of fitting provides a robust method for automated step detection.

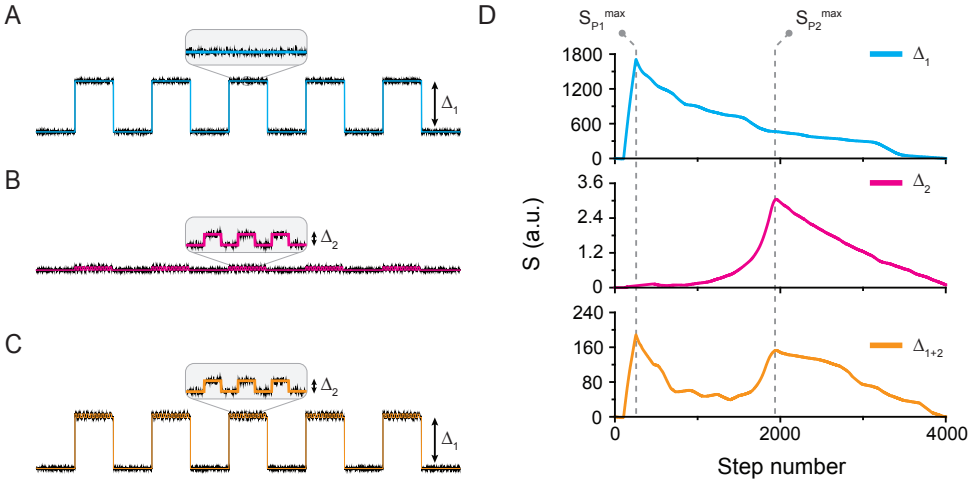


Figure 6.4: Multi-pass step detection to determine the optimal fit

(A) An example trace displaying uniform steps with a size of Δ_1. **(B)** An example trace displaying uniform steps with a size of Δ_2. **(C)** An example trace displaying non-uniform steps with a size of Δ_1 and Δ_2. **(D)** S-curves for the three example traces displayed in [A-C]. The global maximum of peak 1 (S_{P1}) and peak 2 (S_{P2}) are indicated with a dotted grey line. The S-curve for the data set with both large (Δ_1) and small (Δ_2) steps exhibits two peaks. If the small steps of Δ_2 are considered as an increased noise level of Δ_2, the global maximum of peak 1 (S_{P1}^{max}) can be described by $S_{P1}^{max}=1+P_1$, with P_1 is $\Delta_1^2/4(\sigma^2+1/4\Delta_2^2)$. This results in a S-curve with an S^{max} that is located at the same number of iterations as for a data set that exhibits only steps of Δ_1 or Δ_2, albeit lowered. However, if both the steps (Δ_1 and Δ_2) are fitted over the noise, the S-curve shows a secondary peak that can be described by: $S_{P1,2}^{max}=1+P_{1,2}$, with $P_{1,2}=\Delta_{1,2}^2/4\sigma^2$ and $\Delta_{1,2}$ is the weighted average of the large and the small step fractions.

6.3.4 An enhanced algorithm for automated step dectection

For each iteration, the *Stepfinder* algorithm selects an existing plateau (N_w) and splits it into a left (N_L) and right (N_R) plateau (Figure 6.5A). The position of these newly acquired plateaus is strongly dependent on the location of the partition point within N_w (Figure 6.5A). The average position (A) of a plateau (e.g. N_L) for any given location (i) can be described by:

$$A_L = \frac{1}{N_L}\sum_{i=1}^{N_L} x(i)$$

The iterative nature of determining this partition point, requires a substantial amount of computing power and becomes problematic when analyzing large datasets (e.g. >1·10^6 data points) (Figure 6.5B). Previously, *Stepfinder* determined the next partition point of N_w by calculating the χ^2 for all possible locations (i), selecting the step-fit that yields the

largest reduction in χ^2. However, this means that for a dataset with N_0 data points, the algorithm performs N_0^2 single x(i) operations to determine a single partition point. Next, the algorithm would repeat the same cycle to determine the partition point within the newly generated left (N_L) and right (N_R) plateau. With this scheme, it requires $2 \cdot (N_0^2/2)$ x(i) operations to locate the next two partition points. This cycle of partitioning continues plateaus until the algorithm find the user defined number of steps, which roughly scales with $2 \cdot N_0^2$ ((1+½+¼+…) $\cdot N_0^2$) operations per data set. Thereby, the required computing time significantly increases with an increase in the number of data points in a dataset (Figure 6.5B).

To reduce the (i) operations that are required to fit a dataset, we completely re-organized the code and streamlined the iteration process. A strong reduction in the number of required (i) operations can be made by re-using the information that is obtained during the localization of the first partition point. After the algorithm has determined the average (A_w) value of a plateau (N_w), the new version of *Stepfinder* determines the location of both N_L and N_R for x(i), using a single operation. The procedure starts with x(1) that is located at the left side of N_w (Figure 6.5A). The location (A_L) of N_L can be deduced by $A_L(i)=x(i)$, whereas the level of N_R is defined by:

$$A_r(\mathrm{i}) = \frac{(N_w \cdot A_w - x(i))}{(N_w - 1)}$$

This procedure is repeated for the next location (i+1) until each location of N_w is calculated, requiring only N_0 operations per plateau. For a whole data set this roughly scales with $2 \cdot N_0$, which is a gain of a factor of N_0 compared to the previous algorithm. Depending on the size of the analysed dataset, this improvement yields a speed gain of several orders of magnitude (Figure 6.5B). Notably, Additional speed can be gained by only saving a minimal pair of parameters for each step-fit operation.

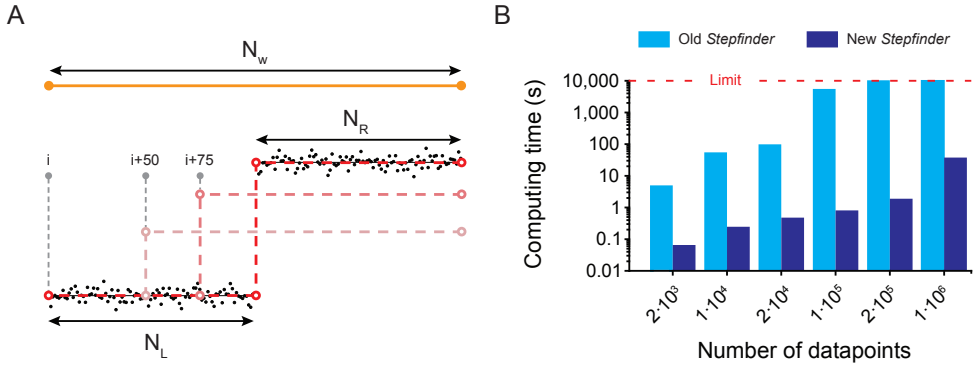


Figure 6.5: An enhanced *Stepfinder* for high-throughput step detection

(A) An example of the iterative nature of the step fit procedure. The existing plateau (N_w, orange line) is partitioned into two new plateaus (N_L and N_R, dark red dotted) at a point that yields the largest reduction in χ^2. To determine this partition point, the algorithm iteratively calculates χ^2 for each data point, starting at i until all data points of N_w have been calculated (e.g. i+50, faded red dotted lines). **(B)** Comparison of the old and new version of the *Stepfinder* algorithm. The algorithms were tested by measuring the computing time of various datasets on a desktop computer. The red dotted line indicates the limit that was set for the computing time.

6.3.5 Step fitting of experimental data

The newly developed automated multi-pass *Stepfinder* algorithm was applied on traces from the CRISPR-associated Cas3 helicase [37–40], which could not be analyzed with the previous version of *Stepfinder*. A detailed description on the experimental procedures and analysis are described in Chapter 3 on page 73. In brief, DNA bound Cas3 molecules were presented with ATP to initiate DNA unwinding. The fluorophores on the DNA substrate were able to report on DNA unwinding though an increase in FRET (Figure 6.6A). Before ATP was added, the labelling positions on the DNA yielded a FRET value that was indistinguishable from the background noise. Upon addition of ATP, a gradual increase in FRET was observed (Figure 6.6B).

The unwinding events of Cas3 were marked by plateaus (Figure 6.6B & Figure 6.6C), suggesting that Cas3 unwinds the DNA in discrete steps. Besides the increase in FRET that reports on unwinding, slipping events, in which the DNA abruptly moves backwards and reanneals (Figure 6.6C) were observed. The unwinding events using the automated multi-pass *Stepfinder* algorithm. The first round of the step fitting yielded a sharp peak in the S-curve (Figure 6.6D), whereas the second round of step fitting yielded a global maximum that was below the threshold. This indicates that the detected steps were in the same order of unity. A histogram of the FRET levels exhibited four equally spaced levels, suggesting that the helicase unwinds the DNA in discrete steps (Figure 6.6E) (Chapter 3 on page 73). This example shows that the enhanced version of the *Stepfinder* algorithm is able to automatically detect steps in baseline type trajectories without any prior knowledge on the number of states in the data.

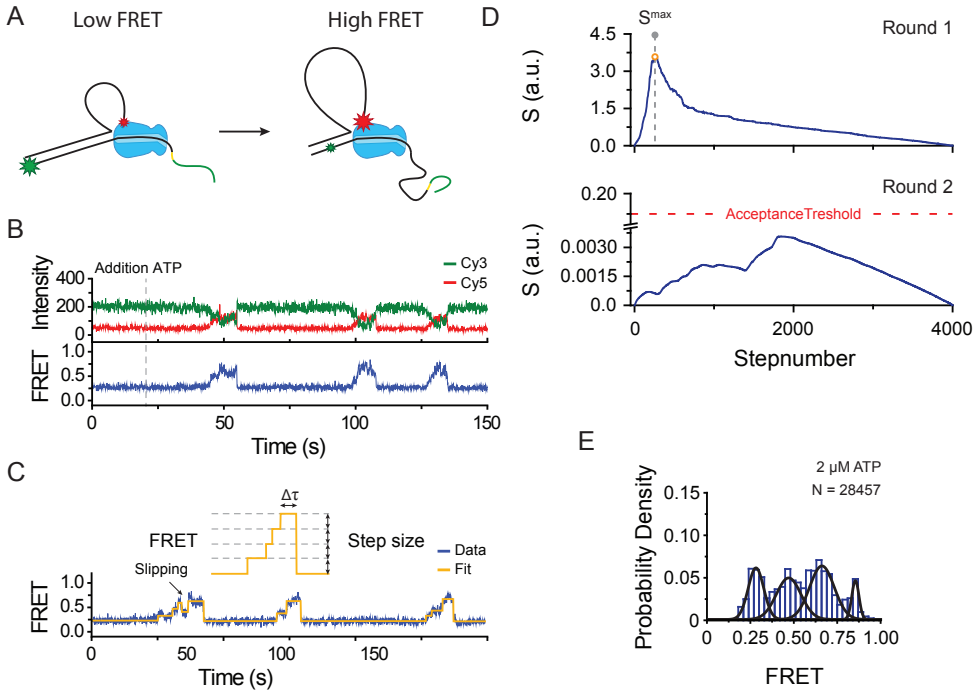


Figure 6.6: Application of the enhanced *Stepfinder* on experimental data

(A) Schematic of loop formation by the CRISPR-associated Cas3 helicase/nuclease protein (blue). The appearance of FRET during loop formation is indicated by the size of the star: low FRET, large green star) or high FRET, large red star. **(B)** A representative time trace of donor (Cy3, green) and acceptor (Cy5, red) fluorescence and corresponding FRET (blue) exhibiting multiple unwinding events. ATP (2 μM) was added at t = 20s (dashed gray line). **(C)** Representative FRET trace (dark blue) fitted with the enhanced *Stepfinder* algorithm (orange). **(D)** S-curve for two rounds of fitting on the dataset with unwinding events of Cas3. The global maximum of the S-curve for round two (bottom) was below the set acceptance threshold and therefore the second round of fitting was not executed. **(E)** Distribution of FRET levels obtained through the step-finder algorithm. Black lines represent a Gaussian fit.

6.4 References

1 C. Joo, M. Fareh, V. Narry Kim, Bringing single-molecule spectroscopy to macromolecular protein complexes. *Trends Biochem. Sci.* **38**, 30–37 (2013).

2 R. A. Forties, M. D. Wang, Minireview Discovering the Power of Single Molecules. *Cell.* **157**, 4–7 (2014).

3 M. F. Juette *et al.*, The bright future of single-molecule fluorescence imaging. *Curr. Opin. Chem. Biol.* **20**, 103–111 (2014).

4 K. C. Neuman, A. Nagy, Single-molecule force spectroscopy : optical tweezers , magnetic tweezers and atomic force microscopy. *Nat. Methods.* **5**, 491–505 (2008).

5 T. Ha, Single-molecule methods leap ahead. *Nat. Methods.* **11**, 1015–1018 (2014).

6 V. Aggarwal, T. Ha, Single-molecule fluorescence microscopy of native macromolecular complexes. *Curr. Opin. Struct. Biol.* **41**, 225–232 (2016).

7 M. Fareh *et al.*, Single-molecule pull-down for investigating protein-nucleic acid interactions. *Methods* (2015).

8 T. R. Blosser *et al.*, Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex. *Mol. Cell.* **58**, 60–70 (2015).

9 S. D. Chandradoss, N. T. Schirle, M. Szczepaniak, I. J. Macrae, C. Joo, A Dynamic Search Process Underlies MicroRNA Targeting. *Cell.* **162**, 96–107 (2015).

10 K. Yeom *et al.*, Single-molecule approach to immunoprecipitated protein complexes: insights intomiRNA uridylation. *EMBO Rep.* **12**, 690–696 (2011).

11 I. F. Gallardo *et al.*, High-Throughput Universal DNA Curtain Arrays for Single-Molecule Fluorescence Imaging. *Langmuir.* **31**, 10310–10317 (2015).

12 B. Sun, M. D. Wang, Single-molecule perspectives on helicase mechanisms and functions. *Crit. Rev. Biochem. Mol. Biol.* **9238**, 1–11 (2015).

13 T. Ha, A. G. Kozlov, T. M. Lohman, Single-molecule views of protein movement on single-stranded DNA. *Annu. Rev. Biophys.* **41**, 295–319 (2012).

14 R. Vlijm *et al.*, Nucleosome Assembly Dynamics Involve Spontaneous Fluctuations in the Handedness of Article Nucleosome Assembly Dynamics Involve Spontaneous Fluctuations in the Handedness of Tetrasomes. *Cell Rep.* **10**, 216–225 (2015).

15 S. J. Heerema, C. Dekker, Graphene nanodevices for DNA sequencing. *Nat. Nanotechnol.* **11**, 127–136 (2016).

16 C. Dekker, Solid-state nanopores. *Nat. Nanotechnol.* **2**, 209–215 (2007).

17 L. Ma, S. L. Cockroft, Biological Nanopores for Single-Molecule Biophysics. *ChemBioChem.* **11**, 25–34 (2010).

18 C. Plesa *et al.*, Fast Translocation of Proteins through Solid State Nanopores. *Nano Lett.* **13**, 11–16 (2013).

19 S. A. Mckinney, C. Joo, T. Ha, Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophys. J.* **91**, 1941–1951 (2006).

20 J. W. J. Kerssemakers *et al.*, Assembly dynamics of microtubules at molecular resolution. *Nature.* **442**, 709–712 (2006).

21 S. A. Mckinney, D. M. J. Lilley, T. Ha, Structural dynamics of individual Holliday junctions. *Nat. Struct. Mol. Biol.* **10**, 93–97 (2003).

22 I. Sgouralis, S. Presse, Biophysical Perspective An Introduction to Infinite HMMs for Single-Molecule Data Analysis. *Biophys. J.*, 2021–2029 (2017).

23 S. C. Kuo, J. Gelles, E. Steuer, M. P. Sheetz, A model for kinesin movement from nanometer-level movements of kinesin and cytoplasmic dynein and force measurements. *J. Cell Sci.* **14**, 135–138 (1991).

24 K. Svoboda, C. F. Schmidt, B. J. Schnapp, S. M. Block, Direct observation of kinesin stepping by optical trapping interferometry. *Nature.* **365** (1993).

25 M. Blanco, N. G. Walter, *Analysis of Complex Single-Molecule FRET Time Trajectories* (Elsevier Inc., ed. 1, 2010; http://dx.doi.org/10.1016/S0076-6879(10)72011-5), vol. 472.

26 T. S. Ferguson, A bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230 (1973).

27 K. E. Hines, A primer on bayesian inference for biophysical systems. *Biophys. J.* **108**, 2103–2113 (2015).

28 K. E. Hines, J. R. Bankston, R. W. Aldrich, Analyzing single-molecule time series via nonparametric bayesian inference. *Biophys. J.* **108**, 540–556 (2015).

29 B. C. Carter, M. Vershinin, S. P. Gross, A comparison of step-detection methods: how well can you do? *Biophys. J.* **94**, 306–319 (2008).

30 S. Myong, M. M. Bruno, A. M. Pyle, T. Ha, Spring-Loaded Mechanism of DNA Unwinding by Hepatitis C Virus NS3 Helicase. *Science.*, 513–517 (2007).

31 R. T. Dame, M. C. Noom, G. J. L. Wuite, Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature.* **444**, 387–390 (2006).

32 S. L. Reck-Peterson *et al.*, Single-Molecule Analysis of Dynein Processivity and Stepping Behavior. *Cell.* **126**, 335–348 (2006).

33 M. A. Beuwer, M. W. J. Prins, P. Zijlstra, Stochastic protein interactions monitored by hundreds of single-molecule plasmonic biosensors. *Nano Lett.* **15**, 3507–3511 (2015).

34 R. Vlijm, J. S. J. Smitshuijzen, A. Lusser, C. Dekker, NAP1-Assisted Nucleosome Assembly on DNA Measured in Real Time by Single-Molecule Magnetic Tweezers. *PLoS One.* **7**, 1–11 (2012).

35 B. T. Harada *et al.*, Stepwise nucleosome translocation by RSC remodeling complexes. *Elife.* **5**, 1–20 (2016).

36 H. Isojima, R. Iino, Y. Niitani, H. Noji, M. Tomishige, Direct observation of intermediate states during the stepping motion of kinesin-1. *Nat. Chem. Biol.* **12**, 290–297 (2016).

37 T. Sinkunas *et al.*, Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).

38 S. Mulepati, S. Bailey, In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA Target. *J. Biol. Chem.* **288**, 22184–22192 (2013).

39 Y. Huo *et al.*, Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**, 771–7 (2014).

40 R. N. Jackson, M. Lavin, J. Carter, B. Wiedenheft, Fitting CRISPR-associated Cas3 into the Helicase Family Tree. *Curr. Opin. Struct. Biol.* **24**, 106–114 (2014).

Summary

In this thesis, we adopted single-molecule fluorescence techniques to investigate how the CRISPR-Cas adaptive immune system of *Escherichia coli* mediates defense against invading DNA viruses. As described in chapter 1, the CRISPR-Cas immune system relies on Clustered regularly interspaced short palindromic repeats (CRISPR) and their CRISPR associated proteins (Cas). Chapter 1 provides a detailed overview of the processes that underlie CRISPR immunity, with a main focus on the type I-E system of *E. coli*. In brief, immunity is conveyed in three distinct steps. In the first step, called adaptation, small fragments of viral DNA are integrated into the genome of *E. coli*, resulting in memory formation. In the second step, called CRISPR RNA biogenesis, the viral fragments are transcribed and processed into short non-coding guide RNA molecules (crRNA). These non-coding RNA molecules assemble with Cas proteins to form RNA-guided effector complexes, commonly referred to as Cascade. In the last step of CRISPR immunity, called interference, Cascade complexes locate a complementary viral DNA target (called protospacer) and flag the target for degradation by the Cas3 protein.

To efficiently recognize viral DNA, the Cascade complex requires several sequence elements. For example, the first eight nucleotides (with exception of the sixth nucleotide) of the protospacer, or ''seed'' region, must be a perfect match for target recognition. Additionally, target recognition requires a trinucleotide protospacer adjacent motif (PAM) that is immediately neighboring the protospacer. The PAM sequence provides an important means to discriminate self from non-self DNA. Only when the PAM and the target DNA sequence elements are present the Cascade complex can flag the viral DNA for destruction by the Cas3 protein.

In chapter 2 we exploited single-molecule Förster Resonance Energy Transfer (FRET), to investigate how the Cascade complex coordinates the interaction between the PAM, seed and protospacer. We show that recognition of PAM, seed and protospacer is tightly controlled, allowing for target recognition with high-fidelity. While Cascade marks canonical targets for destruction, the complex is also involved in a process called priming that results formation of new memory against mutated targets that escape CRISPR immunity. Our data suggests that Cascade exhibits a non-canonical binding mode with low fidelity, that facilitates the recognition of mutated targets. Mutated targets are bound transiently in a PAM and seed-independent manner, which can occur from any segment of the RNA guide. This dual role of the Cascade complex with distinct fidelities underpin robustness to CRISPR-Cas immunity.

Once Cascade has bound a canonical target, the target is flagged for destruction by the Cas3 protein with both helicase and nuclease activities. In chapter 3 we investigated the mechanism of CRISPR interference using single-molecule FRET. We show that the Cascade complex and the Cas3 protein remain tightly associated while Cas3 unwinds the double stranded DNA helix, resulting in loops in the target

DNA. The DNA is reeled in distinctive burst of thee base pairs which each underlie three elementary steps of one nucleotide. Unwinding is highly repetitive, allowing Cas3 to compensate for its intrinsically inefficient nuclease domain. We reveal that the discontinuous helicase properties of Cas3 and its tight interaction with Cascade ensure well controlled degradation of target DNA only.

While prokaryotes exhibit an CRISPR adaptive immunity against invading genetic elements, eukaryotes have evolved an analogous system called RNA interference. Eukaryotic RNA interference uses small non-coding RNA molecules, to silence translation of both viral and endogenous RNA. MicroRNA's (miRNA) are the most abundant class of small RNA molecule in animals and play a critical role in regulating gene expression and cell differentiation. Biogenesis of these microRNA molecules requires multiple maturation steps, that include two endonucleolytic reactions. First, the Drosha protein cleaves a primary miRNA transcript to generate a hairpin shaped precursor microRNA molecule (pre-miRNA). This pre-miRNA is then matured by the Dicer protein into a double stranded miRNA duplex. This mature miRNA duplex is then loaded into the Argonaute protein to form an effector complex called RNA induced silencing complex (RISC).

Given the importance of miRNA molecules in gene regulation, means that the levels of miRNA molecules need to be tightly regulated. Generally, the RNA stability and function is controlled by posttranscriptional modifications that include RNA tailing. One of the most frequently found RNA tailing types is uridylation. Uridylation of RNA molecules is carried out by a group of non-canonical poly(A) polymerases (PAPs), called terminal uridylyl transferases (TUTases or TUTs). In humans, several TUTs have been described, each with distinct substrate specificity and functions. For example, oligo-uridylation of miRNA substates promotes their decay, whereas mono uridylation of specific miRNA substrates promotes their biogenesis. These examples show that TUTs play an essential role in controlling RNA stability through posttranscriptional modifications.

In chapter 4 we used biochemistry, single-molecule, and deep sequencing techniques to elucidate the mechanism by which human TUT7 protein, recognizes and uridylates pre-miRNA molecules of the let-7 family. We found that TUT7 recognizes the overhang structure of the pre-miRNA as a key structural element. By sensing the overhang structure, TUT7 preferentially uridyates pre-miRNA molecules with a truncated 3' end as well as canonical group II pre-miRNAs. Uridylation of these distinct substrates trigger opposing cellular responses. For example, TUT7 restores the functionality of group II pre-miRNAs with a 1-nt 3' through mono-uridylation. This results in a 2-nt 3' overhang that is required for efficient Dicer processing and thereby pre-miRNA biogenesis is facillitated. In contrast, pre-miRNA species with a trimmed 3' end are oligo-uridylated by TUT7 and thereby marked for degradation. Our single-molecule assay further revealed that TUT7 discriminates between substrates by interacting with them at different frequencies. This suggest that TUT7 uses a distributive mode of action for both uridylation pathways. Our study reveals dual roles and mechanisms of uridylation in repair and removal of defective pre-miRNAs.

Chapter 5 describes how single-molecule fluorescence can be combined with various pull-down techniques to obtain data from protein complexes. We provide different strategies and challenges that have to be over-come in order to implement these techniques. To showcase these challenges, we provide four examples of protein complexes that are involved in the biogenesis of RNAi molecules, including the Drosha-DGCR8 complex, the human Dicer-TRBP complex, the Drosophila Dicer 2-Loqs-PD complex and the TUT4-Lin28 complex. The combination of pull-down methods with single-molecule fluorescence allows for real time visualization of the interaction between RNA and protein complexes, with a sub-second and nanometer resolution. This single-molecule pull-down method can be applied on a wide variety of protein complexes that are essential for cellular processes.

Chapter 6, the last chapter of this thesis, describes a fast and automated step detection method for analyzing single-molecule trajectories. The *Stepfinder* algorithm is based on the minimization of chi-squared, which provides a means to determine the variance between the fit and the data. Once *Stepfinder* is loaded with data, the algorithm iteratively executes a series of partition events that allows the algorithm to determine the optimal fit. For each iteration, the quality of the fit is determined by means of a secondary fit, which is called a counter fit. A multi-pass strategy that determines the optimal fit for the data over two rounds, allowing the algorithm to automatically detect steps without any prior knowledge on their size and location. The combination of a user-friendly interface and the automated step-detection of the algorithm provides a robust and "hands-off" fitting procedure that can be executed by someone without programming knowledge.

The results presented in this thesis have contributed to a new level understanding on the molecular mechanisms behind CRISPR immunity and micro RNA biogenesis. Our single-molecule fluorescence approaches have revealed details that otherwise would have been masked by conventional biochemical approaches that average population dynamics. To conclude, we show that the type I-E CRISPR immune response of *E. coli* is highly dynamic, leading to robust immunity against invading mobile genetic elements. These complex dynamics were optimized through evolution, that was driven by the constant arms race between the host and the invaders. Like-wise, evolution has optimized RNA interference pathway in eukaryotes. By repurposing single proteins to convey distinct functional roles within the RNAi pathway, the costs associated with maintaining the RNAi pathway are reduced whereby the fitness of the cells is increased. It will be of interest to further study other CRISPR and RNAi related proteins using single-molecule techniques to obtain a deeper understanding on how the molecular dynamics of these systems drive their biological functions.

Samenvatting

In dit proefschrift hebben wij met enkele molecuul fluorescentie technieken onderzocht hoe het CRISPR-Cas adaptieve immuunsysteem van *Escherichia coli* zich verdedigd tegen invasieve DNA virussen. Zoals beschreven in hoofdstuk 1, maakt CRISPR-Cas immuunsysteem gebruik van een stuk DNA met de naam CRISPR (Clustered regularly interspaced short palindromic repeats) en een set eiwitten genaamd Cas (CRISPR geassocieerd). In hoofdstuk 1 worden de onderliggende moleculaire details van het CRISPR-Cas adaptieve immuunsysteem besproken waarbij de aandacht gevestigd zal zijn op het type I-E CRISPR-Cas systeem van *E. coli*. In het kort, immuniteit van worden verdeeld in drie verschillende stappen. Tijdens de eerste stap, genaamd adaptatie, wordt er een klein stukje van het virale DNA geïntegreerd in het genoom van *E. coli*. Hierdoor wordt er geheugen gevormd tegen het virus. Tijdens de tweede stap, genaamd CRISPR RNA biogenese, wordt het geheugen afgeschreven en verwerkt door de Cas eiwitten. Dit resulteert in korte niet coderende CRISPR RNA-moleculen (crRNA) die samen met de Cas eiwitten gebruikt worden voor het opsporen van het virale DNA in de cel. Het complex van de Cas eiwitten samen met het crRNA-molecuul wordt doorgaans Cascade genoemd. Tijdens de laatste stap, genaamd CRISPR-interferentie, gaat het Cascade complex opzoek naar het virale DNA in de cel. Zodra het Cascade complex een stuk viraal DNA heeft gevonden wordt er een extra eiwit bijgehaald met de naam Cas3 om het virale DNA af te breken.

Om het virale DNA efficiënt te kunnen herkennen maakt het Cascade eiwitcomplex van verschillende DNA-elementen. Zo moeten bijvoorbeeld de eerste acht nucleotide (met exceptie van nucleotide 6) van het virale DNA een perfecte gelijkenis hebben met het crRNA molecuul in het Cascade complex om het virale DNA te kunnen herkennen. Verder moet er direct naast het DNA doel nog een PAM-motief aanwezig zijn. De PAM-sequentie is drie nucleotiden lang en geeft aan of een stuk DNA van een virus of van de bacterie zelf is. Alleen wanneer er een PAM-sequentie aanwezig is en er een match is tussen het virale DNA en de crRNA in het eiwitcomplex kan het Cascade complex aangeven dat het DNA moet worden afgebroken door het Cas3 eiwit.

In hoofdstuk 2 hebben wij gebruik gemaakt van enkele-molecuul FRET (Förster Resonance Energy Transfer) om te onderzoeken hoe het Cascade complex de interactie tussen PAM, seed en de rest van het virale DNA doel coördineert. Wij laten zien dat de herkenning van de PAM tussen PAM, seed en de rest van het doel een stapsgewijs proces is. Hierdoor kan het Cascade eiwitcomplex deze doelwitten met hoge precisie herkennen. Naast dat het Cascade complex betrokken is bij het herkennen en markeren van gebruikelijke doelwitten. Is Cascade ook betrokken bij een proces, met de naam priming, waarbij er nieuw geheugen tegen gemuteerde doelwitten, die normaal gesproken aan het immuunsysteem ontsnappen, wordt gevormd. Onze data suggereert dat Cascade een secundaire niet gebruikelijke bindingsvorm bezit.

Deze bindingsvorm heeft een lage precisie en faciliteert daardoor de herkenning van gemuteerde doelwitten. Deze gemuteerde doelwitten worden kort gebonden in een PAM en seed onafhankelijke manier die van elk segment van het crRNA-molecuul kan voorkomen. Deze twee rollen van het Cascade complex zorgen ervoor dat het CRISPR-Cas immuunsysteem van *E. coli* erg robuust is.

Nadat Cascade een doelwit heeft gevonden, markeert het complex dit doelwit voor degradatie door het Cas3 eiwit. Het Cas3 eiwit bezit twee fundamentele eigenschappen. Zo is het eiwit in staat om de twee strengen waaruit DNA bestaat te ontwinden (helicase) en daarbij te degraderen (nuclease). In hoofdstuk 3 hebben we het mechanisme achter CRISPR-interferentie onderzocht door gebruik te maken van de enkele-molecuul FRET techniek. We laten zien dat het Cascade complex en het Cas3 eiwit met elkaar gebonden blijven terwijl het Cas3 eiwit het DNA ontwindt. Door deze hechte binding tussen het Cascade complex en het Cas3 eiwit worden er DNA lussen gevormd. Daarnaast laten we zien dat het DNA wordt ontwonden door Cas3 in stappen van 3 basen die elk drie onderliggende stappen van 1 nucleotide bevatten. Het ontwinden van het DNA is repetitief, waardoor Cas3 compenseert voor de lage knip activiteit van zijn nuclease domein. We laten zien dat de discontinuïteit in het ontwinden van DNA door het Cas3 eiwit en zijn hechte binding met het Cascade complex er voor zorgen dat alleen het doelwit wordt afgebroken.

Terwijl prokaryoten het CRISPR-Cas adaptieve immuunsysteem bezitten tegen invasieve virussen, hebben eukaryoten een analoog systeem geëvolueerd genaamd RNA interferentie. Eukaryote RNA interferentie is in staat om translatie van viraal en endogeen RNA te stoppen door gebruik te maken van kleine niet coderende RNA moleculen. MicroRNAs (miRNA) zijn de meest gevonden kleine niet coderende RNA moleculen in het dierenrijk en spelen een cruciale rol in de regulatie van gen expressie en cel differentiatie. MicroRNA moleculen worden gegenereerd middels twee knip stappen. De biogenese van miRNA moleculen begint met het knippen van een primair microRNA transcript door het Drosha eiwit. Deze knip stap genereerd een voorloper RNA molecuul (pre-miRNA) die de vorm aanneemt van een haarspeld. Deze zogenaamde pre-miRNA wordt vervolgens geknipt door het Dicer eiwit, waardoor er een dubbel-strengs RNA molecuul ontstaat. Dit dubbel-strengs RNA molecuul wordt vervolgens in het Argonaute eiwit geladen om de translatie van RNA te stoppen.

Omdat miRNA moleculen zo belangrijk zijn in de regulatie van gen expressie, moeten de hoeveelheid miRNA moleculen ook goed gereguleerd worden. Over het algemeen wordt de stabiliteit en de functie van RNA moleculen gereguleerd door het toevoegen van posttranscriptionele modificaties. Een voorbeeld van zo'n modificatie is het toevoegen van een RNA staart. Een van de meest voorkomende RNA staart modificaties is uridylatie. Uridylatie van RNA moleculen wordt uitgevoerd door een groep niet gebruikelijke poly(A)polymerases (PAPs), ook wel terminal uridylyl transferases (TUTases, TUTs) genoemd. In mensen zijn er zeven verschillende TUTs beschreven, die allemaal een eigen substraat specificiteit en functie hebben. Zo zorgt het toevoegen van een uridyl staart aan miRNA moleculen voor de afbraak

van miRNA voorlopers, terwijl het toevoegen van een enkele uridine nucleotide er voor dat de biogenese wordt bevorderd. Deze voorbeelden laten zien dat TUTs een essentiële rol spelen bij het reguleren van de RNA stabiliteit door het toevoegen van posttranscriptionele modificaties.

In hoofdstuk 4 hebben wij gebruik gemaakt van enkele-molecuul fluorescentie, sequencing en biochemische technieken om te onderzoeken hoe het humane TUT7 eiwit zijn pre-let-7 substraten herkent en uridyleerd. Uit het onderzoek is gebleken dat het uiteinde van de RNA haarspeld de bepalende factor is die wordt herkent door het TUT7 eiwit. Hierdoor heeft het TUT7 eiwit een voorkeur voor RNA moleculen met een verkort 3' einde en pre-miRNA moleculen die tot groep II behoren. De uridylatie van deze twee soorten substraten resulteert in verschillend gedrag. Zo wordt het 1-nt 3' uiteinde van groep II pre-miRNA moleculen hersteld door het toevoegen van een enkele uridine groep. Dit resulteert in een pre-miRNA met een 2-nt 3' uiteinde, wat essentieel is voor efficiënte verwerking door het Dicer eiwit. Daarentegen wordt er aan groep II pre-miRNAs met een verkort 3' uiteinde een uridine staart gemaakt. Door deze staart wordt de pre-miRNA gemarkeerd voor degradatie. Onze enkele-molecuul fluorescentie proeven lieten zien dat het substraat bepaald met welke frequentie het TUT7 eiwit een interactie aangaat. Dit suggereert dat TUT7 substraten uridyleerd middels een distributieve modus. Onze studie laat zien dat TUT7 twee rollen en mechanismen voor uridylatie bezit die betrokken zijn in het repareren en verwijderen van defectieve miRNA moleculen.

In hoofdstuk 5 wordt beschreven hoe eiwit complexen bestudeerd kunnen worden door enkele-molecuul fluorescentie te combineren met verschillende immunoprecipitatie technieken. We beschrijven verschillende strategieën en de uitdagingen die overwonnen moeten worden om deze technieken te implementeren. Aan de hand van vier voorbeeld eiwitcomplexen die betrokken zijn bij de biogenese van RNAi (Drosha-DGCR8 complex, human Dicer-TRBP complex, Drosophila Dicer 2-Loqs-PD complex en het TUT4-Lin28 complex), worden deze uitdagingen uitgelegd. De combinatie van immunoprecipitatie en enkele-molecuul technieken zorgen ervoor dat eiwitcomplexen bestudeerd kunnen worden met een sub-seconde en nanometer resolutie. Deze techniek kan worden toegepast op een grote variëteit aan eiwitcomplexen die essentieel zijn voor cellulaire processen.

Hoofdstuk 6, het laatste hoofdstuk van dit proefschrift, beschrijft een snelle en geautomatiseerde detectiemethode om enkele-molecuul data te analyseren. Het *Stepfinder* algoritme is gebaseerd op de minimalisering van de chi-kwadraattoets. De chi-kwadraatstoets is een middel om na te gaan of twee of meer verdelingen van elkaar verschillen. Nadat het algoritme de data heeft geladen, begint er een iteratief proces waarbij er steeds een extra stap aan de fit wordt toegevoegd. Door dit proces is het algoritme in staat de optimale fit voor een data set te bepalen. Na dit iteratieve proces, wordt de kwaliteit van de fit bepaald door een secundaire fit uit te voeren met de naam counterfit. Om de optimale fit te bepalen wordt er gebruik gemaakt van een multi-pass strategie, die de optimale fit bepaald in twee rondes. Deze strategie zorgt ervoor dat het algoritme automatisch de fit kan uitvoeren

zonder enige voorkennis over het aantal stappen en de grote van de stappen in de verdeling. De combinatie van de automatische stap detectie en de vriendelijke gebruiksomgeving zorgen ervoor dat het algoritme gebruikt kan worden door iemand zonder kennis van coderen.

De resultaten die in dit proefschrift worden gepresenteerd dragen bij aan een nieuw niveau van begrip over de moleculaire mechanismen die schuilgaan achter CRISPR-immuniteit en miRNA biogenese. Onze enkele-molecuul aanpak heeft ervoor gezorgd dat details zichtbaar zijn gemaakt die anders gemaskeerd zouden zijn in het populatie gemiddelde van de gebruikelijke biochemische analyses. In conclusie, we laten zien dat het type I-E CRISPR immuunsysteem van *E.coli* een erg dynamisch proces is, dat geoptimaliseerd is voor een robuuste immuun reactie tegen invasieve genetische elementen. Deze complexe dynamica zijn geoptimaliseerd door evolutie die gedreven is door wapenwedloop tussen de bacterie en de virussen. Eveneens heeft evolutie het RNA interferentie pad in eukaryoten geoptimaliseerd. Door eiwitten meerdere rollen te laten vervullen binnen de cel, worden de kosten verlaagd die hiermee gepaard gaan met het onderhouden van het eiwit, waardoor de vitaliteit van de cellen wordt verhoogd. Het zou interessant zijn om andere eiwitten die betrokken zijn bij CRISPR en RNA interferentie te bestuderen met enkele-molecuul technieken. Hierdoor zou men een dieper niveau van begrip kunnen krijgen over hoe de moleculaire dynamica van deze systemen ervoor zorgt dat hun biologische functie wordt vervuld.

Acknowledgements

The past few years have been a blast! It has truly been a great experience that I will cherish for the rest of my life. The fact that I'm writing the last and dearest section of this thesis, is a strange but at the same time a very rewarding feeling. During this period, I have grown scientifically, but more importantly I have grown as a person. In addition to that, I was able to travel the world and make bunch of new friends, what more can I ask for! I feel very fortunate to be surrounded by so many kind-hearted people, that each helped me in their own way. I'm sure that without their help and guidance this thesis would not have existed.

I would like to start by expressing my gratitude to my copromotor and mentor Chirlmin. You were the key player in making this story a success. When I just joined your lab, it immediately struck me that you were a very knowledgeable and pragmatic person. Your humble attitude is reflected in the way you run the lab, providing a nurturing environment that I felt comfortable in and allowed me perform at the best of my ability. You have always given me a lot of freedom and support to explore my ideas and hypotheses, something I should not take for granted. I'm grateful that have always assessed my work critically, keeping me on my toes and simulating me to go the extra mile. Besides the guidance on my projects, I appreciate that you have also involved me in many other aspects of running a lab, which I know will be useful in the future. From time to time you still surprise me how well you know my ways and habits, showing that apart from scientific knowledge you also possess great people skills. You have been a great mentor and I hope I can continue to count on your advice even once I have left Delft.

Cees thank you for agreeing to be my promotor. Despite our limited interaction over the last couple of years, I can see you created a close-knit environment in your lab, where young and enthusiastic scientists flourish. You have been able to translate these same principles throughout the department, creating a vibrant and fun atmosphere to work in. An achievement you can be extremely proud of.

Many thanks to my defence committee, who have sacrificed their precious time to evaluate my PhD thesis. Eui-Jeon Woo, unfortunately we have not interacted much during your last visit to Delft, as I was too busy wrapping up the last chapters of this thesis. Thank you making your way to Delft once again. Blake Wiedenheft, I'm honoured to have a CRISPR pioneer in my committee. We have crossed paths several times at the CRISPR meetings, where you always amazed me with beautiful crystal structures. Johannes Hohlbein, our interaction has been limited, however, the communication we had so far has been joyful. Looking forward to hear your perspective as biophysicist. Marileen Dogterom, thank for keeping our department running and freeing time for this defence in your very busy schedule. Andreas Engel, your endless passion for science is admirable.

Stan, apart from being a member of my defence committee, we have also intensively collaborated for the past four years. You provided a refreshing and different take on the results that we obtained through our single-molecule measu-

rements, pushing our projects forward. Your down to earth attitude makes you very approachable, allowing me to just drop by your office and discuss whatever is on my mind, something I have always enjoyed. Thank you for the continuous support and mentorship both scientifically and personally.

When I just joined BN, it soon became clear that the world of nano-science is drastically different to what I had been exposed to before. With very limited knowledge of physics, I wasn't sure how I was going to get through these four years. Luckily, I received a warm welcome from many of you. Mohamed, Stanley, Michaela, Anna, Ilja, Jetty, Margreet, Mahipal, Zoreh and Pauline I can't thank you enough for making me feel at home right from the start.

Soon after, it became clear that the secretaries and support staff are the lifeline of this department. Dijana, Jolijn, Amanda, Emmylou, Chantal, Liset, Esther, Angela and Marije, thank you for seamlessly running events within our department. I'm sure that BN would fall apart without your support. Many thanks to Anna, Margreet, Pawel, Sacha, Erwin, Andrea, Marek, Theo, Ilja, Esengül, Eve, Anne, Jaco, Theo, Roland, Inge, Jan and Anke, who were always there to answer questions or to lend a helping hand when needed. Sacha special thanks to you, I have always enjoyed our conversations. Jelle thank you for providing our lab with parts for our microscopes. Your experience and perfectionistic attitude are reflected in your solutions.

Thanks to all the members of the Joo lab. I was fortunate enough to be surrounded by a great bunch of colleagues/ friends, that each added their own flavour to the group. We have had many fun parties, drinks and trips, I will miss those for sure.

Tim, I'm extremely grateful that I was able to join you on the CRISPR project. When I joined the project my knowledge on CRISPR was limited, I was impressed & slightly overwhelmed by the complexity of the system. Yet, you had the patience to answer all my questions and guide me through the project. You are an extremely knowledgeable person and I have learned a whole lot from working with you. Maybe the most valuable thing you taught me is to take a step back and think through all the possibilities, instead of just blindly rushing into the next experiment. I'm glad we are still in touch from time-to-time, now that you are overseas. A trend that hopefully continues in the future!

Malwina, thank you for providing loads of support, laughs, coffees and a listening ear over the past years. These few lines are not enough to express my graditude. I really appreciate that you were always there to help me, even if it meant you had to drop you own work. Thank you for putting me on my place, when I was bluntly expressing my opinion. It is reassuring to know that I can count on your support, in good and but also in bad times. I'm happy that you found a new challenge in Portland. I owe you big time for flying over!

Laura, I'm very happy to have you by my side as paranimf during this event. When you joined the lab, I was immediately struck by your enthusiasm and positive attitude. Thank you for spreading your Latino vibes through the lab and introducing

us to aguardiente. It's great that I can discuss anything with you, whether they are personal matters or just cracking jokes. Good luck with the last leg of your PhD, I'm sure the future will hold great things for you and Taylor!

Mohamed, thank you for taking me under your wing when just joined the lab. You were always prepared to lend a helping hand and took your time to explain techniques. It is great to know I can always count on your advice and support, scientifically and personally. Looking back, there are so many great moments to remember, from our trip to Switzerland, our kitesurfing course and the dinners you hosted at your home. Hopefully, you will soon find a place to start your own lab, I'm sure you will be a great PI.

Stanley, thank you for tolerating my Dutch directness in the office. Thanks for the advice during the last few months of my PhD, reassuring me that everything will work out fine. We have always had good banter in the office, which I enjoyed a lot. I'm glad that you managed find a position in Oxford that suits your interests.

Anna, thank you for being at the heart of our lab for so long. You spoiled us by taking care of all the organisational aspects of running the lab, something I took for granted. Now that you have moved to another lab, it becomes clear what a crucial role you played over these years. Many times, the two of us found ourselves alone in the lab, which led to many, laughs, conversations and "complaining" sessions. I really appreciate your down to earth and hands-on mentality.

Viktorija, my fellow CRISPRer and travel buddy! I have many good memories of our abroad adventures. Many thanks for all the laughs, beers and the nice conversations we shared during these years. I appreciate you are always in for a party! Good luck with the last leg of your PhD, I'm sure everything will work out well!

Thijs, I'm glad we ended up in the same office. It is great to see how you have grown and matured over the last year. Before you know it, you will transition to the "grandpa" of the lab. Thanks for all the foosball games, banter, and fun we had. I'm curious to see what the future will hold for you.

Jetty, kudos for battling through your PhD, setting up a lab and the sequencing project was a daunting task. I'm glad that you found a position that suits you at Erasmus. If I ever have patent related questions, I know where to find you!

Sung-hyun, it was great to have you around in the lab. Thanks for all the cups of coffee, scientific advice and good conversations. Hopefully, you will soon start your own lab in Korea. / Pawel, thank you for bringing your no-nonsense attitude to the lab. I appreciate how you always have your heart on your sleeve. I'm happy you managed to find a position in industry! / Mike, you brought a wind of change in the lab, organising many social events. Keep lighting up those coals and light a new tradition of lab barbeques! I'm sure you will soon get a grip on the sequencing project! / Sungchul, the biochemisty guru of the lab! Hopefully, we can push some projects together in the near future. / Seung hwan, Iasonas and Ivo, thank you for all the good conversations and fun both in and outside of the lab.

Michaela and Mathia, to me it kind of feels the two of you were part of the Joo lab all along. Michaela, I really admire how you have ploughed through the last part of your PhD, I know it wasn't easy but you stuck with it till the end! Thank you for all

the fun and support your provided over the last years. I'm glad you found yourself a position that you enjoy in a beautiful and especially warm country. Best of luck to you and Ciro.

Mathia, thank you for being my paranimf and supporting me during my defence. I deeply respect how you have kept your head up high when life threw you a massive curve ball. With such an attitude, you are bound to succeed in life, no matter what you will do. Best of luck with the last period towards your defence.

A big thank you to all the other great people that BN is home to. All the Christmas parties, kavli days and day to day interactions with each of you made this an unforgettable experience. Unfortunately, there is not enough space and time to write each of you a personal message. Yet, I would like to highlight a couple of people: Jacob, I feel fortunate to have worked with you. Your endless enthusiasm for the Stepfinder algorithm is contagious. I admire how you have managed to combine the two things you love, science and art. / Misha, thanks for all the help on bootstrapping, fitting and discussions on CRISPR. Keep bridging the gap between theory and experiments! / Alicia, I'm sure you will find a way to combine synthetic biology with therapeutics. I'm happy we survived our flipped boat! Good luck with the final stages of your PhD. / Sumit and Sam, thanks for all the foosball games. Sumit you always amaze me with your outlandish banter. / Benjamin, thank you for dragging me (and others) along in your training regime. / Carsten, good luck with the remainder of your PhD. / Stephanie, I admire how you kept battling graphene. Thanks for the open and honest conversations we had on our trip to Courchevel. / Louis and Boijk, thanks for the pre- and after work kitesurfing sessions, they help me keeping my mind in order! / Vanessa, kudos for your persistence on your project. Best of luck in Sweden. / Martin, you are one of the most approachable PIs in the department. Thanks for sharing your views on how to run a lab and the advice on statistics. / Ganji, thanks for the discussions on single-molecule FRET and the banter. I'm curious to see what your next step will be! / A big thank you to, Maarten, Zoreh, Pauline, David, Felix, Fabai, Jelmer, Calin, Charl, Adi, Daniel, Becca, Jorine, Orkide, Afshin, Yoones, Seb, Fede, Helena, Nicole, Jonas, Johannes, Lisa, Seungkyu, Sergii, Jochem, Dominic, Nuria, Fabrizio, Anthony, Richard, Kuba, Siddharth, Allard, Artur, Mehran and others, without you the department would not have been the same!

Also, outside of Delft I obtained lots of support from friends and family. Pap, mam ondanks het niet helemaal duidelijk is wat ik allemaal in het lab uitspook hebben jullie mij volop gesteund de afgelopen jaren. Zonder jullie onvoorwaardelijke steun en liefde weet ik niet waar ik was gestrand. Een ding is zeker, dit boekje was er niet geweest. Daan, als klein jongetje keek ik al tegen je op. Ik probeerde je altijd na te doen. Je liet zien dat het mogelijk was om onder aan de ladder te beginnen en vervolgens naar het hoogste niveau te klimmen. Bedankt dat je altijd voor me klaar staat, wat er ook gebeurd. Grootmoeder, ook u wil ik bedanken voor alle steun, ik ben blij dat u deze mijlpaal nog mag meemaken. Robert en Lucie, bedankt dat jullie me altijd zo gesteund en ondersteund hebben. Dit was een extra stimulans mijn best te blijven doen en niet op mijn lauweren te rusten. Robert, dank voor al je goede adviezen, deze waardeer ik altijd erg.

Berend bedankt voor de broodnodige afleiding en steun de afgelopen jaren. De talloze weekenden gevuld met lekker eten, biertjes, muziek, slap geouwehoer en af en toe een serieus gesprek, is iets waar ik altijd weer naar uit kijk. Hopelijk zullen er nog veel van dit soort avonden volgen!

Oleg/ Cheffie, tof om te zien dat het leven in de keuken je zo goed af gaat. Ik ben benieuwd wat de toekomst je zal brengen. Je enthousiasme voor het bakken van brood werkt aanstekelijk. Bedank voor je gegeit, biertjes in de duinen en breek de week "eet bananzas". Een traditie die we hopelijk nog lang vol houden.

Ilja, ook jij hebt ervoor gezorgd dat ik de afgelopen 4 jaar heb overleefd. Dank voor al je steun en rake woordgrappen! Ik ben blij dat ik je al meer dan 20 jaar m'n vriend mag noemen.

Naast de personen hierboven zijn er vele andere die aan dit proefschrift hebben bijgedragen. Helaas is er te weinig tijd en ruimte om iedereen uitgebreid te bedanken. Toch zou ik een aantal mensen kort willen uitlichten: Jan, Ron, Maya, Lucie, Jeroen en Jelle, we hebben afgelopen jaar veel meegemaakt. Jullie steun heeft veel voor mijn betekend in deze turbulente tijd. / Lise a.k.a. Marabira, bedankt voor de winterse avonturen de afgelopen jaren. Hopelijk houdt de traditie stand! / Stefan en Nick, ik kijk altijd weer uit naar onze zaterdag middag chill sessies. / Sander, ook de classic tijden op de Meppelweg hebben bijgedragen aan dit proefschrift. / Joris, thanks voor alle kite sessies, deze zorgde voor de nodige afleiding! / Tim, ik ben blij dat onze vriendschap al meer dan 20 jaar standhoudt. Ondanks we elkaar (te) weinig zien, is het altijd weer vanouds! / Outlet Crew, thanks voor alle classic feestjes. / Dick nogmaals dank voor de hulp met de vormgeving. /

Mierelle ik voel me bevoorrecht dat ik het leven mag delen met zo'n lieve en sterke vrouw. Zonder jouw eindeloze liefde en steun was dit boekje er zeker niet geweest. Bedankt dat je me door dik en dun steunt en zoveel begrip toont, zo sta je voor me klaar ook als ik humeurig thuis kom na een lange dag, geef je aan wanneer het tijd is om te stoppen, en hoor je mijn frustraties aan. Het afgelopen jaar hebben we veel voor onze kiezen gekregen, ik ben trots dat we dit samen hebben doorstaan. Ik hoop dat we nog vele jaren en avonturen mogen delen. Het leven is namelijk een stuk leuker met jou aan mijn zijde. Ik hou van je!

Luuk Loeff
Den Haag, August 2017

Curriculum vitae

Luuk Loeff

14 January 1989	Born in Den Haag, The Netherlands
2001 - 2006	Higher General Secondary Education Aloysius College, Den Haag, The Netherlands
2006 - 2010	B.Sc. in Biology and Medical Laboratory Research Hogeschool Leiden, The Netherlands
2010 - 2012	M.Sc. in Biomolecular Science (*cum laude*) Vrije Universiteit Amsterdam, The Netherlands
2012 - 2017	Ph.D. in Biophysics Title "Microbial Warfare: Illuminating CRISPR adaptive immunity using single-molecule fluorescence" Promotor: Prof. Dr. C. Dekker Co-promotor: Dr. C. Joo Department of Bionanoscience Technical University Delft, The Netherlands

List of publications

8. **L. Loeff**, S. J. J. Brouns & C. Joo, "Single-molecule FRET for CRISPR studies", Bio-Protocols, *In Preparation*

7. **L. Loeff** & C. Joo, "How prokaryotes mediate CRISPR adaptive immunity", Springer Book: "Biophysics of RNA-Protein Interactions", *In Preparation*

6. **L. Loeff***, J. Kerssemakers*, C. Joo & C. Dekker, "A fast and automated step detection method for analysing single-molecule trajectories", *In Preparation*

5. **L. Loeff**, S. J. J. Brouns & C. Joo, "The CRISPR associated Cas3 protein repetitively probes the target DNA with a 1-nt step size", *Under Revision*

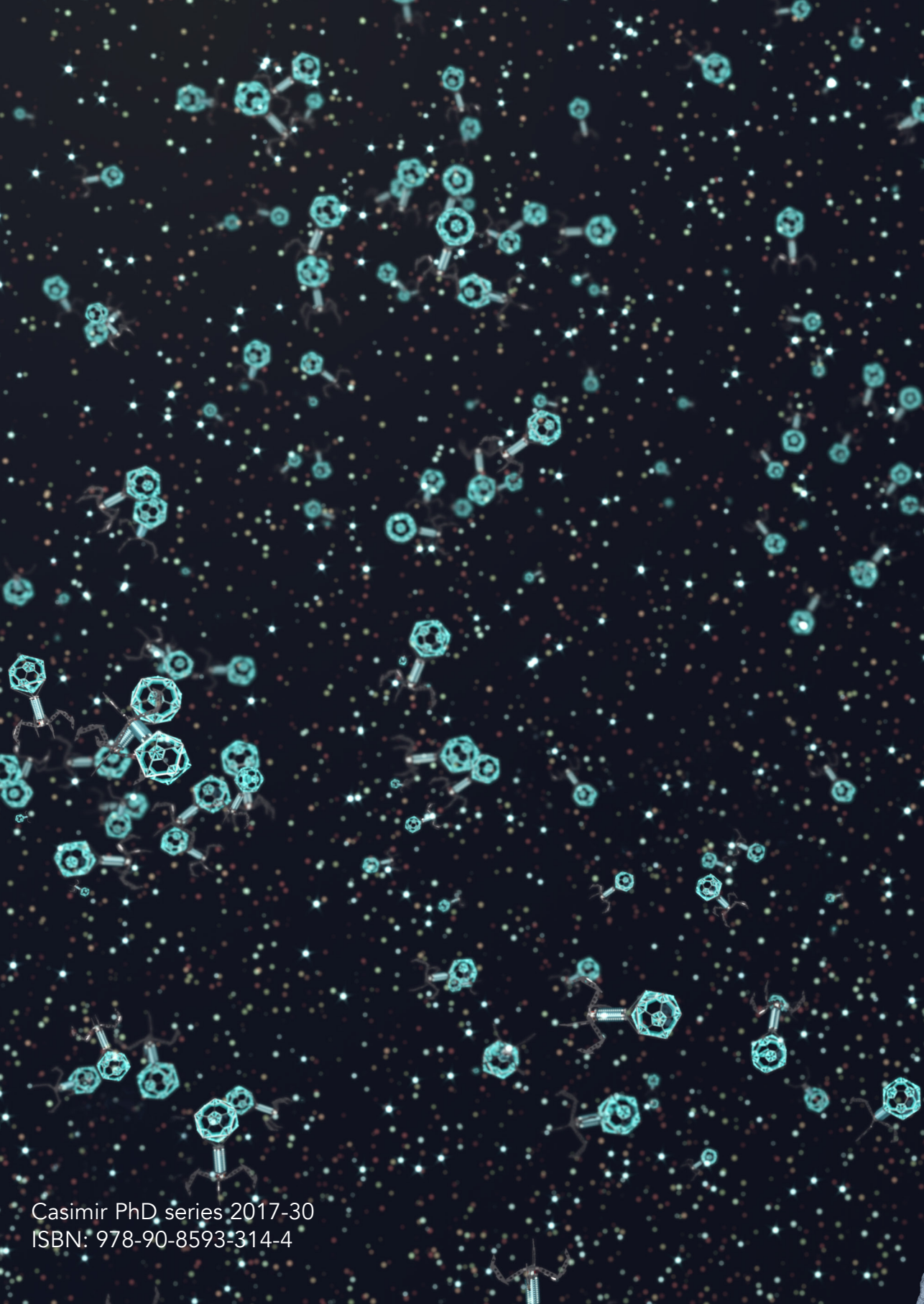
4. M. Fareh, **L. Loeff**, M. Szczepaniak, A. C. Haagsma, K. Yeom & C. Joo, "Single-molecule pull-down for investigating protein–nucleic acid interactions", *Methods*, 105:99-108, (2016)

3. B. Kim*, M. Ha*, **L. Loeff***, H. Chang, D. K. Simanshu, S. Li, M. Fareh, D. J. Patel , C. Joo, & V. N. Kim, "TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms", *EMBO Journal*, 34(13):1801-15, (2015)

2. T. R. Blosser*, **L. Loeff***, E. R. Westra, M. Vlot, T. Künne, M. Sobota, C. Dekker, S. J. J. Brouns & Chirlmin Joo, "Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex", *Molecular Cell*, 58(1):60-70, (2015)

1. N. A. Binai, M. M. M. Bisschops, B. van Breukelen, S. Mohammed, **L.Loeff**, J. T. Pronk, A. J. R. Heck, P. Daran-Lapujade & M. Slijper, "Proteome adaptation of Saccharomyces cerevisiae to severe calorie restriction in retentostat cultures", *Journal of proteome research*, 3(8):3542–3553, (2014)

* Denotes equal contribution



Casimir PhD series 2017-30
ISBN: 978-90-8593-314-4