Imperial College London
Department of Medicine

Single-molecule studies of CRISPR/Cas9 off-target activity

Matthew D. Newton

Thesis submitted for the degree of Doctor of Philosophy
December 2020

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Declaration of Originality

I hereby declare that all text, data, results and analysis contained in this thesis are my own except for the data presented in the first half of Chapter 4, Figures 4.1 - 4.5 and Figure 4.10. This is also highlighted in the relevant figure captions. This data forms the basis of my single-molecule FRET work on the binding stability of hybrid DNA guides as part of a collaborative project with Ben Taylor's group at AstraZeneca. It is included as it provides the vital background into the design and characterisation of the hybrid guides which I then used in my single-molecule assays.

In addition, some of smFRET data included in Chapter 5 Figures 5.8 and 5.10 as well as some of the cleavage assay data in Figures 5.9 and 5.16 were obtained by Shenaz Allyjaun, a Masters student who I was supervising in the lab at the time.

I further declare that all sources, text and figures are appropriately cited and copyright permission for reuse of figures obtained wherever required.

Abstract

The discovery of CRISPR/Cas9 as an easily re-programmable endonuclease has heralded a new era of genetic manipulation. With it comes the prospect of novel gene therapy approaches, and the potential to cure previously un-treatable genetic diseases. However, reports of unintended off-target editing by CRISPR/Cas9 pose a significant hurdle to realising this potential. In this thesis I use single-molecule techniques to investigate this off-target activity of CRISPR/Cas9. First I demonstrate how specificity can be improved by modulating the hybridisation energy of guide and target DNA - through incorporation of DNA bases into the RNA guide molecules. Next I investigate how force applied to target DNA alters Cas9 specificity, demonstrating that disruption of the canonical DNA double-helix structure dramatically increases Cas9 promiscuity, allowing cutting at sites with as many as 50 % of bases mismatched. Finally, I extend this finding by showing that the presence of supercoiling in target DNA also has a significant impact of specificity, with negative supercoiling resulting in the induction of thousands of potential off-target sites across human genomic DNA sequences. Together this work suggests that *in vivo* processes which destabilise the DNA structure have the potential to induce off-target Cas9 activity at previously overlooked sites. I propose that rather than considering Cas9 targeting specificity as an intrinsic property of the protein complex, it is vital that we consider the structural state of DNA it is targeting. These findings will be important for on-going efforts to engineer Cas9 variants with higher specificity, in the endeavour to develop safe and efficient gene therapies.

Acknowledgements

I would first like to acknowledge Prof. David Rueda for taking me into his lab and mentoring me over these past four years. Also thank you to Dr. Benjamin Taylor and Dr. Emanuela Cuomo at AstraZeneca for entrusting me with this project and providing your invaluable knowledge and expertise along the way. It has been a truly great experience working in this lab and on this project. I am also very grateful for the opportunities which have been made available to me in regard to my PhD project but also networking and my scientific development.

Next I would like to thank all the amazing members of the Rueda lab, past and present, who have made the lab such a fun and friendly place to work and have also always enthusiastically provided their help and expertise whenever I needed it. To Shenaz Allyjaun, Ondrej Belan, Dr. Kotryna Bloznelyte, Dr. Martin Billman, Emily Byrd, Dr. Adam Cawte, Elena Crameri, Dr. Todd Fallesen, Maru Fuentes-Perez, Dr. Hailey Gahlon, Dr. Paul Girvan, Dr. Mohammed Ghoneim, Dr. Artur Kaczmarczyk, Dr. Haruki Lino, Federica Marchesin, George Moore, Paul Poudevigne-Durance, Quentin Smith, Dr Sheila Xie and Dr. Tianyi Zhang. In particular Adam, Maru and Paul who showed me the ropes when I first arrived in the lab. Thank you to all those who work at the LMS who have helped me with this project. In particular thank you the the genomics facility, Ivan Andrew, Jaspreet Haywood and Laurence Game. Thank you also to the team at AstraZeneca in Gothenburg who were willing to take a punt on our ideas and helped me with the CIRCLE-seq supercoiling experiments, in particular Pinar Akcakaya and Burcu Bestas.

I would like to thank the team at Lumicks, in particular Rosalie Driessen, Jordi Cabanas-Danés and Willem Peutz who were so welcoming and patient when I first turned up fresh faced in Amsterdam clutching some samples. They were instrumental in getting this project off the ground so I am very grateful for their amazing knowledge and support.

I would also like to thank my year of PhD students who have become a wonderful group of friends, without whom the last four years would have been so much more difficult and less fun. I wish you all the best as we move on to bigger things.

Thank you to Dr. Sveta Makovets for her belief in me, and who taught me how to pipette and gave me the best possible foundation.

Thank you to thank all my friends and family for all their love and support over the past four years. Finally, thank you to the amazing Maddie Hickish for putting up with me during the writing of this thesis and always being so wonderful and loving.

"The power to control our species' genetic future is awesome and terrifying. Deciding how to handle it may be the biggest challenge we have ever faced."

Jennifer Doudna, 2017

'When I described catching living things with light people said, 'Don't exaggerate Ashkin'!'

Arthur Ashkin, 2018

Contents

Abstract **7**

Acknowledgements **9**

1 Introduction **25**

1.1 CRISPR in prokaryotic immunity 25

1.1.1 The discovery of prokaryotic adaptive immunity 25

1.1.2 Classification of CRISPR-Cas Systems 29

1.2 Gene-editing using CRISPR 31

1.2.1 The CRISPR/Cas9 system 32

1.2.2 Gene editing with CRISPR/Cas9 35

1.3 CRISPR/Cas off-target activity 40

1.3.1 Detection of off-target activity 40

1.3.1.1 *In vitro* off-target Detection 42

1.3.1.2 *In vivo* Off-target Detection 43

1.3.1.3 Verification of *in vivo* targets, VIVO 46

1.3.2 Algorithms for off-target prediction 47

1.3.3 Engineering of high fidelity CRISPR/Cas Variants 48

1.3.3.1 Guide engineering 49

1.3.3.2 Protein engineering 50

1.4 Single-molecule studies of CRISPR 53

1.5 Thesis Objectives 60

References 62

2 Background Theory **73**

2.1 The single-molecule approach 73

2.2 TIRF-smFRET 75

2.2.1 TIRF Microscopy 76

2.2.2 Förster resonance energy transfer 78

2.3 Optical tweezers 80

2.3.1 Optical trapping 81

2.3.2 Biophysical properties of DNA 84

2.4 Single-molecule techniques to study CRISPR/Cas9 off-target activity 86

References 87

3 Materials and Methods **92**

3.1 DNA and RNA synthesis, purification and labelling 92

3.1.1 Purification by denaturing polyacrylamide gel electrophoresis, PAGE . . 92

3.1.2 In vitro transcription of tracrRNA 92

3.1.3 Labelling of DNA and RNA 93

3.1.4 HPLC Purification 93

3.2 TIRF smFRET 94
3.2.1 TIRF Microscope 94
3.2.2 Slide preparation 94
3.2.3 Calibration of Cy3 and Cy5 channel 95
3.2.4 smFRET experiments 96
3.2.5 Data acquisition and analysis 96
3.3 Optical tweezer experiments 97
3.3.1 Biotinylation of Phage-λ DNA 97
3.3.1.1 Tortionally unconstrained DNA 97
3.3.1.2 Tortionally constrained DNA 97
3.3.2 Cleaning and passivation of the fludics and flow cell 99
3.3.3 Cas9 optical tweezer experiments 99
3.3.4 Data analysis 99
3.4 CIRCLE-seq with supercoiling 100
3.4.1 Preparation of relaxed and supercoiled genomic DNA circles 100
3.4.1.1 Genomic DNA preparation and fragmentation 100
3.4.1.2 End-repair and A-tailing 101
3.4.1.3 Adapter ligation 101
3.4.1.4 Fragment circularisation 102
3.4.1.5 Negative supercoiling of DNA circles 103
3.4.2 Cas9 *in vitro* digestion 103
3.4.3 Sequencing library preparation 103

3.4.4 Sequencing and Analysis . . . 104

References . . . 105

4 Mapping the sugar dependency for rational generation of a DNA-RNA hybrid-guided Cas9 endonuclease 106

4.1 Abstract . . . 107

4.2 Results . . . 108

4.2.1 Substitution of RNA with DNA in guides . . . 108

4.2.2 crDNA binds tracr:RNA with high affinity . . . 110

4.2.3 Hybrid guides have reduced off-target activity *in vitro* . . . 110

4.2.4 Hybrid guides reduce DNA binding . . . 111

4.2.5 Hybrid guides destabilise R-loop formation . . . 113

4.2.6 Hybrid guides fail to induce efficient target cleavage in cells . . . 119

4.3 Discussion . . . 121

4.4 Supplementary Figures . . . 124

4.5 Supplementary Tables . . . 125

References . . . 125

5 DNA stretching induces CRISPR/Cas9 off-target activity 126

5.1 Abstract . . . 127

5.2 Results . . . 128

5.2.1 Cas9 binds low force stretched DNA tightly and specifically . . . 128

5.2.2 Specifically bound wtCas9 cleaves DNA and remains bound across the DSB 129

5.2.3 DNA stretching induces off-target binding . . . 131

5.2.4 Off-target binding occurs at non-random, crRNA sequence-dependent sites 132

5.2.5 Two DNA strands are required for induced off-target binding 133

5.2.6 Off-target sequence identification 135

5.2.7 DNA bubbles induce Cas9 off-target cleavage 139

5.2.8 In vivo off-target sites require melting for cleavage with Cas9 141

5.3 Discussion 143

5.4 Supplementary Figures 146

References 146

6 DNA supercoiling induces CRISPR/Cas9 off-target activity 151

6.1 Abstract 152

6.2 Results 153

6.2.1 Negatively supercoiling DNA induces Cas9 off-target binding 153

6.2.2 Induced off-target sites correlate with negative supercoiling 155

6.2.3 Supercoiling induced off-targets correspond to force induced off-target sites 158

6.2.4 Force stretching negatively supercoiled DNA induces further off-target sites 161

6.2.5 Negative supercoiling induces off-target cleavage across human genomic DNA 162

6.3 Discussion 167

References 169

7 Conclusion 170

7.1 Thesis Discussion 170

7.2 Future Work 173

References 174

A Appendix **175**

A.1 Off-target Prediction Script 175

A.1.1 createWindows.R 175

A.1.2 findPAMs.R 178

A.1.3 offtargetpredict.Rmd 179

List of Tables

1.1 AAV Serotypes 37

2.1 Typical values of corner frequency, f_0, and trap stiffness for a 4.42 μm diameter bead, at 25 °C based on a viscosity, η, of 1 mPa.s 83

2.2 Parameters of dsDNA and ssDNA 86

3.1 DNA sequences for tracrRNA In vitro transcription 93

3.2 DNA sequences for torsionally constrained DNA 98

4.1 crXNA Sequences 108

4.2 AAVS1 crXNA Sequences 113

4.3 crXNA Sequences 125

5.1 Predicted Off-target Sequences 137

5.2 Previously validated *in vivo* EMX1-1 off-target sites 142

6.1 Predicted Off-target Sequences 160

List of Figures

1.1 **Early observations of tandem repeats**. 27

1.2 **General mechanism of prokayrotic adaptive immunity**. 28

1.3 **Classification of CRISPR/Cas Systems**. 30

1.4 **Double-strand break repair pathways**. 32

1.5 ***Streptococcus pyogenes* CRISPR/Cas9**. 34

1.6 **Comparison of different form of Cas9 for delivery into cells**. 36

1.7 **Non-viral Delivery Methods**. 38

1.8 **Mapping the effect of mismatches on editing efficiencies**. 41

1.9 ***In vitro* genomic methods for detecting off-target activity of CRISPR/Cas**. 43

1.10 ***In vivo* methods for detecting off-target activity of CRISPR/Cas**. . . . 45

1.11 **Engineering Cas9 guides for higher specificity**. 50

1.12 **Engineering of high fidelity CRISPR/Cas Variants**. 52

1.13 **Visualising Cas9 targeting with DNA curtain**. 54

1.14 **Cas9 R-loop formation observed by smFRET**. 56

1.15 **Observation of Cas9 HNH domain conformation changes on binding to on-target and mismatched target DNA by smFRET**. 57

1.16 **Kinetic model of off-target discrimination**. 59

2.1 **Principles of TIRFM**. 76

2.2 **Fluorescent characteristics of Cy3 and Cy5 FRET pair**. 79

2.3 **Principles of optical trapping**. 82

2.4 **DNA force-distance plot**. 84

3.1 **smFRET-TIRFM Slide Construction**. 95

3.2 **Biotinyating DNA**. 98

3.3 **Microfluidics layout for optical tweezer experiments**. 100

4.1 **crXNA *in vitro* cleavage**. 109

4.2 **Systematic substitution of remaining RNA bases within crHyb to generate minimal RNA containing guide, crHybΔ5**. 109

4.3 **crXNA guides bind to Cas9:tracrRNA complex with high affinity**. . . . 110

4.4 **Hybrid guides have reduced off-target activity *in vitro***. 111

4.5 **Hybrid guides have reduced DNA target binding affinity**. 113

4.6 **Hybrid guide smFRET experimental design**. 114

4.7 **Representative smFRET trajectories and time binned histograms for crRNA:tracr:Cas9**. 115

4.8 **crXNA time-binned FRET histograms**. 116

4.9 **Representative smFRET trajectories for crD.Hyb:tracr:Cas9**. 118

4.10 **Induced InDel frequencies of crXNAs in cells**. 120

4.11 **BLI DNA binding assay in the absence of heparin**. 124

5.1 **dCas9 binds DNA tightly and specifically at low force**. 129

5.2 **wtCas9 cleaves DNA in the presence of Mg^{2+}.** 130

5.3 **Force induces off-target binding of Cas9.** 132

5.4 **Off-target binding locations are non-random and sequence dependant.** 134

5.5 **Off-target binding requires both DNA strands.** 135

5.6 **Scoring potential off-target sites and comparison with experimental data.** 136

5.7 **Design of smFRET Cas9 assay.** 138

5.8 **DNA bubbles induce binding at predicted off-target sites.** 139

5.9 **DNA bubbles induce binding at predicted off-target sites.** 140

5.10 **DNA bubbles induce binding at predicted off-target sites.** 141

5.11 **Previously validated EMX1-1 *in vivo* off-target sites require DNA destabilisation for Cas9 cleavage.** 142

5.12 **Cas9 complex binding requires crRNA, tracrRNA and Cas9.** 146

5.13 **Labelling of crRNA with Cy5 does not effect Cas9 cleavage efficiency.** 147

5.14 **Force dependant dwell-time analysis of off-target binding.** 148

5.15 **Off-target binding locations are guide sequence dependant.** 149

5.16 **Off-targets with 6 nt PAM bubbles are not cleaved by Cas9.** 150

6.1 **Negative supercoiling induces Cas9 off-target binding.** 154

6.2 **Induction of off-target sites correlates with degree of negative supercoiling.** 156

6.3 **Off-target sites induced by force are also induced by supercoiling.** . . . 158

6.4 **A simple model predicts supercoiling induce off-target sites.** 159

6.5 **Force stretching DNA induces further off-target binding in the presence of supercoiling**. 161

6.6 **Negative supercoiling induces off-target cleavage across human genomic DNA**. 163

6.7 **Negative supercoiling increases mismatch tolerance**. 164

6.8 **Distribution of mismatches across off-target sites**. 166

7.1 **Theoretical binding energy landscape**. 172

Chapter 1

Introduction

1.1 CRISPR in prokaryotic immunity

1.1.1 The discovery of prokaryotic adaptive immunity

Bacteria and viruses have been engaged in an evolutionary arms race for billions of years. As a result, it is perhaps not surprising that there is an incredible diversity of bacterial immune systems which offer protection against potential infection from bacteriophages. The first line of defence for many bacteria are mechanisms which directly prevent phage adsorption and DNA injection (Josephsen and Neve, 2004). However, if these fail, or in species without these mechanisms, there are further lines of defence directed against the injected foreign DNA. In 1962, restriction enzymes were demonstrated to offer bacteria protection against invasive foreign DNA sequences (Arber and Dussoix, 1962; Dussoix and Arber, 1962). Individual restriction enzymes were subsequently purified and characterised, leading to the discovery of a class of restriction enzymes capable of recognition and cleavage of DNA at specific sequences (Smith and Welcox, 1970; Kelly and Smith, 1970). A vital component of restriction-based bacterial immunity is the ability to distinguish self from non-self, ensuring that only foreign DNA sequences are targeted and that the host genome is left unharmed. For the first restriction enzymes identified this was

shown to be achieved by corresponding methlytransferases, which methylate DNA at any target sites found on the host genome, protecting them from cleavage (Roy and Smith, 1973a; Roy and Smith, 1973b). Over the last 50 years, hundreds of restriction enzymes and corresponding methyltransferases have been identified in both bacteria and also archaea. Together these are capable of recognising and cleaving invading DNA at hundreds of distinct sequences in order to prevent infection.

However, it has now been demonstrated that this is not the only line of defence that bacteria and archaea have against invading DNA. Towards the end of the 1990s, as sequencing of genomic DNA became increasingly feasible, three separate groups reported the presence of interesting repeated sequences in archaea and bacteria (Figure 1.1) (Ishino et al., 1987; Hermans et al., 1991; F. J.M. Mojica, Juez, et al., 1993). Initially termed TREPs (tandem repeats), these differed from other repetitive sequences in that the short sequence of DNA between consecutive repeats was always the exactly the same length, and appeared to show no similarity to any other DNA sequence present in the host genome (F. Mojica et al., 1995). As further prokaryotic genomes were sequenced TREPS, next renamed as SRSRs (short regularly spaced repeats), were identified to be prevalent across diverse lineages (F. J. M. Mojica, Diez-Villasenor, et al., 2000). This finding hinted at an important but unknown biological significance. As the field moved forward these repeats were finally given the name CRISPRs, clustered reguarly interspaced short palindromic repeats, and CRISPR-associated genes were identified, termed *cas* genes (Jansen et al., 2002). Despite many interesting initial hypotheses, the biological function of CRISPRs was not elucidated until the surprise finding of a spacer sequence within an *E. coli* CRISPR array, which appeared to originate from a bacteriophage genome (Francisco J.M. Mojica et al., 2005). This lead to the identification of further bacteriophage sequences within CRISPR arrays, and the discovery that this formed the basic of a prokaryotic adaptive immune system, storing a record of previous infection which prevented future infection by the same invading species (Pourcel et al., 2005; Bolotin et al., 2005; Barrangou et al., 2007).

There are two distinct inital stages to the CRISPR adaptive immune system, adaption and interference (Figure 1.2). During the adaption stage (Figure 1.2 (1.)), spacer sequences are obtained from the DNA of invasive species and added into the CRISPR array. On re-exposure

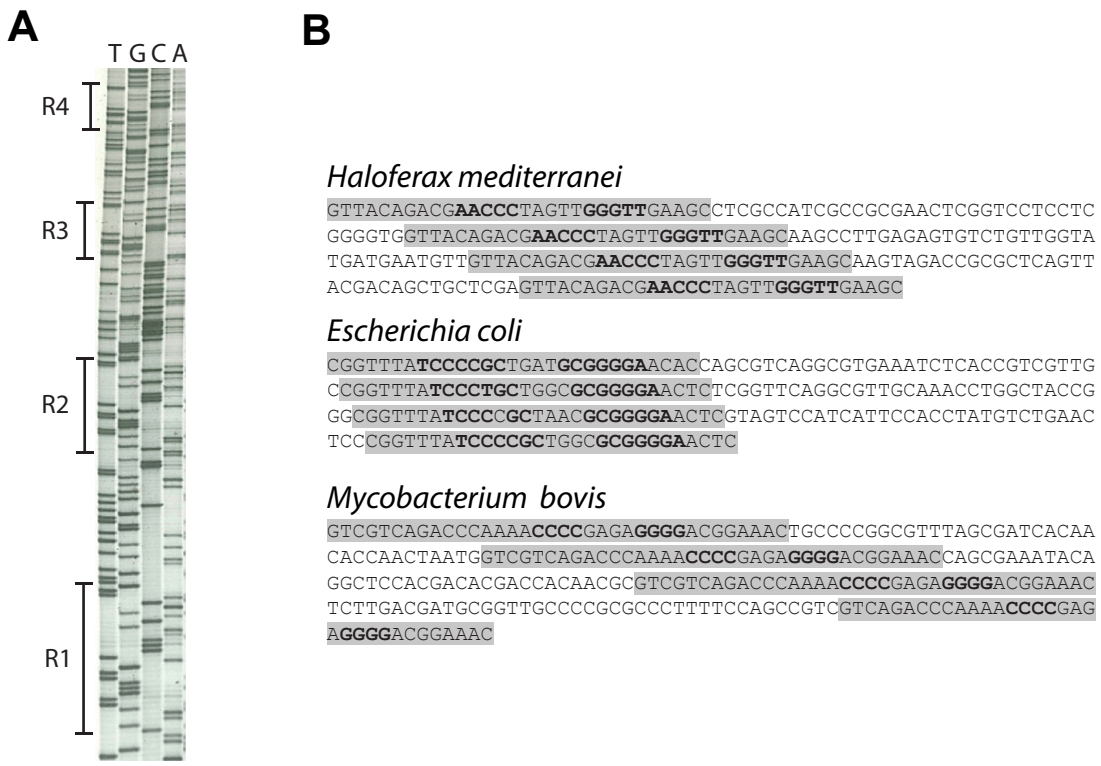


Figure 1.1: Early observations of tandem repeats. (**A**) DNA sequencing gel demonstrating tandem repeats observed in the genome of *Haloferax mediterranei*, repeats labelled R1 - R4. (**B**) Example repeat sequences from *Haloferax mediterranei*, *Escherichia coli* and *Mycobacterium bovis*. Repeats highlighted in grey, inverted repeats in bold. Figure adapted from (F. J. M. Mojica and Rodriguez-Valera, 2016).

to the same invasive DNA, interference occurs preventing reinfection (Figure 1.2 (2.)). The mechanism by which the spacer sequences were able to offer immunity against future infection was not immediately clear. It was initially proposed that immunity was meditated through an RNA interference mechanism, analogous with eukaryotic RNAi, whereby RNA transcribed from CRISPR arrays, crRNA, would interact with invasive transcribed RNA, mediated by the Cas proteins, resulting in RNA degredation or gene silencing (Francisco J.M. Mojica et al., 2005). However, it was demonstrated that crRNA-Cas protein complexes instead directly target and cut double-stranded DNA of invasive plasmids and bacteriophages, though novel CRISPR systems were subsequently identified which do directly target and degrade RNA (Brouns et al., 2008; Garneau et al., 2010; Hale et al., 2009; Gasiunas et al., 2012; Abudayyeh et al., 2016).

It was also noted that, during the adaptation stage, the acquisition of spacer sequences from foreign DNA was not random. When the spacer sequences found in the CRISPR array of a particular species were aligned to the foreign DNA from which they derived, termed the

proto-spacer DNA, a consensus motif was identified immediately upstream (Bolotin et al., 2005; Deveau et al., 2008; F. J.M. Mojica, Díez-Villaseñor, et al., 2009). During the adaption step of CRISPR adaptive immunity, the presence of this proto-spacer adjacent motif, PAM, defines which sequences will be acquired into the spacers of the CRISPR array. Subsequently during CRISPR interference, upon reinfection by the same invading species, this PAM sequence must be present immediately upstream of the target in order for cleavage to occur (Figure 1.2). The requirement for PAM sequences for DNA cleavage enables discrimination between self and non-self, only permitting cleavage of targets on the invading DNA and not of the targets present as spacers in the CRISPR array itself.

Viruses rapidly mutate which can allow them to bypass recognition by the CRISPR system on reinfection. On re-exposure to slightly mutated invading sequences there is an additional third step, priming, utilising the interference and adaption mechanisms, to rapidly update the CRISPR array (Xue et al., 2015).

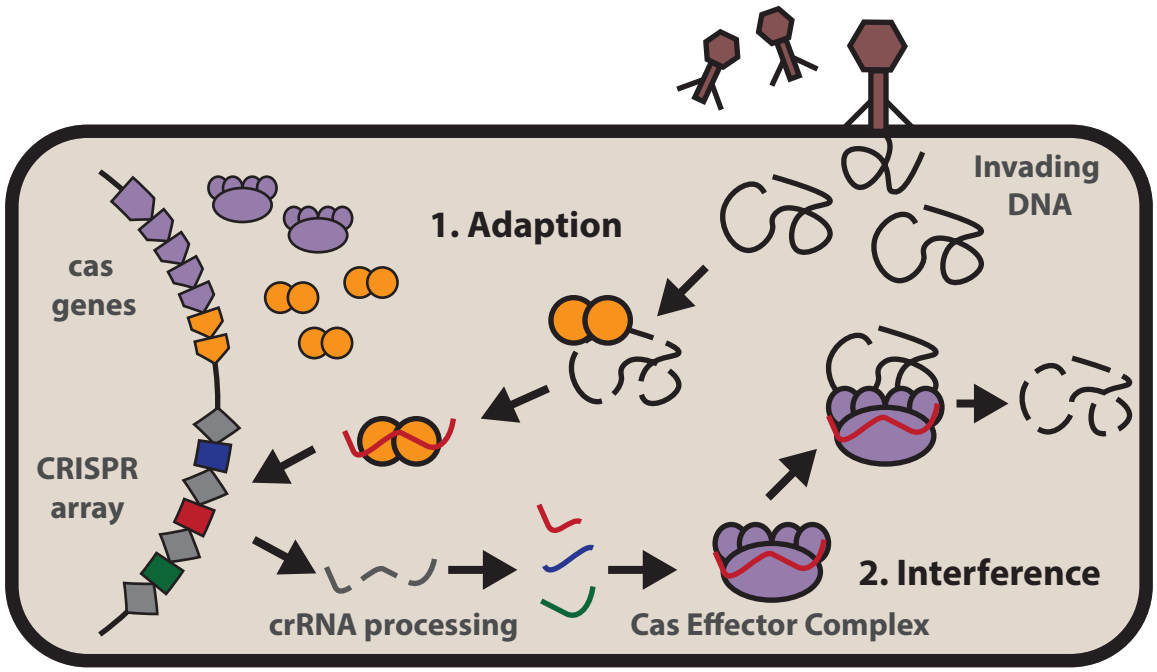


Figure 1.2: General mechanism of prokayrotic adaptive immunity. (**1.**) On first exposure to a foreign DNA sequence cas proteins integrate sequences found downstream of PAM sites into the CRISPR locus. (**2.**) On re-exposure to the same invasive DNA sequence, crRNAs expressed from the CRISPR locus are bound by the cas effector complex which then cuts complementary sequences on the foreign DNA. As the PAM sequence is not present in the CRISPR array the host genome is not targeted. Based on figure from (Patterson et al., 2017)

1.1.2 Classification of CRISPR-Cas Systems

Following the landmark discovery of the prokaryotic adaptive immune system, a huge variety of CRISPR systems have been identified and classified, and the specific mechanisms and sequence requirements of many are now well characterised (Makarova, Haft, et al., 2011; Makarova, Wolf, Alkhnbashi, et al., 2015; Makarova, Wolf, Iranzo, et al., 2020). CRISPR-Cas systems are first separated into two classes based on the identity of the effector complex used during the interference phase. The effector complexes of Class 1 CRISPR-Cas systems comprise multiple Cas proteins bound to the crRNA, which together mediate crRNA binding, target binding and target cleavage. In contrast, Class 2 crRNA-Cas complexes contains a single Cas protein with multiple domains combining all the functions of the multi-protein effector complexes of Class 1 systems (Figure 1.3) (Makarova, Wolf, Iranzo, et al., 2020).

Within these two classes, CRISPR systems are further classified into types, based largely on the structure of the effector complexes. Class 1 has been divided into three such types (I, III and IV) defined by the identity of the cleavage domain (Figure 1.3) (Makarova, Wolf, Alkhnbashi, et al., 2015). Type I, which includes the well studies Cascade complex, is characterised by the presence of Cas3 as a component of the cleavage domain; this nuclease domain is distinct from those in type III, defined by the presence of Cas10, a multisubunit domain with roles in both cleavage and target recognition. The final type, type IV, lacks any such nuclease domain, with some also lacking Cas1 and Cas2 which form the basis of the adaption complex, and have distinct functions not involved in adaptive immunity.

Class 2 has also been subdivided into three types (II, V and VI) based on comparison of the single multidomain effector Cas protein (Figure 1.3) (Makarova, Wolf, Iranzo, et al., 2020). The key distinguishing features between these types are the identities of nuclease domains present - Type II, including Cas9, contains two distinct nuclease domains, HNH and RuvC-like, each responsible for cleaving one stand of the dsDNA target (Gasiunas et al., 2012); Type V, including Cas12, only has a single RuvC domain which sequentially cleaves both of the DNA strands; and Type V1, including Cas13, has two HEPN nuclease domains, which cleave targeted RNA (Figure 1.3).

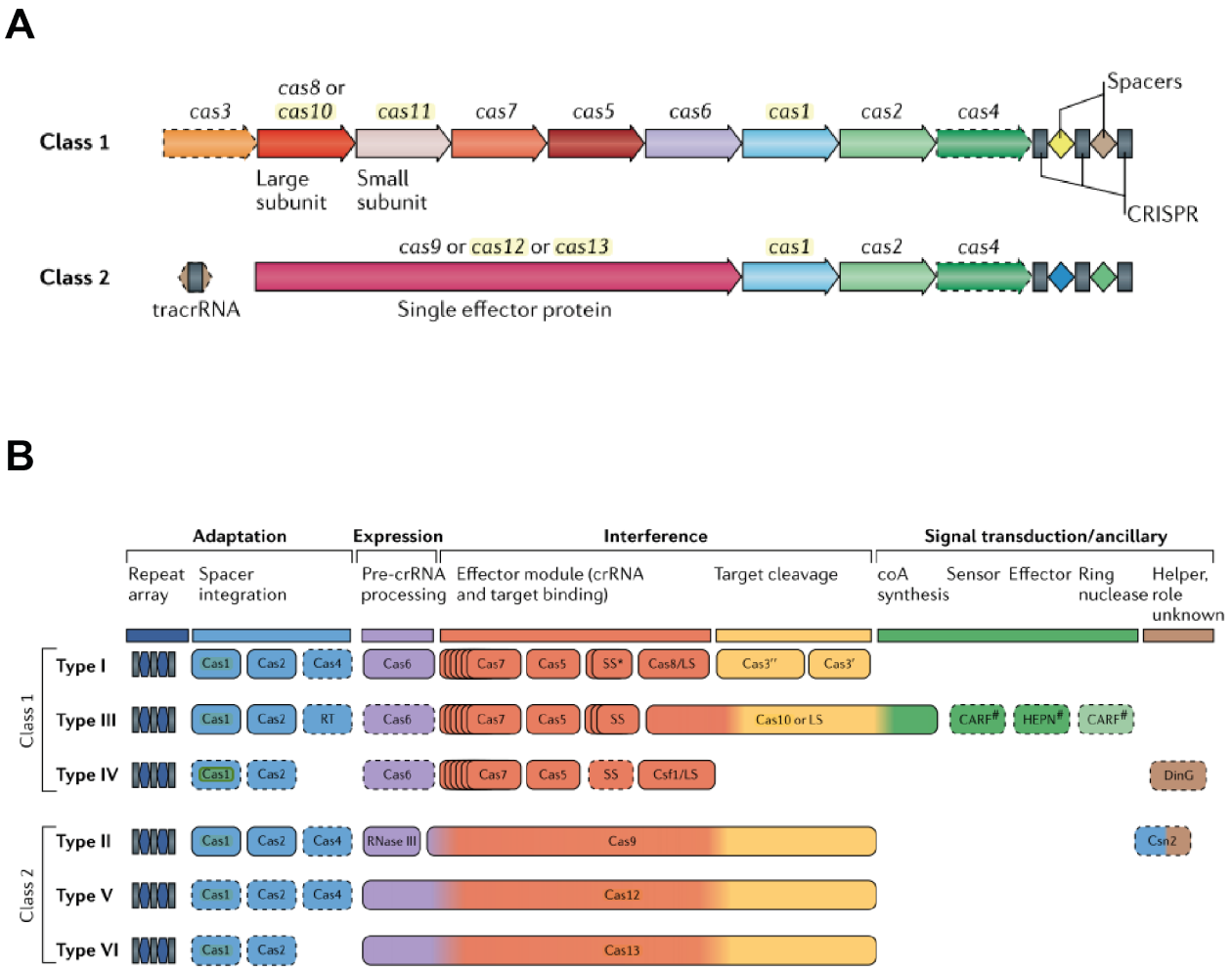


Figure 1.3: Classification of CRISPR/Cas Systems. (**A**) Representation of the two classes of CRISPR systems: Class I effector complexes consist of a multi-protein complex, Class II CRISPR systems consist of a single multi-domain effector protein. (**B**) The two classes are then subdivided into types. Figure adapted from (Makarova, Wolf, Iranzo, et al., 2020).

Each type is further divided into subtypes, primarily by clustering of sequence similarities in the effector domains, to date resulting in a total of 33 distinct CRISPR subtypes across both classes (Makarova, Wolf, Iranzo, et al., 2020)

1.2 Gene-editing using CRISPR

Just as the bacterial arsenal of restriction enzymes has been re-deployed for *in vitro* DNA manipulation, the CRISPR/Cas adaptive immune system offers enormous potential for *in vivo* DNA modification. *In vivo* DNA editing requires the ability to specifically alter a particular genomic DNA sequence in the crowded cellular context. Previous techniques have made use of zinc finger nucleases, ZFNs; or transcription activator-like effector nucleases, TALENs; which can be engineered to recognise and cleave a specific DNA sequence (Y. G. Kim et al., 1996; Urnov et al., 2010; J. Keith Joung and Sander, 2013). Once a double-stranded DNA break has been generated, the cell repairs the damage using one of the two intrinsic DNA damage repair pathways, non-homologous end joining, NHEJ, or homologous recombination (Figure 1.4) (Chapman et al., 2012). In the absence of DNA homologous to the break site NHEJ is typically employed leading to short insertions or deletions, indels, which can be utilised to disrupt the protein coding sequence and generate gene knockouts (Figure 1.4A)(Bibikova et al., 2003; Jeffrey C. Miller et al., 2011). If a piece of DNA with homology across the induced break site is provided, this can be utilised in homologous recombination, allowing the insertion of a novel DNA sequence into the break site (Figure 1.4B) (Bibikova et al., 2003; Jeffrey C. Miller et al., 2011).

Successful gene editing has been reported using targeted induction of double-strand breaks by zing finger proteins and TALENs, followed with repair by either NHEJ or homologous recombination (Bibikova et al., 2003; Urnov et al., 2010; Jeffrey C. Miller et al., 2011; Wood et al., 2011; J. Keith Joung and Sander, 2013). However, the laborious process of re-engineering that is required to target each new sequence represented a significant bottleneck. As the sequence specificity of the CRISPR systems are encoded not in the protein but in the bound 'guide' crRNA sequence, their discovery posed the possibility of an easily re-programmable sequence specific endonuclease.

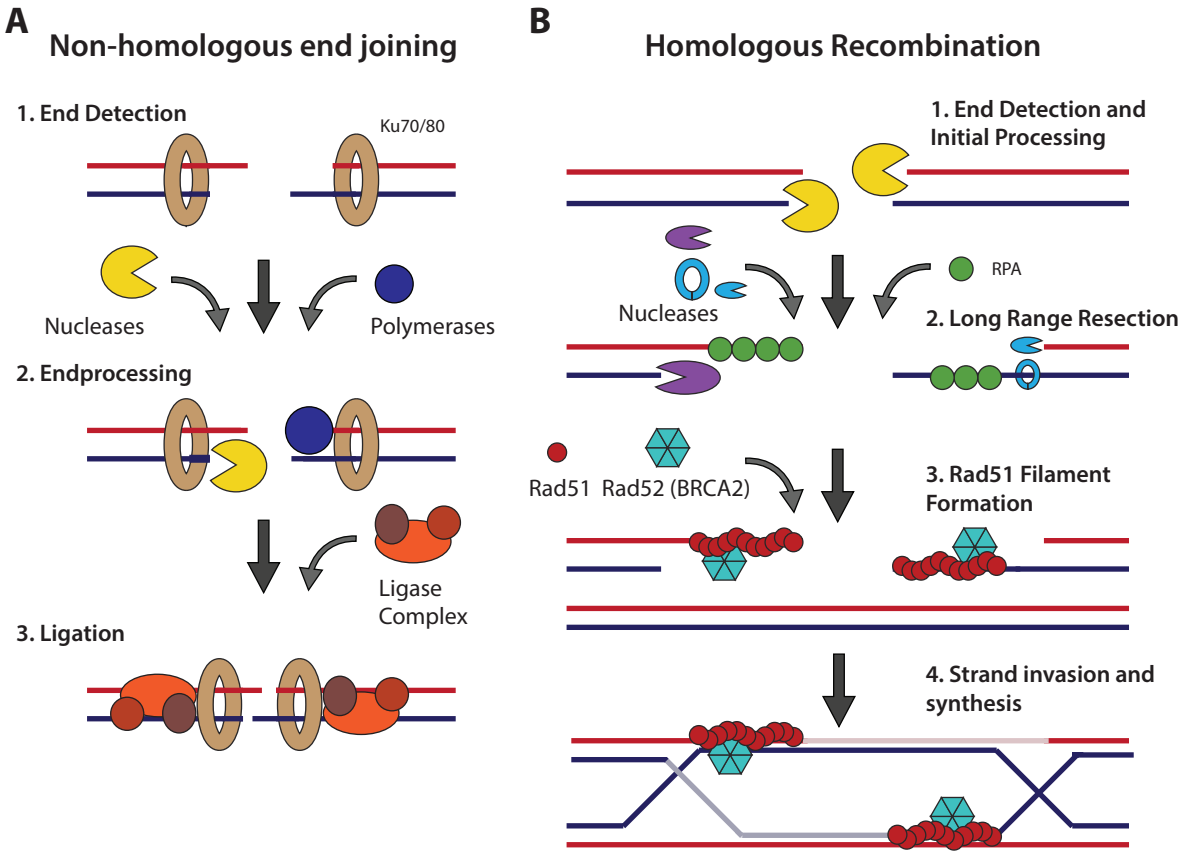


Figure 1.4: Double-strand break repair pathways. (**A**) Non-homologous end joining repair pathway. 1. Double-strand break is recognised by the Ku70/80 complex. 2. Nucleases and polymerases are recruited, repairing the ends to make them compatible for ligation, often resulting in short distance resection. 3. The ligase complex XRCC4/XLF ligates the ends, repairing the DSB. (**B**) Homologous recombination repair pathway. 1. Double-strand break is recognised by MRN complex which undergoes short distance resection. 2. Nucleases are recruited which further resect the DNA around the DSB, resulting in long 5' ssDNA overhangs which are coated with RPA. 3. Rad51 filaments are formed on the overhangs, replacing the bound RPA, aided by additional proteins such at Rad52. 4. Rad51 catalyses strand invasion of the overhangs into a homologous template dsDNA sequence. DNA synthesis occurs using this 'template', restoring the lost sequence and repairing the DSB.

1.2.1 The CRISPR/Cas9 system

This potential for a CRISPR based gene editing system was first demonstrated with the Cas9 protein from *Streptococcus pyogenes* (Jinek et al., 2012). *S. pyogenes* has a functional Class II, Type II CRISPR system, where the interference effector complex consists of a single multidomain protein, Cas9, capable of target recognition and dsDNA cleavage. As such it presented the tantalising possibility of a simple single-protein programmable DNA endonuclease.

The *S. pyogenes* Cas9 associated CRISPR array consists of 7 inverted repeats separated by 30 nt spacer sequences derived from phage sequences (Figure 1.5A) (Deltcheva et al., 2011; Le Rhun et al., 2019). As well as the Cas proteins, two precursor RNAs are transcribed from

the CRISPR locus, the long pre-crRNA from the CRISPR array itself, containing all the spacer sequences separated by repeats, and upstream a trans-activating CRISPR RNA, tracrRNA (Figure 1.5A)(Deltcheva et al., 2011). Two steps of RNA processing are then required, first the tracrRNA anneals to complementary sequences in the repeat sequences of the pre-crRNA, whereby endogenous RNaseIII cuts within the duplexed RNA. Second, the released individual 66 nt pre-crRNAs, annealed to the now mature 75 nt tracrRNA, are bound by the Cas9 protein, which further processes the pre-crRNA, cutting towards the 5' end to generate the mature 39-42 nt crRNA. This active crRNA-tracrRNA-Cas9 complex can bind and cleave dsDNA complementary to the 20 nt spacer sequence present in the crRNA (Deltcheva et al., 2011; Le Rhun et al., 2019). The breakthrough for potential gene-editing, was to demonstrate that the purified Cas9 protein binds directly to purified mature crRNA-tracrRNA and cuts dsDNA *in vitro*, and that it could be targeted to cleave different DNA sequences simply by changing the 20 nt spacer derived sequence in the crRNA (Jinek et al., 2012). The system was simplified even further to work with a single chimeric guide RNA, sgRNA, by connecting the crRNA and tracrRNA with a short loop sequence (Jinek et al., 2012).

Cas9 has two distinct nuclease domains, a hallmark of all type II CRISPR systems, each responsible for cutting one of the two DNA strands (Figure 1.5B and C). The HNH domain cuts the target-strand, the DNA strand complementary to the crRNA, at a site three bases from the PAM sequence; the RuvC-like domain cuts the non-target strand directly opposite the HNH cut site, generating a blunt ended double-strand break (Figure 1.5D). As with other CRISPR systems, Cas9 requires a protospacer-adjacent motif site, PAM site, immediately adjacent to the target site to recognise and cleave the DNA. As discussed previously, this enables discrimination between self and non-self, preventing cutting of the CRISPR locus which contains the target DNA sequence but lacks the PAM sites, and ensuring only invading DNA from which the spacer sequences were acquired will be cut. For spCas9 the canonical PAM sequence is 5'-NGG-3', however alternatives can be tolerated as will be discussed later (see Section 1.3). Whereas the recognition of the DNA target sequence is mediated by DNA:RNA base pairing with the crRNA, the recognition of the PAM site is mediated by a protein-DNA interaction (Figure 1.5E). Cas9 first binds the PAM site and then forms a DNA:RNA hybrid between the target

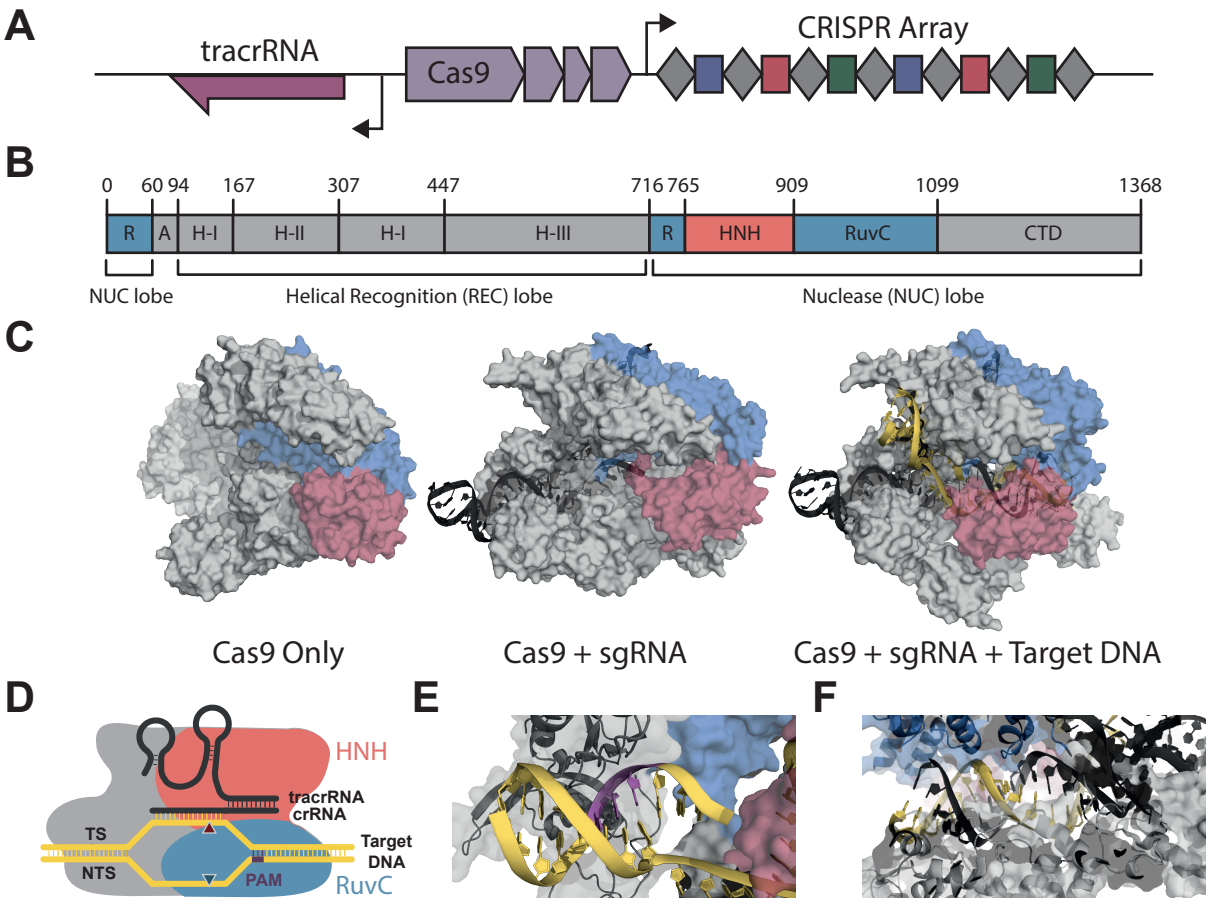


Figure 1.5: Streptococcus pyogenes CRISPR/Cas9. (**A**) Layout of the *Streptococcus pyogenes* Cas9 CRISPR locus. Adapted from Le Rhun et al., 2019 (**B**) Domain layout of SpCas9, HNH nuclease domain (red), RuvC nuclease domain (blue). Adapted from Jiang et al., 2015 (**C**) Crystal structures of Cas9 alone (PDB:4CMP, **JinekStructuresActivation**), with sgRNA (pdb:4ZT0, Jiang et al., 2015) and with sgRNA bound to DNA (PDB:4UN3, Anders et al., 2014). HNH nuclease domain highlighted in red and RuvC nuclease domain highlighted in blue, sgRNA in black and target DNA in yellow. (**D**) Schematic of Cas9 bound to target DNA. HNH and RuvC nuclease domain cut sites labelled with red and blue triangles respectively. (**E**) Section of crystal structure of Cas9 with sgRNA (no shown) bound to target DNA (yellow) highlighting recognition of PAM sequence (purple) by the REC domain (grey) (PDB:4UN3, Anders et al., 2014). (**F**) Section of crystal structure of Cas9 with sgRNA (black) bound to target DNA (yellow) highlighting interactions of REC domain (grey), HNH domain (red) and RuvC domain (blue) with the sgRNA:DNA hybrid (PDB:4UN3, Anders et al., 2014).

DNA strand and the crRNA, displacing the non-target DNA strand to generate an R-loop (Figure 1.5C-F). While the PAM site is predominantly recognised by the REC-domain (Figure 1.5E), the DNA:RNA hybrid interacts with the REC, HNH and RuvC domains, which together mediate target cleavage (Figure 1.5F). Importantly, the protein interactions with the DNA:RNA hybrid are with the nucleic acid backbone and not the individual bases. As such, apart from the PAM requirement, sequence specificity is entirely mediated by the base pairing between the crRNA and target DNA (Figure 1.5F).

1.2.2 Gene editing with CRISPR/Cas9

Within a matter of months from the seminal reports of Cas9 as a re-programmable RNA-guided DNA endonuclease, two separate groups demonstrated genome editing in human cells (Cong et al., 2013; Mali et al., 2013). In both instances the Cas9 gene was codon optimised for expression in human cells, fused to nuclear localisation sequences, and cloned into mammalian expression vectors. This construct was then transfected into 293T, K562 or induced-pluripotent stem cells, along with vectors expressing either the pre-crRNA and tracrRNA components (Cong et al., 2013) or sgRNA constructs (Mali et al., 2013) targeting specific genomic loci. Editing of the specific target site was observed in all cases, though efficiencies varied depending on cell type and guide expression system. This has subsequently heralded a revolution in genetic editing both of cell cultures, *in vitro*, and at the level of organisms, using either *ex vivo* or *in vivo* approaches. CRISPR/Cas based methods have opened the door to genetic manipulation of organisms that were previously inaccessible.

There are several approaches to gene editing with Cas9, depending on the required application (Lino et al., 2018). For genetic modification of populations of cultured cells, transfection-based approaches are still mostly used and give very high editing efficiencies (Figure 1.6A and B). Cas9 can be delivered in DNA, mRNA or ribonucleoprotein (RNP) form (Figure 1.6A), with transfection typically achieved by nucleofection (Mali et al., 2013), electroporation (Ding et al., 2013) or lipofectamine treatment (Figure 1.6B) (Cong et al., 2013). DNA-based delivery typically consists of one or multiple plasmids encoding the Cas9 protein and sgRNA. After transfection the plasmids must first reach the nucleus, followed by transcription, translation in the cytoplasm, and nuclear import, before gene editing can occur (Figure 1.6B) (Cong et al., 2013; Mali et al., 2013). In contrast, direct delivery of the Cas9 mRNA allows immediate translation on entering the cytoplasm. Direct transfection of the active RNP sgRNA:Cas9 bypasses even the translation step and offers the fastest editing kinetics (Figure 1.6B) (S. Kim et al., 2014; X. Liang et al., 2015). However, another consideration is the stability or the construct and resulting intracelluar concentration of Cas9. DNA-based delivery results in a prolonged, relatively stable, expression of Cas9, increasing the on-target editing efficiency but with the potential increased risk of unwanted mutagenic effects from prolonged exposure (Figure 1.6B X. Liang et al., 2015). In

contrast, RNP delivery is the shorted lived and allows the most direct control of the Cas9 concentration administered (Figure 1.6B) (S. Kim et al., 2014; X. Liang et al., 2015).

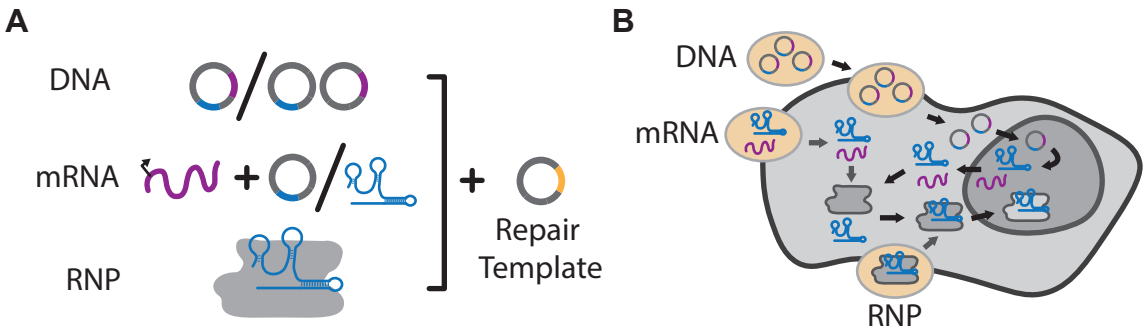


Figure 1.6: Comparison of different form of Cas9 for delivery into cells. (**A**) Cas9 can be delivered as a plasmid(s), RNA or RNP. (**B**) Choice of delivery approach effects editing kinetics, intra-cellular Cas9 concentration and exposure times.

For *ex vivo* applications with small numbers of cells, for example embryonic editing, physical injection remains the most effective method of delivery (H. Wang et al., 2013; Sander and J Keith Joung, 2014; Raveux et al., 2017) This is the approach used in attempts at human germ line editing (Tang et al., 2017; Fogarty et al., 2017; P. Liang et al., 2015; Doudna, 2020). Putting aside the not insignificant ethical concerns regarding human germ line genetic editing, further work is required to improve efficiency as well as to understand and reduce unpredictable outcomes, such as off-target editing, loss of heterozygosity and mosaicisms. (Kang et al., 2016)

Viral-based vectors are the preferred delivery method for ZFNs and TALENs and have also been applied to Cas9 *in vitro* and *in vivo* delivery. The editing machinery is packaged as DNA into the viral particles which are then used to infect cells, delivering the DNA cargo. For ZFNs and TALENs the use of adeno-associated virus (AAV) as the viral vector has been particularly promising and has been successfully utilised with Cas9 (Long et al., 2016). A significant limitation to this approach is the capacity of the viral capsid which, while technically capable of holding the sgRNA:Cas9 system, leaves little room for additional markers, tags or a DNA template for repair. Alternative approaches include separate packaging of the sgRNA and Cas9 protein (Swiech et al., 2015), or even splitting Cas9 between two capsids (Truong et al., 2015; Chew et al., 2016). Alternatively, AAV delivered *in vivo* gene-editing has also been demonstrated utilising a closely related but smaller Cas9 protein, *Staphylococcus aureus* Cas9,

allowing additional space for accessory sequences (Ran et al., 2015). AAV based methods also have the advantage of an additional level of control by utilising specific AAV serotypes, known to preferentially infect particular tissue types (Table 1.1) (Wu et al., 2006; Lisowski et al., 2015). This, in combination with tissue specific promoters, has been used for targeted Cas9 editing to specific organs. Examples include muscle (Long et al., 2016), liver (Ran et al., 2015) and brain (Swiech et al., 2015). Alternative viral vectors include adenovirus and lentivirus, both of which offer significantly larger capsid capacities than AAV making them well suited to deliver all the required components of Cas9 simultaneously (Maddalo et al., 2014; Platt et al., 2014). Initial use of adenovirus vectors posed issues with the triggering of immune responses (S. Ahi et al., 2011), however these have largely been addressed by removal of many of the viral genes (Palmer and Ng, 2003; Brunetti-Pierri and Ng, 2017). Lentivirus vectors also posed challenges as their retroviral activity involves integration into the host genome, which could have unintended mutagenic consequences. However non-integrating lentivirus vectors have now been developed which address this issue and have been demonstrated as a viable method for Cas9 delivery (Apolonia et al., 2007; Choi et al., 2016).

Table 1.1: AAV Serotypes

Target Tissue	Optimal Serotype
Liver	AAV8, AAV9
Skeletal Muscle	AAV1, AAV6, AAV7, AAV8, AAV9
CNS	AAV1, AAV4, AAV5
Photoreceptor Cells	AAV5
Lung	AAV9
Heart	AAV8
Pancreas	AAV8
Kidney	AAV2

Table modified from (Wu et al., 2006)

Promising results have also be found with non-viral delivery methods for *in vivo* application. Transfection using lipofectamine is already a well-established technique in isolated cells and has also been demonstrated for *in vivo* Cas9 delivery (Zuris et al., 2015). Lipofectamine is a positively charged synthetic lipid capable of forming lipid vesicles around negatively charged molecules such as nucleic acids or the Cas9 RNP complex, and delivering them across cell lipid membranes (Figure 1.7A). Injection of these vesicles *in vivo* has been demonstrated to result in

effecient Cas9 uptake and editing (Zuris et al., 2015). An intriguing alternative to the lipid vesicle is the DNAnanochew, a self assembling DNA 'ball' which can be functionalised with sgRNA:Cas9 complexes. This structure, when coated with the positively charged polyethylenimine, PEI, can be transported into cells and escape the endosome, delivering the Cas9 cargo (Figure 1.7B)(Sun et al., 2015). Gold nanoparticles, AuNPs, have been demonstrated to be readily taken up into cells and can be functionalised to bind DNA molecules, as well as displaying electrostatic interactions with proteins (K. Lee et al., 2017). This makes them strong candidates for RNP Cas9 and DNA delivery. In combination with an outer coating of silica and positively charged polymer, PAsp(DET), to facilitate escape from the endosome, AuNP delivery of Cas9 has been demonstrated with high levels of gene editing in *in vivo* mouse models (Figure 1.7C) (K. Lee et al., 2017).

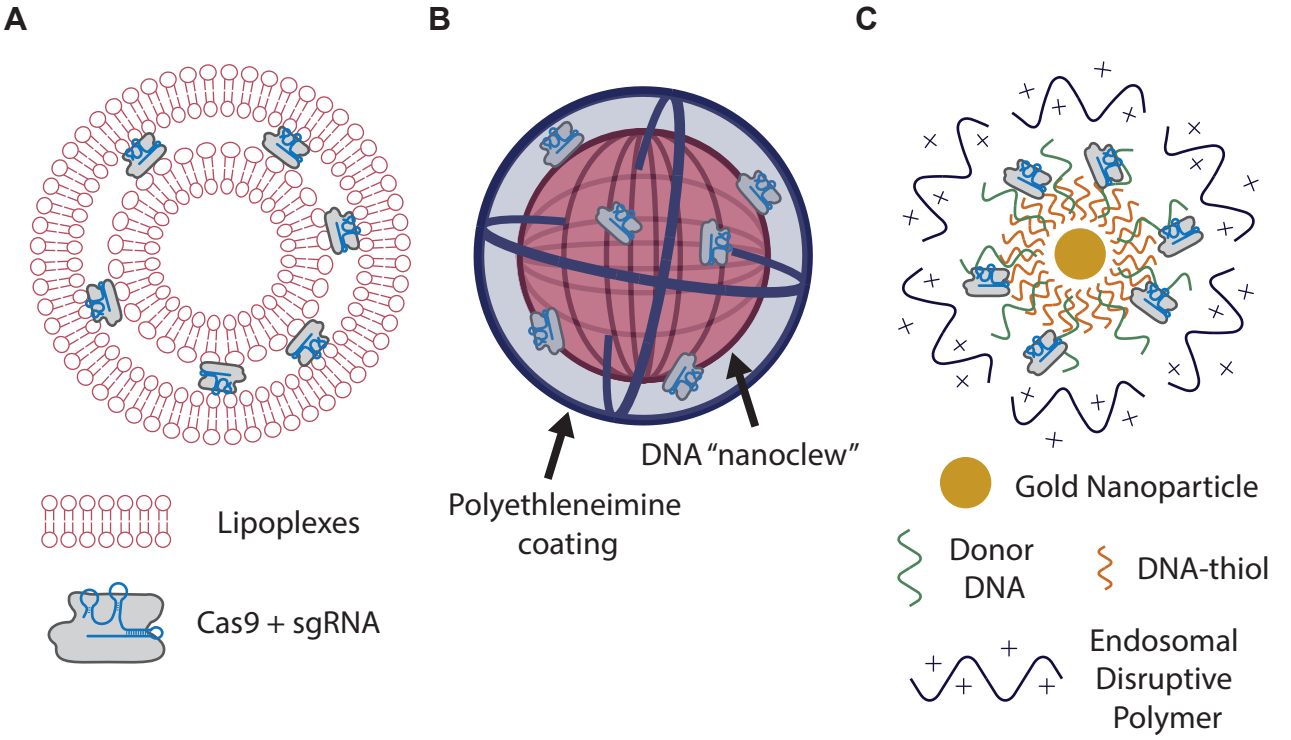


Figure 1.7: Non-viral Delivery Methods. (**A**) Lipofectamine lipid vesicles containing Cas9 complexes can deliver cargo across lipid membranes into the cytoplasm of cells, (Zuris et al., 2015). (**B**) DNA nanoclew, an internal self-assembling DNA ball is functionalised with Cas9 complexes. Coating with polyethleneimine enables transport into cells, (Sun et al., 2015) (**C**) Gold nanoparticles, can be functionalised with Cas9 complexes can be taken up by cells. The addition of an outer polymer coating aids escape from the endosome and delivery into the cytoplasm, (K. Lee et al., 2017)

These recent breakthroughs with CRISPR/Cas gene editing and delivery systems have significantly advanced efforts to develop human gene therapies, with some already being trialled.

However, whilst it is already an incredibly powerful research tool, the jump to it becoming a mainstream therapeutic is significant, with questions of efficiency and specificity paramount when considering *in vivo* somatic human editing.

1.3 CRISPR/Cas off-target activity

In the first demonstrations of CRISPR as an adaptive immune system, there was evidence that mismatches between the guide and target could be tolerated and still lead to target degradation (Garneau et al., 2010; Deveau et al., 2008). The question of specificity became paramount after successful gene editing was demonstrated with *Streptococcus pyogenes* Cas9 (Jinek et al., 2012). The tolerance of even a few mismatches within the 20 nt target sequence could significantly hinder the potential for therapeutic applications; any guide designed to a target unique site in a large genome would be predicted to have tens to hundreds of potential off-target sites. However, conflicting reports began to emerge, with some suggesting high levels of promiscuous off-target cutting at sites with multiple mismatches, while other showing highly specific activity with no discernible off-target effects. It quickly became apparent that factors such as the locus being targeted, locations of mismatches in the sequences, cell type and method of delivery could all dramatically effect the extent of off-target activity.

1.3.1 Detection of off-target activity

The first approaches to investigate off-target activity were based on the assumption that potential off-targets would be sites with sequences similar to the on-target site. The tolerance of mismatches between the guide and target was first interrogated by focusing on a small number of genomic locations and quantifying how introducing mismatches in the guide RNA impact editing efficiencies in cells at these particular sites (Figure 1.8A) (Fu, Foden, et al., 2013; Hsu et al., 2013). Furthermore the presence of 'bulges' between the guide and DNA can also be tolerated in some instances (Figure 1.8B) (Y. Lin et al., 2014). These can be either be DNA bulges, formed within the target DNA, or RNA bulges, formed within the guide RNA (Y. Lin et al., 2014). In an alternative *in vitro* approach, the off-targeting of individual guides were explored using large libraries of potential DNA target sequences. This was combined with methods to enrich for cut fragments and next-generation sequencing to determine which sequences were cut (Pattanayak et al., 2013). Together these approaches revealed that while mismatches in the 10-12 nts proximal to the PAM site dramatically reduced target cleavage, multiple mismatches

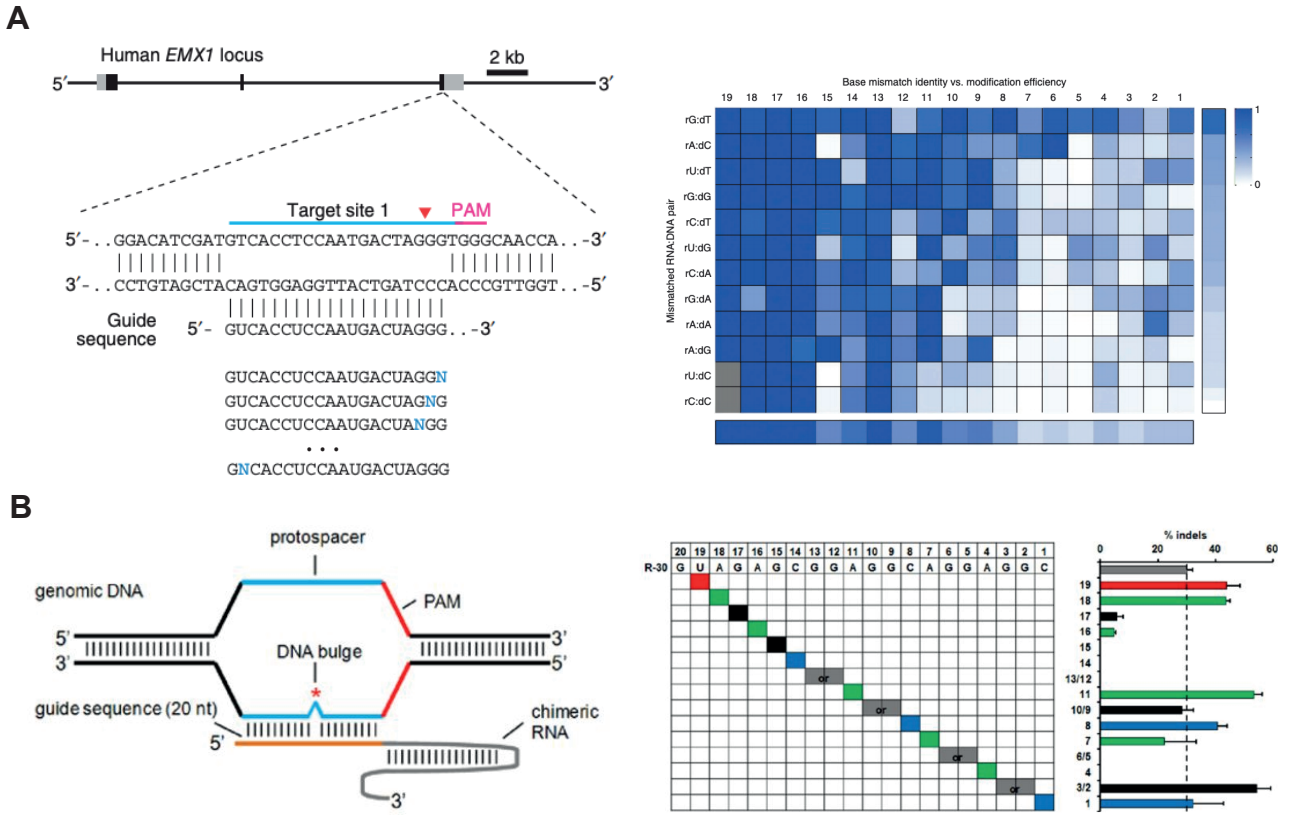


Figure 1.8: Mapping the effect of mismatches on editing efficiencies. (**A**) Single mismatches are inserted across the sgRNA sequence to test the editing efficiencies at different positions, left. Resulting indel frequencies for different sgRNA mismatch position and types, right. Mismatches within the 'seed' region (position 1-8) are less well tolerated than those at the PAM distal end. (Hsu et al., 2013) (**B**) sgRNA sequences targeting a DNA sequence are designed with missing bases, such that base pairing of the sgRNA along the length of the target requires formation of a DNA bulge, left. Cleavage efficiency within cells is measured with guide RNAs inducing DNA bulges at each site along the target, right. The tolerance of DNA bulges is location specific, with high cleavage efficiencies maintained with bulges immediately adjacent to the PAM (1-3), around the central positions (7-10), and at the PAM distal sites (18-19). (Y. Lin et al., 2014)

at the PAM distal end of the target could be tolerated. Furthermore, it was found that some mutations to the PAM sequence could be also tolerated (Hsu et al., 2013). This lead to the definition of a ∼10 nt required 'seed' region of complementarity proximal to the PAM site, and a broadening of the PAM sequence to 5'-NRG-3', accommodating the non-canonical 5'-NAG-3'.

These observations were used to identify candidate off-target sites present within the human genome for a given guide sequence. Following Cas9 editing, these potential off-target locations could then be individually sequenced to identify if any mutations had been induced at these sites (Fu, Foden, et al., 2013; Hsu et al., 2013). Together these methods demonstrated that off-target editing can occur *in vivo* at sequences containing multiple mismatches, with non-canonical PAM sequences and even containing 4 bp 'bulges' (Y. Lin et al., 2014).

Although informative, such approaches require initial assumptions in shortlisting genomic off-target locations, and were not sufficiently high-throughput to validate all potential sites. New technologies have been developed for unbiased detection of off-target activity both *in vitro* and *in vivo*. The highest sensitivity comes from *in vitro* approaches, however methods to detect off-target sites *in vivo* are also important as Cas9 specificity is affected by the cellular context and *in vitro* approaches alone cannot predict *in vivo* outcomes. *In vivo* methods are based around detecting the presence of DSBs in cells, either by pull down of cut DNA ends, identification of translocation events, incorporation of a 'tag' DNA sequence across the break or by localisation of DSB repair machinery (Figure 1.10) (Ran et al., 2015; Frock et al., 2015; Tsai, Zheng, et al., 2015; Wienert et al., 2019).

1.3.1.1 *In vitro* off-target Detection

The first genomic *in vitro* method of off-target detection was Digenome-seq (D. Kim et al., 2015). Purified genomic DNA is incubated with Cas9, then fragmented and sequenced by NGS. After aligning reads to the known genomic sequence, sites that have been cut are identified by the presence of reads all terminating at the same location. As there is no enrichment of sequences that have been cut, the sensitivity is limited by the error rate of NGS, and requires high read numbers to get good sensitivity. However, this method can detect off-target sites with frequencies down to ∼0.1% (D. Kim et al., 2015). Selective enrichment and identification of tagged genomic DNA ends by sequencing, SITE-Seq, has increased sensitivity and requires lower numbers of reads than Digenome-seq due to the the addition of a step enriching for cut DNA by ligation of biotinylated DNA to cut end and biotin pull-down (Cameron et al., 2017).

Currently the most sensitive *in vitro* method is CIRCLE-seq (Figure 1.9B) (Tsai, Nguyen, et al., 2017). DNA is purified from the cell type of interest, fragmented to ∼300 bp and these fragments circularised to generate a library of DNA circles with sequences covering the entire genome. All non-circular DNA is degraded from the library followed by incubation with sgRNA:Cas9 targeting a particular site of interest. Any circle containing a sequence that is cut by sgRNA:Cas9 will be linearised. Adapters for next-generation sequencing are then ligated to these linear fragments and the library is sequenced. As only the cut fragments get sequenced

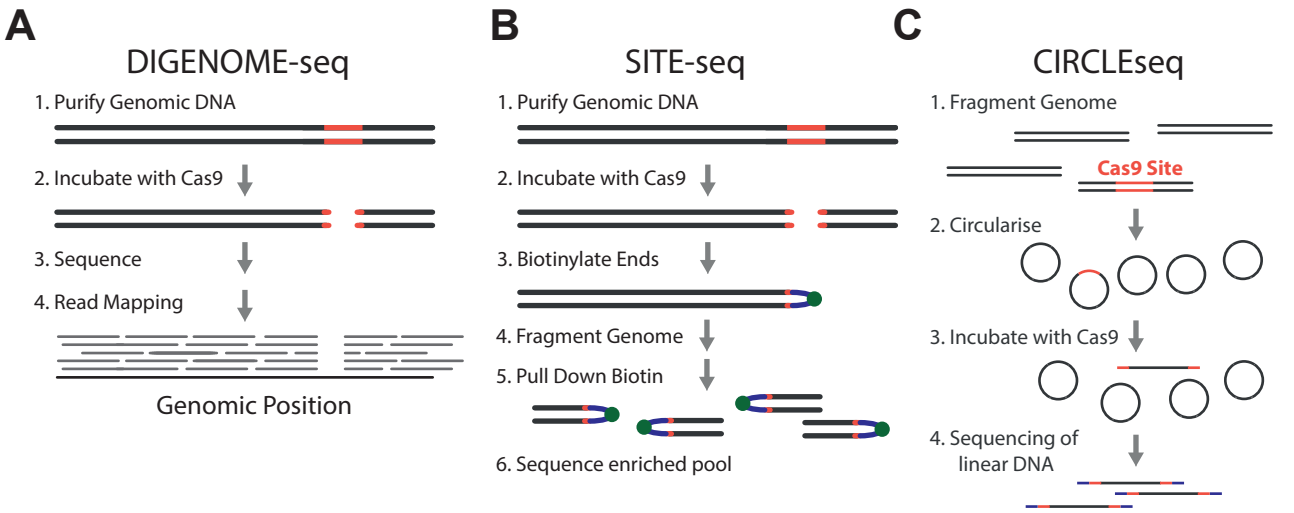


Figure 1.9: In vitro genomic methods for detecting off-target activity of CRISPR/Cas. (**A**) Digenome-seq, genomic DNA is isolated and purified from cells (1.) and then incubated with Cas9 complex (2.) to induce cutting at on-target and any off-target sites. The DNA is then sequenced and any induced DSBs identified where mapped reads all terminate at the same location (3.,4.). (**B**) SITE-seq, genomic DNA is purified from cells (1.) and incubated with Cas9 to induce cutting at on-target and any off-target sites (2.). Cut ends are labelled with biotinylated adapters before fragmentation of the DNA (3., 4.). Pull down of biotinylated fragments enriches for DNA sequences cut by Cas9 (5.) which are then sequenced to identify sites of cutting (6.). (**C**) CIRCLE-seq, genomic DNA is purified from cells and fragmented (1.). Fragments are circularised and any remaining linear DNA is degraded (2.). Circles are incubated with Cas9, any circles which are cut will be linearised (3.). Linear DNA fragments are then sequenced to identify locations of Cas9 cutting (4.).

the background is extremely low compared to other *in vitro* NGS techniques and off-target sites can be detected that would be well below the sensitivity of other methods (Figure 1.9C) (Tsai, Nguyen, et al., 2017; D. Kim et al., 2015; Frock et al., 2015). However, it is still unclear what the false positive rate of this method is, as even the most sensitive methods of off-target detection in cells have a limit of ∼0.1%. Although CIRCLE-seq is considered an unbiased method of detection, the bioinformatic analysis of CIRCLE-seq data does require knowledge of the guide sequence when peak calling to identify potential site, and there is a parameter of number of allowed mismatches in a detected off-target. It is therefore possible that unexpected sites would be missed.

1.3.1.2 *In vivo* Off-target Detection

In the case of *In vivo* off-target detection, other factors besides just sequence can determine whether Cas9 will cut at an off-target site. For example it has been demonstrated that the chromatin environment effects Cas9 efficiency, with highly chromatinised regions being less

accessible (Jensen et al., 2017). Furthermore Cas9 concentrations are typically much lower that those used in *in vitro* methods. As such it is currently assumed that the off-targets detected by CIRCLE-seq is an exhaustive list of all potential sites and contain many false positives.

One of the first approaches applied *in vivo* was BLESS (breaks labelling, enrichment on streptavidin and next-generation sequencing) (Figure 1.10A). Edited cells are fixed and the nuclei isolated, followed by *in situ* labelling of DBSs with biotinylated linker DNA. The genomic DNA is then extracted and fragmented and the biotinylated fragments purified and sequenced (Crosetto et al., 2013; Ran et al., 2015). Although BLESS has been successfully applied to detect off-target sites in cells and *in vivo* in mice, a significant limitation of this approach is it can only detect DBSs present at the moment the cells are fixed - any repaired sites will not be detected. Furthermore DNA breaks can also be generated during the preparation and extraction of nuclei, increasing the background signal and reducing sensitivity. In a related method, BLISS, the DSB labelling is performed *in situ* on the cells fixed to a microscope slide, reducing the potential for inducing DSBs and also reducing the required number of input cells (Yan et al., 2017). Recently an alternative approach of *in vivo* DSB detection has been developed, based on ChiP-seq of the repair protein Mre11, which is localised to DSBs early in the repair pathway (Figure 1.10B) (Wienert et al., 2019). In this method, termed DISCOVER-seq, edited cultured cells or tissue samples are first fixed with formaldehyde to cross-link bound proteins to the DNA, the nuclei are extracted and the protein bound DNA extracted and fragmented. The Mre11 bound fragments are then isolated by pull-down with specific antibodies and the DNA purified and sequenced to generate genomic profiles of Mre11 localisation (Wienert et al., 2019). Distinctive Mre11 localisation peaks around DSBs allow the cut site to be determined with single base resolution. Although also limited to only detect unrepaired sites, DISCOVER-seq has much higher sensitivity that BLESS based approaches, and was demonstrated to be able to detect sites with indel frequencies of $\sim$0.3%. It also provides additional information as to the degree of resection that occurs in either direction at each detected site - which is important for understanding possible repair outcomes.

Alternatively, sites of repaired DBSs can be identified in edited cultured cells using either LAM-NTGTS (linear amplification–mediated high-throughput, genome-wide, translocation

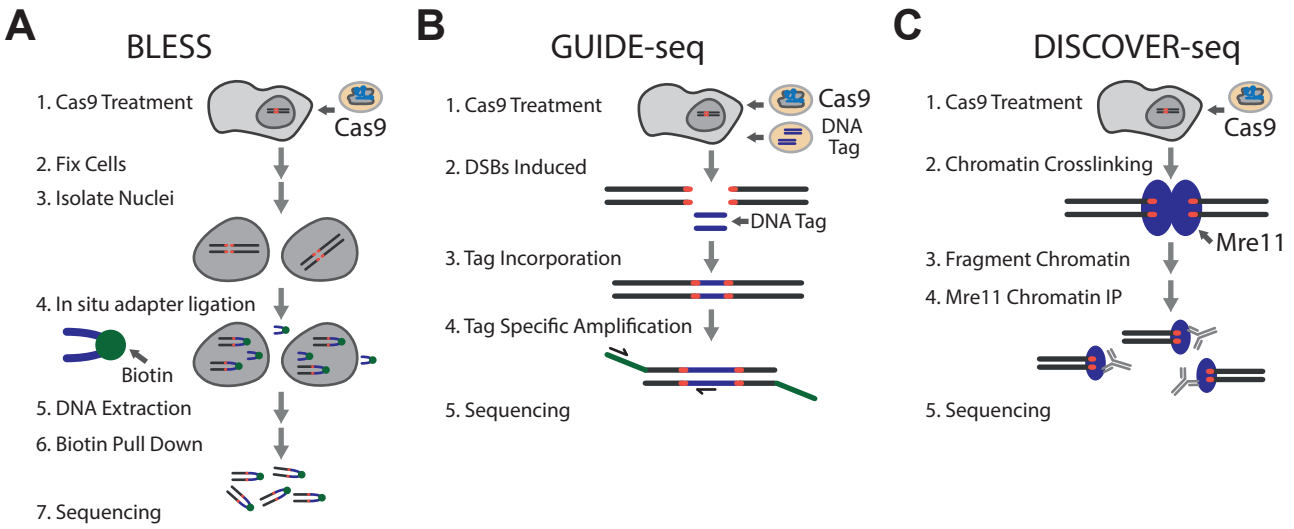


Figure 1.10: In vivo methods for detecting off-target activity of CRISPR/Cas. (**A**) BLESS, Cells are treated with Cas9 to induce editing (1.). At a defined time point the cells are fixed, preserving any DSBs present at that time point (2.), and the nuclei isolated (3.). Biotinylated adapters are ligated to any DSBs in situ in the nuclei (4.). DNA is extracted and purified from the nuclei (5.). Biotin pull down is performed to enrich for DNA with ligated adapters (6.), before final sequencing to identify cut sequences (7.). (**B**) GUIDE-seq, Cells are treated with Cas9 and a DNA tag (1.). At sites of DSBs the DNA tag is incorporated during repair (2.,3.). DNA is isolated and purified before Tag specific amplification of fragments containing incorporated tags (4.). Amplified sequences are sequenced to identify sites of DSBs where the Tag was incorporated (5.). (**C**) DISCOVER-seq, Cells are treated with Cas9 to induce editing (1.). Sites of DSBs recruit Mre11, cells are treated with crosslinker to fix DNA bound proteins to the DNA (2.). Chromatin is fragmented (3.) and chromatin immunoprecipitation performed against Mre11 to pull down DNA fragments with Mre11 bound at sites of DSBs (4.). DNA is purified and sequenced to identify on and off-target Cas9 cleavage sites (5.).

sequencing) or GUIDE-seq. In LAM-NTGTS, translocations that occur due to DSB repair between the on-target site and any off-target sites are detected. Cells are edited and the genomic DNA extracted, followed by linear-amplification PCR from the on-target site with a biotinylated primer. The biotinylated amplified sequences are then purified and sequenced to identify any translocations that have occurred between the on-target cleavage site any any DSBs induces at off-target sites. LAM-NTGTs was used to identify off-targets of four guide RNA designed to target the RAG1 gene. One guide resulted in 33 detected translocation events, and another guide two such events, the other two guides displayed no translocation event. This again demonstrates that off-targets are highly guide sequence dependant. While this approach demonstrates that Cas9 editing can induce translocations, the majority of DBSs are repaired locally and these would not be detected. Furthermore, the probability of two DSBs becoming joined is affected by the spacial localisation of the chromosomes within the nucleus and as such translocation efficiency between two sites is not a direct metric of the frequency of off-target

cutting. GUIDE-seq is an alternative method to detect the locations of repaired DSBs (1.10C). A 'tag' double-stranded oligodeoxynucliotide, dsODN, is transfected alongside Cas9 RNP which can be incorporated into DSBs by NHEJ. The genomic DNA is then extracted and sequences containing the incorporated dsODN specifically amplified using a primers specific to the 'tag', followed by next-generation sequenced (Tsai, Zheng, et al., 2015). When thirteen different guide sequences were tested in two different cell types, HEK293 and U2OS, as many as 150 off-target sites were identified with some guides, whereas others displayed no detectable off-targets. This once again demonstrates the highly variable levels of promiscuity of different guides. Importantly, there were also significant differences in the frequencies of off-target editing between the two different cell types, clearly demonstrating that sequence alone is only one factor which effects locations of off-target activity. GUIDE-seq is currently the most sensitive method available for detection of off-targets in cultured cells, it is able to detect sites with indel frequencies <0.1%. However, unlike DISCOVER-seq, GUIDE-seq cannot be used to directly detect off-targets *in vivo*.

1.3.1.3 Verification of *in vivo* targets, VIVO

The most comprehensive protocol of *in vivo* off-target detection is termed VIVO, verification of *in vivo* targets. CIRCLE-seq is first used to detect all possible off-targets for a given guide and genome, followed by targeted amplicon sequencing to determine whether any of these sites were actually cut *in vivo* (Akcakaya et al., 2018). This method has been applied to detection of induced off-targets in mice. With the use of a promiscuous guide, deliberately designed based on predictions of a large number of potential off-target sites across the mouse genome, VIVO resulted in the detection of over 2000 off-target at the first *in vitro* CIRCLE-seq detection step. From this list, a small number of sites were selected with varying readcounts and between 1-6 mismatches, for targeted amplicon sequencing of *in vivo* edited mouse livers. Of the 45 sites tested, the presence of Cas9 induced indels were identified at 19 sites with upto 3 mismatches. The frequency of indel formation correlated well with readcounts from the CIRCLE-seq analysis, however indels were still detected at some sites with low readcounts. The use of VIVO clearly demonstrates the potential for *in vivo* off-target editing with Cas9 with promiscuous guide

design. In contrast, when used to detect potential off-targets from a guide designed to minimise off-target activity, VIVO detected only 182 off-targets at the CIRCLE-seq stage, none off which showed indel formation *in vivo* by targeted amplicon sequencing. This demonstrates the importance of guide design in reducing the potential of off-target activity.

While this approach is the most sensitive to date for *in vivo* detection, the sensitivity is still limited by the detection limit of next-generation sequencing, and has an estimated sensitivity of 0.13%. When considering applying Cas9 editing of millions of cells, off-targets sites with cutting frequencies lower than 0.1% could still result in indel formation in thousands of cells.

1.3.2 Algorithms for off-target prediction

Various algorithms have been developed to predict potential off-targets of a given guide, based on the data obtained from the previously described methods of off-target detection, with the aim of improving specificity by careful guide design. The first of these were based on searching the genome of interest for all potential sites with sequence similarity to the on-target, allowing the user to define the parameters such as numbers of tolerated mismatches, bulges, and use of non-canonical PAM sequences (Stemmer et al., 2015; Xiao et al., 2014; Bae et al., 2014). Identified off-targets are scored based on the position and number of mismatches they contain and the PAM site used. For CCTop, the seed region was handled by giving higher penalties for mismatches closer to the PAM site (Stemmer et al., 2015). CasOT looks separately at the number of mismatches present in the seed and non-seed in determining off-target scores (Xiao et al., 2014). Although useful for comparison of the specificity of potential guide sequences, the choice of parameters is largely arbitrary with little empirical data supporting which predicted sites are likely to be cut *in vivo*. Other approaches incorporate cutting frequency determination scores, CFDs. The CFDs are per base penalty scores of the relative cutting frequency of guides containing each potential mismatch at each location. The CFDs were calculated empirically from cutting data in cells, and have improved as larger datasets of off-targets have became available (Hsu et al., 2013). These were used to score predicted off-targets and showed good results at distinguishing false-positives (Haeussler et al., 2016). However, they performed best

only for off-target sites with relatively high editing frequencies >1%.

Machine learning approaches have further improved the ability to remove false-positives from initial shortlists of off-target identified by genomic search (J. Lin and Wong, 2018; Listgarten et al., 2018). Models are trained on the existing data-sets of validated off-targets and can then be used to predict off-target sites for untested candidate guide sequences. These perform better at prediction of *bona fide* off-target sites, however the features used to make the predictions are not accessible and so they do not give any new mechanistic insight into what makes a particular sequence an off-target. In alternative machine learning approaches, predefined features such as GC content and mismatch location and identity are used to train predictive models (Abadi et al., 2017; Peng et al., 2018). The trained models outperform the traditional algorithms, whilst also revealing features such as mismatch distribution and identity, which influence off-target activity. For example, interesting features such as the off-target DNA enthalpy and DNA groove width at the PAM site were demonstrated to be important predictors (Abadi et al., 2017). Attempts have also been made to incorporate chromatin accessibility data into the prediction algorithms, based on the finding that chromatin state at potential off-target sites can effect editing efficiencies (R. Singh et al., 2015).

1.3.3 Engineering of high fidelity CRISPR/Cas Variants

A clearer understanding of the potential of off-target activity of Cas9 has increased specificity by more intelligent guide design. However, this reduces the potential number of sites which can be safely targeted and there are also concerns about off-target activity below the current level of detection, which could still be very problematic for therapeutic *in vivo* editing. As a result, there has been a concerted effort to develop Cas9 systems with higher intrinsic specificity. Many approaches to improve specificity are based around the 'excess energy' hypothesis; the idea that binding energy of Cas9 is higher than required at on-target sites and as such could be reduced without reducing on-target activity. This reduction in binding energy may however prevent stable binding to off-target sites which are already partially destabilised due to mismatches between the guide and DNA.

1.3.3.1 Guide engineering

One of the simplest approaches has been to modify the sgRNA to reduce the annealing energy between the RNA and DNA target. Truncation of the 5' end of the guide reduces activity at off-target site while retaining efficient on-target cleavage (Fu, Sander, et al., 2014). An alternative approach to reduce the binding energy of the complex is to engineer DNA bases into the RNA guide, as RNA-DNA duplexes are more stable than DNA-DNA duplexes (Figure 1.11A). While this approach shows promise, there have been conflicting reports as to which positions can be replaced with DNA while still retaining *in vivo* activity (Rueda et al., 2017; Yin et al., 2018; Kartje et al., 2018; H. Y. Kim et al., 2019). Furthermore, although there is evidence of reduced off-target editing, it is still not clear whether this is simply due to an overall reduced activity. Our work on this approach will be discussed in Chapter 4.

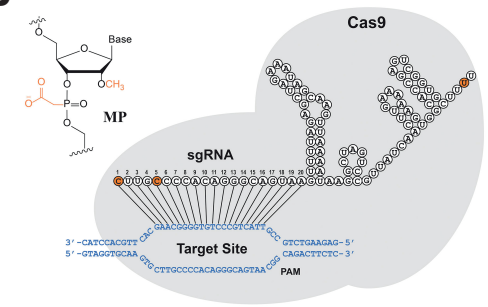
The hybridisation energy of the RNA-DNA duplex can also be tuned by incorporation of modified nucleotides. By adding 2-O-methyl-3-phosphonoacetate, MP, modified nucleotides at specific locations, the hybridisation energy is reduced and there is an improved discrimination between on-target and off-target sites (Figure 1.11B) (Ryan et al., 2018). Interestingly, the use of bridged nucleic acids, which increase RNA-DNA hybridisation by encouraging formation of the A-form helix, have also be demonstrated to increase specificity (Figure 1.11C) (Rahdar et al., 2015; Cromwell et al., 2018). While this is counter-intuitive in respect to the excess energy hypothesis, their increase in specificity may come from an increased energetic cost of the presence of mismatches in the DNA-RNA duplex which disrupt the helix. Use of these modified nucleotides can also increase stability in cells by protecting against endonucleases (Rahdar et al., 2015; Ryan et al., 2018; Cromwell et al., 2018).

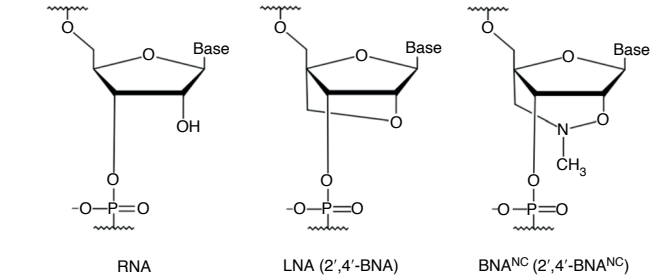
Recently, an alternative approach to was proposed whereby an hairpin structure is added to the 5' end of the guide. For Cas9 R-loop formation at a potential target site, this hairpin must be disrupted. This adds an additional energy barrier, which reduces activity at sites with mismatches while retaining on-target activity (Figure 1.11D) (Kocak et al., 2019).

A

Name	Sequence
	|---Guide Region---| |-----cr Repeat------|
	....5....10...15...20 ...25...30...35...40.
crRNA	**agcacugcacgccguaggguc guuuuagagcuaugcuguuuug**
crDNA	AGCACTGCACGCCGUAGGUC GUUUTAGAGCTATGCTGTTTTG
crHyb	**a**GCACTGCACGCCG**ua**GG**u**C G**uuu**TAGAGCTATGCTGTTTTG

RNA bases in lowercase red, DNA bases uppercase black





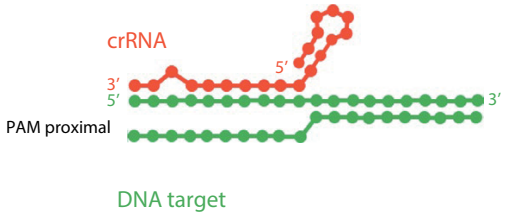


Figure 1.11: Engineering Cas9 guides for higher specificity. (**A**) DNA hybrid guides, Cas9 does not tolerate full replacement of RNA bases in the crRNA with DNA (crDNA) however a hybrid guide, containing only 7 RNA bases, exhibits high *in vitro* cleavage efficency. (Rueda et al., 2017). Our work on development of DNA containig hybrid guides with be discussed in 4. (**B**) Incorporation of 2-O-methyl-3-phosphonoacetate bases (MPs, top left) at the 5' and 3' ends of the crRNA protect against endonuclease degradation, internal MP bases a specific sites, e.g. position 5 shown, can increase on-target versus off-target Cas9 specificity (Ryan et al., 2018). (**C**) Incorporation of locked (LNA) and bridged (BNA) nucleic acids, have been demonstrate to alter Cas9 specificity (Rahdar et al., 2015; Cromwell et al., 2018) (**D**) Engineering of a hairpin at the 5' PAM distal end of the crRNA increases mismatch discrimination due to the increased energy barrier of hairpin unfolding required for stable binding (Kocak et al., 2019).

1.3.3.2 Protein engineering

Two SpCas9 varieties, enhanced specificity SpCas9, eSpCas9; and high-fidelity Cas9, HF-Cas9; have been created by introducing targeted mutations to residues involved in DNA binding (Figure 1.12) (Slaymaker et al., 2016; Kleinstiver et al., 2016). Although both variants show a reduction in off-target activity, there are also reports of reduced on-target efficiency. This suggests with these systems it is still a trade-off between specificity and activity (Vakulskas et al., 2018). Conflicting data suggests that these mutants actually retain comparable binding affinities and that increased specificity is instead due to altered conformational dynamics (Chen et al., 2017). On this basis, a further high-specificity variant, Hypa-Cas9, was generated by introducing targeted mutations, N692A/M694A/Q695A/H698A, in the REC3 domain, a key conformation change of which is thought to be required for cleavage (Chen et al., 2017). Mechanistic studies

suggest the increased specificity comes from a reduced cleavage rate when bound to a potential target, biasing activity to sites bound with high affinity; this effect has been described for other specific endonucleases suggesting another general approach to reducing off-target activity (Jeffrey C Miller et al., 2019). As well as targeted approaches, random screens and directed evolution methods have been applied to identify new variants. A bacterial screen for SpCas9 variants with reduced off-target activity but which retain high on-target activity identified a promising Cas9-R691A mutant, renamed HiFi-Cas9 (Vakulskas et al., 2018). A similar screen in yeast, targeting mutations in the REC3 domain, identified a variant, evoCas9, containing four mutations, M495V/Y515N/K526E/R661Q, which retained efficient on-target cleavage with reduced off-target activity (Casini et al., 2018). In addition, directed evolution approaches have also successfully generated high specificity variants such as Sniper-Cas9 and xCas9, which also have a broader range of tolerated PAM sequences (J. K. Lee et al., 2018; Hu et al., 2018).

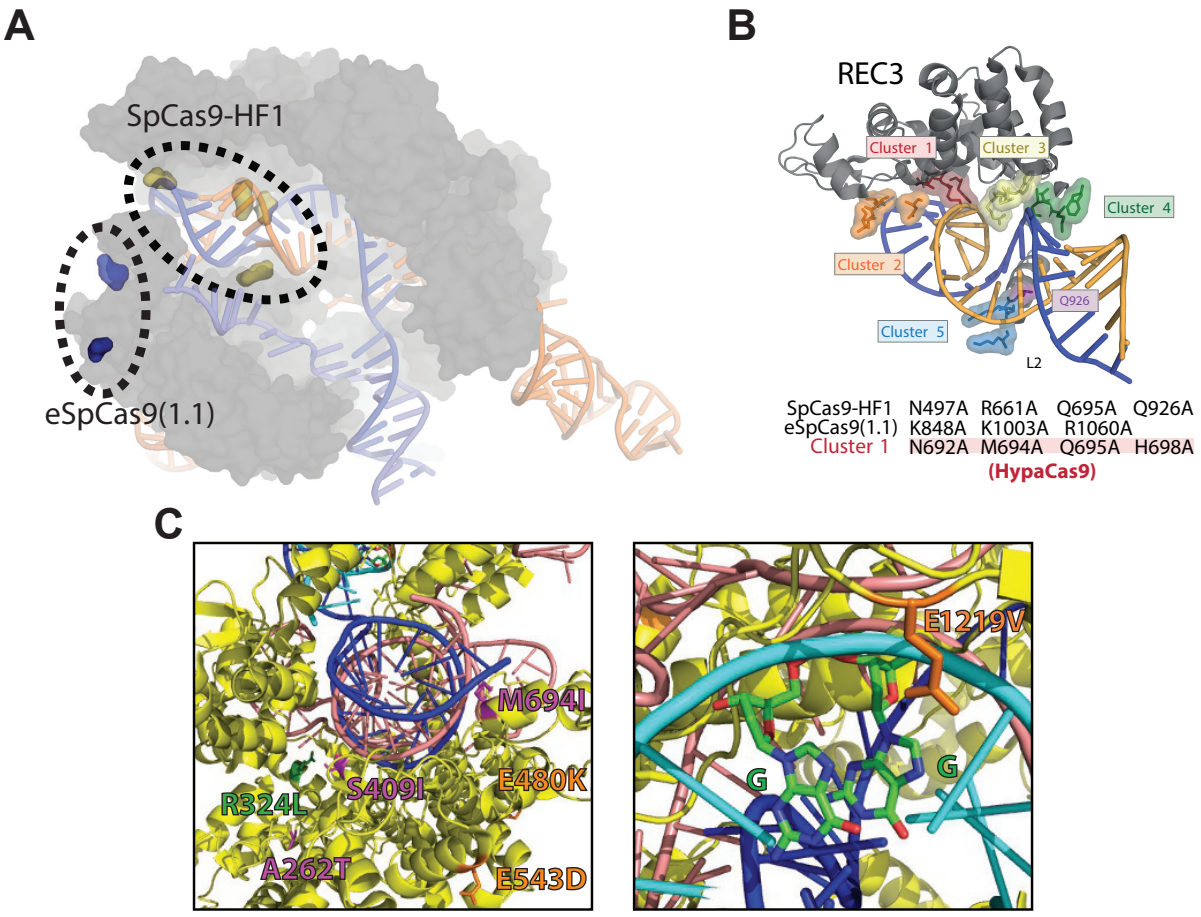


Figure 1.12: Engineering of high fidelity CRISPR/Cas Variants. (**A**) Location of mutated residues in the high fidelity Cas9 variants SpCas9-HF1 and eSpCas9 highlighted on the structure of Cas9 bound to sgRNA and target DNA (PDB: 5F9R) (Figure adapted from Chen et al., 2017). For Cas9-HF1, mutations were selected to disrupt target-strand DNA interactions (Kleinstiver et al., 2016). For eCas9 mutations were selected to effect non-target strand DNA interaction (Slaymaker et al., 2016). (**B**) Locations of clusters mutated in the REC3 domain during engineering of a high fidelity Cas9. Mutations in cluster 1 were identified to retain on-target activity while increasing specificity leading to the development of HypaCas9. HypaCas9 mutated residues in comparison to residues mutated in SpCas9-HF1 and eSpCas9. (Figure adapted from Chen et al., 2017) (**C**) Structure highlighting mutated residues selected by directed evolution to create xCas9 with increased specificity and tolerance of alternative PAM sequence. Left, mutations involved in DNA:RNA hybrid interactions thought to be responsible for increased specificity. Right, mutation hypothesised to result in broader PAM sequence tolerance. Figure adapted from Hu et al., 2018.

1.4 Single-molecule studies of CRISPR

A detailed mechanistic understanding of Cas9 target recognition and cleavage is crucial to development of safe and efficient gene editing. Structural work has provided key snapshots of the complexes activity and aided targeted engineering of higher fidelity variants. Single-molecule methods have been used extensively to observe target search, both *in vitro* and *in vivo*; interrogate conformational changes involved in assembly, binding and cleavage; and to understand mechanistically the features that effect specificity.

The first single molecule work explored the target search and binding of Cas9. Purified Cas9 was labelled with quantum dots and visualised using DNA curtains with TIRF microscopy (See 2.1) (Sternberg, Redding, et al., 2014). In this approach, a microscope slide is prepared with chromium 'barriers' and then coated with a lipid bilayer. Phage-λ DNA is tethered to the lipid bilayer by biotin-streptavin interactions and, by applying flow across the slide, teathered DNA diffusing through the bilayer is trapped by the barriers, generating DNA curtains (Figure 1.13A) (Fazio et al., 2008). Labelled Cas9 is then flowed across the DNA curtains and DNA:protein interactions visualised by TIRF microscopy. In the absence of bound crRNA:tracr, *apo*-Cas9, Cas9 displayed non-specific DNA interactions across λ-DNA (Figure 1.13B) (Sternberg, Redding, et al., 2014). However, when targeted with guide for a particular sequence highly efficient binding at the expected location was observed (Figure 1.13B, C). Stable interactions were also observed at off-target sites, though at a much lower frequency, and appeared to correlate with local PAM site density (Figure 1.13D). To visualise the target search, similar experiments were performed in the absence of flow using double-tethered DNA curtains. No evidence of 1D diffusion was observed, suggesting a 3D diffusion mechanism of target search with rapid binding an dissociation and only stable binding at the on-target site (Sternberg, Redding, et al., 2014). Interestingly, using a smFRET approach (See 2.2.2), 1D diffusion has been described, though only over relatively short distances (Globyte et al., 2019). This suggests that while Cas9 does not undergo large scale DNA scanning through 1D diffusion, on DNA interaction it can scan locally to sample potential binding sites.

To directly measure the binding and dissociation kinetics, smFRET TIRF approaches have

been used (See 2.2.2) (D. Singh, Sternberg, et al., 2016). Target DNA was labelled with Cy3 close to the PAM site and immobilised on microscope slide. The crRNA:tracr:Cas9 complex, labelled with Cy5 close to the 3' end of the crRNA, was then flowed over the DNA. Cas9 binding was observed by FRET between the labelled crRNA and target DNA and detailed kinetic information extracted by looking at dwell times in the presence and absence of mismatches (D. Singh, Sternberg, et al., 2016). In the presence of PAM proximal mismatches, binding was still observed however there was rapid dissociation compared to perfectly matched sequences or mismatches at the PAM distal end of the target.

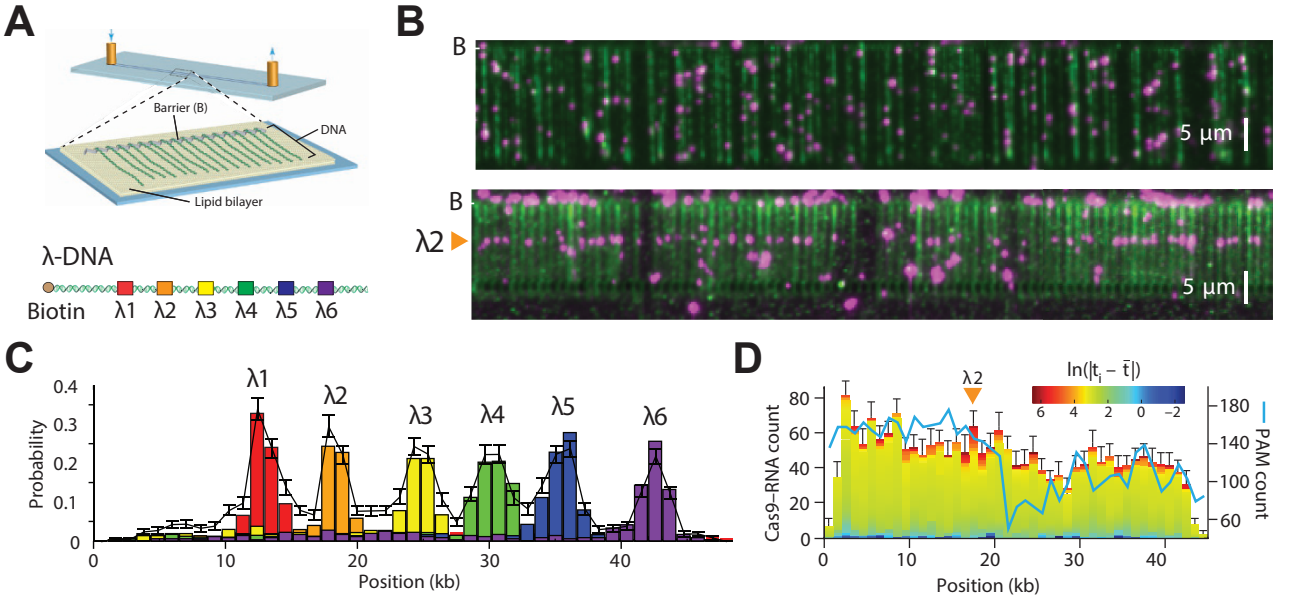


Figure 1.13: Visualising Cas9 targeting with DNA curtain. (**A**) Schematic of DNA curtains experimental design. Phage-λ DNA is caught by a barrier across the lipid bilayer coated flow cell, resulting in the formation of a curtain of DNA molecules, top. crRNA guides are designed to target 6 unique sites across the λ genome (λ1-λ6), bottom. (**B**) Top, TIRFM image of λ-DNA curtains (YOYO1 labelled, green) in the presence of quantum dot labelled apo-dCas9 (purple), in the absence of guide RNA. Bottom, DNA curtain TIRFM image in presence of λ2-crRNA:dCas9 (purple). λ2-crRNA:dCas9 is observed localising to the λ2 target site (orange arrow). (**C**) Probability distribution of binding events across λ-DNA for dCas9 complexed with different crRNA guides (1-6). For each guide a well defined binding peak is observed at the known DNA target site. (**D**) Distribution of all observed λ2-crRNA:dCas9 binding events across λDNA overlayed with PAM site counts. Figure adapted from Sternberg, Redding, et al., 2014.

The mechanism of R-loop formation was first investigated for the closely related *Streptococcus thermophilus* Cas9, StCas9, and type I *Streptococcus thermophilus* Cascade CRISPR system. Using magnetic tweezers, the DNA unwinding associated with R-loop formation could be directly measured for single-molecule binding events (Szczelkun et al., 2014). By using non-canonical PAM sequence they demonstrated a role of the PAM site in the initial formation of the R-loop and the effect of mismatches between the guide and target DNA reducing the overall

R-loop stability (Szczelkun et al., 2014). Additional work on the Cascade complex suggested a directional R-loop formation, starting at the PAM site and moving across the target sequence (Rutkauskas et al., 2015). However, Cascade R-loop formation has also been shown to occur by alternative non-directional routes, particularly in the presence of mismatches (Blosser et al., 2015; Dillard et al., 2018). Directional R-loop of has also demonstrated for Cas9 with the use of a smFRET TIRF assay, and unlike for Cascade this appears to be the only pathway. By placing FRET fluorophore pairs on the target DNA and guide RNA R-loop formation is directly observed as an increase in FRET as the RNA anneals with the DNA from the PAM proximal to PAM distal end of the target (Figure 1.14A-C) (Lim et al., 2016). The presence of mismatches prevents full R-loop formation (Figure 1.14D). These studies demonstrate the role of R-loop formation in stable binding of the Cas9 complex, and how the presence of mismatches can prevent activity due to rapid dissociation of the complex. However, it does not capture the actual target cleavage event.

Structural studies have demonstrated that large conformational changes occur between the apo-Cas9, crRNA:tracr:Cas9 and DNA bound state (Figure 1.5C) (**JinekStructuresActivation**; Jiang et al., 2015; Nishimasu et al., 2014; Anders et al., 2014). In particular, the REC domain undergoes a dramatic rearrangement from *apo*-Cas9 to guide bound Cas9 and the two cleavage domains, HNH and RuvC, undergo significant movement to reach the cleavage competent conformation (Nishimasu et al., 2014). Using FRET pairs appropriately placed on the relevant domains, the movement of HNH can be monitored (Sternberg, Lafrance, et al., 2015; Dagdas et al., 2017; Osuka et al., 2018). In bulk and single-molecule FRET experiments, a conformation change of the NHN domain is observed on binding to an on-target site but not in the presence of off-target DNA which is not stably bound by Cas9 (Sternberg, Lafrance, et al., 2015; Dagdas et al., 2017). This is coupled to movement of the second nuclease domain, RuvC, suggesting that there is allosteric regulation of the RuvC domain through HNH. This rearrangement of the nuclease domains does not occur when bound to off-target DNA that cannot be cleaved by Cas9, suggesting that the regulation of HNH provides a second level of off-target discrimination, uncoupling stable binding of the target with cleavage activity. From analysis of the single-molecule FRET trajectories, three states can be resolved: bound RNA only (R), intermediate

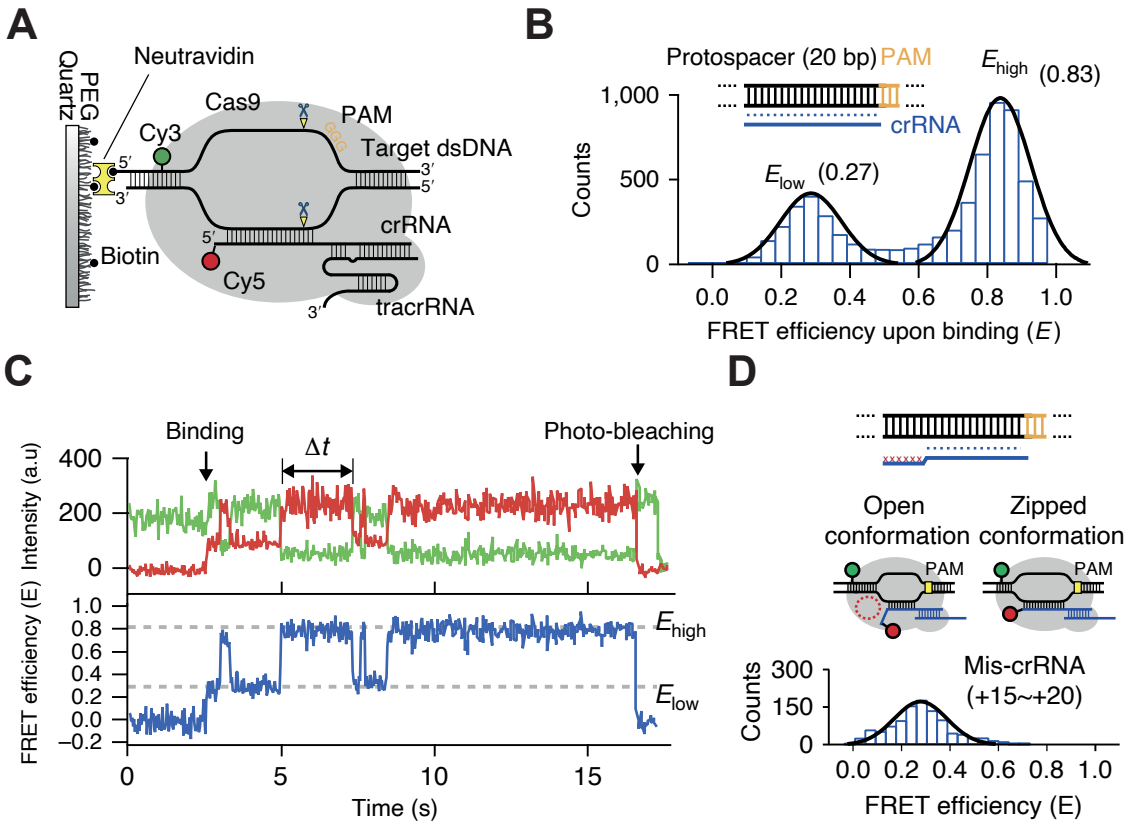


Figure 1.14: Cas9 R-loop formation observed by smFRET. (**A**) Experimental set-up, DNA is attached to PEGylated quartz slide by biotin-neutravidin interaction, target DNA is labelled with Cy3 at the and crRNA labelled with Cy5 at the 5' end. On binding and R-loop formation FRET occurs between the Cy3 and Cy5. (**B**) FRET histogram of Cas9 binding to target DNA, a population of low and high FRET states are observed. (**C**) Intensity and corresponding FRET trajectory showing R-loop formation. On Cas9 binding a low FRET state is observed (0.3) which transitions to a high FRET state (0.8) on full R-loop formation. Transitions between the two states are observed before photo-bleaching of the fluorophores. (**D**) FRET histogram showing the effect of mismatches, preventing R-loop formation and resulting in observation of only the low FRET state. Figure adapted from Lim et al., 2016.

(I) and docked (D). Mismatches that inhibit cleavage limit transitions to between the R and I state, whereas binding to an on-target site results in a relatively stable occupancy of the D state. For intermediate mismatched off-targets, which can be cleaved by Cas9 but at lower efficiencies than on-target sites, the HNH domain displays transitions between the I and D state, with occupancy of the D state correlating with cleavage efficiency (Dagdas et al., 2017). Together these studies suggest the presence of a so called 'conformational checkpoint', whereby HNH must transition to the docked state for DNA cleavage, which provides additional discrimination of off-target sites, preventing cleavage even in the presence of mismatches that accommodate stable R-loop formation. These methods have also been applied to understand how high-fidelity variants improve off-target discrimination. The eCas9 and HF-Cas9 variants, generated by

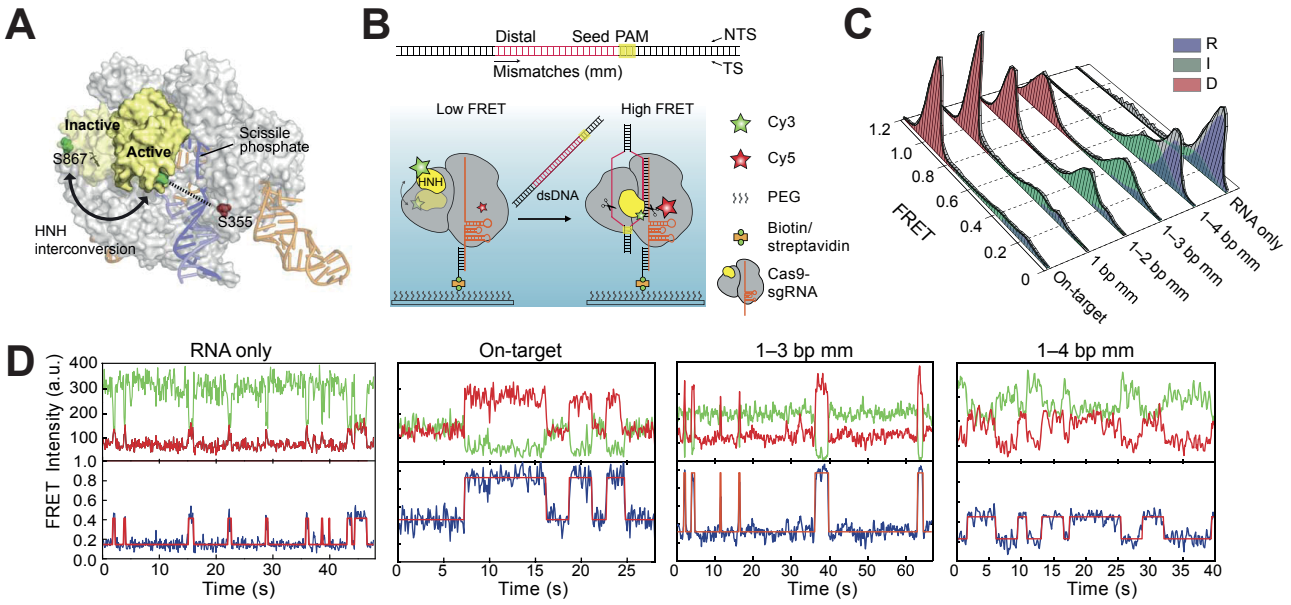


Figure 1.15: Observation of Cas9 HNH domain conformation changes on binding to on-target and mismatched target DNA by smFRET. (**A**) Diagram of labelling configurations for FRET experiments. The HNH domain is labelled with the Cy3 donor at position S867 and residue S355 labelled with the Cy5 acceptor to monitor conformational changes between the inactive and active state of HNH, left. Cas9 is immobilised on the microscope slide by biotin-streptavidin interactions using a biotinylated linker which anneal to the end of the sgRNA, right. In the absence of DNA a low FRET state is expected, on binding to DNA the movement of the HNH domain results in a high FRET state. (**C**) FRET histograms for Cas9 with RNA only, bound to on-target DNA and bound to DNA contain different numbers of mismatches. With RNA only, a single population of low FRET states is observed, R. On binding on-target DNA a single population of high FRET states is observed, D. On binding mismatched DNA a third intermediate FRET state is observed, I. (**D**) Example fluorescence intensity and corresponding FRET trajectories for Cas9 with RNA only, bound to on-target DNA and bound to DNA containing mismatches. Figure adapted from Dagdas et al., 2017.

targeted mutagenisis to destabilise DNA-protein binding, have both been studied using the complimentary approaches of observing R-loop formation and the cleavage associated HNH conformational changes (Chen et al., 2017; D. Singh, Y. Wang, et al., 2018). Furthermore, by looking at R-loop formation, it has been shown that the high fidelity variants are more sensitive to PAM distal mismatches, spending less time in the unwound state or failing to form the R-loop at all when there are greater than 3 mismatches at the PAM distal end of the target (D. Singh, Y. Wang, et al., 2018). This is despite stable binding of all variants to DNA when only 10 PAM proximal matching bases are present. When instead looking at the conformation of the HNH domain of wtCas9 it was shown that even with only 3 PAM promimal mismatches, which still tolerates R-loop formation, HNH spends significant time in the inactive intermediate state (I). The high-fidelity variants are even more sensitive, with even a single mismatch resulting in the HNH domain remaining almost entirely in the inactive conformation (Chen et al., 2017).

These results have been used to argue against a purely thermodynamic model of mismatch sensing through R-loop formation, in favour of a 'conformational checkpoint' where Cas9 can sense PAM distal mismatches. However, direct detection of the opening of the DNA target by single-molecule FRET suggests off-target discrimination could still be explained by the thermodynamics of R-loop formation. When a FRET pair was placed the across the DNA target it was shown, in the absense of mismatches, that binding and R-loop formation results in stable opening of the DNA. However, even 1-3 mismatches, where previously stable formation of the R-loop was proposed based on the alternative fluorophore configuration, resulted in reduced opening of the PAM distal DNA and transitions between open and annealed states (D. Singh, Y. Wang, et al., 2018). Furthermore, constructs containing up to 4 PAM distal mismatches but with this region pre-melted, by inserted non-complementary bases on the non-target strand, recover stable opening of the DNA (D. Singh, Y. Wang, et al., 2018). This suggests the HNH conformational change may be sensitive to full unwinding of the target rather that an intrinsic sensing of the correct base pairing between guide and target. For the high fidelity variants, eCas9 and HF-Cas9, stable DNA opening was more sensitive to mismatches than for wtCas9 (D. Singh, Y. Wang, et al., 2018). This was also demonstrated for the second generation Hypa-Cas9 variant and, to a lesser extent, also for Sniper-Cas9 (Okafor et al., 2019).

This important mechanistic detail of Cas9 mismatch sensing is still not fully resolved, however comprehensive kinetic models have been developed which perform well in explaining the observed mismatch sensitivities of Cas9 (Klein et al., 2018; Eslami-Mossallam et al., 2020). To describe R-loop formation and cleavage, a model was devised incorporating four kinetic bias parameters: PAM binding, Δ_{PAM}; extension through a matched base, Δ_C, extension through a mismatched base, Δ_I, and cleavage once the R-loop has formed, Δ_{clv} (Figure 1.16A). This model recapitulated empirical data on binding and cleavage efficiencies based on number and distribution of mismatches across the target (Klein et al., 2018; Eslami-Mossallam et al., 2020). The calculated energy landscapes demonstrate that, in the absence of mismatches, initial PAM binding and formation of the first 7 bps of the R-loop is the rate limiting step (Figure 1.16B). Furthermore, a local minimum defines an intermediate state of formation of an 8-12 bp, whereby stable Cas9 binding occurs before full formation of the R-loop as required for cleavage (Figure

1.16B). An important implication of this model is that, due to the directional progression across the target sequence during R-loop formation, from the PAM proximal to PAM distal end, encountering a mismatch shifts the energy landscape of all subsequent positions (Figure 1.16C,D). As such, a mismatch in the seed region results in an increase in the energy barrier of the rate limiting step of reaching the intermediate R-loop state (Figure 1.16C) reducing Cas9 target binding and cleavage. The overall downward slope of the energy landscape allows for better tolerance of mismatches encountered towards the PAM distal end. Interestingly, this model also predicts that a mismatch towards the PAM distal end of the target can create a new rate limiting step of formation of the end of the R-loop, trapping Cas9 in a bound state unable to induce cleavage (Figure 1.16D). This can be reconciled with the previously discussed idea of a conformational checkpoint for Cas9 cleavage, giving a thermodynamic explanation, and demonstrating that this checkpoint could be better described as an energy barrier.

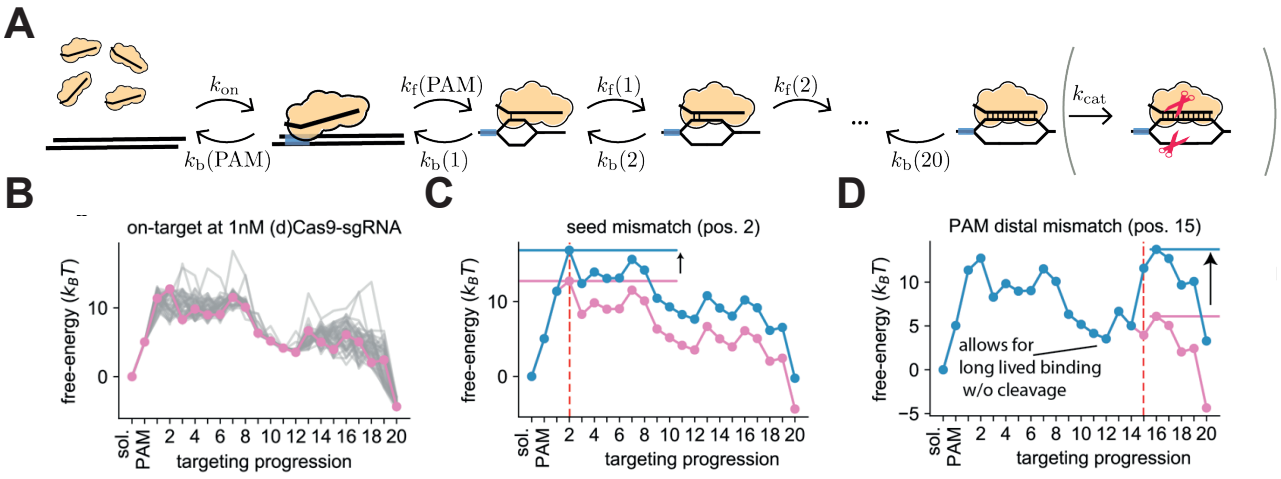


Figure 1.16: Kinetic model of off-target discrimination. (**A**) Kinetic model parameters (**B**) Calculated free energy landscape for Cas9 binding and R-loop formation, where sol. is Cas9 in solution before binding, PAM is Cas9 bound to PAM site, and targeting progression the basewise formation of the R-loop. The rate limiting step is the transition from unbound to formation of an 8-12 bp R-loop. (**C**) Energy landscapes demonstrating effect of single mismatch position within the seed region. This increases the energy barrier of the rate limiting step of formation of an 8-12 bp R-loop. (**D**) Energy landscapes demonstrating effect of a single mismatch at the PAM distal end. Formation of the end of the R-loop becomes the rate limiting step, creating a local minimum of Cas9 bound but without cleaving DNA. Figure adapted from Eslami-Mossallam et al., 2020.

1.5 Thesis Objectives

CRISPR systems, and in particular CRISPR/Cas9, have already been demonstrated to be a powerful tool for genetic manipulation both in the context of biological research but also with huge therapeutic potential. However, the potential of off-target activity clearly poses a significant hurdle. While methods for detection of off-target sites both *in vitro* and *in vivo* have been developed, these are still limited in their sensitivity and, when considering application of Cas9 in complex organisms of millions of cells, there are concerns as to the risks associated with lower frequency off-target events. One approach to solve this problem has been to try and engineer higher fidelity CRISPR/Cas variants by modifications to the protein and RNA, either through random mutagenisis approaches or by rational design based of structural studies. The use of single-molecule techniques have allowed detailed dissection of some of the key mechanistic steps in Cas9 binding and cleavage which further inform design of higher fidelity systems. However, key questions as to the mechanism of on- versus off-target site discrimination are still unanswered. In particular, the distinction between rejection of off-target sites due to reduced R-loop stability, or the ability for Cas9 to detect the presence of mismatches between the guide and DNA target, a protein mediated 'checkpoint'. Additionally, how does altering the base pairing energy between guide and target affect sequence discrimination? It is largely unknown how other factors present in the cellular context influence Cas9 specificity. For instance how do changes to DNA structure affect activity?

In this thesis, I will present my work addressing these open questions with the aim of furthering our understanding of the basis of Cas9 target interrogation, and in particular the factors that can affect Cas9 specificity. I will first discuss my collaborative project done with Benjamin Taylor's group in the Discovery Sciences department at AstraZeneca, in which we looked at the effect of replacing RNA bases within the crRNA and tracrRNA guides with DNA bases. For this projected, I performed single-molecule Cas9 binding smFRET experiments to investigate R-loop formation with the DNA/RNA hybrid guides that our collaborators had developed, in order to understand how RNA-DNA and DNA-DNA base pairing influences target discrimination and Cas9 cleavage activity (See Chapter 4, Rueda et al., 2017). The remainder of the thesis

will be dedicated to my work on how structural changes to the target DNA alters Cas9 activity. Firstly, I will discuss how force stretching DNA affects specificity and discrimination of on- and off-target sites based on my work using optical tweezers with confocal microscopy and smFRET to track single molecules of Cas9 binding to on- and off-target sites (See Chapter 5, Newton et al., 2019). Secondly, I will present results on the effect of negative supercoiling in target DNA on Cas9 activity, again using a single-molecule optical tweezers approach, and also through development of a novel next-generation sequencing based technique for off-target detection in the presence of supercoiling (See Chapter 6).

References

Abadi, Shiran et al. (Oct. 2017). "A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action". *PLOS Computational Biology* 13.10. Ed. by Han Xu, e1005807.

Abudayyeh, Omar O et al. (2016). "C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector". *Science* 353.6299.

Akcakaya, Pinar et al. (Sept. 2018). "In vivo CRISPR editing with no detectable genome-wide off-target mutations". *Nature* 561.7723, pp. 416–419.

Anders, Carolin et al. (Sept. 2014). "Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease". *Nature* 513.7519, pp. 569–573.

Apolonia, Luis et al. (Nov. 2007). "Stable Gene Transfer to Muscle Using Non-integrating Lentiviral Vectors". *Molecular Therapy* 15.11, pp. 1947–1954.

Arber, Werner and Daisy Dussoix (1962). "Host specificity of DNA produced by Escherichia coli: I. Host controlled modification of bacteriophage λ". *Journal of Molecular Biology* 5.1, pp. 18–36.

Bae, Sangsu, Jeongbin Park, and J.-S. Kim (May 2014). "Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases". *Bioinformatics* 30.10, pp. 1473–1475.

Barrangou, R. et al. (Mar. 2007). "CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes". *Science* 315.5819, pp. 1709–1712.

Bibikova, Marina et al. (May 2003). *Enhancing gene targeting with designed zinc finger nucleases.*

Blosser, Timothy R. et al. (Apr. 2015). "Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex". *Molecular Cell* 58.1, pp. 60–70.

Bolotin, Alexander et al. (2005). "Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin". *Microbiology* 151.8, pp. 2551–2561.

Brouns, S. J. J. et al. (Aug. 2008). "Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes". *Science* 321.5891, pp. 960–964.

Brunetti-Pierri, Nicola and Philip Ng (Oct. 2017). "Gene therapy with helper-dependent adenoviral vectors: lessons from studies in large animal models". *Virus Genes* 53.5, pp. 684–691.

Cameron, Peter et al. (June 2017). "Mapping the genomic landscape of CRISPR–Cas9 cleavage". *Nature Methods* 14.6, pp. 600–606.

Casini, Antonio et al. (2018). "A highly specific SpCas9 variant is identified by in vivo screening in yeast". *Nature Biotechnology* 36.3, pp. 265–271.

Chapman, J. Ross, Martin R.G. Taylor, and Simon J. Boulton (Aug. 2012). "Playing the End Game: DNA Double-Strand Break Repair Pathway Choice". *Molecular Cell* 47.4, pp. 497–510.

Chen, Janice S. et al. (Oct. 2017). "Enhanced proofreading governs CRISPR–Cas9 targeting accuracy". *Nature* 550.7676, pp. 407–410.

Chew, Wei Leong et al. (Oct. 2016). "A multifunctional AAV–CRISPR–Cas9 and its host response". *Nature Methods* 13.10, pp. 868–874.

Choi, J. G. et al. (July 2016). "Lentivirus pre-packed with Cas9 protein for safer gene editing". *Gene Therapy* 23.7, pp. 627–633.

Cong, Le et al. (2013). "Multiplex genome engineering using CRISPR/Cas systems". *Science* 339.6121, pp. 819–823.

Cromwell, Christopher R et al. (2018). "Incorporation of bridged nucleic acids into CRISPR RNAs improves Cas9 endonuclease specificity". *Nature Communications* 9.1.

Crosetto, Nicola et al. (2013). "Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing". *Nature Methods* 10.4, pp. 361–365.

Dagdas, Yavuz S et al. (Aug. 2017). "A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9". *Science Advances* 3.8, eaao0027.

Deltcheva, Elitza et al. (Mar. 2011). "CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III". *Nature* 471.7340, pp. 602–607.

Deveau, Hélène et al. (2008). "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus". *Journal of Bacteriology* 190.4, pp. 1390–1400.

Dillard, Kaylee E et al. (2018). "Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex". *Cell* 175.4, pp. 934–946.

Ding, Qiurong et al. (Apr. 2013). "Enhanced Efficiency of Human Pluripotent Stem Cell Genome Editing through Replacing TALENs with CRISPRs". *Cell Stem Cell* 12.4, pp. 393–394.

Doudna, Jennifer A (Feb. 2020). "The promise and challenge of therapeutic genome editing". *Nature* 578.7794, pp. 229–236.

Dussoix, Daisy and Werner Arber (1962). "Host specificity of DNA produced by Escherichia coli: II. Control over acceptance of DNA from infecting phage λ". *Journal of Molecular Biology* 5.1, pp. 37–49.

Eslami-Mossallam, Behrouz et al. (July 2020). "A kinetic model improves off-target predictions and reveals the physical basis of Sp Cas9 fidelity". *bioRxiv*, p. 2020.05.21.108613.

Fazio, Teresa et al. (Sept. 2008). "DNA Curtains and Nanoscale Curtain Rods: High-Throughput Tools for Single Molecule Imaging". *Langmuir* 24.18, pp. 10524–10531.

Fogarty, Norah M. E. et al. (Oct. 2017). "Genome editing reveals a role for OCT4 in human embryogenesis". *Nature* 550.7674, pp. 67–73.

Frock, Richard L et al. (2015). "Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases". *Nature Biotechnology* 33.2, pp. 179–188.

Fu, Yanfang, Jennifer A Foden, et al. (Sept. 2013). "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells". *Nature Biotechnology* 31.9, pp. 822–826.

Fu, Yanfang, Jeffry D Sander, et al. (Mar. 2014). "Improving CRISPR-Cas nuclease specificity using truncated guide RNAs". *Nature Biotechnology* 32.3, pp. 279–284.

Garneau, Josiane E et al. (2010). "The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA". *Nature* 468.7320, pp. 67–71.

Gasiunas, Giedrius et al. (Sept. 2012). "Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria". *Proceedings of the National Academy of Sciences* 109.39, E2579–E2586.

Globyte, Viktorija et al. (Feb. 2019). "CRISPR /Cas9 searches for a protospacer adjacent motif by lateral diffusion". *The EMBO Journal* 38.4, e99466.

Haeussler, Maximilian et al. (Dec. 2016). "Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR". *Genome Biology* 17.1, p. 148.

Hale, Caryn R et al. (2009). "RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex". *Cell* 139.5, pp. 945–956.

Hermans, P. W.M. et al. (1991). "Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains". *Infection and Immunity* 59.8, pp. 2695–2705.

Hsu, Patrick D. et al. (Sept. 2013). "DNA targeting specificity of RNA-guided Cas9 nucleases". *Nature Biotechnology* 31.9, pp. 827–832.

Hu, Johnny H. et al. (Apr. 2018). "Evolved Cas9 variants with broad PAM compatibility and high DNA specificity". *Nature* 556.7699, pp. 57–63.

Ishino, Yoshizumi et al. (1987). "Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product." *Journal of Bacteriology* 169.12, pp. 5429–5433.

Jansen, Ruud. et al. (Mar. 2002). "Identification of genes that are associated with DNA repeats in prokaryotes". *Molecular Microbiology* 43.6, pp. 1565–1575.

Jensen, Kristopher Torp et al. (July 2017). "Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency". *FEBS Letters* 591.13, pp. 1892–1901.

Jiang, Fuguo et al. (June 2015). "A Cas9-guide RNA complex preorganized for target DNA recognition". *Science* 348.6242, pp. 1477–1481.

Jinek, Martin et al. (Aug. 2012). "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity". *Science* 337.6096, pp. 816–821.

Josephsen, Jytte and Horst Neve (Jan. 2004). "Bacteriophage and antiphage mechanisms of lactic acid bacteria". *Lactic Acid Bacteria Microbiological and Functional Aspects, Third Edition: Revised and Expanded*. CRC Press, pp. 295–350.

Joung, J. Keith and Jeffry D Sander (Jan. 2013). "TALENs: a widely applicable technology for targeted genome editing". *Nature Reviews Molecular Cell Biology* 14.1, pp. 49–55.

Kang, Xiangjin et al. (May 2016). "Introducing precise genetic modifications into human 3PN embryos by CRISPR/Cas-mediated genome editing". *Journal of Assisted Reproduction and Genetics* 33.5, pp. 581–588.

Kartje, Zachary J et al. (May 2018). "Chimeric Guides Probe and Enhance Cas9 Biochemical Activity". *Biochemistry* 57.21, pp. 3027–3031.

Kelly, Thomas J and Hamilton O. Smith (July 1970). "A restriction enzyme from Hemophilus influenzae. II. Base sequence of the recognition site". *Journal of Molecular Biology* 51.2, pp. 393–409.

Kim, Daesik et al. (2015). "Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells". *Nature Methods* 12.3, pp. 237–243.

Kim, Hyo Young et al. (2019). "Chimeric crRNAs with 19 DNA residues in the guide region show the retained DNA cleavage activity of Cas9 with potential to improve the specificity". *Chemical Communications* 55.24, pp. 3552–3555.

Kim, S. et al. (June 2014). "Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins". *Genome Research* 24.6, pp. 1012–1019.

Kim, Yang Gyun, Jooyeun Cha, and Srinivasan Chandrasegaran (1996). "Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain". *Proceedings of the National Academy of Sciences of the United States of America* 93.3, pp. 1156–1160.

Klein, Misha et al. (Feb. 2018). "Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules". *Cell Reports* 22.6, pp. 1413–1423.

Kleinstiver, Benjamin P et al. (Jan. 2016). "High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects". *Nature* 529.7587, pp. 490–495.

Kocak, D. Dewran et al. (2019). "Increasing the specificity of CRISPR systems with engineered RNA secondary structures". *Nature Biotechnology* 37.6, pp. 657–666.

Le Rhun, Anaïs et al. (Apr. 2019). "CRISPR-Cas in Streptococcus pyogenes". *RNA Biology* 16.4, pp. 380–389.

Lee, Jungjoon K. et al. (Dec. 2018). "Directed evolution of CRISPR-Cas9 to increase its specificity". *Nature Communications* 9.1, p. 3048.

Lee, Kunwoo et al. (2017). "Nanoparticle delivery of Cas9 ribonucleoprotein and donor DNA in vivo induces homology-directed DNA repair". *Nature Biomedical Engineering* 1.11, pp. 889–901.

Liang, Puping et al. (May 2015). "CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes". *Protein & Cell* 6.5, pp. 363–372.

Liang, Xiquan et al. (Aug. 2015). "Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection". *Journal of Biotechnology* 208, pp. 44–53.

Lim, Youngbin et al. (Nov. 2016). "Structural roles of guide RNAs in the nuclease activity of Cas9 endonuclease". *Nature Communications* 7, p. 13350.

Lin, Jiecong and Ka Chun Wong (2018). "Off-target predictions in CRISPR-Cas9 gene editing using deep learning". *Bioinformatics*. Vol. 34. 17, pp. i656–i663.

Lin, Yanni et al. (2014). "CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences". *Nucleic Acids Research* 42.11, pp. 7473–7485.

Lino, Christopher A et al. (Jan. 2018). "Delivering CRISPR: a review of the challenges and approaches". *Drug Delivery* 25.1, pp. 1234–1257.

Lisowski, Leszek, Szun Szun Tay, and Ian Edward Alexander (Oct. 2015). "Adeno-associated virus serotypes for gene therapeutics". *Current Opinion in Pharmacology* 24, pp. 59–67.

Listgarten, Jennifer et al. (Jan. 2018). "Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs". *Nature Biomedical Engineering* 2.1, pp. 38–47.

Long, Chengzu et al. (Jan. 2016). "Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy". *Science* 351.6271, pp. 400–403.

Maddalo, Danilo et al. (2014). "In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system". *Nature* 516.7531, pp. 423–428.

Makarova, Kira S, Daniel H Haft, et al. (June 2011). "Evolution and classification of the CRISPR–Cas systems". *Nature Reviews Microbiology* 9.6, pp. 467–477.

Makarova, Kira S, Yuri I Wolf, Omer S Alkhnbashi, et al. (Nov. 2015). "An updated evolutionary classification of CRISPR–Cas systems". *Nature Reviews Microbiology* 13.11, pp. 722–736.

Makarova, Kira S, Yuri I Wolf, Jaime Iranzo, et al. (Feb. 2020). "Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants". *Nature Reviews Microbiology* 18.2, pp. 67–83.

Mali, P. et al. (Feb. 2013). "RNA-Guided Human Genome Engineering via Cas9". *Science* 339.6121, pp. 823–826.

Miller, Jeffrey C et al. (2019). "Enhancing gene editing specificity by attenuating DNA cleavage kinetics". *Nature Biotechnology* 37.8, pp. 945–952.

Miller, Jeffrey C. et al. (Feb. 2011). "A TALE nuclease architecture for efficient genome editing". *Nature Biotechnology* 29.2, pp. 143–148.

Mojica, F. J.M., C. Díez-Villaseñor, et al. (2009). "Short motif sequences determine the targets of the prokaryotic CRISPR defence system". *Microbiology* 155.3, pp. 733–740.

Mojica, F. J.M., G. Juez, and F. Rodriguez-Valera (Aug. 1993). "Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites". *Molecular Microbiology* 9.3, pp. 613–621.

Mojica, F.J.M. et al. (July 1995). "Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning". *Molecular Microbiology* 17.1, pp. 85–93.

Mojica, Francisco J. M., Cesar Diez-Villasenor, et al. (Apr. 2000). "Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria". *Molecular Microbiology* 36.1, pp. 244–246.

Mojica, Francisco J. M. and Francisco Rodriguez-Valera (Sept. 2016). "The discovery of CRISPR in archaea and bacteria". *The FEBS Journal* 283.17, pp. 3162–3169.

Mojica, Francisco J.M. et al. (2005). "Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements". *Journal of Molecular Evolution* 60.2, pp. 174–182.

Newton, Matthew D. et al. (Mar. 2019). "DNA stretching induces Cas9 off-target activity". *Nature Structural & Molecular Biology* 26.3, pp. 185–192.

Nishimasu, Hiroshi et al. (Feb. 2014). "Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA". *Cell* 156.5, pp. 935–949.

Okafor, Ikenna C et al. (2019). "Single molecule analysis of effects of non-canonical guide RNAs and specificity-enhancing mutations on Cas9-induced DNA unwinding". *Nucleic Acids Research* 47.22, pp. 11880–11888.

Osuka, Saki et al. (May 2018). "Real-time observation of flexible domain movements in CRISPR–Cas9". *The EMBO Journal* 37.10.

Palmer, Donna and Philip Ng (Nov. 2003). "Improved system for helper-dependent adenoviral vector production". *Molecular Therapy* 8.5, pp. 846–852.

Pattanayak, Vikram et al. (Sept. 2013). "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity". *Nature Biotechnology* 31.9, pp. 839–843.

Patterson, Adrian G, Mariya S Yevstigneyeva, and Peter C Fineran (June 2017). "Regulation of CRISPR–Cas adaptive immune systems". *Current Opinion in Microbiology* 37, pp. 1–7.

Peng, Hui et al. (2018). "Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions". *Bioinformatics*. Vol. 34. 17, pp. i757–i765.

Platt, Randall J et al. (2014). "CRISPR-Cas9 knockin mice for genome editing and cancer modeling". *Cell* 159.2, pp. 440–455.

Pourcel, C, G Salvignol, and Gilles Vergnaud (2005). "CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies". *Microbiology* 151.3, pp. 653–663.

Rahdar, Meghdad et al. (Dec. 2015). "Synthetic CRISPR RNA-Cas9–guided genome editing in human cells". *Proceedings of the National Academy of Sciences* 112.51, E7110–E7117.

Ran, F Ann et al. (Apr. 2015). "In vivo genome editing using Staphylococcus aureus Cas9". *Nature* 520.7546, pp. 186–191.

Raveux, Aurélien, Sandrine Vandormael-Pournin, and Michel Cohen-Tannoudji (2017). "Optimization of the production of knock-in alleles by CRISPR/Cas9 microinjection into the mouse zygote". *Scientific Reports* 7.

Roy, Paul H. and Hamilton O. Smith (1973a). "DNA methylases of Hemophilus influenzae Rd. I. Purification and properties". *Journal of Molecular Biology* 81.4, pp. 427–444.

— (1973b). "DNA methylases of Hemophilus influenzae Rd. II. Partial recognition site base sequences". *Journal of Molecular Biology* 81.4, pp. 445–459.

Rueda, Fernando Orden et al. (Dec. 2017). "Mapping the sugar dependency for rational generation of a DNA-RNA hybrid-guided Cas9 endonuclease". *Nature Communications* 8.1, p. 1610.

Rutkauskas, Marius et al. (2015). "Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection". *Cell Reports* 10.9, pp. 1534–1543.

Ryan, Daniel E et al. (2018). "Improving CRISPR-Cas specificity with chemical modifications in single-guide RNAs". *Nucleic Acids Research* 46.2, pp. 792–803.

S. Ahi, Yadvinder, Dinesh S. Bangari, and Suresh K. Mittal (Aug. 2011). "Adenoviral Vector Immunity: Its Implications and Circumvention Strategies". *Current Gene Therapy* 11.4, pp. 307–320.

Sander, Jeffry D and J Keith Joung (Apr. 2014). "CRISPR-Cas systems for editing, regulating and targeting genomes". *Nature Biotechnology* 32.4, pp. 347–355.

Singh, Digvijay, Samuel H. Sternberg, et al. (Sept. 2016). "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Nature Communications* 7, p. 12778.

Singh, Digvijay, Yanbo Wang, et al. (Apr. 2018). "Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis". *Nature Structural & Molecular Biology* 25.4, pp. 347–354.

Singh, Ritambhara et al. (Oct. 2015). "Cas9-chromatin binding information enables more accurate CRISPR off-target prediction". *Nucleic Acids Research* 43.18, e118–e118.

Slaymaker, Ian M. et al. (Jan. 2016). "Rationally engineered Cas9 nucleases with improved specificity". *Science* 351.6268, pp. 84–88.

Smith, Hamilton O. and K.W. Welcox (July 1970). "A Restriction enzyme from Hemophilus influenzae". *Journal of Molecular Biology* 51.2, pp. 379–391.

Stemmer, Manuel et al. (Apr. 2015). "CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool". *PLOS ONE* 10.4. Ed. by Stefan Maas, e0124633.

Sternberg, Samuel H., Benjamin Lafrance, et al. (2015). "Conformational control of DNA target cleavage by CRISPR-Cas9". *Nature* 527.7576, pp. 110–113.

Sternberg, Samuel H., Sy Redding, et al. (Mar. 2014). "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9". *Nature* 507.7490, pp. 62–67.

Sun, Wujin et al. (Oct. 2015). "Self-Assembled DNA Nanoclews for the Efficient Delivery of CRISPR-Cas9 for Genome Editing". *Angewandte Chemie International Edition* 54.41, pp. 12029–12033.

Swiech, Lukasz et al. (Jan. 2015). "In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9". *Nature Biotechnology* 33.1, pp. 102–106.

Szczelkun, Mark D. et al. (July 2014). "Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes". *Proceedings of the National Academy of Sciences of the United States of America* 111.27, pp. 9798–9803.

Tang, Lichun et al. (June 2017). "CRISPR/Cas9-mediated gene editing in human zygotes using Cas9 protein". *Molecular Genetics and Genomics* 292.3, pp. 525–533.

Truong, Dong-Jiunn Jeffery et al. (July 2015). "Development of an intein-mediated split–Cas9 system for gene therapy". *Nucleic Acids Research* 43.13, pp. 6450–6458.

Tsai, Shengdar Q, Nhu T Nguyen, et al. (May 2017). "CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets". *Nature Methods* 14.6, pp. 607–614.

Tsai, Shengdar Q, Zongli Zheng, et al. (Mar. 2015). "GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases". *Nature Biotechnology* 33.2, pp. 187–197.

Urnov, Fyodor D et al. (Sept. 2010). "Genome editing with engineered zinc finger nucleases". *Nature Reviews Genetics* 11.9, pp. 636–646.

Vakulskas, Christopher A et al. (Aug. 2018). "A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells". *Nature Medicine* 24.8, pp. 1216–1224.

Wang, Haoyi et al. (May 2013). "One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering". *Cell* 153.4, pp. 910–918.

Wienert, Beeke et al. (2019). "Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq". *Science* 364.6437, pp. 286–289.

Wood, Andrew J et al. (July 2011). "Targeted Genome Editing Across Species Using ZFNs and TALENs". *Science* 333.6040, pp. 307–307.

Wu, Zhijian, Aravind Asokan, and R Jude Samulski (Sept. 2006). "Adeno-associated Virus Serotypes: Vector Toolkit for Human Gene Therapy". *Molecular Therapy* 14.3, pp. 316–327.

Xiao, An et al. (Apr. 2014). "CasOT: a genome-wide Cas9/gRNA off-target searching tool". *Bioinformatics* 30.8, pp. 1180–1182.

Xue, Chaoyou et al. (Dec. 2015). "CRISPR interference and priming varies with individual spacer sequences". *Nucleic Acids Research* 43.22, pp. 10831–10847.

Yan, Winston X et al. (2017). "BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks". *Nature Communications* 8.

Yin, Hao et al. (Mar. 2018). "Partial DNA-guided Cas9 enables genome editing with reduced off-target activity". *Nature Chemical Biology* 14.3, pp. 311–316.

Zuris, John A et al. (Jan. 2015). "Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo". *Nature Biotechnology* 33.1, pp. 73–80.

Chapter 2

Background Theory

2.1 The single-molecule approach

For the most part, biological processes have been studied by ensemble methods, by observing the average activity of thousands of molecules to understand the overall activity of a system. However, this can be a simplification of the heterogeneity present in biological processes. Without observing molecules individually, the distribution around the mean is unknown. A single population of molecules with tightly distributed activities could give the same mean activity as a two or more populations, some of which are highly active and some with very low activity. As such, ensemble averaging obscures important activities if they occur in only a small fraction of the population, or for a short period of time. As an analogy, if one was to look at the average activity of a large group of surfers at a beach, one might conclude nothing much is happening; on average the surfers appear to be merely bobbing up and down, floating at the same spot. However, if you observe each surfer separately, you realise that every so often someone will take off, catching a wave and riding it into the beach, before paddling back out to join the group. With single molecule observation, these important events can be captured, even if they only occur transiently. Furthermore, by observing the route individual molecules take to an end point, intermediate states and mechanistic steps can be identified.

A variety of techniques are now available for single-molecule studies. While significant progress has been made with *in vivo* single-molecule methods, the greatest mechanistic detail is still only achievable with *in vitro* methods, and this will be the focus of discussion here. *In vitro* single-molecule methods can be broadly categorised as fluorescence or force based. Fluorescence-based approaches can be used to image and track fluorescently-labelled molecules of interest, for example motor proteins interacting with actin filaments or microtubules, or DNA binding proteins interacting with DNA. The major challenge is to reach the signal to noise ratio required to observe fluorescence from a single-molecule. However, this can be achieved by using total internal reflection microscopy, TIRM (See Section 2.1), or confocal scanning microscopy (See Section 2.3). In addition, the use of single-molecule Förster resonance energy transfer, smFRET, enables direct monitoring of conformational changes within a molecule, or the interactions between two molecules (See Section 2.2). Force-based single-molecule techniques include magnetic and optical tweezers, where forces can be applied and measured on individual molecules tethered to beads, which are then manipulated with either magnetic fields or optical traps (See Section 2.3.1). Optical tweezers more easily lend themselves to the addition of fluroescence imaging (See Section 2.3), whereas magnetic tweezers have the additional benefit of easily allowing the application of torque to the molecule by rotation of the magnetic field. Atomic force microscopy, AFM, can also be used to apply and measure forces in a similar way to optical and magnetic tweezers, using a cantilever tip to interact with the sample. High-resolution AFM has been used to image biological samples, providing structural information of individual molecules and recently high-speed AFM has allowed real-time imaging of protein dynamics. Often overlooked as a single-molecule technique, cryogenic electron microscopy, cryo-EM, is now routinely used to solve protein structures at Angstrom resolution. Although obtaining high-resolution requires averaging images of thousands of particles, cryo-EM is at its heart a single-molecule method. During particle classification, images can be sorted into sub-populations which enables identification of distinct conformations present within the population. Furthermore, the development of time-resolved cryo-EM enables the changes in these conformations to be observed over time.

Two key single-molecule techniques were used in this work, single-molecule Förster resonance energy transfer, smFRET, utilising total internal reflection microscopy, TIRF; and optical

tweezers with fluorescence confocal microscopy. The following sections provide a brief history and background theory of both techniques for the study of biological systems.

2.2 TIRF-smFRET

Single-molecule FRET is a fluorescence-based technique, requiring the detection of photons emitted from single fluorophores. In order to reach the level of single-molecule observation by fluorescence, the first challenge was to reach a sufficiently high signal-to-noise ratio that a single molecule in a biological sample could be detected. In the earliest examples of single molecule experiments, this was achieved by attaching hundreds of fluorophores to single antibody molecules, in order to obtain enough signal to detect them (Hirschfeld, 1976). An alternative solution to this problem is to dramatically reduce the background signal in the sample; one way this has been achieved is the use of total-internal reflection microscopy (TIRF, See Section 2.2.1). This enabled observation of single-labelled molecules of myosin and ATP turnover for the first time (Funatsu et al., 1995). In parallel, it was demonstrated that Förster resonance energy transfer (FRET, See Section 2.2.2) could be used to measure the interaction between two molecules (Ha et al., 1996) or conformational changes within a molecule (Taekjip Ha, Ting, et al., 1999; Taekjip Ha, Zhuang, et al., 1999). The combination of TIRF microscopy with FRET enabled single-molecule detection of conformation states over hundreds of seconds, with one of the earliest examples being the observation of multiple folding pathways of ribozyme molecules (Zhuang, 2000). Over the past two decades, TIRF microscopy combined with smFRET, TIRF-smFRET, has become a well established technique providing mechanistic information on a broad range of biological systems including DNA replication (Christian et al., 2009), transcription (Kapanidis et al., 2006), RNA folding (Bokinsky et al., 2003; Rueda and Walter, 2005; Alemán et al., 2008; Zhao and Rueda, 2009), splicing (Steiner et al., 2008; Warnasooriya and Rueda, 2014; Karunatilaka and Rueda, 2009), mRNA translation (Munro et al., 2007; Blanchard et al., 2004), nucleosome remodelling (Willhoft et al., 2018) and CRISPR/Cas9 activity as previously discussed (See Section 1.4) (Lim et al., 2016; Singh, Sternberg, et al., 2016; Dagdas et al., 2017; Chen et al., 2017; Singh, Y. Wang, et al., 2018).

2.2.1 TIRF Microscopy

Total internal reflection microscopy, TIRFM, is a method used to achieve localised excitation of fluorophores very close to the surface of the microscope slide, thereby reducing background fluorescence. This approach has been used to image cells for a wide variety of biological studies, however it is also vital for surface-based single-molecules studies to reach the signal-to-noise ratio required for single molecule detection.

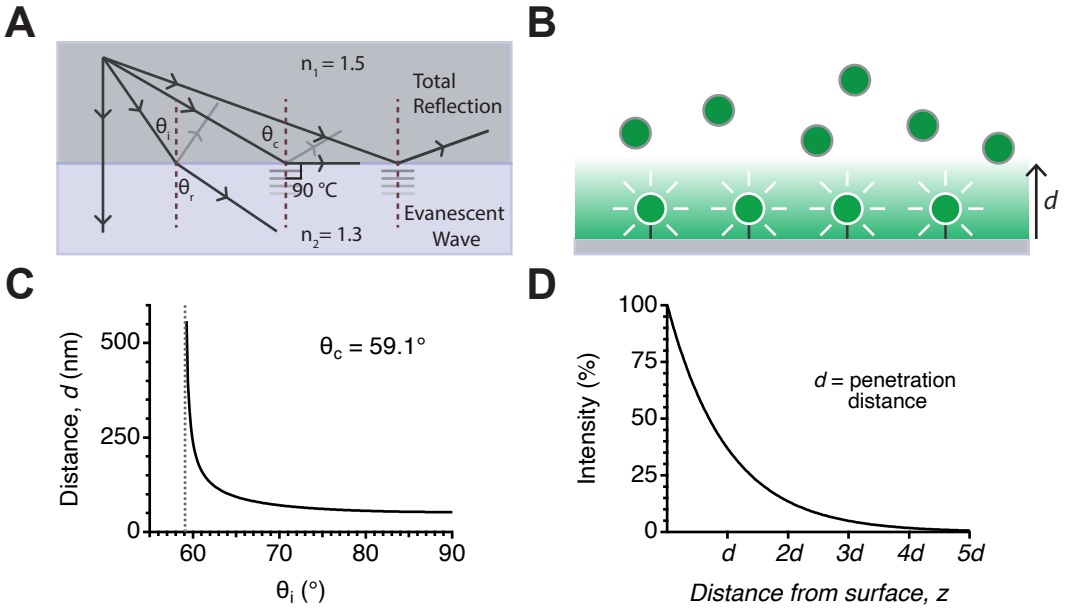


Figure 2.1: Principles of TIRFM. (**A**) Refraction of light passing from a medium of high refractive index, n_1 to a medium low refractive index, n_2, illustrating the angle of incidence, θ_i, and angle of refraction, θ_r. If the angle of incidence, θ_i, is greater than the critical angle, θ_c, total reflection occurs, generating a evanescent wave which penetrates a short distance into the second medium. (**B**) TIR enables selective excitation of fluorophores close to the surface, reducing the background signal coming from freely diffusing fluorophores. (**C**) The penetration distance of the evanescent field, *d*, is at its maximum at the critical angle, θ_c, and decreases rapidly at higher angles (See Equation 2.3). $\lambda_0 = 532$ nm, $n_1 = 1.553$, $n - 2 = 1.333$. (**D**) Evanescent field intensity as a function of distance from the boundary surface, *z*, where *d* is penetration distance.

TIRFM is based around the principle of generating an evanescent field at the interface between the slide and the sample (Axelrod et al., 1983). When light hits a boundary between two substances with different refractive indices, the light is refracted where the angle of refraction, θ_r, is dependant on the angle of incidence, θ_i, and the refractive index of the two media, n_1 and n_2 (Figure 2.1A). This is described by Snell's law with the following relationship:

$$\mathrm{n}_1 \times \theta_\mathrm{i} = \mathrm{n}_2 \times \theta_\mathrm{r} \tag{2.1}$$

As the angle of refraction, θ_r, reaches 90° the light is no longer refracted into the second medium. By setting θ_r to 90° the critical angle of incidence, θ_c can be calculated:

$$\theta_{\mathrm{c}} = \sin^{-1}\left(\frac{\mathrm{n}_2}{\mathrm{n}_1}\right) \tag{2.2}$$

At $\theta_i > \theta_c$ total internal reflection occurs, whereby the incident light is entirely reflected by the boundary (Figure 2.1A) (Axelrod et al., 1983). However, this also generates an electromagnetic field which penetrates a short distance into the second medium, the evanescent field. The distance of penetration of the evanescent field, d, (Figure 2.1B), is dependant on the two refractive indexes, the angle of incidence, and also the wavelength of incident light, λ_0:

$$\mathrm{d} = \frac{\lambda_0}{4\pi}(\mathrm{n}_{\mathrm{i}}^2 \times \sin^2\theta_{\mathrm{i}} - \mathrm{n}_{\mathrm{r}}^2)^{-\frac{1}{2}} \tag{2.3}$$

For example, for a wavelength of 532 nm with refractive indicies of $n_1 = 1.553$ for a typical quartz slide and $n_2 = 1.333$ for water, the penetration distance varies from 560 - 55 nm as the angle of incidence increases beyond the critical angle (Figure 2.1C) (Axelrod et al., 1983).

For an evanescent field with a defined penetration distance, d, the intensity, I, decreases rapidly as the distance from the boundary, z increases (Figure 2.1D) with the relationship:

$$\mathrm{I} = \mathrm{I}_0 \times \mathrm{e}^{-\left(\frac{\mathrm{z}}{\mathrm{d}}\right)} \tag{2.4}$$

Where I_0 is the initial intensity.

By tuning the angle of incidence and light intensity, fluorophores close to the surface are selectively excited without exciting freely diffusing fluorophores, dramatically reducing the background signal and allowing imaging of single molecules in biological samples (Figure 2.1B) (Axelrod et al., 1983).

2.2.2 Förster resonance energy transfer

Förster resonance energy transfer, FRET, is a process of energy transfer between two fluorescent molecules. If two fluorophores with partially overlapping emission and excitation spectra are within close proximity of each other, 20 - 100 Å, then excitation of one, the donor, will result in direct energy transfer to the other, the acceptor (Figure 2.2). Importantly, this energy transfer does not involve radiative emission from the donor fluorophore, and can be highly efficient. The donor is excited to a higher energy state, from which it would normally decay and release photons at its characteristic emission wavelength. However, due to FRET, some of this energy is transferred, exciting the acceptor which then decays to release photos at the acceptor emission wavelength. The efficiency of energy transfer, E_t, is highly dependant on the distance between the two fluorophores as described by the Förster equation:

$$\mathrm{E_{FRET}} = \frac{\mathrm{R_0^6}}{\mathrm{R^6 \ + \ R_0^6}} \tag{2.5}$$

Where R is the distance between the fluorophores and R_0 is the distance at which 50% energy transfer takes place. R_0 is dependant on the intrinsic properties of a donor and acceptor pair but also on their relative orientation and can be calulated from the following equation:

$$\mathrm{R_0 \ = \ 9.725 \times 10^3 (Q_D \times J(\lambda) \times n^{-4} \times \kappa^2)^{\frac{1}{6}}} \tag{2.6}$$

Where Q_D is the quantum yield of the donor alone, $J(\lambda)$ is the overlap between the emission and excitation spectra of the donor and acceptor (Figure 2.7A), n^{-4} is the refractive index of the media and κ^2 is a measure of the relative orientation of the fluorophores. Although measurement of κ^2 is challenging (Norman et al., 2000), assuming random fluorophore orientation, it can be approximated to $\kappa^2 = \frac{2}{3}$. For example, for the commonly used donor acceptor pair of Cy3 and Cy5 in water, and assuming random fluorophore orientation, $R_0 = 58$ Å. Given this R_0 value, for Cy3 and Cy5 the FRET efficiency goes from 1 to 0 in the distance range of $10 - 100$ Å (Figure 2.2B), making FRET a powerful approach to measure biological molecular distances. By directly measuring the donor intensity, I_D, and acceptor intensities, I_A, the FRET efficiency,

E_t, can be calculated using:

$$\mathrm{E_F RET} = \frac{\mathrm{I_A}}{\mathrm{I_D} + \mathrm{I_A}} \tag{2.7}$$

For a known R_O the measured E_t can be put back into Equation 2.5 to calculate the distance between the two fluorophores. However often absolute distances are not required and typically the E_t is used directly.

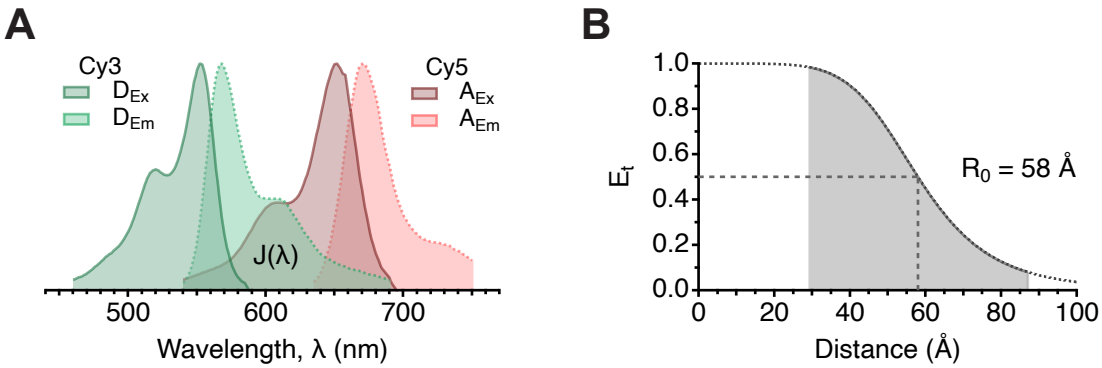


Figure 2.2: Fluorescent characteristics of Cy3 and Cy5 FRET pair. (**A**) Cy3, donor (D), and Cy5, acceptor (A), excitation ($_{Ex}$) and emission ($_{Em}$) spectra. Overlap between D_{Em} and A_{Ex} gives J(λ). (**B**) Calculated FRET efficiency (E_t) for Cy3 Cy5 donor and acceptor fluorophore pair separated by a distance of 0 to 100 Å based on an R_0 of 58 Å. Shaded region demonstrates usable distance range of distinguishable FRET change, between $0.5\mathrm{R}_0$ and $1.5\mathrm{R}_0$.

FRET measurements can be performed in bulk to give an average FRET value for a population of molecules, however the real power of this method was realised when it was demonstrated that it was possible to measure the state of molecules individually (Ha et al., 1996). Single-molecule FRET, smFRET, can be performed using a confocal set-up and dilute sample, individual molecules can then be detected diffusing through the small confocal volume and the FRET values calculated to identify the distribution of conformations within the population. This approach gives a snapshot of each molecule, and is useful for exploring the potential conformations and average occupancy of each state. However, it does not give information as to transitions between states. Alternatively, to follow the conformational changes of individual molecules over time, a surface based assay can be used where molecules are immobilised on a glass slide. Within a single field of view, hundreds of molecules can be imaged and changes in FRET measured over time. This allows direct measurement of the transition rates and observation of multiple steps allowing reconstitution of complex multistep conformational changes. In order to reach

the high signal-to-noise ratio required to image individual molecules total-internal reflection microscopy, TIRFM (See Section 2.2.1), is typically used to only excite molecules very close to the surface; this reduces the background fluorescence from freely diffusing molecules.

2.3 Optical tweezers

The basis of optical trapping is the ability to use highly focused light to trap small particles. This was first demonstrated by Arthur Ashkin (A. Ashkin, 1970; Ashkin et al., 1986) (See Section 2.3.1), and would later lead to two Nobel prizes: the 1997 Nobel Prize in Physics awarded to Steven Chu, Claude Cohen-Tannoudji and William D. Phillips for the "development of methods to cool and trap atoms with laser light", and the 2018 Nobel Prize in awarded to Arthur Ashkin for "the optical tweezers and their application to biological systems". He later demonstrated that this approach could be used to study biological processes *in vivo*, measuring the forces of mitochondrial transport inside an intact cell (A. Ashkin et al., 1990). Subsequent work utilised the high sensitivity of optical trapping to measure forces exerted by single motor proteins, such as kinesin (Steven M. Block et al., 1990) and myosin (Finer et al., 1994) offering unique insight into their mechanisms of force generation. At the same time, the biophysical properties of DNA were being studied, first by magnetic tweezers (S. Smith et al., 1992; C. Bustamante et al., 1994) and later using optical tweezers (S. B. Smith et al., 1996; M. D. Wang et al., 1997), (See Section 2.3.2).

Force spectroscopy experiments using optical tweezers have continued to provide breakthroughs in our understanding of the mechanisms of motor proteins and properties of DNA but have also allowed measurements of protein folding energy landscapes (Liu et al., 2017; Carlos Bustamante et al., 2020); the activity of DNA and RNA polymerases (Herbert et al., 2008; Hodges et al., 2009; Naufer et al., 2017); DNA nucleosome wrapping (Bennink et al., 2001); properties and mechanosensing of the intracellular cytoskeleton, cell membranes and the extracellular matrix (Hu et al., 2017); and ligand-receptor binding (Weisel, 2003) - to highlight only a few. Furthermore, *in vivo* application have now been demonstrated in multicellular model organisms such as zebrafish, where measurements of blood flow and stimulation of hearing organs have

been performed with optical tweezers (Harlepp et al., 2017; I. A. Favre-Bulle et al., 2019; I. Favre-Bulle et al., 2020).

Some of the earliest DNA stretching experiments also combined the use of fluorescence imaging, allowing the visualisation of individual DNA molecules stained with ethidium bromide (S. Smith et al., 1992). Wide field fluorescence microscopy has been utilised to image DNA-protein interactions such as DNA unwinding by RecBCD (Bianco et al., 2001), RecA and Rad51 filament formation (Galletto et al., 2006; Mameren et al., 2006) and diffusion of restriction enzymes (Biebricher et al., 2009). However, due to the high background signal from wide-field illumination, imaging of single-molecules is challenging (Heller, Tjalle P. Hoekstra, et al., 2014b). The use of TIRF microscopy in combination with optical tweezers has be used to reduce the background fluorescence and enable observation of single molecules of RNA polymerase interacting with DNA in real-time (Harada et al., 1999). However, optical trapping and TIRF are not naturally compatible due to the requirement to be close to the slide surface for imaging and to hold the trapped beads off the surface for stable trapping (Heller, Tjalle P. Hoekstra, et al., 2014b). More recently the use of confocal microscopy has enabled imaging of single-molecule DNA-protein interactions in real-time, even in the presence of a relatively high concentration of labelled complex (20 - 100 nM) (Comstock et al., 2011; Heller, Sitters, et al., 2013; Heller, Tjalle P. Hoekstra, et al., 2014b).

2.3.1 Optical trapping

Photons carry momentum which can be transferred to particles and used to stably trap them close to the focus of a highly focused laser (A. Ashkin, 1970; Ashkin et al., 1986). This force is generally considered in two components, the scattering force and the gradient force (Neuman and Steven M Block, 2004). The scattering force comes from the photons hitting the particle and being scattered, effectively pushing the particle in the same direction as the laser propagation (Figure 2.3A). The change in momentum of the photons results in an opposing change in momentum of the particle by Newton's third law, as the laser is highly directional this resolves to a net force in the direction of the beam. The gradient force results from photons passing

through the particle. If the particle has a higher refractive index than the surrounding media the photons are refracted; this change in momentum exerts a force on the particle in the opposite direction (Figure 2.3A). At the laser focus there is a very high intensity gradient, as such a particle slightly shifted from the beam centre will experience a net force back towards the centre of the beam, trapping it (Figure 2.3B). For stable trapping, the the gradient force must be greater than the scattering to prevent the scattering force pushing the particle out of the focus. This is achieved using very high numerical aperture objectives to focus the beam. Despite this, the scattering force means the particle position is still slightly displaced from the focus along the beam path (Figure 2.3A). While spherical beads are typically used, non-spherical objects can also be trapped. However, accurate measurement of displacement and forces is more challenging with asymmetric particles.

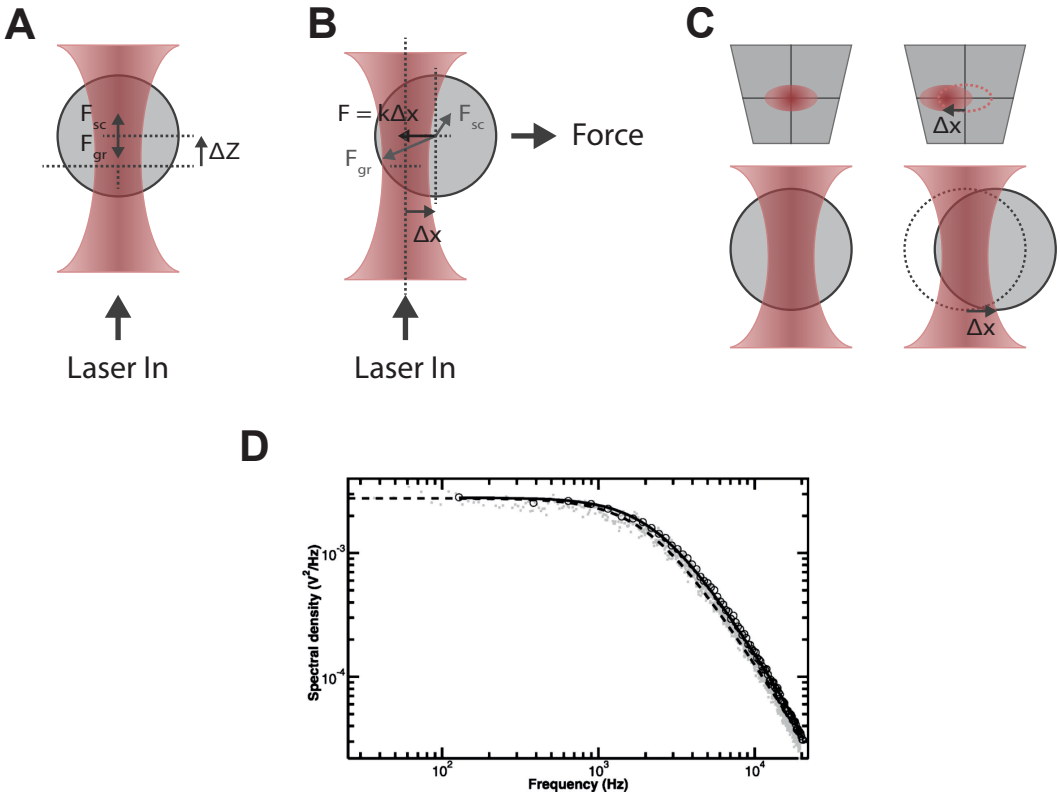


Figure 2.3: Principles of optical trapping. (**A**) Scattering and gradient forces (F_sc and F_gr) acting on a particle in centre of an optical trap without any applied external force. (Heller, Tjalle P Hoekstra, et al., 2014a) (**B**) Scattering and gradient forces (F_sc and F_gr) acting on a particle deflected from the centre of an optical trap by an any applied external force. (**C**) Quadrant photodiode, QPD, displacement detection. Displacement of the particle from the centre of the trapping laser causes a deflection of the laser path which is detected on a QPD (**D**) Calculating the corner frequency, f_0, for force calibration (Neuman and Steven M Block, 2004).

Once trapped, the force, F, required to displace the bead by a specific distance, x, follows Hookes law, with the trap acting as a Hookean spring:

$$\mathrm{F} = -\alpha \times \mathrm{x} \tag{2.8}$$

Where α is the trap stiffness.

The particle displacement is typically measured by detecting the deflection of the trapping laser using a quadrant photodiode, QPD (Figure 2.3C). This allows high frequency measurement of the position of the particle in 2-dimensions relative to the beam centre and is used to calibrate the trap stiffness and to then calculate forces on the particle.

The trap stiffness depends on the size and refractive index of the particle, the refractive index of the surrounding medium, and the wavelength, intensity and focus of the laser. It is typically determined by calibration based on measuring the Brownian motion of a particle of known size in the trap when no external forces are being applied (Neuman and Steven M Block, 2004). This is based on the principle that the higher the stiffness of the trap the more the high frequency movements will dominate. A power spectrum is measured, representing the Brownain motion at different frequencies, and fit to determine the rolloff or corner frequency (Figure 2.3D). The corner frequency, f_0 is then used to calculate the trap stiffness, α, using the formula:

$$\alpha = 12\pi^2 \times \mathrm{f_0} \times \eta \times \mathrm{r} \tag{2.9}$$

Where η is the viscosity of the medium and r is the radius of the bead. Typical values for a 4.5 μm and 2 μm bead are given in Table 2.1. Once the trap stiffness has been calibrated forces applied to the bead can be accurately measured with high temporal resolution.

Bead Diameter (μm)	f_0 (Hz)	Trap Stiffness (pN nm^{-1})
4.42	720	0.19

Table 2.1: Typical values of corner frequency, f_0, and trap stiffness for a 4.42 μm diameter bead, at 25 °C based on a viscosity, η, of 1 mPa.s

2.3.2 Biophysical properties of DNA

Extensive work has also been done on probing the properties of DNA and DNA:protein interactions, and this will be the focus here. DNA molecules can be held between a surface, or fixed point, and a trapped bead or between two individually trapped beads. By moving the bead the DNA can be stretched, and the force on the DNA and corresponding extension can be measured. In the first instance this was used to explore the properties of DNA molecules alone (S. Smith et al., 1992; Perkins, D. Smith, et al., 1994; C. Bustamante et al., 1994; Perkins, Quake, et al., 1994; Marko and Siggia, 1995; M. D. Wang et al., 1997; S. B. Smith et al., 1996).

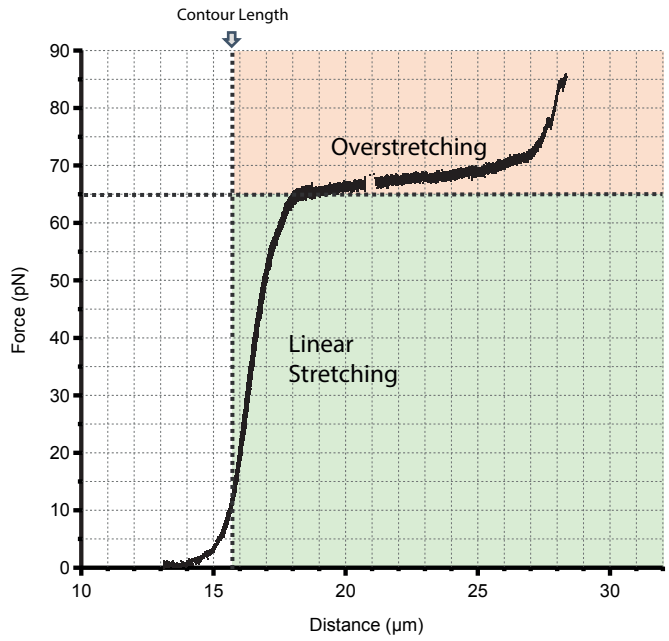


Figure 2.4: DNA force-distance plot. Force-distance plot obtained from stretching a molecule of λ-DNA from 0 - 85 pN. Once the distance reaches the contour length of the DNA the molecule begins to stretch linearly with distance, linear stretching. At high forces (65 pN) the DNA begins to melt, overstretching.

Stretching a single DNA molecule between two beads reveals a characteristic force extension curve (Figure 2.4). When the distance between the beads is significantly lower than the contour length, L_c, of the DNA the molecule is randomly disordered. As the distance approaches the contour length, the DNA is straightened out, with energy required to overcome this loss of entropy (Figure 2.4). As the distance exceeds the contour length, the molecule enters an enthapic phase where the DNA molecule is being stretched, with small changes in distance resulting in a large increase in force (Figure 2.4). Finally at very high forces it enters the overstretching

regime, where the base pairing between the two DNA strands start to breakdown and the DNA melts, transitioning from double to single stranded DNA (Figure 2.4). At low force, < 5 pN, the behaviour of the DNA can be described using the Worm-like chain model, WLC. This models DNA as a polymer with relatively high stiffness over short distances but which is intrinsically flexible over large distances. The local stiffness is defined by the persistence length, L_p, and describes the distance scale over which is molecule remains roughly straight. Within this low force regime the length of the DNA, L_{DNA}, as a function of force, F, can be fit to the basic WLC formula to obtain values of the L_c and L_p:

$$\mathrm{L_{DNA}} = \mathrm{L_c}\left(1 - \frac{1}{2}\sqrt{\frac{\mathrm{k_B T}}{\mathrm{F} \times \mathrm{L_p}}}\right) \tag{2.10}$$

Where k_B is the Boltzman constant, 4.11 pN.nM, and T is the temperature in kelvin.

However this does not take into account the stretching of the DNA molecule, and fails to fully describe the force extension curve, especially above forces of 5 pN. The extension due to DNA stretching at a given force is defined by the stretching modulus, S, of the DNA:

$$\Delta\mathrm{L} = \mathrm{L_c} \times \frac{\mathrm{F}}{\mathrm{S}} \tag{2.11}$$

Incorporating this into the WLC formula gives the extensible WLC, eWLC, which fits well to whole force extension curve below the overstretching transition (Odijk, 1995; M. D. Wang et al., 1997):

$$\mathrm{L_{DNA}} = \mathrm{L_c}\left(1 - \frac{1}{2}\sqrt{\frac{\mathrm{k_B T}}{\mathrm{F} \times \mathrm{L_p}}} + \frac{\mathrm{F}}{\mathrm{S}}\right) \tag{2.12}$$

This has been further extended to better describe the region close to the overstretching regime, >35 pN, by taking into account twisting of the DNA helix (Gross et al., 2011), however this is not required for the analysis of this work.

Beyond the overstretching regime, the DNA is fully melted and becomes single-stranded. Single-stranded DNA, ssDNA, is far more flexible than dsDNA and as such is modelled as a freely jointed chain, FJC, whereby each neighbouring base is able to more freely relative to each other.

At low forces, this model is not sufficient as it does not consider potential base pairing within the relaxed molecule. However, at forces >15 pN the FJC, with the addition of parameters to account for backbone stretching, can be used (M. D. Wang et al., 1997):

$$\mathrm{L_{DNA}} = \mathrm{L_c}\left(\coth\left(\frac{2\times \mathrm{F}\times \mathrm{L_p}}{\mathrm{k_B T}}\right) - \left(\frac{\mathrm{k_B T}}{2\times \mathrm{F}\times \mathrm{L_p}}\right)\right)\times\left(1+\frac{\mathrm{F}}{\mathrm{S}}\right) \tag{2.13}$$

Where L_c is the contour length, L_p is the persistance length, and S is the stretch modulus. Typical parameters for dsDNA and ssDNA in 150 mM NaCl at pH 7 are shown in Table 2.2.

	L_c (nm bp^{-1})	L_p (nm)	S (pN)	
dsDNA	0.34	50	1400	(M. D. Wang et al., 1997)
ssDNA	0.42	10	800	(M. D. Wang et al., 1997)

Table 2.2: Parameters of dsDNA and ssDNA

2.4 Single-molecule techniques to study CRISPR/Cas9 off-target activity

As previously discussed, there are many open questions surrounding CRISPR/Cas9 off-target activity, specifically the detail of on- versus off-target site discrimination (See Section 1.3). However, despite the potentially detrimental results of off-target activity, off-target events are comparatively rare and interactions of Cas9 at off-target sites may be transient. As such traditional bulk experiments, and even the best next-generation sequencing approaches, are limited to detection of only the most prevalent off-target sites and may fail to detect low frequency sites and short lived off-target interactions. The use of single-molecule technique enables such events to be detected. The use of smFRET-TIRF (See Section 2.2) assays allows even transient interactions with potential target sites to be observed, and can be used to probe intermediate steps in Cas9 binding and cleavage to dissect mechanistic detail of off-target discrimination. In addition, the use of optical tweezers with confocal microscopy (See Section 2.3) enables the study of target search of Cas9 across kilobase length DNA sequences and detection of transient off-target binding interactions. Optical tweezers also enable the mechanical pertubation of the

DNA structure and investigation of the effect of this on Cas9 activity. In the work presented in this thesis, I use single-molecule techniques in conjunction with bulk assays and novel next-generation sequencing approaches (See Chapter 3, Methods) to further our understanding of the factors that effect Cas9 off-target activity, with the goal that such understanding can lead to improved specificity.

References

Alemán, Elvin A., Rajan Lamichhane, and David Rueda (Dec. 2008). "Exploring RNA folding one molecule at a time". *Current Opinion in Chemical Biology* 12.6, pp. 647–654.

Ashkin, A et al. (May 1986). "Observation of a single-beam gradient force optical trap for dielectric particles". *Optics Letters* 11.5, p. 288.

Ashkin, A. (Jan. 1970). "Acceleration and Trapping of Particles by Radiation Pressure". *Physical Review Letters* 24.4, pp. 156–159.

Ashkin, A. et al. (Nov. 1990). "Force generation of organelle transport measured in vivo by an infrared laser trap". *Nature* 348.6299, pp. 346–348.

Axelrod, Daniel, Nancy L. Thompson, and Thomas P. Burghardt (Jan. 1983). "Total internal reflection fluorescent microscopy". *Journal of Microscopy* 129.1, pp. 19–28.

Bennink, Martin L. et al. (July 2001). "Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers." *Nature structural biology* 8.7, pp. 606–10.

Bianco, Piero R. et al. (Jan. 2001). "Processive translocation and DNA unwinding by individual RecBCD enzyme molecules". *Nature* 409.6818, pp. 374–378.

Biebricher, Andreas et al. (Apr. 2009). "Tracking of Single Quantum Dot Labeled EcoRV Sliding along DNA Manipulated by Double Optical Tweezers". *Biophysical Journal* 96.8, pp. L50–L52.

Blanchard, Scott C. et al. (Aug. 2004). "tRNA dynamics on the ribosome during translation". *Proceedings of the National Academy of Sciences* 101.35, pp. 12893–12898.

Block, Steven M., Lawrence S. B. Goldstein, and Bruce J. Schnapp (Nov. 1990). "Bead movement by single kinesin molecules studied with optical tweezers". *Nature* 348.6299, pp. 348–352.

Bokinsky, Gregory et al. (Aug. 2003). "Single-molecule transition-state analysis of RNA folding". *Proceedings of the National Academy of Sciences* 100.16, pp. 9302–9307.

Bustamante, C. et al. (Sept. 1994). "Entropic elasticity of lambda-phage DNA". *Science* 265.5178, pp. 1599–1600.

Bustamante, Carlos et al. (June 2020). "Single-Molecule Studies of Protein Folding with Optical Tweezers". *Annual Review of Biochemistry* 89.1, pp. 443–470.

Chen, Janice S. et al. (Oct. 2017). "Enhanced proofreading governs CRISPR–Cas9 targeting accuracy". *Nature* 550.7676, pp. 407–410.

Christian, Thomas D., Louis J. Romano, and David Rueda (Dec. 2009). "Single-molecule measurements of synthesis by DNA polymerase with base-pair resolution". *Proceedings of the National Academy of Sciences* 106.50, pp. 21109–21114.

Comstock, Matthew J., Taekjip Ha, and Yann R. Chemla (Apr. 2011). "Ultrahigh-resolution optical trap with single-fluorophore sensitivity". *Nature Methods* 8.4, pp. 335–340.

Dagdas, Yavuz S et al. (Aug. 2017). "A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9". *Science Advances* 3.8, eaao0027.

Favre-Bulle, Itia et al. (June 2020). "Sound generation in zebrafish with Bio-Opto-Acoustics (BOA)". *bioRxiv*, p. 2020.06.09.143362.

Favre-Bulle, Itia A. et al. (May 2019). "Optical trapping in vivo: theory, practice, and applications". *Nanophotonics* 8.6, pp. 1023–1040.

Finer, Jeffrey T., Robert M. Simmons, and James A. Spudich (Mar. 1994). "Single myosin molecule mechanics: piconewton forces and nanometre steps". *Nature* 368.6467, pp. 113–119.

Funatsu, Takashi et al. (1995). "Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution". *Nature* 374.6522, pp. 555–559.

Galletto, Roberto et al. (Oct. 2006). "Direct observation of individual RecA filaments assembling on single DNA molecules". *Nature* 443.7113, pp. 875–878.

Gross, Peter et al. (Sept. 2011). "Quantifying how DNA stretches, melts and changes twist under tension". *Nature Physics* 7.9, pp. 731–736.

Ha, T et al. (1996). "Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor". *Proceedings of the National Academy of Sciences of the United States of America* 93.13, pp. 6264–6268.

Ha, Taekjip, Alice Y Ting, et al. (Feb. 1999). "Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism". *Proceedings of the National Academy of Sciences* 96.3, pp. 893–898.

Ha, Taekjip, Xiaowei Zhuang, et al. (Aug. 1999). "Ligand-induced conformational changes observed in single RNA molecules". *Proceedings of the National Academy of Sciences* 96.16, pp. 9077–9082.

Harada, Yoshie et al. (Feb. 1999). "Single-Molecule Imaging of RNA Polymerase-DNA Interactions in Real Time". *Biophysical Journal* 76.2, pp. 709–715.

Harlepp, Sébastien et al. (Nov. 2017). "Hemodynamic forces can be accurately measured in vivo with optical tweezers". *Molecular Biology of the Cell* 28.23. Ed. by Manuel Théry, pp. 3252–3260.

Heller, Iddo, Tjalle P Hoekstra, et al. (Mar. 2014a). "Optical Tweezers Analysis of DNA–Protein Complexes". *Chemical Reviews* 114.6, pp. 3087–3119.

— (Mar. 2014b). "Optical Tweezers Analysis of DNA–Protein Complexes". *Chemical Reviews* 114.6, pp. 3087–3119.

Heller, Iddo, Gerrit Sitters, et al. (Sept. 2013). "STED nanoscopy combined with optical tweezers reveals protein dynamics on densely covered DNA". *Nature Methods* 10.9, pp. 910–916.

Herbert, Kristina M., William J. Greenleaf, and Steven M. Block (June 2008). "Single-Molecule Studies of RNA Polymerase: Motoring Along". *Annual Review of Biochemistry* 77.1, pp. 149–176.

Hirschfeld, T (Dec. 1976). "Optical microscopic observation of single small molecules". *Applied Optics* 15.12, p. 2965.

Hodges, Courtney et al. (July 2009). "Nucleosomal Fluctuations Govern the Transcription Dynamics of RNA Polymerase II". *Science* 325.5940, pp. 626–628.

Hu, Xian et al. (July 2017). "Molecular stretching modulates mechanosensing pathways". *Protein Science* 26.7, pp. 1337–1351.

Kapanidis, Achillefs N. et al. (Nov. 2006). "Initial Transcription by RNA Polymerase Proceeds Through a DNA-Scrunching Mechanism". *Science* 314.5802, pp. 1144–1147.

Karunatilaka, Krishanthi S and David Rueda (July 2009). "Single-molecule fluorescence studies of RNA: A decade's progress". *Chemical Physics Letters* 476.1-3, pp. 1–10.

Lim, Youngbin et al. (Nov. 2016). "Structural roles of guide RNAs in the nuclease activity of Cas9 endonuclease". *Nature Communications* 7, p. 13350.

Liu, Kaixian et al. (July 2017). "The ribosome destabilizes native and non-native structures in a nascent multidomain protein". *Protein Science* 26.7, pp. 1439–1451.

Mameren, Joost van et al. (Oct. 2006). "Dissecting Elastic Heterogeneity along DNA Molecules Coated Partly with Rad51 Using Concurrent Fluorescence Microscopy and Optical Tweezers". *Biophysical Journal* 91.8, pp. L78–L80.

Marko, John F and Eric D Siggia (1995). "Stretching DNA". *Macromolecules* 28.26, pp. 8759–8770.

Munro, James B. et al. (Feb. 2007). "Identification of Two Distinct Hybrid State Intermediates on the Ribosome". *Molecular Cell* 25.4, pp. 505–517.

Naufer, M. Nabuan et al. (July 2017). "Single-molecule mechanochemical characterization of E. coli pol III core catalytic activity". *Protein Science* 26.7, pp. 1413–1426.

Neuman, Keir C and Steven M Block (Sept. 2004). "Optical trapping". *Review of Scientific Instruments* 75.9, pp. 2787–2809.

Norman, David G et al. (May 2000). "Location of Cyanine-3 on Double-Stranded DNA: Importance for Fluorescence Resonance Energy Transfer Studies †". *Biochemistry* 39.21, pp. 6317–6324.

Odijk, Theo (Sept. 1995). "Stiff Chains and Filaments under Tension". *Macromolecules* 28.20, pp. 7016–7018.

Perkins, T., Quake, et al. (May 1994). "Relaxation of a single DNA molecule observed by optical microscopy". *Science* 264.5160, pp. 822–826.

Perkins, T., D. Smith, and Steven Chu (May 1994). "Direct observation of tube-like motion of a single polymer chain". *Science* 264.5160, pp. 819–822.

Rueda, David and NilsG. Walter (Dec. 2005). "Single Molecule Fluorescence Control for Nanotechnology". *Journal of Nanoscience and Nanotechnology* 5.12, pp. 1990–2000.

Singh, Digvijay, Samuel H. Sternberg, et al. (Sept. 2016). "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Nature Communications* 7, p. 12778.

Singh, Digvijay, Yanbo Wang, et al. (Apr. 2018). "Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis". *Nature Structural & Molecular Biology* 25.4, pp. 347–354.

Smith, S, Laura Finzi, and Carlos Bustamante (Nov. 1992). "Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads". *Science* 258.5085, pp. 1122–1126.

Smith, Steven B., Yujia Cui, and Carlos Bustamante (Feb. 1996). "Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules". *Science* 271.5250, pp. 795–799.

Steiner, Miriam et al. (Sept. 2008). "Single-molecule studies of group II intron ribozymes". *Proceedings of the National Academy of Sciences* 105.37, pp. 13853–13858.

Wang, Michelle D et al. (1997). "Stretching DNA with optical tweezers". *Biophysical Journal* 72.3, pp. 1335–1346.

Warnasooriya, Chandani and David Rueda (Aug. 2014). "Single-molecule fluorescence-based studies on the dynamics, assembly and catalytic mechanism of the spliceosome". *Biochemical Society Transactions* 42.4, pp. 1211–1218.

Weisel, J (Apr. 2003). "Protein–protein unbinding induced by force: single-molecule studies". *Current Opinion in Structural Biology* 13.2, pp. 227–235.

Willhoft, Oliver et al. (Oct. 2018). "Structure and dynamics of the yeast SWR1-nucleosome complex". *Science* 362.6411, eaat7716.

Zhao, Rui and David Rueda (Oct. 2009). "RNA folding dynamics by single-molecule fluorescence resonance energy transfer". *Methods* 49.2, pp. 112–117.

Zhuang, Xiaowei (June 2000). "A Single-Molecule Study of RNA Catalysis and Folding". *Science* 288.5473, pp. 2048–2051.

Chapter 3

Materials and Methods

3.1 DNA and RNA synthesis, purification and labelling

3.1.1 Purification by denaturing polyacrylamide gel electrophoresis, PAGE

The integrity of all DNA and RNA for single molecule experiments were first checked by denaturing polyacrylamide gel electrophoresis (8 M urea, 18 % polyacrylamide:bis-acrylamide in 1 x TBE) with gels run at 20 W for 4 hours. Where additional bands below the full length product were observed, samples were purified by denaturing PAGE with extraction by crush and soak in elution buffer (40 mM Tris-HCl pH 8, 480 mM NaOAc pH 5, 0.5 mM EDTA) followed by ethanol precipitation.

3.1.2 In vitro transcription of tracrRNA

The tracrRNA was produced by *in vitro* transcription. DNA containing the tracrRNA sequence and T7 trascription start site (Table 3.1) was amplified by PCR with Q5 High-Fidelity DNA Polymerase (NEB, M0491S). PCR reaction product, 40 μl, was added to *in vitro* transcription reaction mix (110 μl; 1x T7 transcription buffer, 36 mM $MgCl_2$, 1 mM DTT, 500U T7 Polymerase

(NEB M0251S), 0.15U *E. coli* IPP (NEB, M0361S), 5 mM NTPs) and incubated overnight at 37 °C. Following *in vitro* transcription the RNA was purified by denaturing PAGE (See Section 3.1.1)

Table 3.1: DNA sequences for tracrRNA In vitro transcription

ID	Name	Sequence 5'-3'
MDN0001	T7 trRNA T1	ATGTA ATACG ACTCA CTATA GGACA GCATA GCAAG TTAAA ATAAG GCTAG TCCGT TATCA ACTTG AAAAA GTGGC ACCGA GTCGG TGCTT TTT
MDN0002	T7 trRNA T2	AAAAA GCACC GACTC GGTGC CACTT TTTCA AGTTG ATAAC GGACT AGCCT TATTT TAACT TGCTA TGCTG TCCTA TAGTG AGTCG TATTA CAT
MDN0003a	T7 F Primer	ATGTA ATACG ACTCA CTATA GG
MDN003b	trRNA Rev primer	AAAAA GCACC GACTC GGTG

3.1.3 Labelling of DNA and RNA

For in-house labelling of DNA and RNA, oligos were ordered from IDT with modified amino-linker bases (iAmMC6T, IDT) at the required sites for labelling and resuspended in ddH_2O. Mono-reactive Cy3 or Cy5 dye (GE Healthcare, PA23001) was resuspended in DMSO (14 μl, 99%, SIGMA). Resuspended oligos (50 μl, 100 μM) were incubated in $NaHCO_3$ buffer (43 μl, 0.5 M, pH 9) with resuspended dye (7 μl) for at least two hours at 4°C. Unreacted dye was removed by ethanol precipitation, resuspending in ddH_2O (50 μl).

3.1.4 HPLC Purification

After labelling, labelled DNA and RNA was purified by HPLC (Discovery BIO Wide Pore C8 5μm column, 568323-U) in TEAA buffer (50 mM acetonitrile, 2.8 mM triethylamine, adjusted to pH 7 with glacial acetic acid) to remove the unlabelled population. Collected fractions containing the labelled population were vacuum dried and finally purified by ethanol precipitation. Where required, labelled oligos were further purified by denaturing PAGE (See 3.1.1).

3.2 TIRF smFRET

3.2.1 TIRF Microscope

All smFRET experiments were performed on a home built prism-based total internal reflection microscope, TIRM, built around and Olympus IX71 microscope with 532 nm excitation (Vortran Stradus, US) and 542 nm long pass emission filter (Chroma 542LP ET). The acceptor and donor emission is split with a beam splitter containing a 635 nm long pass dichroic mirror (Chroma T635LPXR) and detected on an Andor iXon Ultra EMCCD camera.

3.2.2 Slide preparation

Holes were drilled in ultra-low fluorescence quartz slides with diamond tipped drill bits, in either a two channel or four channel configuration (Figure 3.1A). Slides were cleaned with a paste of Alconox powder detergent by vigorous scrubbing followed by rinsing with ddH_2O, and two rounds of scrubbing with first ethanol and then ddH_2O for 1 minute each. Slides were then boiled in batches of eight in basic piranha solution (ddH_2O, 150 ml; H_2O_2, 30 ml; NH_4OH, 30 ml) for 40 minutes. The slides were removed and rinsed with ddH_2O and flamed briefly with Bunsen burner to pyrolyse any remaining organic residues. The cleaned slides and new coverslips were placed in coplin jars and sonicated in KOH (1 M) for 30 minutes, rinsed with ddH_2O and methanol and then sonicated in methanol for 30 minutes.

For aminopropylsilation of the surfaces, slides and coverslips were placed in coplin jars and incubated in aminopropylsilation solution (aminopropylsilane (Vectabond), 1 ml; glacial acetic acid, 5 ml; methanol, 100 ml, 100 %) for 10 minutes followed by a 2 minute sonication and a further 10 minute incubation. Slides and coverslips were rinsed with methanol, ddH_2O, then again with methanol and dried with N_2. For PEGylation slides were placed in PEGylation chamber above a reservoir of water to maintain humidity. The PEGylation reaction solution was prepared by adding biotinylated PEG (7 mg, SVA, MW 3400) and unbiotinylated mPEG (80 mg, MW 5000) to bicarbonate buffer (380 μl, 84 mg $NaHCO_3$ in 10 ml ddH_2O). For each

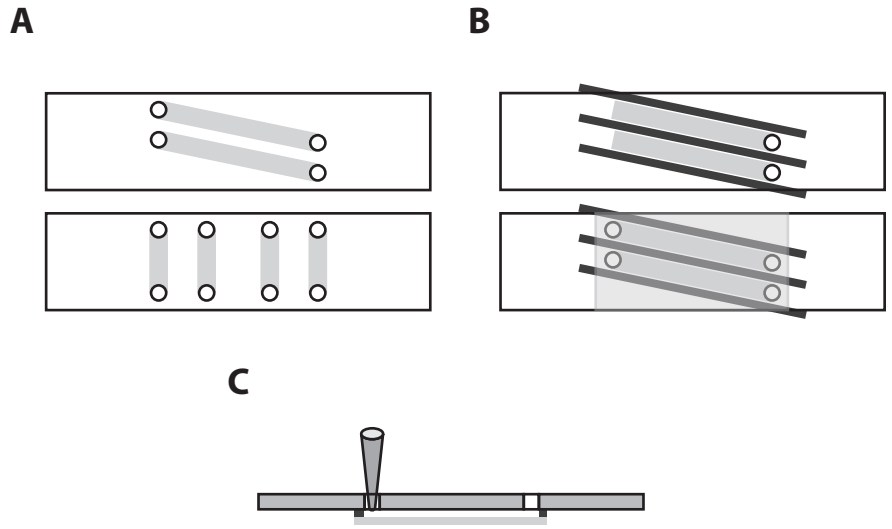


Figure 3.1: smFRET-TIRFM Slide Construction. (**A**) Hole layouts for 2 and 4 channel configuration (**B**) Slide construction (**C**) Slide injection

slide 80 μl PEGylation reaction solution was added in the centre, the coverslip placed over the top and slide incubated overnight.

Following PEGylation slides and coverslips were rinsed thoroughly with ddH_2O, dried with N_2 and stored together in 50 ml falcon tubes with PEGylated surfaces facing outwards. Tubes were filled with N_2, vacuum sealed in polythene bags and stored at -20°C for upto 3 months. Before construction sealed slides stored at -20°C were left to return to room temperature before opening. Two layers of cut doublesided sticky tape was used to create 4 mm wide channels between inlet and outlet holes on the PEGylated slide surface and coverslip stuck down firmly on top, PEGylated side down, with epoxy glue used to seal around the edges.

3.2.3 Calibration of Cy3 and Cy5 channel

For calibration of the mapping of the Cy3 and Cy5 channels for smFRET TIRF experiments, a slide was prepared containing a low density of fluorescent beads with emission in both the Cy3 and Cy5 channels (FluoSpheres Carboxylate-Modified Microspheres, 0.2 μm, red fluorescent (580/605), 2 % w/v). An ultra-low fluorescence quartz slide was cleaned with ddH_2O, methanol, again with ddH_2O and dried with N_2. A single layer of doublesided tape was used to create a 2 cm wide channel, with a clean coverslip attached on top. Diluted beads were incubated in the

channel for 15 minutes before sealing with epoxy glue. For each new day of smFRET TIRF experiments a short video was recorded with > 100 beads in the field of view and used to map the Cy3 and Cy5 signals when extracting smFRET traces. (Zhao and Rueda, 2009)

3.2.4 smFRET experiments

Constructed slide chambers were passivated and functionalised by incubation with 0.2 mg ml^{-1} BSA in T_{50} buffer (50 mM Tris-HCl, pH 7.0, 50 mM NaCl) for 10 minutes, followed by a wash with T_{50} and incubation with Neutravidin (1 mg ml^{-1} in T_{50} for 10 minutes. This was washed out with T_{50} then replaced with imaging buffer (100 mM NaCl, 50 mM Tris-HCl pH 8, 1mM $MgCl^2$, 0.2 mg ml^{-1} BSA, 125 mM PCA, 10 μM PCD and saturating trolox) ready for sample injection. Cy3 labelled DNA for immobilisation was diluted to 12 pM before injection into the slide chamber and incubated for 10 minutes before washing with imaging buffer. DNA was imaged on TIRF microscope to check density before injection of Cas9 complex (200 μl, 10 nM Cy5-crRNA:tracrRNA:Cas9 in imaging buffer) and data collection.

3.2.5 Data acquisition and analysis

Unless otherwise stated, videos were recorded with a 30 ms exposure time and EM gain of 600×. Custom IDL scripts were used for alignment mapping of the Cy3 and Cy5 channels, background subtraction, molecule detection and extraction of single-molecule trajectories. Single-molecule trajectories were analysed using custom MATLAB or R scripts. Donor to acceptor channel leak was estimated to be 20% and subtracted from acceptor intensity before FRET calculation using Equation 2.7.

3.3 Optical tweezer experiments

3.3.1 Biotinylation of Phage-λ DNA

3.3.1.1 Tortionally unconstrained DNA

For the majority of experiments biotinylated λ-phage DNA, λ-DNA, was prepared by filling in the 5' 12-bp overhangs with biotinylated nucleotides (Figure 3.2A) (Gross et al., 2010). Per 50 μl reaction: in a PCR tube, Lambda DNA (21 μl, 500 μg, Life Technologies Limited SD0011) was combined with B-dCTP (10 μl 40 μM Life Technologies Limited 19518018), B-dATP (10 μl, 40 μM, Life Technologies Limited 19524016), dTTP (0.5 μl, 10 mM,), dGTP (0.5 μl, 10 mM,), NEB Buffer 2 (5 μl, 10X, NEB), ddH2O (3 μl), and finally Klenow Pol (0.05 μl, 50U//μl, NEB M0212M). Pipetting of any solutions containing λ-DNA was done slowly and carefully with wide bore pipettes to reduce mechanical nicking of the DNA. This was transfered to a PCR machine and run with the following program: 37°C, 30 minutes; 70°C, 15 minutes; 4°C, ∞. This was then purified using QIAquick PCR Purification Kit (QIAGEN 28104), eluting in 60 μl elution buffer and quantified using a BioDrop μLite+. Final concentration was typically 30 μg/ml. Biotinylated DNA was stored at 4°C for $>$ 1 month or aliquots flash frozen in LN_2 and stored at -20°C, freeze thawing cycles were minimised to prevent nicking.

3.3.1.2 Tortionally constrained DNA

For experiments with tortionally constrained DNA for supercoiling experiments, caps containing multiple biotins modifications were ligated to either end of λ-DNA (Figure 3.2B, Table 3.2) (King et al., 2019). The caps were purchased from IDT with HPLC purification and resuspended in ddH_2O at 100 μM.

First the two caps and λ-DNA were 5' phosphorylated. For the DNA, T4 DNA Ligase Buffer (5.5 μl, 10X,) and polynucleotide kinase (0.5 μl, 10 U/μl) was added to λ-DNA (50 μl, 500 μg, Life Technologies Limited SD0011). For each cap (2 μl, 100 μM), deionized water (15.5 μl),

Table 3.2: DNA sequences for torsionally constrained DNA

ID	Name	Sequence 5'-3'
MDNXXXX	End Cap #1	AGGTCGCCGCCCGGAGTTGAACG**TT**T**T**T**T**ACGTTCAACTCC
MDNXXXX	End Cap #2	GGGCGGCGACCTCAAGTTGGACAA**TT**T**TT**TGTCCAACTTG

Modified biotinylated nucleotides (iBiodT) highlighted in **bold**.

T4 DNA Ligase Buffer (2 μl, 10X) and polynucleotide kinase (0.5 μl, 10 U/μl) was added. All three reactions were incubated at 37°C for 1 hr.

Next, End Cap #1 was ligated to the λ-DNA. The phophorylated λ-DNA (56 μl) phospohorylated End Cap #1 (1 mul, 10 μM) were added to T4 DNA Ligase Buffer (50 μl, 10X), deionized water (400 μl). To facilitate annealing this mixture was incubated at 80°C for 5 minutes then placed directly on ice. T4 DNA ligase (4 μl, 400 U/μl) was added and the reaction incubated at 16°C for 12 hours. Phospohorylated End Cap #2 (10 mul, 10 μM) was then added and the reaction mixture again incubated at 80°C for 5 minutes and cooled rapidly on ice before incubation at 16°C for 12 hours. The final ligated DNA construct was stored at 4°C for upto 2 weeks or purified by ethanol precipitation and resuspended in TE for longer term storage at 4°C.

For all solutions containing λ-DNA, slow pipetting with wide bore pipette tips of and minimal mechanical mixing is vital to avoid nicking of the DNA.

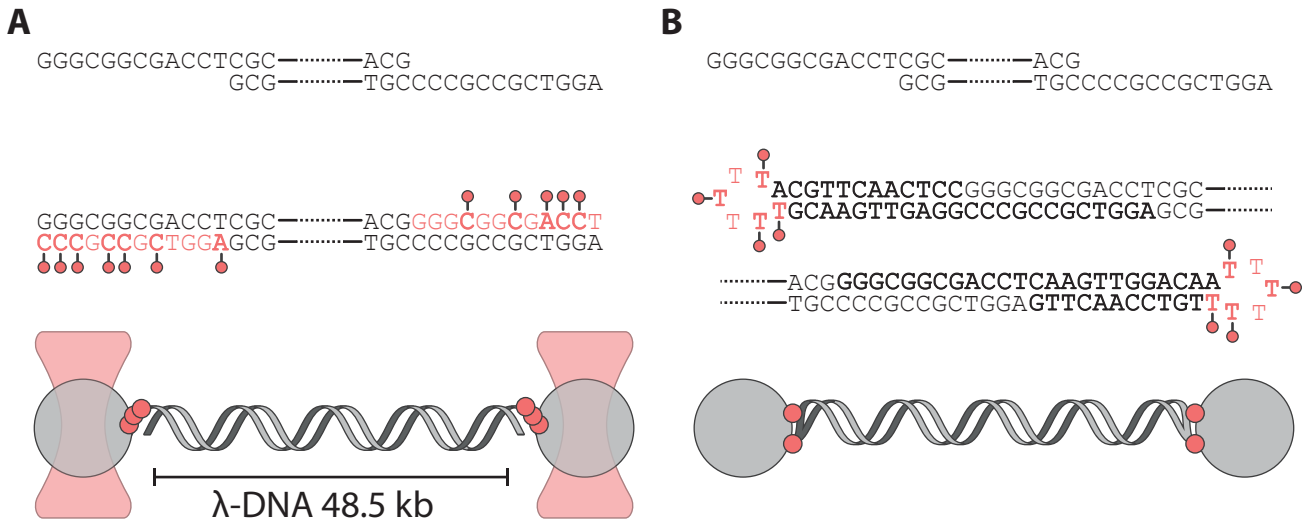


Figure 3.2: Biotinyating DNA. (**A**) Unconstrained λ-DNA tethered between two trapped beads. Positions of biotin shown with red circles. (**B**) Tortionally constrained λ-DNA. Positions of biotin shown with red circles.

3.3.2 Cleaning and passivation of the fludics and flow cell

The syringes, tubing and flow cell was cleaned with bleach (1 ml, 5%, NaClO), followed by ddH_2O (1 ml), sodium thiosulphate (1 ml, 100 mM in PBS) and finally ddH_2O, each flowed over approximately 10-15 minutes. All channels were passivated with Pluronic F128 (1 ml, 0.5% w/v in PBS) flowed over 30 minutes and protein channels further passivated with BSA (1 ml, 0.1% w/v in PBS) flowed over 30 minutes.

3.3.3 Cas9 optical tweezer experiments

Experiments were performed in a micrfluidics flow cell (Figure 3.3). SPHERO 4.5 μm Streptavidin Coated polystyrene beads were diluted to 0.005% w/v and flowed in channel 1 (Figure 3.3, **(1.)**). Biotinylated λ-DNA was diluted to 30 ng/ml and flowed in channel 2 (Figure 3.3, **(2.)**). The two part guide was annealed by mixing crRNA and tracrRNA in annealing buffer (100 mM NaCl, 50 mM Tris-HCl pH 8), heating to 90°C for 1 minute, and cooled at room temperature for 10 minutes before being put on ice. Cas9 was complexed with annealed guide in a 1:1 ratio at 1 uM for 10 minutes at room temperature before storing on ice. For experiments with dCas9, imaging buffer (100 mM NaCl, 50 mM Tris-HCl pH 8, 1mM $MgCl_2$, 0.2 mg $ml^{-}1$ BSA, 125 mM PCA, 10 μM PCD and saturating trolox) was flowed in channel 3 (Figure 3.3, **(3.)**)and crRNA:tracr:Cas9 (1 - 10 nM) in imaging buffer flowed in channel 4 (Figure 3.3, **(4.)**). For cutting experiments with wtCas9, $MgCl_2$ was omitted and EDTA added (1 mM) for channels 3 and 4, and cutting buffer with 10 mM $MgCl_2$ put in channel 5 (Figure 3.3 **(5.)**). Before beginning experiments, the protein channel was further passivated with crRNA:tracrRNA:Cas9 complex by flowing 0.5 ml over 30 minutes.

3.3.4 Data analysis

All data was exported as '.h5' files and analysed in python using the 'lumicks.pylake' python package and custom scripts. Images were extracted as .tiff files and either analysed with python of using FIJI to extract intensity profiles and quantify fluorescence. To map intensity signals back to the known DNA sequence the position of the beads and known confocal pixel size was

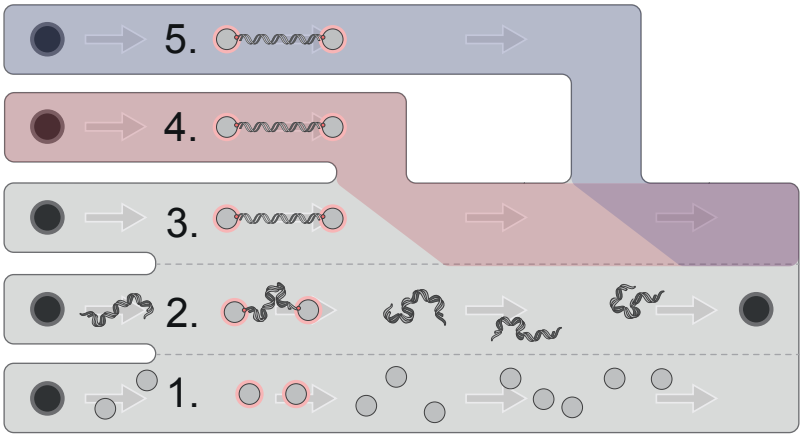


Figure 3.3: Microfluidics layout for optical tweezer experiments. (**1.**) Bead Channel (**2.**) DNA Channel (**3.**) Buffer Channel (**4.**) Protein Channel (**5.**) Additional Buffer/Protein Channel

used to determine the pixels corresponding to the DNA molecule. These were then linearly scaled along the length of the DNA sequence and peak positions accurately determined by multipeak gaussian fitting using Igor Pro from Wavemetrics.

3.4 CIRCLE-seq with supercoiling

3.4.1 Preparation of relaxed and supercoiled genomic DNA circles

3.4.1.1 Genomic DNA preparation and fragmentation

Preparation of DNA circles was performed as previously described (Tsai et al., 2017). HEK293T cells were grown on 150 mm plates in 50 ml media to ∼90 % confluence. Genomic DNA was extracted and purified with Qiagen Gentra Puregene Cell Kit (Cat. 158388). Purified genomic DNA was quantified by Qubit using high sensitivity dsDNA kit (Qubit dsDNA HS Assay Kit, Q32851), approximately 600 μg of genomic DNA was obtained for each 150 mm plate of cells. Purified DNA was mechanically sheared to an average length of 1000 bp in covaris 130 μl tubes with 10 μg DNA per tube using the following protocol: water temperature 7 °C. Further size selection cleanup was performed with AMPureXP Size Selection Magnetic beads using a 0.5

x bead wash and resuspended in low EDTA TE. Fragmented purified DNA was quantified by Qubit (Qubit dsDNA HS Assay Kit, Q32851) and size distribution analysed on a bioanlalyser to confirm correct distribution of fragment sizes around 1000 bp. For 600 μg input DNA before fragmentation approximately 240 μg was recovered.

3.4.1.2 End-repair and A-tailing

End-repair was then performed on purified fragmented genomic DNA (5 μg DNA, 7 μl KAPA End Repair Buffer 10X, 5 μl KAPA End Repair Enzyme Mix in final volume of 70 μl; 20 °C) for 30 minutes). For each end-repair reaction 5 μg of input DNA was used and for each condition to be tested by CIRCLE-seq. Five reactions, 25 μg, of input DNA is required, so for four conditions, e.g. with/without Cas9 and with/without supercoiling, 100 μg of input DNA is required. Post end-repair clean-up was performed with a 1.4x AMPure XP Bead ratio, 100 μl per 70 μl reaction, and DNA eluted in 42 μl of 1x TE buffer per reaction.

Next A-tailing was performed on end-repaired DNA fragments to add a 3' single adenine base for adapter ligation (42 μl end repaired DNA, 5 μl KAPA A-tailing Buffer 10x, 3 μl KAPA A-tailing Enzyme in a final volume of 50 μl; 30 °C for 30 minutes). Clean-up was performed with a 1.8x AMPure XP Bead ratio, with elution in 30 μl TE buffer per reaction and without removal of the beads.

3.4.1.3 Adapter ligation

The CIRCLE-seq hairpin adapter was then ligated to the A-tailed DNA fragments. First the adapter was annealed in the absence of the genomic DNA to allow formation of the hairpin structure (oSQT1288 Adapter at μM, 95 °C 5 minutes, -1 °C per minute for 70 minutes then hold at 4 °C. For each adapter ligation reaction, 5 μl of annealed adapter is required. The annealed adapter was then ligated to the A-tailed DNA (30 μl A-tailed DNA with AMPure beads still present, 5 μl annealed hairpin adapter at 40 μM, 10 μl KAPA ligation buffer 5x, 5 μl KAPA T4 DNA ligase; 20 °C for 1 hour). Clean-up was performed by adding 0.8x PEG/NaCl SPRI solution, 40 μl for each 50 μl adapter ligation reaction, to remove un-ligated adapters,

and eluted in 40 μl TE buffer with AMPure XP beads removed. All reactions were then pooled and quantified by Qubit HS dsDNA kit and on fragment analyzer to check concentration and average fragment size. A slight increase in fragment size is observed due to the ligated adapter. The adapter ligated DNA was next treated with exonulceases to remove any unligated fragments and adapters (12 μl adpater ligated DNA, 5 μl exonuclease I reaction buffer 5x, 4 μl Lambda Exonuclease at 5 U/μl, 1 μl *E. coli* Exonuclease I at 20 U/μl in 50 μl final reaction volume; 37 °C for 1 hour, 70 °C for 10 minutes). Clean-up was performed with 0.8x AMPure XP bead ratio and eluted in 40 μl 1x TE without removal of the beads.

3.4.1.4 Fragment circularisation

The adapter ligated DNA fragments were then treated with USER enzyme to open up the adapter hairpin for subsequent circularisation (40 μl exonuclease treated DNA with beads, 5 μl 10x T4 DNA ligase buffer, 3 μl USER enzyme at 1 U/μl, 2 μl T4 polynucleotide kinase at 10 U/μl; 37 °C for 1 hour. Clean-up was performed with 1.8x PEG/NaCl SPRI Solution and eluted in 40 μl 1x TE buffer with beads removed. All reactions were then pooled and quantified by Qubit HS dsDNA kit and on fragment analyzer to check concentration and average fragment size.

The USER treated DNA fragments were then circularised (500 ng USER treated DNA, 10 μl 10x T4 DNA Ligase Buffer, 2 μl DNA Ligase at 400 U/μl in a final volume of 100 μl; 16 °C for 16 hours). Clean-up was performed with 0.6x AMPure XP beads and eluted in 13.3 μl 1x TE with beads removed. Circularised DNA was treated with Plasmid-Safe ATP-dependent DNase to remove any non-circular fragments (38 μl circularised DNA, 5 μl 10x plasmid-safe reaction buffer, 2 μl ATP at 25 mM, 5 μl plasmid-safe ATP-dependant DNase at 10 U/μl; 37 °C for 1 hour, 70 °C for 30 minutes). Clean-up was performed with 1x AMPure XP beads and eluted in 60 μl 1x TE with beads removed. All reactions were then pooled and quantified by Qubit HS dsDNA kit and on fragment analyzer to check concentration and average fragment size.

3.4.1.5 Negative supercoiling of DNA circles

After preparation of the library of DNA circles one half was negatively supercoiled with DNA gyrase (1 μg DNA circles, 80 μl gyrase reaction buffer, 2 μl DNA gyrase at 5U/μl, made up to a final volume of 400 μl; 37 °C for 1 hour). For the negative control of non-supercoiled circles the same reaction was performed with the gyrase omitted. Clean-up was performed with 2x AMPure XP beads eluted in 165 μl in 1x TE with beads removed. For each sample, 1 μg of input DNA circles were used in one 400 μl reaction, e.g. to test one guide sequence with and without supercoiling and with and without Cas9 cleavage 4 x 1 μg each in 400 μl was required.

3.4.2 Cas9 *in vitro* digestion

Supercoiled and non-supercoiled prepared DNA circles were treated with Cas9 complexed with the guide of interest. First the tracrRNA and crRNA were annealed (5 μl tracrRNA at 10 μM, 5 μl crRNA at 10 μM, 5 μl Cas9 annealing buffer; 95 °C for 2 minutes, room temperature for $>$ 10 minutes and then placed on ice). Next the Cas9 nuclease was complexed with the annealed guides (10 μl 10x Cas9 reaction buffer, 9 μl SpCas9 Nuclease at 1 μM, 3 μl annealed crRNA:tracrRNA at 3.3 μM in 45 μl; held at room temperature for 10 minutes) before addition of the DNA circles and the initiation of the cleavage reaction (45 μl complexed Cas9 mix, 250 ng DNA circles in a total volume of 100 μl; 37 °C for 1 hour). For the negative controls without Cas9 the same protocol was followed with the Cas9 and annealed guide replaced with ddH_2O. Clean-up was performed with 1x AMPure XP beads eluted in 42 μl in 1x TE with beads removed.

3.4.3 Sequencing library preparation

The samples were then prepared by standard KAPA library preparation for NEBNext sequencing. Any circles linearized by Cas9 cleavage, as well as control samples, were first A-tailed (42 μl DNA sample, 5 μl 10x KAPA A-tailing buffer, 3 μl KAPA A-tailing enzyme; 30 °C for 30 minutes) with 1x ratio of AMPure XP beads and elution in 25 μl without removing the beads. Next the NEBNext sequencing adaptor was ligated to the fragments (25 μl A-tailed DNA, 10 μl

5x KAPA ligation buffer, 5 μl KAPA T4 DNA ligase 10 μl NEBNext Illumina adaptor at 15 μM (E7601A); 20 °C for 1 hour) and clean-up with 1x ratio of PEG/NaCl SPRI solution and elution in 47 μl 1x TE without removal of the beads. The adapter was then cut by addition of USER enzyme (3 μl USER enzyme, 47 μl adaptor ligated DNA with beads; 37 °C for 30 minutes) and clean-up with 0.7x ratio of PEG/NaCl SPRI solution and elution in 20 μl of 1x TE buffer with beads removed. The final DNA concentration of each sample was then measured with Qubit HS dsDNA kit. Samples were indexed with compatible combinations of NEBNext i5 and i7 indexing primers (25 μl 2x KAPA HiFi HotStart ReadyMix, 5 μl NEBNext i5 Primer at 10 μM, 5 μl NEBNext i7 Primer at 10 μM, 20 ng USER enzyme treated DNA in a total volume of 50 μl; 98 °C for 45 s, 22 cycles of (98 °C for 15 s, 65 °C for 30 s, 72 °C for 45 seconds), 72 °C for 2 minutes, hold at 4 °C) and clean-up with 0.7x ratio of AMPure XP beads and eluted in 30 μl 1x TE buffer. The concentration of indexed libraries were measured with Qubit BR dsDNA kit and bioanalyser run to check fragment size.

3.4.4 Sequencing and Analysis

Paired-end 150 bp read length sequencing was performed on an Illumina MiSeq with all samples from one experiment multiplexed. The sample was spiked with 10 % PhiX due to potential low sequence diversity of the sample. De-multiplexed FastQC files were then analysed using the CIRCLE-seq analysis software with custom parameters (Tsai et al., 2017).

References

Gross, Peter et al. (2010). "Combining Optical Tweezers, Single-Molecule Fluorescence Microscopy, and Microfluidics for Studies of DNA-Protein Interactions". *Methods in Enzymology*. Vol. 475. C, pp. 427–453.

King, Graeme A et al. (Dec. 2019). "Supercoiling DNA optically". *Proceedings of the National Academy of Sciences* 116.52, pp. 26534–26539.

Tsai, Shengdar Q et al. (May 2017). "CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets". *Nature Methods* 14.6, pp. 607–614.

Zhao, Rui and David Rueda (Oct. 2009). "RNA folding dynamics by single-molecule fluorescence resonance energy transfer". *Methods* 49.2, pp. 112–117.

Chapter 4

Mapping the sugar dependency for rational generation of a DNA-RNA hybrid-guided Cas9 endonuclease

4.1 Abstract

Efforts to rationally engineer higher specificity CRISPR Cas9 variants have been largely based around the 'excess energy hypothesis', that there is an excess on-target Cas9 binding energy. For Cas9 to bind at off-target sites in the presence of mismatches, this excess binding energy is required. As such, reducing the binding energy overall may not impact on-target activity, but could preferentially destabilise off-target activity and increase Cas9 specificity. One potential target for modulating the binding energy is the energy of hybridisation between the Cas9 guide and target DNA.

The following chapter describes a collaborative project between myself and Benjamin Taylor's group at AstraZeneca. Members of the Taylor group have developed a series of hybrid guides, containing DNA bases in place of the canonical guide RNA bases. Using a series of biophysical assays they demonstrated that these hybrid guides can bind the Cas9 complex with high affinity, and direct it for high efficiency *in vitro* target cleavage. Based on the reduced hybridisation energy of DNA-DNA base-pairing compared to RNA-DNA base pairing, we hypothesised that such guides may exhibit improved specificity. They demonstrated that such hybrid guides are better at discriminating off-targets containing single mismatches. Furthermore, as predicted the presence of DNA bases in the guide results in reduced DNA binding affinity in bulk assays.

In order to understand this reduced affinity in more detail we performed single-molecule FRET experiments to probe target binding and R-loop formation of these hybrid guides. We demonstrate that while these hybrid guides can bind target DNA, the high FRET state associated with full R-loop formation is destabilised. This could explain the overall reduced DNA binding affinity but also the improved discrimination of sites containing mismatches.

4.2 Results

4.2.1 Substitution of RNA with DNA in guides

To investigate whether Cas9 can accommodate DNA instead of RNA in the crRNA, a fully RNA guide, crRNA, and equivalent fully DNA guide, crDNA, were synthesised targeting a site in a 790 bp dsDNA fragment encoding enhanced GFP, eGFP (Table 4.1). First, *in vitro* cleavage was tested; the crRNA guide resulted in efficient specific cleavage of the dsDNA fragment, however no cleavage was observed with the crDNA guide (Figure 4.1A). We hypothesised that the lack of activity with the crDNA guide could be due to the loss of interactions between the Cas9 protein and the 2'-hydroxly groups absent in the DNA guide compared to the RNA guide. Based on existing crystal structures of crRNA:tracr:Cas9 bound to target DNA (Nishimasu et al., 2014), seven ribose specific 2'-hydroxly interactions were identified between the guide and Cas9. To test whether these are required for activity, a third DNA hybrid guide was synthesised containing only these seven RNA bases to restore these interactions, crHyb (Table 4.1). The crHyb guide exhibited efficient and specific dsDNA target cleavage (Figure 4.1A). This result was further confirmed with guides targeting two additional sites. For all targets tested, efficient *in vitro* cleavage was observed with crRNA and crHyb guides, with no cleavage detected with the crDNA guide (Figure 4.1B). This demonstrates that only seven RNA residues are required for Cas9 activity *in vitro*.

Table 4.1: crXNA Sequences

Name	Sequence
	|---Guide Region---| |-----cr Repeat------|
	....5....10...15...20 ...25...30...35...40.
crRNA	**agcacugcacgccguaggucguuuuagagcuaugcuguuuug**
crDNA	AGCACTGCACGCCGUAGGUCGUUUTAGAGCTATGCTGTTTTG
crHyb	**a**GCACTGCACGCCG**ua**GG**u**C G**uuu**TAGAGCTATGCTGTTTTG
crHybΔ5	**a**GCACTGCACGCCG**ua**GGTC GT**uu**TAGAGCTATGCTGTTTTG

RNA bases in lowercase **red**, DNA bases uppercase black

To further investigate the critical interactions required for activity, each of the seven RNA residues in the crHyb were individually substituted for DNA and tested for *in vitro* cleavage

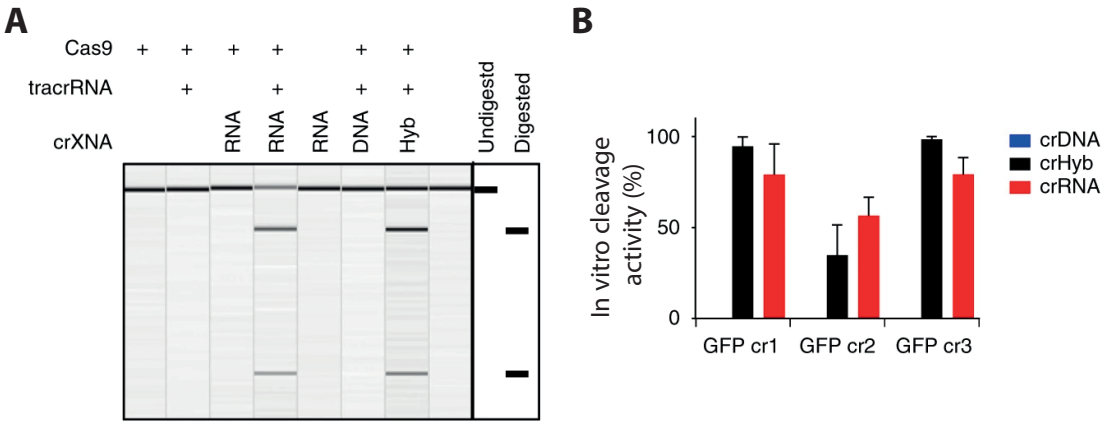


Figure 4.1: crXNA in vitro cleavage. (**A**) Target eGFP DNA was incubated with crXNA Cas9 complexes, cleavage products were resolved by gel electrophoresis. Expected cleavage products demonstrated on the right. (**B**) Relative *in vitro* cleavage efficiencies with crRNA (red), crDNA (blue) and crHyb (black) guides targeting three different sites in eGPF DNA. Error bars display SD, n = 3-7. Experiments performed by Fernando Orden Rueda and Michal Bista.

activity. Interestingly, four of the residues, 1, 15, 19 and 22, could be individually substituted for DNA and retain efficient cleavage activity (Figure 4.2). Furthermore, it was found that both residues 19 and 22 could be substituted for DNA at the same time without a loss of activity, resulting in a guide containing only 5 RNA residues, crHybΔ5 (1, 15, 16, 23 and 24) (Figure 4.2 and Table 4.1).

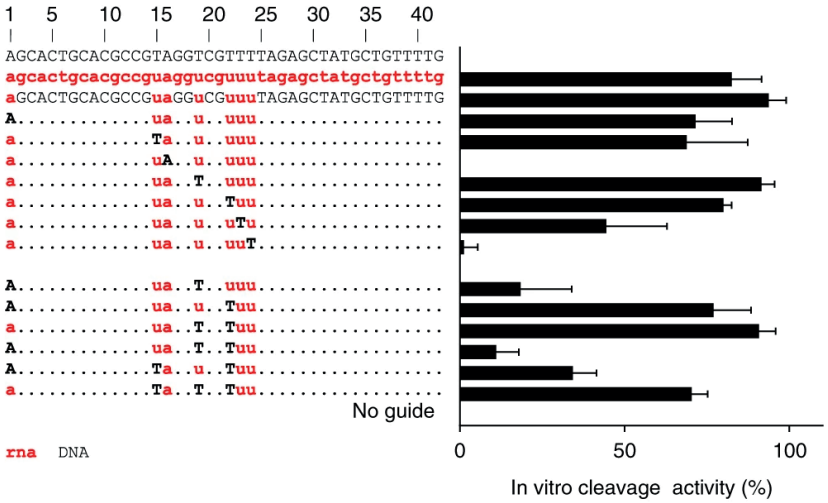


Figure 4.2: Systematic substitution of remaining RNA bases within crHyb to generate minimal RNA containing guide, crHybΔ5. Each of the remaining RNA bases present in the crHyb were individually substituted for DNA bases and cleavage activity assessed by *in vitro* cleavage assay, top. Based on results from single base replacements pairs of RNA bases were tested to obtain the minimal RNA containing guide, crHybΔ5, bottom. Experiments performed by Fernando Orden Rueda and Michal Bista.

4.2.2 crDNA binds tracr:RNA with high affinity

To understand whether the loss of activity of the crDNA guide was due to an inability to bind the Cas9 complex, binding kinetics were measured using surface plasma resonance, SPR. First, dCas9 was immobilised on the surface and complexed with tracr RNA. Next, the crXNA was added to measure formation of the full crXNA:tracr:Cas9 complex (Figure 4.3A). Binding of all three crXNAs was observed, crRNA bound with the highest affinity (K_d = 81 nm, t_2 = 240 s) however crDNA still bound with sub nanomolar affinity (K_d = 584 nm, t_2 = 42 s) (Figure 4.3B). This suggests Cas9 can still form a highly stable complex with a fully DNA guide.

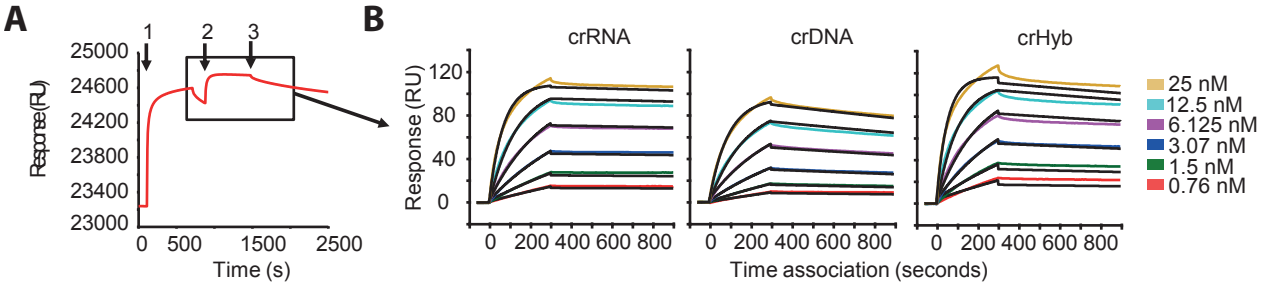


Figure 4.3: crXNA guides bind to Cas9:tracrRNA complex with high affinity. **(A)** Overview of SPR protocol. First, dCas9 was immobilised on the SPR chip. Next an excess (100 nM) of tracrRNA was injected to saturate bound dCas9 (1). Then the crXNA was injected to measure binding affinity (2). Finally a wash step was performed to remove any unbound crXNA, and dissociation of bound crXNA from the dCas9:tracrRNA complex association measured (3). **(B)** SPR trajectories of crRNA, crDNA and crHyb dCas9:tracrRNA binding and dissociation and crXNA concentrations ranging from 0.76 nM - 25 nM. Experiments performed by Fernando Orden Rueda and Michal Bista.

4.2.3 Hybrid guides have reduced off-target activity *in vitro*

Previous work has suggested that reducing the binding energy of the Cas9 complex with DNA can increase specificity (See section 1.3.3). Due to the reduced stability of DNA:DNA compared to RNA:DNA hybrids it was hypothesised that using guides containing DNA bases could 'tune' the binding energy of the complex to reduce off-target activity. To test this hypothesis, crRNA and crHyb guides were generated containing a single C to A substitution at residue 5, resulting in a single mismatch between the guide and DNA target at the PAM-distal end, crXNAC5A (Supplementary Table 4.3). For the fully RNA guide, the presence of the mismatch resulted in a 40% reduction in *in vitro* cleavage activity (Figure 4.4A). However, for the hybrid guide, the presence of a single mismatch almost entirely prevented cleavage, with less than 4% cleavage

activity detected compared to the perfectly matching crHyb guide (Figure 4.4A). This supports the idea that incorporating DNA bases into the guide can increase Cas9 specificity.

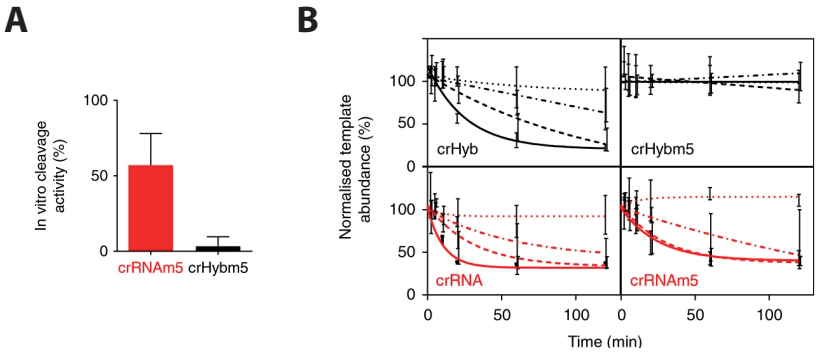


Figure 4.4: Hybrid guides have reduced off-target activity in vitro. (**A**) Analysis of relative cleavage efficiencies in the presence of single mismatch versus fully complementary crRNA and crHyb guides. crHyb displays greater mismatch sensitivity. Error bars SD, n = 3-13. (**B**) qPCR target depletion assay for crRNA and crHyb with and without mismatch, crRNAm5 and crHybm5. With the fully complementary guides crHyb displays slower target depletion than the crRNA guide. In the presence of a single mismatch, cleavage rate is reduced for crRNA but no cleavage is observed with the crHyb guide. Graphs show normalised template abundance (%) over time, with Cas9 protein at 5 nM (solid line), 2.5 nM (long dash line), 1.25 nM (dot dash line) or 0.625 nM (dotted line). Error bars show SD of 3–4 independent measurements. Experiments performed by Fernando Orden Rueda and Michal Bista.

One explanation for this apparent increased specificity is an overall reduction in activity of the crHyb Cas9 complexes. To test this, time-course cleavage assays were performed with a range of protein concentrations using a qPCR target depletion assay. The assay is designed such that cleavage of the target DNA by Cas9 prevents PCR amplification, with qPCR used to calculate the relative abundance of uncut intact targets at each time-point. Target cleavage was slower with the crHyb:tracr:Cas9 complex compared to the crRNA:tracr:Cas9 complex (Figure 4.4B). Efficient target cleavage was observed with the crRNA guide containing a single mismatch, although there was a reduction in rate compared to the perfectly matched guide. Strikingly, no cleavage was detected at any time point or concentration with the crHyb guide containing a single mismatch. This suggest that while the presence of DNA in the guide does slow down the rate of Cas9 cleavage overall, it also dramatically increases the sensitivity to mismatches.

4.2.4 Hybrid guides reduce DNA binding

The observed reduced cleavage rates with crHyb could be explained by a reduced affinity for the target DNA. To test this, the kinetics of target binding were measured using BioLayer

Interferometry, BLI, and SwitchSENSE assays. In both assays, target DNA containing a sequence from the human AAVS1 locus was immobilised, and binding of freely diffusing Cas9 complex with the corresponding guide sequence measured. For BLI, the DNA target is immobilised on a sensing tip which is then 'dippped' into a solution contain the crXNA:tracr:Cas9 complex; binding results in a slight increase in tip thickness which is detected by changes in the diffraction pattern. The tip is then moved to a buffer only solution, allowing the dissociation to be measured. In the SwitchSENSE assay, target DNA with a reporter fluorophore is immobilised on a gold surface and an alternating voltage applied, causing the DNA to oscillate towards and away from the surface. On addition of Cas9 complex, binding to the target is detected by a change in the speed of oscillation due to the increase in hydrodynamic drag when Cas9 is bound to the DNA. With the crRNA:tracr:Cas9 complex, both methods detected high affinity binding to the target, observed by a rapid drop in the normalised dynamic response in the SwitchSENSE assay and increase in the BLI response (Figure 4.5A and B). No binding was observed with crDNA guided Cas9 in either assay, suggesting the lack of cleavage activity is due to an inability for the complex to bind the target site. Using the SwitchSENSE assay, binding of the crHyb:tracr:Cas9 complex was detected but with a slower on rate than the crRNA complex, demonstrating a lower affinity for the target (Figure 4.5A). However, almost no binding of the same crHyb:tracr:Cas9 complex was detected with the BLI assay (Figure 4.5B).

To try to understand whether the reduced affinity with crHyb is due to the presence of DNA bases in the region which directly pairs with the target DNA, or in the region which interacts with the tracrRNA, two additional hybrid guides were tested. In the first, crD.Hyb, RNA bases were used for the entire cr repeat region with DNA bases left in the guide region as with crHyb (Table 4.2). For the second, crR.Hyb, RNA was used for the entire guide region and only 13 DNA residues used at the 3' end of the cr repeat region (Table 4.2). In the SwitchSENSE assay crD.Hyb displayed an increased affintiy compared to crHyb, with an on rate close to that of the crRNA guide (Figure 4.5A). However in the BLI assay crD.Hyb dispalyed very weak binding, as was also seen for the crHyb guide with this method (Figure 4.5B). In contrast the crR.Hyb recovered high affinity binding in the BLI assay, suggesting that DNA target binding in this assay is very sensitive to DNA bases in the guide region (Figure 4.5B). Together this data

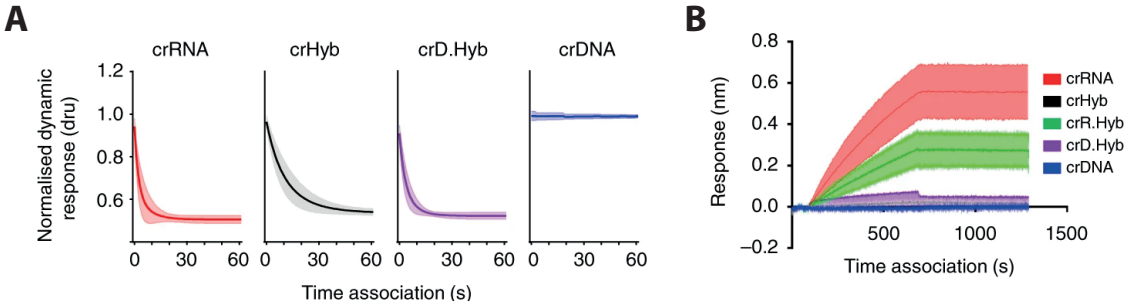


Figure 4.5: Hybrid guides have reduced DNA target binding affinity. (**A**) SwitchSENSE DNA binding assay. High affinity crRNA guided Cas9 complex binding to the target DNA is observed as a rapid decrease in the normalised dynamic response. In contrast no DNA binding is detected with the crDNA guide. crHyb displays an intermediate binding affinity, with crD.Hyb showing a response comparable to the crRNA. Shaded areas represent standard deviation, n = 3. (**B**) BLI DNA binding assay. High affinity binding of crRNA guided Cas9 to target DNA is observed by a rapid response in the switchSENSE plot, red. No binding of the crDNA guided complex is observed, with no response detected, blue. No binding is detected with the crHyb, black, and only very weak interaction detected with crD.Hyb complex, purple. The crR.Hyb complex shows intermediate affinity, green. Shaded regions represent standard deviation, n = 4. Experiments performed by Fernando Orden Rueda and Michal Bista.

suggests that the lack of cleavage with the crDNA guides can be explained by an inability of the complex to bind DNA. crHyb complexes appear to bind with a much lower affinity, which is affected by the presence of DNA bases in both the guide and cr repeat region.

Table 4.2: AAVS1 crXNA Sequences

Name	Sequence
	....5....10...15...20 ...25...30...35...40.
crRNA	**gccaguagccagccccgucc guuuuagagcuaugcuguuuug**
crDNA	GCCAGTAGCCAGCCCCGTCC GTTTTAGAGCTATGCTGTTTTG
crHyb	**g**CCAGTAGCCAGCC**cc**GT**c**C G**uuu**TAGAGCTATGCTGTTTTG
crD.Hyb	**g**CCAGTAGCCAGCC**cc**GT**c**C **guuuuagagcuaugcuguuuug**
crR.Hyb	**gccaguagccagccccgucc guuuuagag**CTATGCTGTTTTG

RNA bases in lowercase red, DNA bases uppercase black

4.2.5 Hybrid guides destabilise R-loop formation

Together, the data of crXNA tracr:Cas9 complex binding, DNA target binding and *in vitro* cleavage, suggests that the hybrid guides are bound by the Cas9 complex and can direct the complex to specifically bind target DNA and reach a catalytically competent conformation allowing cleavage. The reduced target binding affinity observed with the crHyb guide in

comparison to the fully RNA crRNA could be due to a reduced stability of the DNA:DNA base paring at the target, or a more general loss of activity of the complex. To investigate this we used a previously described single-molecule Forster resonance energy transfer, smFRET, assay to Cas9 binding and R-loop formation. Target DNA containing the previously characterised AAVS1 target sequence, and labelled with Cy3, was immobilised by biotin-streptavidin interactions on the surface of a microscope slide and imaged on our home-built total internal reflection microscope, TIRF (Figure 4.6A). The crXNA:tracr:dCas9 complex was labelled with Cy5 on the 5' end of the crXNA and injected into the slide. On binding to the target DNA, FRET occurs between the donor (Cy3) on the DNA, and acceptor (Cy5) on the crXNA, the magnitude of which provides information on the conformational state and extent of R-loop formation (Figure 4.6B and C).

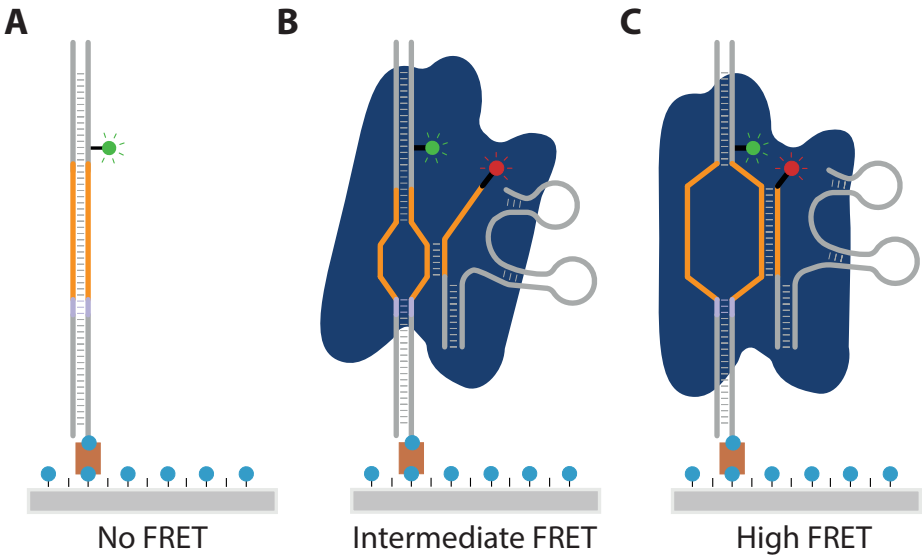


Figure 4.6: Hybrid guide smFRET experimental design. Experimental design of the smFRET Cas9 binding assay. DNA containing the target sequence (orange) and PAM site (lilac) is labelled with Cy3 donor (green) at the tethered to the surface by biotin-streptavidin interactions. Cas9 complex is labelled with Cy5 acceptor (red) on the end of the crRNA. (**A**) In the absence of Cas9 binding, no acceptor is present and there is no FRET. (**B**) On Cas9 binding and partial R-loop formation FRET takes place between the donor and acceptor give a intermediate FRET value. (**C**) On full R-loop formation the donor and acceptor are in very close proximity, resulting in high FRET.

We first tested the fullly RNA crRNA guide. following injection of the crRNA:tracrRNA:dCas9 complex, the majority of single-molecule trajectories (66%, n = 60/91) reach a stable high-FRET state (∼0.95) which, in traces where the initial binding event is observed, transition through a transient intermediate FRET conformation (∼0.5, arrow) with an average dwell time

of 310 ± 40 μs (n = 30, Figure 4.7A). Based on previous studies, we assigned these trajectories to initial binding in an open conformation (∼0.5 FRET, arrow) followed by zipping into a full heteroduplex conformation (∼0.95). A single-molecule time-binned FRET histogram shows that these molecules predominantly occupy the high-FRET conformation, as expected (Figure 4.7B).

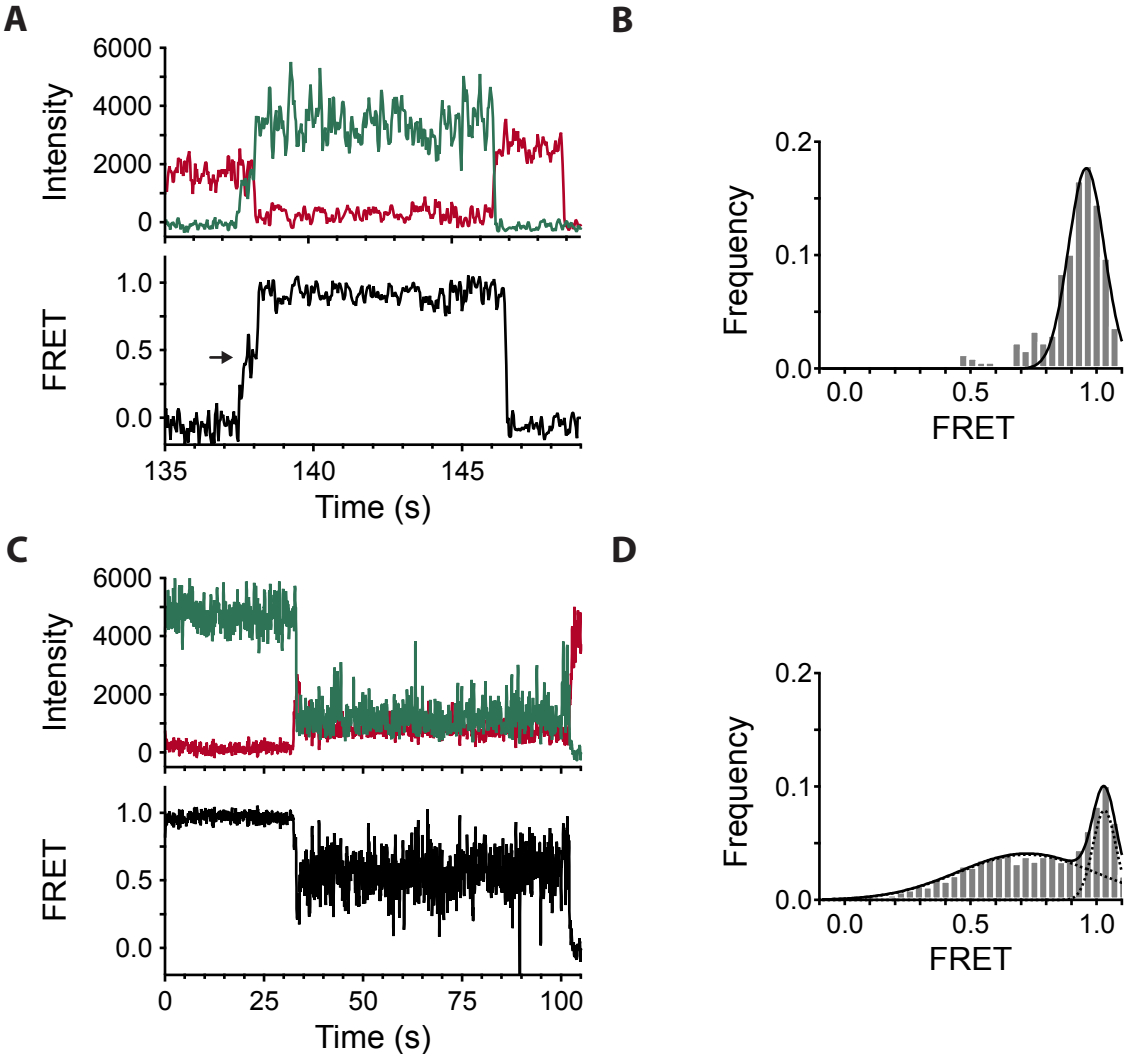


Figure 4.7: Representative smFRET trajectories and time binned histograms for crRNA:tracr:Cas9. (**A**) Cas9 binding and stable R-loop formation. Top, donor Cy3 (green) and acceptor Cy5 (red) intensities. Bottom, calculated FRET. (**B**) Time-binned FRET histogram of trajectory in A, fit with single Gaussian. (**C**) Transition from stable high FRET state to dynamic state. Top, donor Cy3 (green) and acceptor Cy5 (red) intensities. Bottom, calculated FRET. (**D**) Time-binned FRET histogram of trajectory in C, fit with double Gaussian.

Interestingly, a second population of trajectories (34%, n = 31/91), reach the high-FRET conformation but then transition irreversibly into a dynamic ensemble of states characterised by rapid FRET fluctuations from ∼0.4 to ∼0.9 FRET (Figure 4.7C). The corresponding time-

binned single-molecule histogram shows a narrow peak centred on ∼1.0 and a broad peak at ∼0.70 (Figure 4.7D). This population may correspond to molecules that either unzip from the fully duplexed conformation and repeatedly attempt to return to it, or a protein conformational change that affects the local environment of the FRET pair. This population of molecules may not have been observed in previous studies, due to our work being at near-physiological [Mg^{2+}], 1 mM, rather than 10 mM in previous studies . A time-binned FRET histogram of 91 trajectories with crRNA shows that in the presence of a fully RNA guide the complex predominantly exists in the fully heteroduplex conformation (Figure 4.8, crRNA), consistent with the tight binding previously observed.

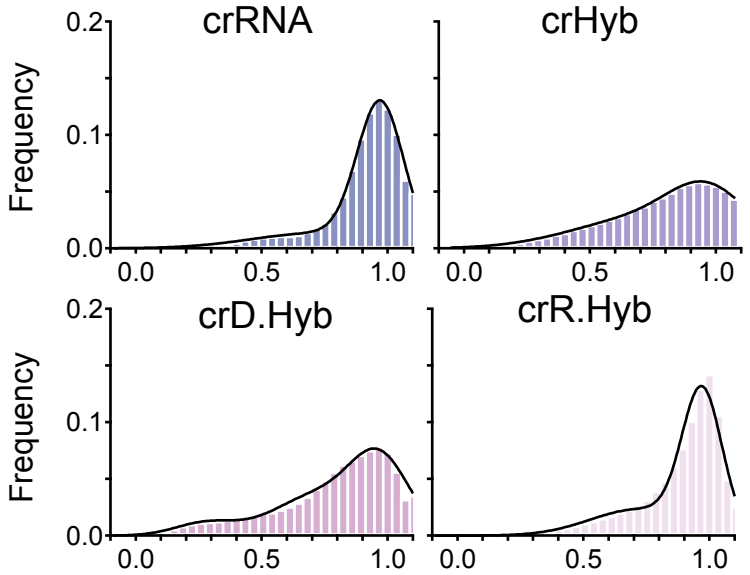


Figure 4.8: crXNA time-binned FRET histograms. Time-binned FRET histograms of all binding events detected for crRNA (n = 91), crHyb (n = 57), crD.Hyb (n = 197) and crR.Hyb (n = 172).

We next tested the hybrid guide, crHyb, in the same assay. Compared to crRNA, crHyb resulted in eightfold fewer overall binding events, consistent with the lower binding affinity observed by BLI and SwitchSENSE. Of the binding events detected (n = 57), we observed a fourfold decrease of stable high-FRET trajectories (17.5%) in favour of traces demonstrating the dynamic broad mid-FRET state (70%), and a minor (12.5%) previously unobserved low-FRET (∼0.2) population. The corresponding time-binned FRET histogram of all trajectories, shows a significant decrease in the high-FRET population and a broadening of the mid-FRET distribution (Figure 4.8, crHyb). This indicates that the presence of DNA bases in crHyb

hinders the formation of the active high-FRET conformation. These results are in agreement with the perturbed binding and cleavage kinetics observed (Figures 4.4 and 4.5). Taken together, the data with dCas9:tracrRNA:crHyb confirm the reduced binding affinity for target DNA and suggest that, in concordance with the detected *in vitro* cleavage activity, it can reach the active conformation required for cleavage but that this state has reduced stability. Furthermore, the reduced stability of this state could explain the increased specificity observed (Figure 4.4), with the complex more sensitive to the presence of mismatches which further destabilise the active conformation.

To test whether the low levels of crHyb-binding observed arises from the presence of DNA bases in the guide region or the cr repeat region, we measured the binding dynamics of the Cas9:tracrRNA:crD.Hyb complex (Figure 4.9). This complex has a fully RNA cr repear region but almost entirely DNA guide region, with only the 4 RNA bases previously demonstrated to be required for activity (Table 4.2). Interestingly, a partial (3-fold) recovery of the number of binding events was observed compared to crHyb, but still (2.6-fold) less binding than crRNA, also in agreement with the SwitchSENSE measurements. However, the fraction of molecules in each state was comparable to those observed with crHyb. Only 18.5% of trajectories maintained the stable high-FRET state (Figure 4.9A), while 53% display the dynamic broad mid-FRET state (Figure 4.9B and C) and the remaining 27% reach the 0.2 FRET state (Figure 4.9D). The resulting time-binned FRET histogram (Figure 4.8, crD.Hyb, n = 197) shows only a minimal recovery of the high-FRET peak, in comparison to crHyb and crRNA, and a minor new peak at ∼0.2 FRET. These data indicate that while crD.Hyb allows the Cas9 complex to efficiently bind target DNA, the active high FRET conformation remains unstable.

We similarly tested the crR.Hyb in which on the final 12 bases of the cr repeat region are replaced with DNA (Table 4.2). The dCas9:tracrRNA:crR.Hyb complex displays binding comparable to the full RNA guide complex, in agreement with the BLI and SwitchSENSE measurements (Figure 4.8). The fraction of trajectories reaching the stable high- FRET is 54% (n = 93/172) and a time-binned histogram shows an almost identical distribution to that of crRNA (Figure 4.8). Together these data indicate that the crR.Hyb dCas9 complex adopts the stable high-FRET conformation with similar efficiency as the crRNA complex.

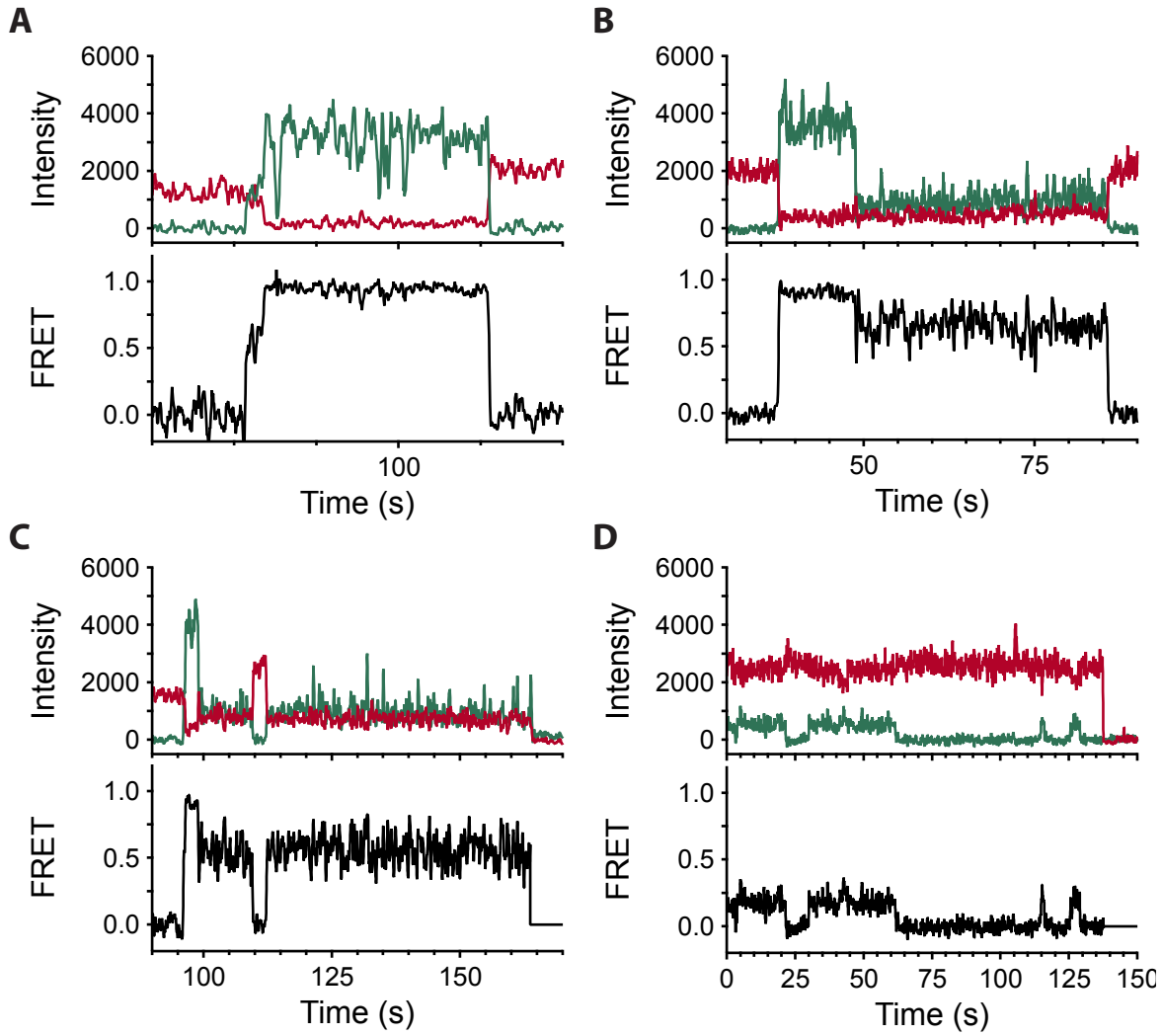


Figure 4.9: Representative smFRET trajectories for crD.Hyb:tracr:Cas9. (**A**) Cas9 binding and stable R-loop formation. (**B**) Transition from stable high FRET state to dynamic state. (**C**) Transition from stable high FRET state to dynamic state. (**D**) Transient repeat binding in a low FRET state. Top, donor Cy3 (green) and acceptor Cy5 (red) intensities. Bottom, calculated FRET.

Overall, the single-molecule results suggest that DNA in both the cr repeat region and the guide region can affect the ability of the complex to bind and transition from the 'open' to stably bound active state. The presence of DNA in the guide region results first in an overall reduced binding affinity, and second, when binding occurs, in a reduced stability of the high-FRET state in favour of the dynamic mid-FRET state. This suggests that the presence of high DNA load in the guide region destabilises this state resulting in rapid transitions between the two. Specific comparison of the crHyb and crD.Hyb guides demonstrates that in combination with DNA in guide region, DNA in the cr repeat region results in a further reduction in stable binding overall, and reduces the proportion of molecules which reach a stably bound active

state. However, the high activity of crR.Hyb, comparable to that of crRNA, demonstrates that Cas9 can accommodate DNA in the cr repeat region with a fully RNA guide region.

4.2.6 Hybrid guides fail to induce efficient target cleavage in cells

Following the *in vitro* characterisation of hybrid guides we next tested whether they are active in mammalian cells. A reporter assay was developed with Cas9 targeting the AAVS1 locus. HEK293 cells were transfected with preformed crXNA:tracr:Cas9 complexes by electroportation. Repair of induced double-strand breaks results in *de novo* mutations in the form of insertions and deletions, indels. Cells were harvested 5 days after transfection and indels detected by PCR and sanger sequencing of the target site, followed by deconvolution using TIDE analysis to calculate mutation frequencies. Transfection with the crRNA:tracr:Cas9 complex resulted in an indel frequency of 60% at the AAVS1 locus, demonstrating efficient targeting and cleavage (Figure 4.10). As expected from the *in vitro* experiments, no indels were detected with the crDNA:tracr:Cas9 complex (Figure 4.10). We next tested the crHyb guided complex, no indels were detected despite the efficient *in vitro* cleavage demonstrated previously. The crD.Hyb also failed to induce target cleavage (Figure 4.10). Indels were detected with the crR.Hyb guided complex, however with a much lower frequency (15.6%) than with the crRNA guide (60%) (Figure 4.10). This suggests that the lower binding affinity of Cas9 complexes with the DNA containing guides may prevent efficient cleavage in the cellular environment. Alternatively these complexes may have reduced stability in cells.

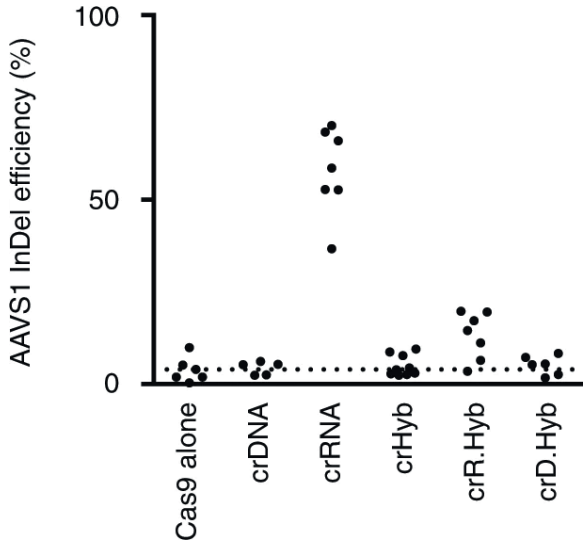


Figure 4.10: Induced InDel frequencies of crXNAs in cells. Quantification of InDels induced at AAVS1 locus 5 days after transfection with either Cas9 alone or Cas9 in complex with the crDNA, crRNA, crHyb, crR.Hyb and crD.Hyb guides.

4.3 Discussion

Since the discovery of the *Streptococcus pyogenes* Cas9 system as an RNA guided endonucleases it has been applied extensively for genetic manipulation both *in vitro* and *in vivo*. However, observations of off-target activity of Cas9 have lead to efforts to improve specificity by protein engineering of Cas9 systems. Due to the high stability of Cas9 on-target, rational design has typically been based around the excess energy model; a reduction in the Cas9 protein binding affinity limits off-target interactions while maintaining on-target activity (Slaymaker et al., 2016; Kleinstiver et al., 2016). As the high affinity of Cas9 to target DNA comes primarily from the formation of the DNA:RNA duplex, we hypothesised that replacing RNA bases within the guide with DNA bases could allow fine tuning of the binding affinity, due to the lower stability of DNA:DNA duplexes. Here, we demonstrate that Cas9 can form a complex with a fully DNA guide, however this complex is not active for cleavage (Figure 4.1). By rational design we identify 7 bases that make important RNA specific contacts with the Cas9 protein, and demonstrate that restoring these sites to RNA in the crDNA guide recovers on-target cleavage activity (Figure 4.1). Further systematic substitution of each of the 7 remaining RNA bases with DNA results in a function guide, crHybΔ5, containing only 5 RNA bases (Figure 4.2. *In vitro*, Cas9 on-target cleavage efficiency with the new crHyb, containing 7 RNA bases, is comparable to cleavage with crRNA, although slightly slower (Figure 4.4). There is significant variation in the absolute cleavage efficiency between different guide and target sequences for crRNA and crHyb, suggesting that sequence not only influences Cas9 activity, but also the tolerance for DNA bases in the guide. In the presence of a single mismatch between the guide and target, cleavage efficiency with crHyb was reduced to $< 4\%$ (Figure 4.4). In contrast, efficient cleavage ($> 60\%$) was still observed for crRNA, although at a slower rate than cleavage of the target without any mismatches (Figure 4.4). This demonstrates that the presence of DNA bases in the guide sequence reduces off-target activity while retaining on-target activity in vitro, consistent with the excess energy hypothesis for rational design of higher fidelity Cas9 complexes.

We have further demonstrated that the presence of DNA bases in the guide reduces DNA affinity using both BLI and SwitchSENSE assays. However these two approaches yielded dramatically

different values of DNA affinity (Figure 4.5). The SwitchSENSE assay results were consistent with the *in vitro* cleavage assay data. For crDNA no binding was detected, consistent with it's inability to cleave target DNA, whereas, crHyb displays an intermediate binding affinity compared to crRNA, consistent with it's ability to cleavage on-target but loss of activity in the presence of a mismatch. Interestingly, retaining the RNA bases in the cr repeat region increased affinity, as demonstrated with the crD.Hyb experiments. In contrast, in the BLI assays no binding was detected with either crDNA or crHyb, and only a very low level detected with crD.Hyb. High-affinity binding was observed with crRNA as expected, and an intermediate level of binding detected with a mostly RNA guide, crR.Hyb, containing only 13 DNA bases at the 3' end. A possible explanation for the different results from the BLI and SwitchSENSE assays is the requirement for the use of heparin in the BLI assay to reduce non-specific interactions of the Apo-Cas9 complex with the probe. While binding of the crRNA and crR.Hyb Cas9 complexes are still detected in the presence of heparin, binding by the crHyb and crD.Hyb complexes may be inhibited (Figure 4.5, Supplementary Figure 4.11). Furthermore, when measuring Cas9 complex dissociation in the absence of heparin there was a large retention of bound complex for crHyb and crD.Hyb compared to the apo-Cas9 which rapidly dissociated (Supplementary Figure 4.11). This suggests that the binding of the hybrid guides is specific on-target binding.

Using a single-molecule FRET assay we were able to observe stable target binding of the crRNA:tracrRNA:dCas9 complex. The majority of molecules reach a stable high FRET state indicative of full R-loop formation. Although the binding of the crHyb guided complex was dramatically reduced, molecules were observed reaching the high FRET state, consistent with the ability for the crHyb guide to cleave target DNA *in vitro*. However, the emergence of a mid-FRET state population suggests that the presence of DNA bases in the guide destabilises the R-loop formation. This would explain the reduced binding affinity, but could also explain the improved specificity. The presence of mismatches between the guide and target would be expected to further destabilise the R-loop, which may prevent the complex reaching the cleavage competent conformation. It is evident that the incorporation of DNA bases in the cr repeat region also has an effect of the activty of the Cas9 complex. When all cr repeat region bases in the crHyb were restored to RNA, crD.Hyb, we observed an increase in the number of binding

events in the smFRET assays. However the complexes still displayed reduced stability of the high FRET state compared to the crRNA guide, suggesting that this is due to the presence of DNA bases in the guide region. This supports the hypothesis that the replacement of RNA bases with DNA could specifically alter R-loop stability and improve discrimination of mismatches which would not be tolerated in an already less stable R-loop.

Based on all the *in vitro* characterisation of the hybrid guides it was surprising that efficient target cleavage was not observed in cells. One explanation of this is that the reduced affinity of these guides proves more detrimental to cleavage activity *in vivo*. For example the addition of heparin in the BLI assay significantly affected the Cas9 interactions with target DNA, suggesting that under these conditions non-specific Cas9 interactions may have a large impact on target affinity. Furthermore, in the cellular context the majority of DNA present does not contain the target site whereas in these assays only short DNA sequences were used. While reduced affinity may still permit cleavage *in vitro*, *in vivo* it may be detrimental to activity. Furthermore, the stability of these hybrid guides *in vivo* is unknown. If they are recognised as RNA-DNA hybrids or single-stranded DNA they may be degraded by cellular exo- and endonucleases. The use of locked-nucleic acids, LNA, or bridged nucleic acids, BNA, has been suggested to improve the stability of the Cas9 guide RNAs (Rahdar et al., 2015; Cromwell et al., 2018; Ryan et al., 2018) and could be tested in this context for their ability to prevent guide degradation.

Overall, we demonstrate that Cas9 can accommodate an almost entirely DNA guide while retaining *in vitro* cleavage activity and with improved specificity. The use of such hybrid guides reduces the stability of R-loop formation, likely due to a reduced hybridisation energy. This further supports the idea that modulation of the hybridisation energy between the guide and target DNA can be used to improve mismatch discrimination. We propose that such approaches could be informative to the development of higher fidelity Cas9 complexes with reduced off-target activities.

4.4 Supplementary Figures

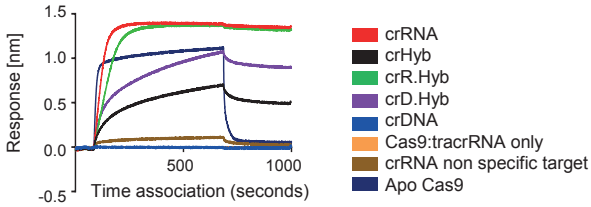


Figure 4.11: BLI DNA binding assay in the absence of heparin. In the absence of heparin, non-specific interaction of the Apo Cas9 complex is observed in the BLI assay. After the wash step rapid dissociation of the Apo-Cas9 is observed whereas the crRNA, crR.Hyb, crD.Hyb and crHyb complex dissociate slowly - confirming specific interaction of these complexes to the target DNA.

4.5 Supplementary Tables

Table 4.3: crXNA Sequences

Name	Sequence
	....5....10...15...20 ...25...30...35...40.
	|---Guide Region---| |-----cr Repeat------|
crRNA 1	**agcacugcacgccguaggucguuuuagagcuaugcuguuuug**
crDNA 1	AGCACTGCACGCCGTAGGTCGTTTTAGAGCTATGCTGTTTTG
crHyb 1	**a**GCACTGCACGCCG**ua**GG**u**C G**uuu**TAGAGCTATGCTGTTTTG
crRNA 2	**caaguucagcguguccggcgguuuuagagcuaugcuguuuug**
crDNA 2	CAAGTTCAGCGTGTCCGGCGGTTTTAGAGCTATGCTGTTTTG
crHyb 2	**c**AAGTTCAGCGTGT**cc**GG**c**G G**uuu**TAGAGCTATGCTGTTTTG
crRNA 3	**gagcuggacggcgacguaaaguuuuagagcuaugcuguuuug**
crDNA 3	GAGCTGGACGGCGACGTAAAGTTTTAGAGCTATGCTGTTTTG
crHyb 3	**g**AGCTGGACGGCGA**cg**TA**a**A G**uuu**TAGAGCTATGCTGTTTTG
crHybΔ5	**a**GCACTGCACGCCG**ua**GGTC GT**uu**TAGAGCTATGCTGTTTTG
crRNAC5A	**agca****a****ugcacgccguaggucguuuuagagcuaugcuguuuug**
crHybC5A	**a**GCA**A**TGCACGCCG**ua**GG**u**C G**uuu**TAGAGCTATGCTGTTTTG

RNA bases in lowercase **red**, DNA bases uppercase black

References

Cromwell, Christopher R et al. (2018). "Incorporation of bridged nucleic acids into CRISPR RNAs improves Cas9 endonuclease specificity". *Nature Communications* 9.1.

Kleinstiver, Benjamin P et al. (Jan. 2016). "High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects". *Nature* 529.7587, pp. 490–495.

Nishimasu, Hiroshi et al. (Feb. 2014). "Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA". *Cell* 156.5, pp. 935–949.

Rahdar, Meghdad et al. (Dec. 2015). "Synthetic CRISPR RNA-Cas9–guided genome editing in human cells". *Proceedings of the National Academy of Sciences* 112.51, E7110–E7117.

Ryan, Daniel E et al. (2018). "Improving CRISPR-Cas specificity with chemical modifications in single-guide RNAs". *Nucleic Acids Research* 46.2, pp. 792–803.

Slaymaker, Ian M. et al. (Jan. 2016). "Rationally engineered Cas9 nucleases with improved specificity". *Science* 351.6268, pp. 84–88.

Chapter 5

DNA stretching induces CRISPR/Cas9 off-target activity

5.1 Abstract

In the pursuit of improved CRISPR/Cas9 specificity and development higher fidelity variants, key mechanistic details must be understood in regard to of on- versus off-target site discrimination. To this end we have developed a single-molecule optical tweezers assay, combined with confocal microscopy, enabling real-time observation of single-molecule Cas9 on-target and off-target binding.

In this chapter I will present my work investigating the effect of mechanically stretching DNA on Cas9 specificity. On low force stretched DNA we observe highly specific on-target binding and cleavage with no off-target interactions. However, as the DNA is force stretched off-target binding is induced, with repeated binding observed at the same off-target locations. These induced off-target sites are guide sequence specific. Using a single-molecule FRET and bulk cleavage assays we further demonstrate that destabilisation of the DNA, through introduction of bubbles, induces binding and cleavage at these force induced off-target sites with as many as 10 mismatches, as well as at previously identified *in vivo* off-target sites. This work demonstrates that duplex destabilisation dramatically reduces Cas9 specificity. As such we propose that cellular process such as supercoiling, transcription, and replication could reveal such 'cryptic' off-target sites.

5.2 Results

5.2.1 Cas9 binds low force stretched DNA tightly and specifically

To monitor the interaction of *Streptococcus pyogenes* Cas9, herin Cas9, with DNA in real time and at single-molecule resolution, we combined optical tweezers with confocal fluorescence microscopy and microfluidics (See Sections 2.3 and 3.3). Cas9 complexes were assembled with a 5'-Cy3-labeled crRNA, unlabelled tracrRNA, and catalytically inactive dCas9 (Figure 5.1A). A previously characterised crRNA sequence was selected that targets Cas9 to a unique site on λ-DNA (λ2, at 18 kb, Sternberg et al., 2014). The λ2 site is located asymmetrically, one third of the way along the λ-DNA, enabling the orientation of the DNA to be determined if on-target binding is observed; this is important as the DNA can tether between the beads in either orientation. Experiments were performed using a microfludic flow cell (Figure 5.1A). First, we trapped two beads (4.5 μm streptavidin coated polystyrene beads) and caught a single piece of biotinylated λ-DNA, before moving to the buffer only channel where a force extension curve was recorded from 0-65 pN to confirm that only a single piece of intact DNA of the expected length was tethered between the beads. For these experiments the DNA is tortionally unconstrained, it is tethered by biotin-streptavidin interactions on only one strand at each end, resulting in the characteristic overstretching plateau at 65 pN (See Section 3.2). The DNA was then moved into the channel containing the labelled dCas9 complex and held at a force of 5 pN (Figure 5.1A). This force is sufficient to extend the DNA to its contour length (16 μm), and allow confocal imaging, but does not alter the base pairing of the two strands.

In the presence of labelled λ2-crRNA:tracrRNA:dCas9 complex, a single binding event was observed at the expected on-target location (Figure 5.1B), which remained bound for $>$ 30 minutes (n=5). This is consistent with previous observations that the Cas9 complex binds target DNA with very high affinity. No binding was observed with labelled crRNA alone, crRNA:tracrRNA or crRNA–dCas9 complex, confirming that these events represent stable binding of the holoenzyme complex (Supplementary Figure 5.12A-C). Furthermore, no binding was observed with a 5'-Cy3-crRNA lacking a target site (nt-crRNA) (Supplementary Figure 5.12,D). To further validate the on-target specificity, we added a second Cas9 complex labelled

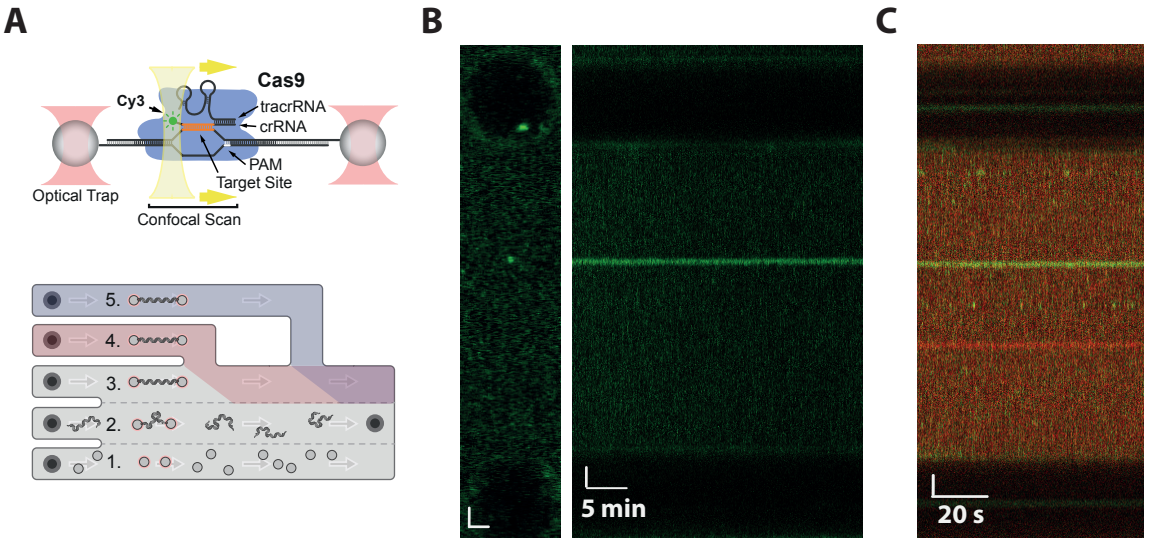


Figure 5.1: dCas9 binds DNA tightly and specifically at low force. (**A**) Top, diagram showing the experimental set-up. λ-DNA is held between optically trapped beads. Cas9 complex is labelled with Cy3 on the λ2-crRNA, designed to target a site at 18 kb. Bottom, layout of the set-up of the microfluidics: 1. Bead channel, 2. DNA channel, 3. Buffer only channel, 4. Cas9 channel. (**B**) Left, confocal 2D image of λ-DNA held between two beads at 5 pN with 10 nM Cy3-λ2:tracrRNA:dCas9 complex. A single molecule of Cas9 is seen bound to the expected target site (green). Right, kymograph generated from repeated 1D scan along the line of held DNA, Cas9 remains bound to the target site for $>$ 30 minutes, with no binding observed at any other sites. (**C**) λ-DNA held between two beads at 5 pN with 10 nM Cy3-λ2:tracrRNA:dCas9 and 10 nM Cy5-λ4:tracrRNA:dCas9. The λ2 complex (green) and λ4 complex (red) are seen bound to their respective target sites, no stable binding is observed at any other sites.

instead with Cy5, and targeting a distinct site (λ4 at 30.5 kb). In the presence of both dCas9 complexes, we observed clear specific binding to their respective targets (Figure 5.1,C). None of the crRNAs exhibited any off-target binding within the experimental time resolution (250 ms). These data confirm that dCas9 binds on-target sites on low force-stretched DNA with high affinity, and shows no off-target binding.

5.2.2 Specifically bound wtCas9 cleaves DNA and remains bound across the DSB

Next, we tested whether the Cas9 complex could cleave force-stretched λ-DNA. We assembled the complex with Cy3 labelled λ2-crRNA, tracrRNA, and catalytically competent wild type Cas9, wtCas9, in the presence of EDTA, to prevent cleavage. Fluorescent labelling did not affect wtCas9 activity in control bulk cleavage assays (Supplementary Figure 5.13). The wtCas9 complex was observed bound its specific on-target site, as expected, and catalysis was then activated by moving the DNA to a protein-free channel containing 10 mM Mg^{2+} (Figure 5.2A).

At low force (5 pN) cleavage was not observed within 10 s (n = 11). However, on gradually increasing the force from 5 to 55 pN, DNA rupture occurred at an average force of 40 ± 1 pN (mean ± s.e.m., n=11, Figure 5.2A and B). This suggests that Cas9 remained tightly bound across the DSB post-cleavage, consistent with previous observations of delayed recognition of Cas9 induced DSBs in cells (REFS). While cleavage may occur rapidly in the presence of Mg^{2+} after specific binding at low force, substantial mechanical force is required for disassembly of the post-cleavage complex. It has been suggested that on dissociation of the two ends of the DSB, Cas9 remains preferentially bound to the PAM-proximal side (REF). In a few instances, where photobleaching had not occurred before cleavage, flow-stretching the DNA revealed the tagged wtCas9 complex still bound to the cleaved strand, suggesting it does not fully dissociate even after the two ends of the DSB come apart (Figure 5.2C). These results demonstrate that the Cas9 complex is active under our experimental conditions, and confirms previous observations that the complex holds together cleaved DNA ends with high affinity, giving the first estimates of the mechanical strength of the post-cleavage complex.

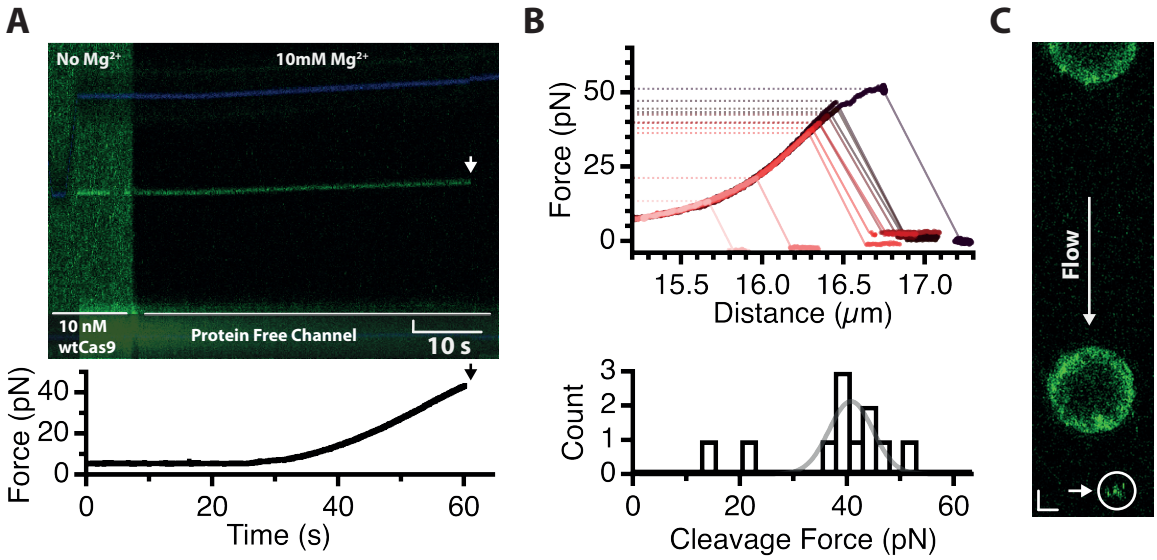


Figure 5.2: wtCas9 cleaves DNA in the presence of Mg^{2+}. (**A**) Kymograph, top, and corresponding force measurements, bottom, of wtCas9 bound to specific target in the absence and presence of Mg^{2+} followed by force ramp. Cleavage is observed at 60 s, at a force of 43 pN, as a loss of both the fluorescent signal and force between the beads. (**B**) Force-extension curves showing DNA cleavage by Cas9, top, and histogram of cleavage forces, bottom. n = 11. (**C**) Confocal 2D image of flow stretched DNA post-cleavage. wtCas9 can be seen bound to DNA attached to bottom bead.

5.2.3 DNA stretching induces off-target binding

Next, we investigated how increasing the tension on the DNA affects Cas9 binding. Using the Cy3-labeled λ2-crRNA:tracrRNA:dCas9 complex, we obtained an on-target binding event at low force, 5 pN, before increasing the force to 20 pN. At the higher force, we immediately started to see additional binding events at off-target locations (Figure 5.3A). Reducing the force to 5pN results in dissociation of all off-target-bound complexes from the DNA, with only the on-target Cas9 complex remaining bound. To determine the location of off-target binding events, we time-binned the kymographs, and mapped the inter-bead distance to the λ-DNA sequence (Figure 5.3B). The resulting intensity profiles at 20 pN show the on-target peak at the expected site, and six off-target peaks distributed over the DNA (Figure 5.3B, 20 pN). Increasing the force to 30, 40 and 50pN results in more off-target binding events across λ-DNA (Figure 5.3B).

To understand the force dependence of the off-target binding kinetics, we measured the dwell times for hundreds of λ2-crRNA:tracrRNA:dCas9 off-target binding events across λ-DNA held at forces from 20-50pN, and constructed dwell-time histograms (Figure 5.3C, top, and Supplementary Figure 5.14). The resulting cumulative probability distributions reveal a fast and a slow population of off-target dissociation events, except for 20pN, which only exhibits the fast population (Figure 5.3C, bottom, and Supplementary Figure 5.14). Interestingly as the force is increased, the lifetimes of the two states remain approximately constant (τ^{fast}=9.0 ± 0.6s and τ^{slow} = 61 ± 8 s, mean ± s.e.m., Figure 5.3D, top). However, at higher forces the relative amplitudes of the fast and slow states is altered, with an increased proportion of molecule exhibiting the slow dissociation lifetime (Figure 5.3D, bottom). This suggests that Cas9 binds off-targets in two binding modes, a transient (short-lived) and a meta-stable (longer-lived) mode, the latter becoming dominant at higher forces. Together, these data show that stretching the DNA induces frequent off-target Cas9 binding in a force-dependent fashion at previously unobserved sites.

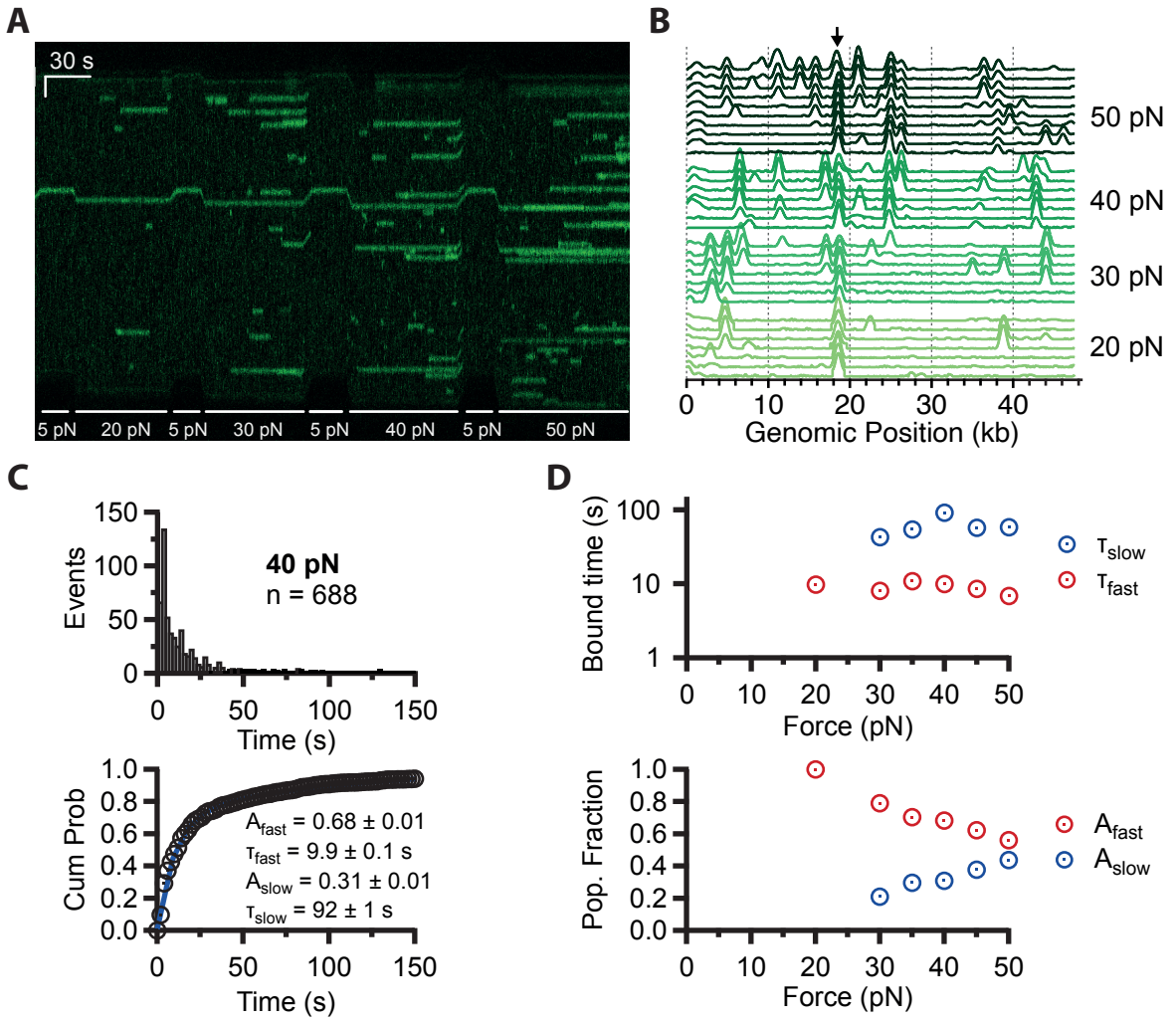


Figure 5.3: Force induces off-target binding of Cas9. (**A**) Kymograph of λ-DNA in the presence of 10 nM Cy3-λ2:tracrRNA:dCas9 complex (green) held at forces between 5 and 50 pN. At forces $>$ 20 pN off-target binding events are observed. (**B**) Time-binned intensity plots of Cas9 binding at forces $>$ 20 pN, mapped to λ-DNA, showing genomic positions of force induced off-target binding. (**C**) Dwell time histogram, top, and cumulative probability distribution, bottom, for off-target binding events of λ2-crRNA induced at 40 pN. (**D**) Off-target binding lifetimes, top, and amplitudes, bottom, as a function of force.

5.2.4 Off-target binding occurs at non-random, crRNA sequence-dependent sites

Interestingly, the same off-target locations appear to be bound multiple times, suggesting an underlying sequence dependence of the induced sites (Figure 5.3). To further test this observation, a single DNA molecule was pulled three times and held at 40 pN. In the resulting kymograph and time-binned intensity plot (Figure 5.15A) we observed repeated binding to the same genomic locations. We first considered whether the local density of PAM sites, and/or GC content, of the underlying DNA could explain the distribution of induced off-target sites as

had been previously suggested (Sternberg et al., 2014). However, locations of induced off-target binding appear to be highly specific, being observed at discrete peaks spread across the length of the DNA, and do not appear to directly correlate with PAM site density or GC content (Supplementary Figure 5.15A). We next asked whether the binding locations of these off-target sites are guide sequence dependant. Using two additional guide sequences, λ4-crRNA and nt-crRNA, off-target binding events were induced on stretching λ-DNA to 40 pN (Figure 5.4C). On comparison of the locations of induced off-target events for the three crRNA guide sequences, λ2, λ4, and nt, we saw strong correlations between repeat experiments with the same guide but no correlation in locations between the different guides (Supplementary Figure 5.15B-E). The overall binding distributions show unique sites of induced off-target binding for each guide (Figure 5.4D). This demonstrates that these force induced off-target events are non-random and crRNA sequence-dependent, suggesting they are in part mediated by base pairing between the DNA and crRNA.

5.2.5 Two DNA strands are required for induced off-target binding

One possible explanation for force-induced binding is the formation of nicks or short single-stranded regions of DNA during stretching. To investigate this we performed competition assays with GFP-labelled ssDNA-binding human replication protein A, hRPA. DNA was stretched from 5 to 65 pN in the presence of GFP-hRPA and Cy3-nt-crRNA:tracrRNA:Cas9 (Figure 5.5A). At forces below 60 pN, RPA was only observed binding to a small number of nicks in the DNA, and does not compete with the observed binding of Cas9 (Figure 5.5A). At forces $>$ 65 pN, large regions of ssDNA begin to propogate from nicks, with Cas9 apparently excluded from binding to these regions (Figure 5.5A).

The inability for Cas9 to stably bind at ssDNA regions was further demonstrated using a high concentration of nt-crRNA:tracrRNA:Cas9 in the absence of GFP-hRPA. As the DNA was stretched from 5 to 60 pN, Cas9 binding is induced across the DNA (Figure 5.5B). However at forces $>$ 60 pN Cas9 binding decreases, and as large regions of ssDNA form Cas9 is again excluded from these regions (Figure 5.5B). After extensive overstretching of the DNA, some

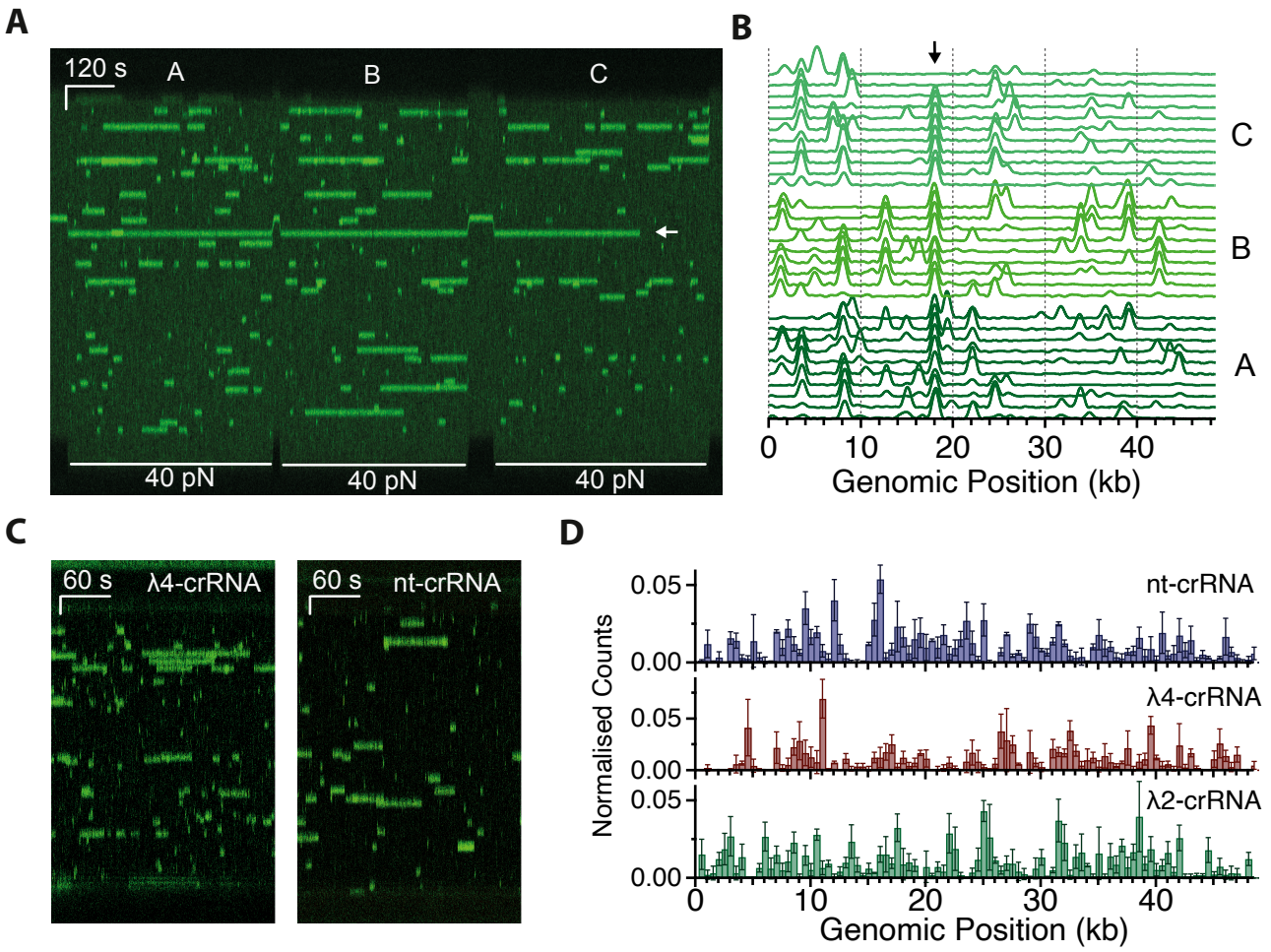


Figure 5.4: Off-target binding locations are non-random and sequence dependant. (**A**) Kymograph of λ-DNA in the presence of 10 nM Cy3-λ2-crRNA:tracrRNA:dCas9 complex (green) pulled three times (A-C) to 40 pN and held for approximately 5 minutes before relaxing back to 5 pN. (**B**) Time-binned intensity plots of Cas9 binding during each of the 40 pN pulls (A-C), mapped to λ-DNA, showing genomic positions of force induced off-target binding. Arrow, λ-2 on-target at 18.2 kb. (**C**) Representative kymographs of λ-DNA held at 40 pN in the presence of either 10 nM Cy3-λ4-crRNA:tracrRNA:dCas9 complex, left, or Cy3-nt-crRNA:tracrRNA:dCas9, right. (**D**) Count normalised histograms of location of off-target binding events on λ-DNA for the three guides tested (λ2, n = 1,234; λ4, n = 797; and nt (no target), n = 1,950). Bin size = 500 nucleotides, error bars, + s.d.

regions remain single-stranded even after reducing the force to 5 pN, however Cas9 does not bind to ssDNA even under low tension. This suggests that induced binding is not caused by the presence of nicks or large regions of ssDNA, and that both DNA strands are required. It has been observed previously that although Cas9 can bind to a target with single-stranded target sequence, a double-stranded PAM site is still required (REF).

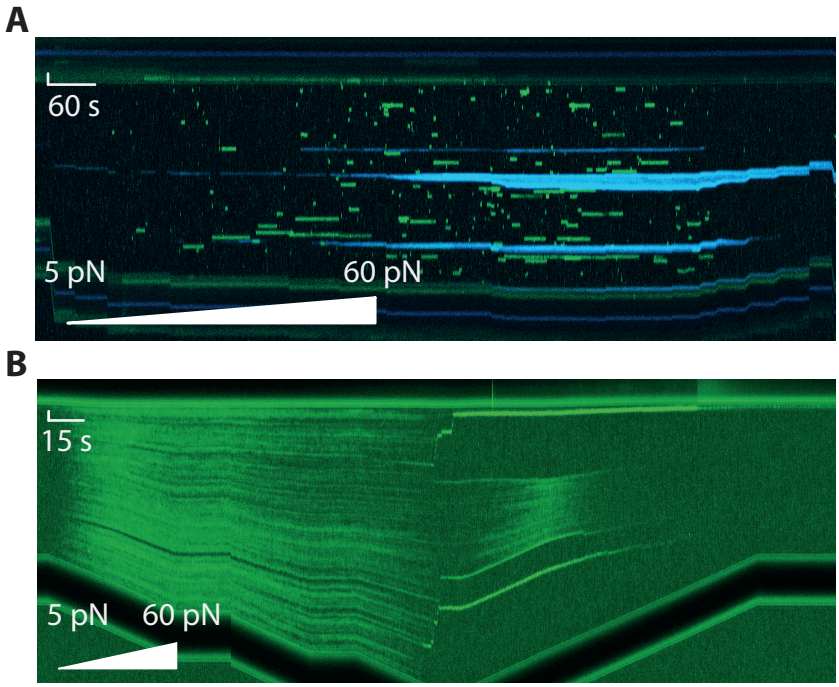


Figure 5.5: Off-target binding requires both DNA strands. (**A**) Kymograph of λ-DNA in the presence of 5 nM hGFP (blue) and 10 nM Cy3-nt-crRNA:tracrRNA:dCas9 complex (green) pulled from 5 - 65 pN. RPA can be seen first binding to nicks, and then to ssDNA which propogates from nicks, as the DNA is overstretched. Cas9 binding does not colocalise with RPA. (**B**) Kymograph of λ-DNA in the presence of 100 nM Cy3-nt-crRNA:tracrRNA:dCas9 complex (green) pulled from 5 - 60 pN and then significantly overstretched to generate large regions of ssDNA. ssDNA regions at high force, and those that remain as the force is reduced, occlude Cas9 binding.

5.2.6 Off-target sequence identification

Based on our observations that the locations of off-target sites are crRNA sequence dependant, we next asked if Cas9 is binding to sites with sequences similar to the on-target. We developed a simple model for scoring all potential binding sites across the phage-λgenome based on the number of matching bases, presence of a PAM site, and taking into account thermodynamic considerations such as wobble base pairing, nearest neighbour effects and weighting of matches in the previously described 'seed region'. First the entire λ-genome was scanned for potential sites adjacent to a canonical PAM sequence (NGG) or the alternative PAM sequence (NAG). Within these selected sites each base was compared to the on-target sequence and assigned a score, a perfect match was given +8, a mismatch 0, and a mismatch resulting in a wobble base pair given +4. Next the nearest neighbours were considered with a bonus of +2 given for each base-paired neighbour and +1 given for wobble paired neighbours (Figure 5.6A). To take into account the previous observations of increased importance of matching bases in the seed region,

here defined as the first 8 nucleotides next to the PAM site, scores for this region were doubled. Finally, sites containing the canonical PAM as opposed to the alternative PAM were given an additional bonus of +20. This gives every potential site a score between 0 and 350, where 0 is an entirely mismatched site with the alternative PAM sequence, and 350 is the on-target site (Figure 5.6A).

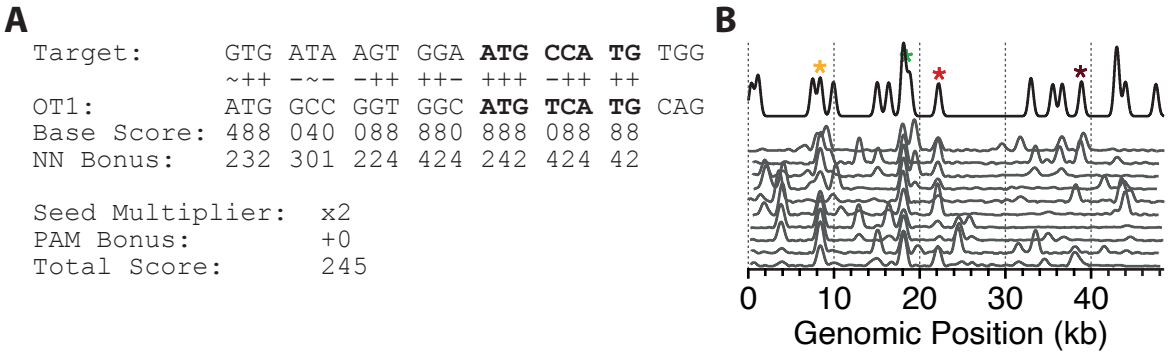


Figure 5.6: Scoring potential off-target sites and comparison with experimental data. (**A**) Scoring example for OT1 site. +, base-paired, +8; -, no base-pair, 0; ∼, resulting in wobble base-pair, +4. Nearest neighbour (NN) bonus, (**B**) Predicted profile of off-target binding peaks for λ2-crRNA target, top, aligned with experimental data of off-target binding for λ2-crRNA at 40 pN.

As this scoring system is largely arbitrary is was devised only as a means of ranking all the potential sites to identify a list of the most likely sites to be bound, which could then be further tested. For the λ2 guide sequence, taking only the sequences with scores great than or equal to 225, 18 sites were identified, predicted to be the most likely to be bound by Cas9 (Table 5.1). These sites were compared to binding data for the λ2-crRNA Cas9 complex binding to DNA stretched at 40 pN (Figure 5.6B). Of the 17 off-target sites predicted, over 50% had apparent corresponding binding peaks in the experimental data. Three of these sites were selected for further analysis (Table 5.1).

In order to validate these sites, we used a previously described single-molecule FRET assay to monitor binding and R-loop formation (Figure 5.7A, Rueda et al., 2017). The on-target and three selected off-target sequences where labelled with Cy3 and immobilised on the surface. On addition of the Cy5-λ2-crRNA:tracrRNA:dCas9 complex, binding results in a FRET signal between the Cy3 (donor) and Cy5 (acceptor). In this assay, full R-loop formation on binding to

Table 5.1: Predicted Off-target Sequences

Sequence (5' - 3')	Mismatches	Score	Position (nt)	Strand	PAM
TCATAAAATGGTATGCCGAA	10	226	340	+	NGG
GTGCTTTTTGCCATACCACG	8	252	1143	-	NGG
CTGATGCACTTCACGCCACG	10	242	7492	+	NGG
ATGGCCGGTGGCATGTCATG	7	245	8378	+	NAG
CTGCGGATGGTTATGCCACA	10	232	9933	-	NGG
ATATGAATTACAGCGCCATC	10	231	15054	-	NGG
CCCATCCGCGCTACGGCATG	10	230	16359	+	NGG
GTGATAAGTGGAATGCCATG	0	350	18084	+	NGG
AATATCTGCCGAATGCCGTG	8	255	18841	-	NGG
GCATCTCGTTGAAGACCATC	10	228	22180	+	NGG
CGATCATTTGCAACGCCACG	10	244	32979	-	NGG
ATTCAAAGCAGAAGGCTTTG	9	225	35528	-	NGG
GCGATATATTAAAGGTCTTA	9	226	36628	-	NGG
GTTTTACGCAGAATGGCAAG	7	234	38877	-	NAG
CTAATCAGTTCGACGCGATA	9	238	42914	-	NGG
TTTTTACTGAGAACGTCATG	9	238	43071	-	NGG
CTGATGTGATGGCTGCTATG	8	232	43993	+	NGG
ATCATCAACAAAACACAAGG	11	225	47560	-	NGG

on-target sites results in stable high FRET state (Figure 5.7B).

Although the overstretching experiments demonstrate that Cas9 is not binding to fully single-stranded DNA, we hypothesised that the reduction in energy required to melt DNA at high tension could induce off-target binding. To mimic the high force stretched DNA from the optical tweezers experiments we created a pre-melted bubble substrate, whereby the non-target strand is non-complementary to the target strand. The non-target strand sequences were substituted with polythymine, or with cystosine where thymine would result in a basepair.

We first tested the fully hybridised on-target DNA construct, nearly half of the molecules reached the high FRET state (46%, n = 2510, FRET = 0.95 ± 0.02, mean ± s.d.) demonstrating stable binding and full R-loop formation (Figure 5.8A). We next tested the full-bubble on-target construct, similar levels of binding were observed (53 %, n = 2986), reaching the same high FRET state (FRET = 0.96 ± 0.02), as with the fully hybridised construct (Figure 5.8A). This is consistent with the lack of sequence specific interactions between the non-target strand and Cas9. We next tested the first off-target construct (Off-target 1, OT1; Figure 5.8A and

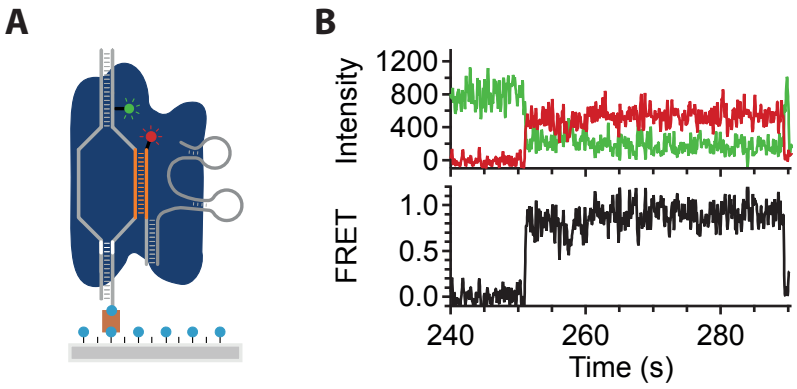


Figure 5.7: Design of smFRET Cas9 assay. (**A**) Design of smFRET Cas9 binding and R-loop formation experiments. Target DNA, labelled with Cy3 (green) on the target strand, is immobilised on the microscope slide. Binding and R-loop formation of dCas9 complex labelled with Cy5 (red) results in FRET. Target site (orange) and PAM site (white). (**B**) Example single-molecule trajectory for Cy5-λ2-crRNA dCas9 complex binding to the Cy3 labelled λ2 on-target site, top. Corresponding calculated FRET trajectory. On-target binding results in high-FRET state.

Table 5.1). Consistent with the lack of off-target binding to low force stretched DNA, no binding to OT1 was detected with the fully hybridised construct (Figure 5.8A). However, the corresponding pre-melted full-bubble OT1 construct restores a high levels of binding (68 %, n = 893) in a high FRET state comparable to the on-target site (0.93 ± 0.2) (Figure 5.8A). This supports the hypothesis that a reduction in the target DNA melting energy can induce Cas9 binding to previously unbound off-target sites. We further test two additional off-target sequences (Off-target 2, OT2; and Off-target 3, OT3). Again, no binding was observed with the double-stranded constructs. However, the pre-melted bubble construct restores binding (OT2 = 78 %, n = 1,069; OT3 = 33 %, n = 2,062) (Figure 5.8A). No binding to the full-bubble constructs was observed with Cy5-crRNA:tracrRNA in the absence of Cas9, demonstrating that this binding is protein mediated. Interestingly we observed conformational differences in the bound state of the Cas9 complex between the three off-targets. While a high-FRET state was observed for OT1 (0.93 ± 0.02), comparable to on-target binding, OT3 reached a lower FRET state (0.85 ± 0.2) and OT2 appeared to be in dynamic equilibrium between two states (0.5 ± 0.2 and 0.84 ± 0.02) (Figure 5.8A). These likely result from incomplete or unstable R-loop formation, and could be due to differences in the number and position of mismatches within the target. Interestingly, for OT1 and OT2 a small 6 nucleotide PAM proximal bubble was sufficient to allow Cas9 binding with a low FRET state, as predicted due to incomplete R-loop formation

(OT1, FRET = 0.23 ± 0.02, 52%, n = 2466; and OT2, FRET = 0.27 ± 0.02, 75%, n = 3704) (Figure 5.8B). No binding was observed for OT3 (n = 626) (Figure 5.8B). Taken together this data validates multiple sites predicted by the scoring model, and demonstrates that pre-melting of the target can induce binding and previously unobserved off-target sequences.

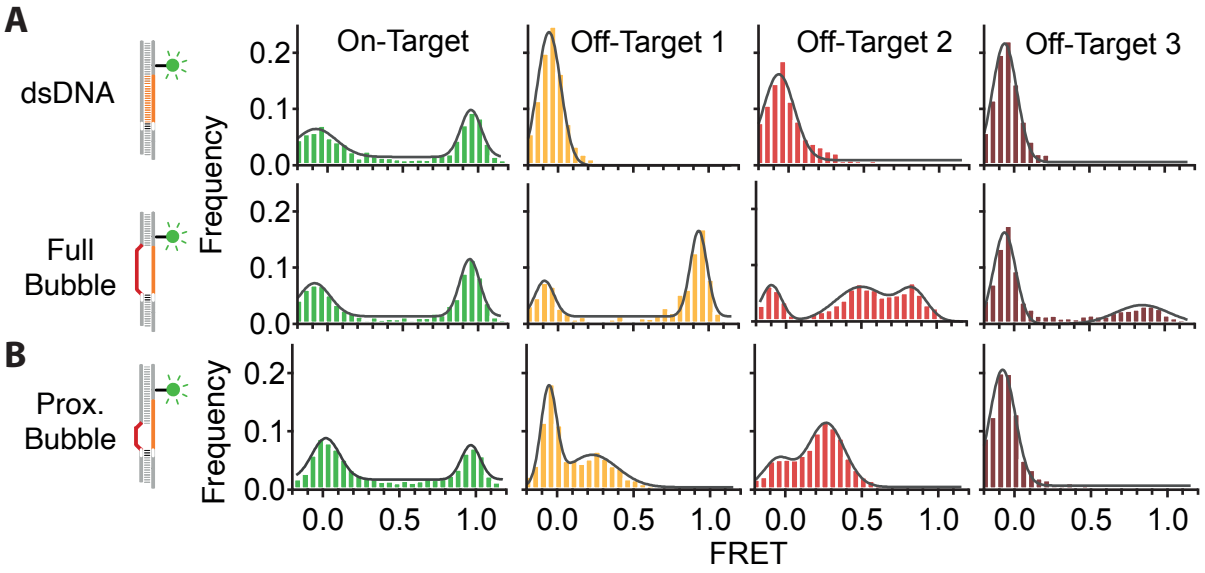


Figure 5.8: DNA bubbles induce binding at predicted off-target sites. (**A**) FRET histograms of Cas9 binding to dsDNA, top, and bubble constructs, bottom, of on- and off-target sequences. For dsDNA, binding is only observed for the on-target sequence in a high FRET state. For full-bubble constructs, Cas9 binds to the on- and off-target sequences, reaching high-FRET states. (**B**) FRET histograms of Cas9 binding to 6 nt PAM proximal bubble constructs of on- and off-target sequences. Cas9 binds OT1 and OT2 in low FRET state but does not reach the high FRET state observed for on-target binding. No binding is observed for OT3. Peaks fit with gaussian distributions (grey)

5.2.7 DNA bubbles induce Cas9 off-target cleavage

We next tested whether these identified off-target sites could be cleaved by Cas9. The labelled dsDNA and full-bubble constructs were incubated with Cy5-crRNA:tracrRNA:wtCas9 for 1 hour at 37 °C before visualisation by denaturing PAGE and fluorescence detection. As expected, the on-target sequence was efficiently cleaved (Figure 5.9A). For the three off-target sequences, Cas9 was unable to cleave the fully-hybridised constructs, however cleavage was observed for all three of the full bubble constructs (OT1-OT3) (Figure 5.9A). Importantly no cleavage was observed in the absence of Mg^{2+}, or when an alternative control guide sequence, λ-4, was used (Figure 5.9A). The cleavage efficiency of the three off-target full-bubble constructs ranged from 18-57% (Figure 5.9B). Surprisingly the highest cleavage efficiency (57 %) was observed for OT2

despite it containing the most mismatches (10/20 mismatched bases) and despite the instability of the high FRET state (Figure 5.8). This suggests that even transient formation of the full R-loop is sufficient for Cas9 cleavage and that, where DNA is pre-melted, Cas9 is significantly more promiscuous than previously reported, capable of binding and cleaving sites with as many as 10 mismatches.

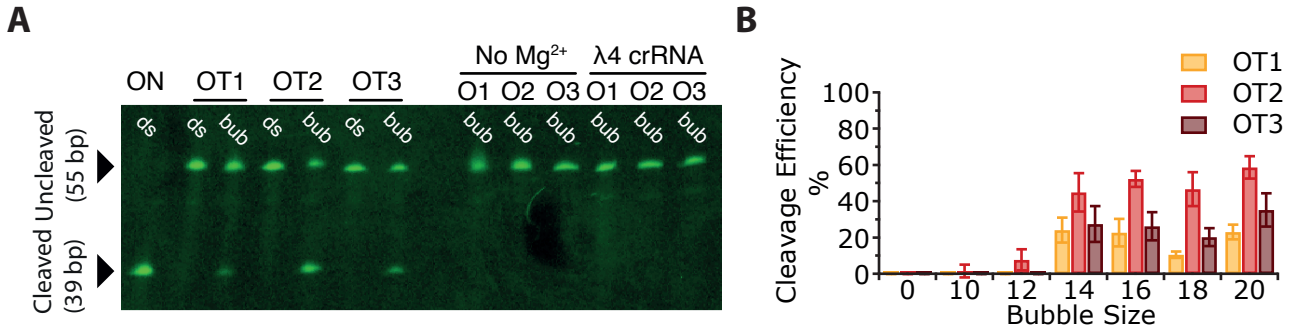


Figure 5.9: DNA bubbles induce binding at predicted off-target sites. (**A**) Cleavage assay for double-stranded (ds) and bubbled (bub) target DNA, visualised by denaturing PAGE and fluorescent imaging. Cas9 only cleaves bubbled DNA off-targets. No cleavage of these sites is observed in the absence of Mg^{2+} or using the alternative λ4-crRNA. (**B**) Percentage cleavage efficiencies of constructs with different bubble sizes for the three off-target sequences (OT1-OT3).

We next explored the extent of target site melting required to induce Cas9 activity at these sites. We measured off-target cleavage efficiencies for PAM proximal bubbles ranging from 0 - 20 nucleotides. For the three off-target sites tested, a bubble of $\geq$ 14 nt was required for efficient cleavage (Figure 5.9B). Increasing the bubble size beyond 14 nucleotides did not further increase cleavage efficiency (Figure 5.9B). Although our single-molecule FRET experiments demonstrate that Cas9 can bind OT1 and OT2 with only a 6 nt PAM proximal bubble (Figure 5.8), no cleavage was detected with these constructs (Supplementary Figure 5.16). This is consistent with the low FRET state observed representing incomplete R-loop formation, and a requirement for at least transient formation of the full R-loop for cleavage. Furthermore, no cleavage was observed for a 6 nucleotide bubble in the middle of the target or at the PAM distal end for any of the off-targets (Supplementary Figure 5.16).

To further investigate the role of the non-target strand in Cas9 off-target activity we next tested single-stranded DNA targets. No stable binding was observed to the fully single-stranded off-target sites, however binding was partially restored in the presence of a double-stranded

PAM site (Figure 5.10). For the on-target, binding and cleavage was observed for both the fully single-stranded target and double-stranded PAM site target (Figure 5.10). For the off-target sites a low level of cleavage was observed for the double-stranded PAM, however cleavage was less efficient than with the full-bubble construct. This demonstrates that the non-target strand has a role in binding and cleavage of Cas9 to these off-target sites.

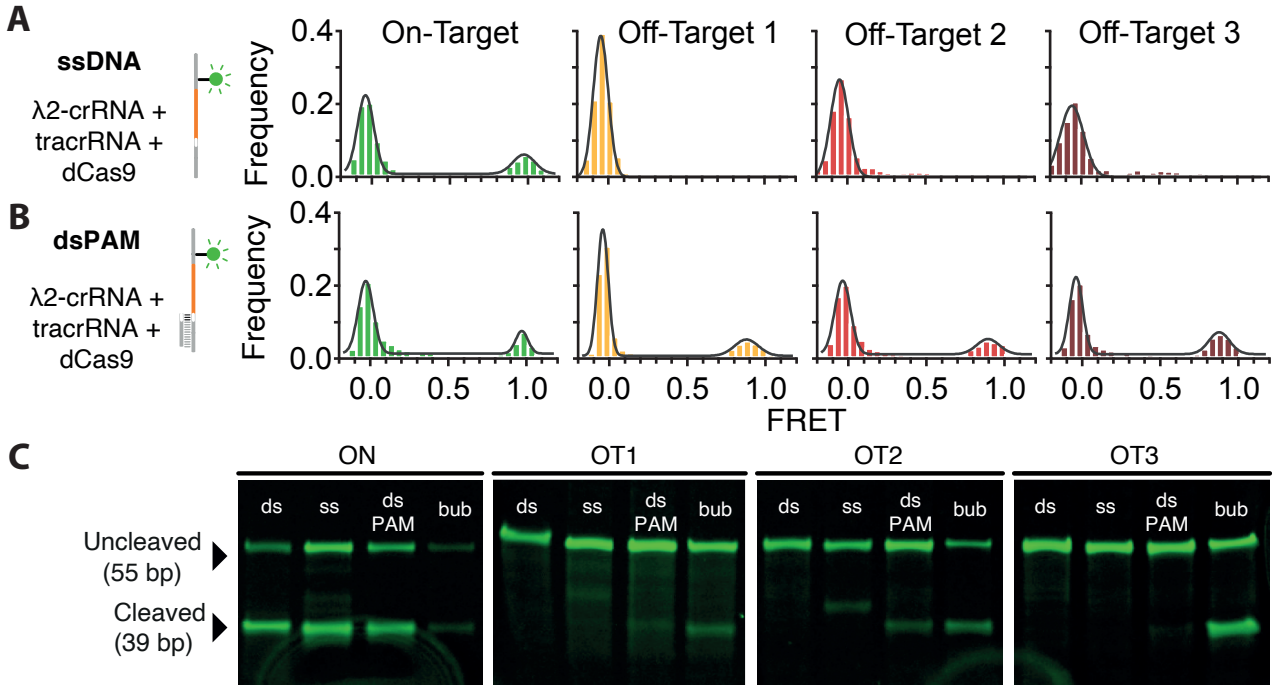


Figure 5.10: DNA bubbles induce binding at predicted off-target sites. (**A**) FRET histograms of λ2 Cas9 complex binding to ssDNA on- and off-targets. The high FRET state is only observed on the on-target site. (**B**) FRET histograms of λ2 Cas9 complex binding to ssDNA constructs with double-stranded PAM sequences. A small population of high FRET Cas9 binding state is observed for both the on- and off-target sequences. (**C**) Cleavage assays for double-stranded (ds), single-stranded (ss), double-stranded PAM (dsPAM) and full bubble (bub) target DNA for the on-target and off-target constructs. Visualised by denaturing PAGE and fluorescent imaging. All on-target constructs are cut with high efficiency, for the off-targets the highest cleavage efficiency is seen with the full bubble construct.

5.2.8 In vivo off-target sites require melting for cleavage with Cas9

Based on these findings we hypothesised that *in vivo* cellular processes that destabilise the DNA structure could reveal these 'cryptic' off-target sites. To test this we selected four previously identified off-target sites of the EMX1-1 guide which had been detected by CIRCLE-seq and validated *in vivo* (Table 5.2). In bulk assays no cleavage was detected after incubation with an 100 times molar excess of EMX1-crRNA:tracrRNA:wtCas9 for 1 hour at 37 °C (Figure 5.11).

However, pre-melting the target by introduction of a full bubble resulted in cleavage of all four off-targets with efficiencies of upto 60 % (Figure 5.11). This provides evidence that *in vivo* off-target editing of previously identified site requires destabilisation of the target DNA.

Table 5.2: Previously validated *in vivo* EMX1-1 off-target sites

Name	Genomic Location (chromosome:nt)	Sequence (5' - 3')	PAM
ON	2:73160981-7316100	CCC-TTCTTCTTCTGCTCGGACTC	GGG
OT11:4	11:443747931-43747954	CCT-TTCTTCCTTTGCTCGGGCTT	AGG
OT3:63	3:63468110-63468133	CCC-TTCTTCTCCTGCTCCAACTT	GGG
OT2:21	2:218378101-218378124	CCT-TTATTCTCCTGCTTAGACTC	AGG
OT7:11	7:111475442-111475465	CCA-TTCTTTTTGTGCTTGGGCTC	TGG

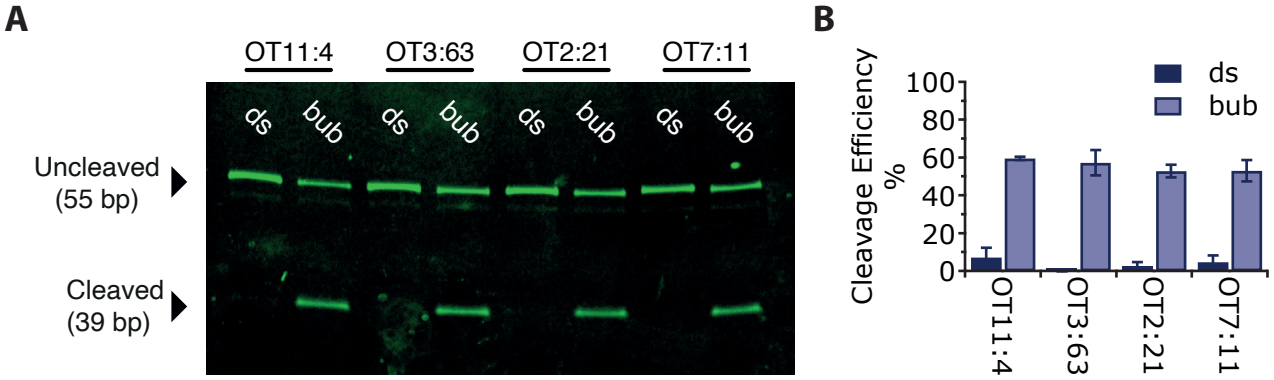


Figure 5.11: Previously validated EMX1-1 in vivo off-target sites require DNA destabilisation for Cas9 cleavage. (**A**) Cleavage assay for double-stranded (ds) and full bubble (bub) EMX1 off-target sites. Visualised by denaturing PAGE and fluorescence imaging. Cas9 only cleaves the full bubble constructs. (**B**) Quantification of cleavage efficiencies. All four off-target sites are cut with high efficiency when the DNA contains the full bubble.

5.3 Discussion

The ease with which CRISPR/Cas9 can be programmed to target a specific sequence has resulted in its wide scale adoption for gene editing and to regulate gene expression. Successful use in cells and *in vivo* paves the way for potential therapeutic application with great promise for treatment of previously untreatable genetic diseases. However, reports of unanticipated off-target activity of the Cas9 complex pose a significant barrier, especially for any clinical applications. As a result it has become vital to gain a mechanistic understanding of factors that effect Cas9 on- and off-target binding and cleavage in order to improve off-target site prediction and increase specificity. Here we have developed a single-molecule assay, combining optical tweezers with fluorescence confocal microscopy, to visualise on- and off-target interactions of Cas9 complexes with a single piece of DNA. When DNA is held at low force, Cas9 complexes bind tightly and specifically to the expected on-target site with no detectable off-target activity (Figure 5.1. However, as the DNA is stretched off-target binding events are induced in a highly force dependant manner, with higher forces resulting in an increase in the number of long-lived off-target binding events (Figure 5.3). The locations of induced off-target binding are guide sequence dependant, with different guides resulting in unique distributions of off-target events (Figure **??**. Using a simple scoring model we identify off-target sequences being bound (Figure 5.6), and test them using single-molecule FRET assays and bulk cleavage assays (Figures 5.8 and 5.9).

We have demonstrated that destabilisation of the DNA duplex dramatically reduces Cas9 specificity, with binding and cleavage observed even in the presence of 10 mismatches (Figure 5.9). Based on these findings, we hypothesise that cellular processes which disrupt the DNA duplex, such as he presence of negative supercoiling, DNA replication or transcription, could expose 'cryptic' off-target sites. To test this we selected four previously identified off-target sites which have been previously validated *in vivo* (Figure 5.11). Strikingly, no cleavage of these site was detected *in vitro* when the target DNA was double-stranded, even in the presence of a large molar excess of Cas9 complex (Figure 5.11). However, when these sites were destabilised with the presence of a bubble, efficient target cleavage was observed (Figure 5.11). This provides strong

evidence for the requirement of target duplex destabilisation for cleavage of some off-targets *in vivo*.

It is currently not well known what mechanical forces DNA is exposed to inside the nucleus of a cell. As such it is unclear whether the DNA regularly experiences the magnitude of force shown here to induce off-target binding of Cas9. However, it has been demonstrated that during anaphase mechanical breaks in the DNA can be induced (Jannink et al., 1996). This would typically require forces in excess of 100 pN, suggesting that the forces used in this work are in a physiologically relevant range. Furthermore, the use of DNA bubbles to induce off-target activity suggests that it is the reduction in energy required for DNA melting, rather than an specific requirement for DNA to be under tension, that leads to the loss of Cas9 specificity. This is consistent with previous work which has demonstrated that DNA duplex melting and R-loop formation are the rate-limiting steps in Cas9 binding and cleavage (Klein et al., 2018; Gong et al., 2018). As such, lowering of the energy barrier for DNA melting enables stable R-loop formation even when multiple mismatches are present between the guide and target. This, combined with the data obtained with known *in vivo* off-target sites, indicates that DNA destabilisation by other cellular process could be an important consideration when thinking about Cas9 specificity.

Although our simple scoring model enabled the identification of the sequences of some off-target sites, it failed to predict many of the observed force induced off-target locations. It was also a biased approach, based on existing Cas9 off-target rules. In order to generate an unbiased picture of Cas9 sequence specificity on force stretched DNA existing high-throughput off-target analysis protocol could be adapted with DNA destabilisation achieved through the incorporation of DNA bubbles into target sequence libraries or the addition of DNA destabilising agents such as intercalators. This would not only give us a better understanding of the mechanism of this force induced loss of specificity but would aid the development of algorithms to predict potential off-target sites.

It was particularly striking to observe high efficiency binding and cleavage of off-target sites with as many as 10 mismatches between the crRNA and DNA. Under such conditions Cas9

cleavage could hardly be described to be specific if 50 % of bases in potential target can be mismatched and cleavage still observed. When considering the number of potential off-targets in the human genome, allowing up to 10 mismatches would result in tens of thousands of possible sites. Even if only a small fraction of these site are active due to the influence of cellular processes this could have dangerous consequences when considering gene editing on the level of whole organisms or specific tissues. Furthermore, active off-target sites could be more likely to fall within transcriptionally active regions of the genome, increasing the probability of negative consequences of off-target editing.

Taken together the work presented in this chapter demonstrates that DNA tension is a previously unconsidered factor influencing Cas9 specificity. This could have potential implications for the development of safe gene editing therapeutics using CRISPR/Cas9.

5.4 Supplementary Figures

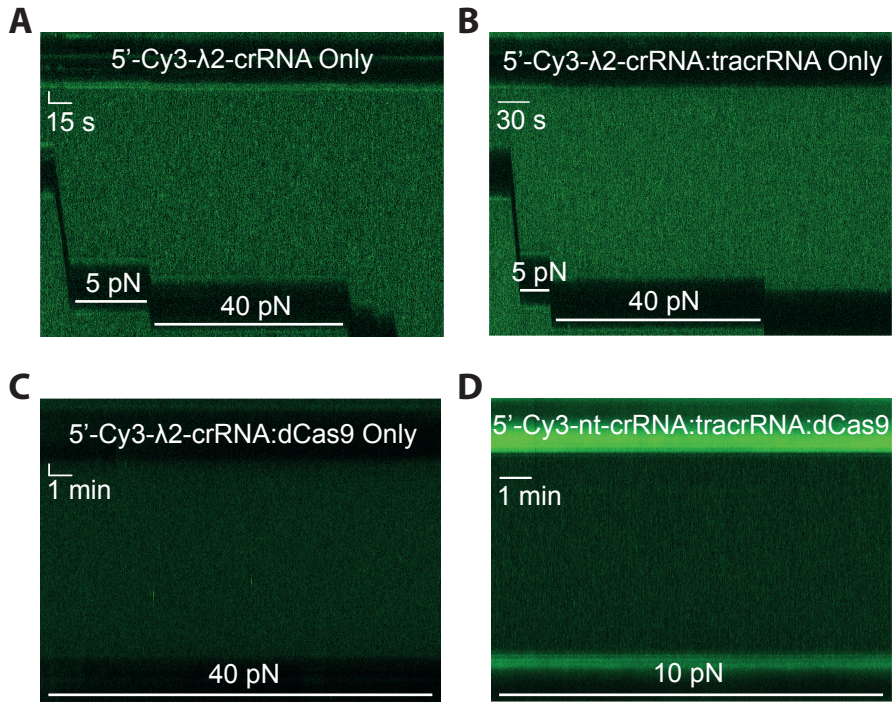


Figure 5.12: Cas9 complex binding requires crRNA, tracrRNA and Cas9. (**A**) Kymograph of λ-DNA held at forces of 5 pN and 40 pN in the presence of 10 nM Cy3-λ2-crRNA only, with no tracrRNA or Cas9 protein. Cas9 does not bind at on-target or off-target sites at low or high force. (**B**) Kymograph of λ-DNA held at forces of 5 pN and 40 pN in the presence of 10 nM Cy3-λ2-crRNA and 10 nM tracrRNA, with no Cas9 protein. Cas9 does not bind at on-target or off-target sites at low or high force. (**C**) Kymograph of λ-DNA held at 40 pN in the presence of 10 nM Cy3-λ2-crRNA and 10 nM Cas9 protein, without the tracrRNA. Cas9 does not bind at on-target or off-target sites. (**D**) Kymograph of λ-DNA held at 10 pN in the presence of 10 nM Cy3-nt-crRNA:tracrRNA:dCas9, the no-target (nt) crRNA guide has no target site on λ-DNA. This Cas9 complex does not bind λ-DNA at low force.

References

Gong, Shanzhong et al. (Jan. 2018). "DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity". *Cell Reports* 22.2, pp. 359–371.

Jannink, Gerard, Bertrand Duplantier, and Jean-Louis Sikorav (1996). *Forces on Chromosomal DNA during Anaphase*. Tech. rep., pp. 451–465.

Klein, Misha et al. (Feb. 2018). "Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules". *Cell Reports* 22.6, pp. 1413–1423.

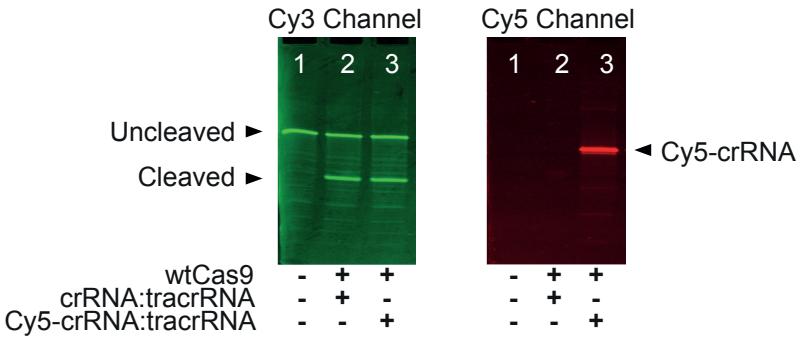


Figure 5.13: Labelling of crRNA with Cy5 does not effect Cas9 cleavage efficiency. Cleavage of Cy3 labelled $\lambda 2$ target DNA with unlabelled or 5'-Cy5 labelled crRNA. Cleavage products, and labelled crRNA, visualised by denaturing PAGE and fluorescence imaging of Cy3, left, and Cy5, right. No cleavage of target DNA is observed with Cas9 alone, in the presence of $\lambda 2$-crRNA:tracrRNA:Cas9 efficient cleavage is observed, cleavage is unaffected by 5'-Cy5 crRNA labelling.

Rueda, Fernando Orden et al. (Dec. 2017). "Mapping the sugar dependency for rational generation of a DNA-RNA hybrid-guided Cas9 endonuclease". *Nature Communications* 8.1, p. 1610.

Sternberg, Samuel H. et al. (Mar. 2014). "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9". *Nature* 507.7490, pp. 62–67.

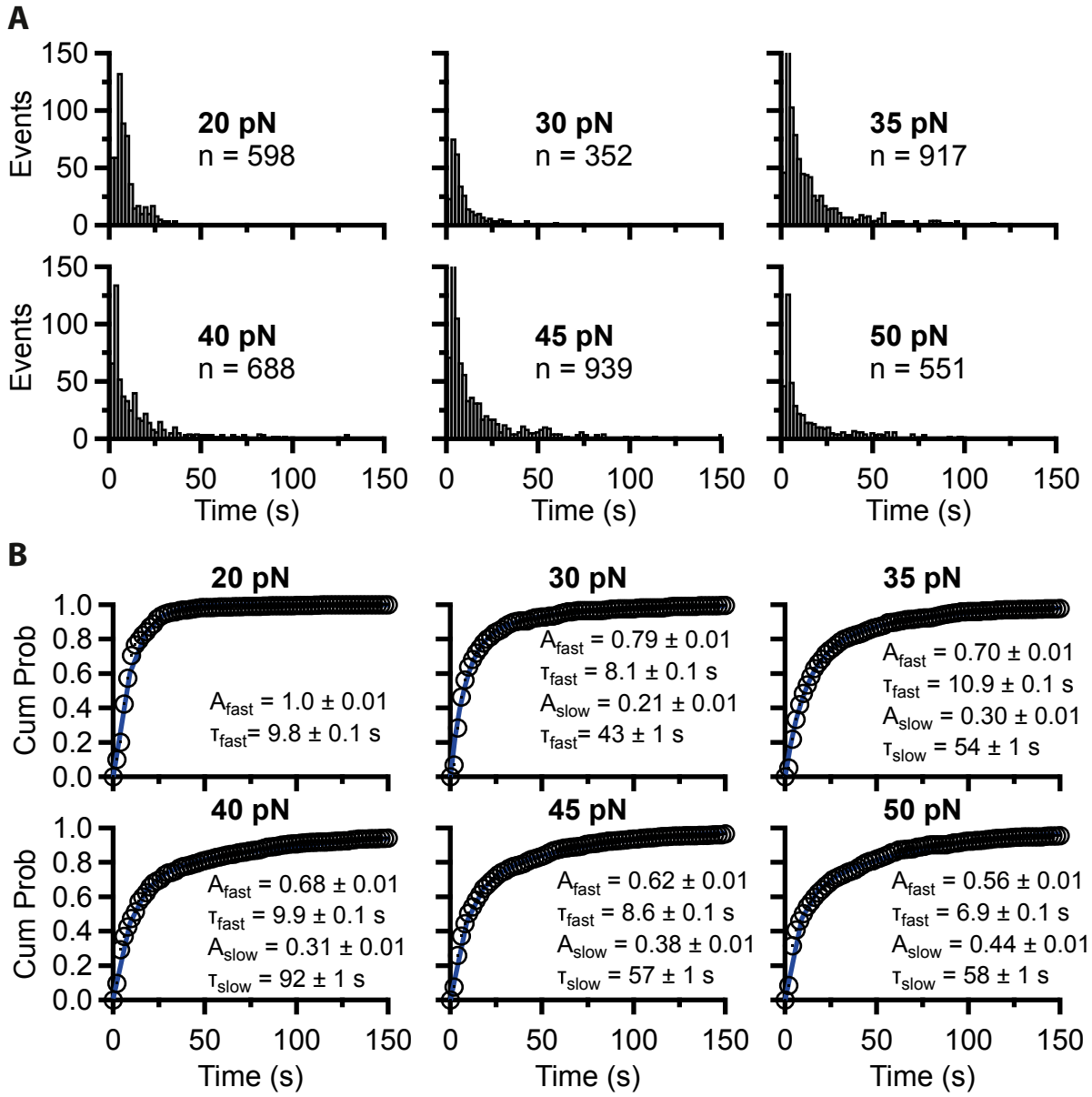


Figure 5.14: Force dependant dwell-time analysis of off-target binding. (**A**) Dwell time histograms of force induced λ2-crRNA off-target binding events for forces from 20 - 50 pN. (**B**) Cumulative probability distributions of dwell times for force induced λ2-crRNA off-target binding events from 20 - 50 pN. Fit with double exponential fits to obtain time constants, τ, and amplitudes, A, for the two populations. Data at 20 pN fit with single exponential.

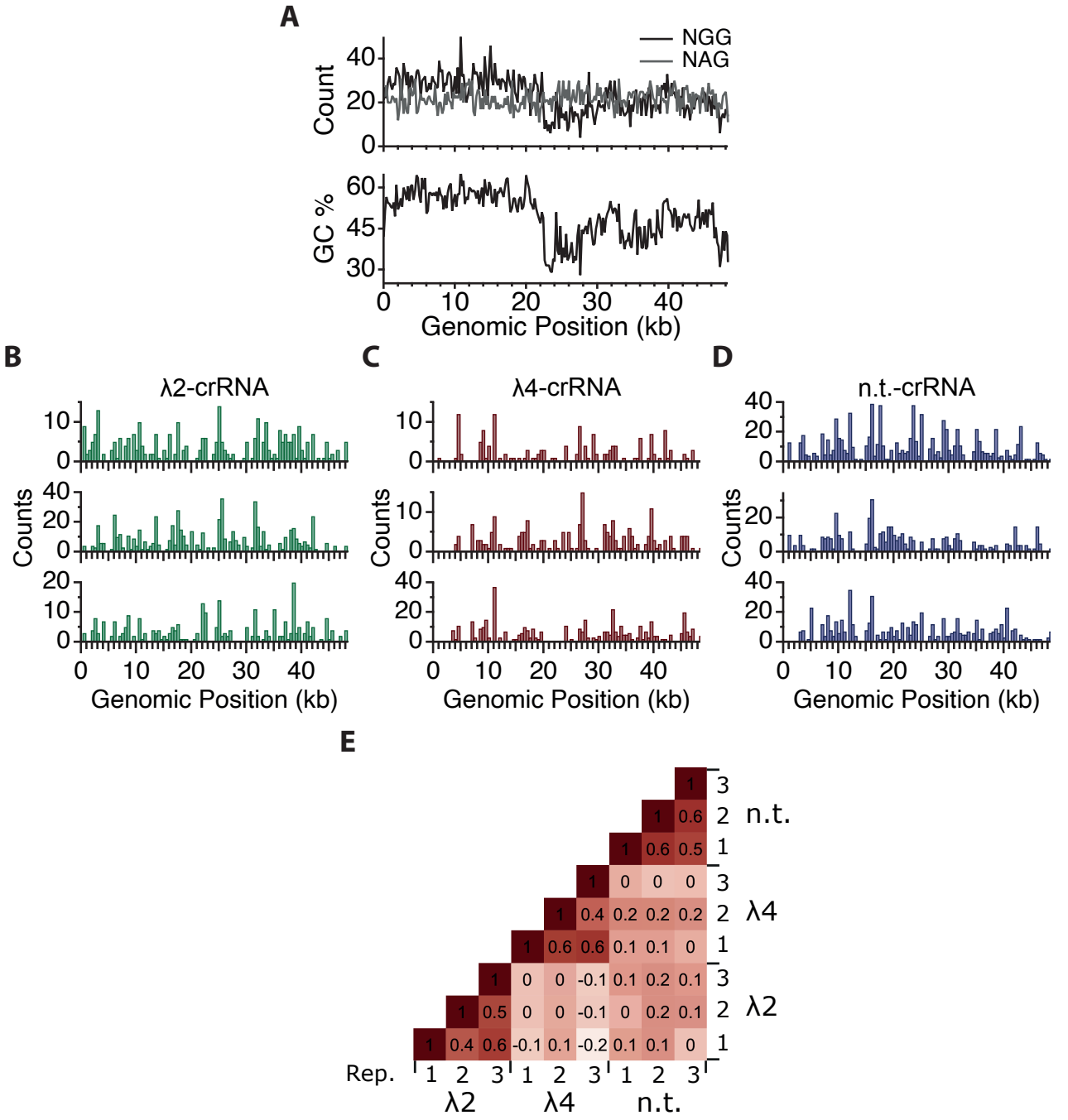


Figure 5.15: Off-target binding locations are guide sequence dependant. (**A**) Top, distribution of canonical (NGG, dark grey) and non-canonical (NAG, light grey) PAM sites across phage-λ genome. Bottom, distribution of GC content across the phage-λ genome. The first half of the genome (0 - 21 kb) is comparatively GC rich compared to the second half of the genome (21 - 48 kb) which has a lower GC content. (**B**) Histograms of location of off-target binding events for Cy3-λ2-crRNA:tracrRNA:dCas9 complex on λ-DNA for three different DNA molecules. crRNA, n = 278, 645 and 311 (**C**) Histogram of location of off-target binding events for Cy3-λ4-crRNA:tracrRNA:dCas9 complex on λ-DNA for three different DNA molecules.n = 437, 210 and 150 (**D**) Histogram of location of off-target binding events for Cy3-nt-crRNA:tracrRNA:dCas9 complex on λ-DNAfor three different DNA molecules. n = 605, 876 and 469 (**E**) Pearson's correlation analysis of distribution of induced off-target binding at 40 pN from count normalised histograms. Strong correlations (dark red) in binding locations are observed between repeat experiments with the same guide crRNA (Rep. 1-3). No correlation is observed between data obtained with different guide crRNA sequences (λ2, λ4 and n.t.), demonstrating that locations of force induced off-targets are crRNA sequence dependant.

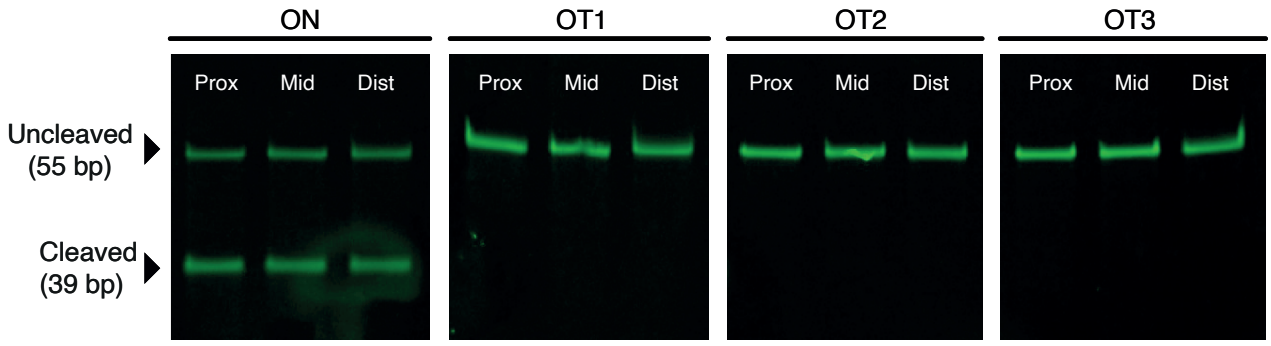


Figure 5.16: Off-targets with 6 nt PAM bubbles are not cleaved by Cas9. Cleavage assay for off-target constructs with 6 nt bubbles in PAM proximal, middle or distal locations, visualised by denaturing PAGE and fluorescent imaging. Cas9 only cleaves on-target sequences, no off-target cleavage is observed with a 6 nt bubble at any of the locations tested.

Chapter 6

DNA supercoiling induces CRISPR/Cas9 off-target activity

6.1 Abstract

We have previously demonstrated that force stretching DNA dramatically reduces CRISPR/Cas9 specificity, inducing off-target activity at sites with as many as 10 mismatching bases (See Chapter 5). Such a dramatic loss of Cas9 specificity could be hugely problematic in the context of therapeutic gene editing. Although it is unclear whether such forces can occur within cells, based on our hypothesis that this loss of specificity is driven by the reduction in the energy barrier for DNA melting, we reason that other biological processes which destabilise the DNA duplex could have a similar effect. One potential candidate is negative supercoiling. Underwound, negatively supercoiled, DNA molecules have a reduced DNA duplex stability, with extreme levels of supercoiling resulting in spontaneous formation of melted DNA bubbles (Jacob et al., 1974; Sheinin et al., 2011; King et al., 2019). As such we reasoned that negatively supercoiling DNA may also induce Cas9 off-target activity.

In this chapter, I will present my work on the effect of negative supercoiling on Cas9 specificity. Using optical tweezers to negatively supercoil DNA, I demonstrate that supercoiling induces Cas9 off-target binding at low force at the same sites as observed for force stretching DNA. The number of induced off-target events correlates with the degree of negative supercoiling. An increased number of off-target events can be induced by force stretching the negatively supercoiled DNA. We further investigate the potential for negative supercoiling to induced off-target activity across human genomic DNA. Using a modified CIRCLE-seq protocol, we observe thousands of novel off-target sites induced in the presence of negative supercoiling. This work demonstrates that Cas9 specificity can be modulated by the structural state of target DNA. This offers new insight into the mechanism of on- versus off-target discrimination could inform development of high specificity Cas9 systems.

6.2 Results

6.2.1 Negatively supercoiling DNA induces Cas9 off-target binding

To investigate the effect of DNA supercoiling on Cas9 specificity, we first used a single-molecule approach, combining optical tweezers with confocal microscopy to monitor Cas9 binding to λ-DNA in real-time, as previously described (5). Experiments were performed in a microfluidic flow cell enabling easy capture and tethering of single DNA molecules between optically trapped beads and rapid movement between channels separated by the laminar flow (Figure 6.1A). We first optically trap two polystyrene beads (4.5 μm , streptavidin coated particles, Figure 6.1A, 1.) before moving through the flow cell to the DNA channel, where we catch a single molecule of torsionally constrained λ-DNA (Figure 6.1A, 2.). Next, we move to the buffer channel (Figure 6.1A, 3.) where the DNA is negatively supercoiled by repeated over-extension and torsional relaxation until the required degree of supercoiling is induced (Figure 6.1B and C). On over-extension of the torsionally constrained DNA (Forces $>$ $\sim$115 pN), one of the biotin interactions between the bead and DNA can momentarily release resulting is a release of torsional strain through unwinding of the DNA helix (Figure 6.1B, 2.). As the DNA relaxes the biotin interactions re-attach, constraining the DNA in an underwound, negatively supercoiled state (Figure 6.1B, 3.). This process can be monitored by changes in the force-distance curves (Figure 6.1C). In the absence of any supercoiling, extension of torsionally constrained DNA gives a characteristic force-distance profile with a overstretching plateau at $\sim$ 120 pN (Figure 6.1C, 1.). On holding the DNA in an overstretched state (Figure 6.1C, 2.), negative supercoiling can be induced. This is observed as a permanent increase in the DNA extension at forces $>$ 50 pN (Figure 6.1C, 3.). The degree of negative supercoiling, σ, is described as the fractional change in the DNA linking number, Lk, and calculated by the following relationship:

$$\sigma = \frac{\mathrm{L_k} - \mathrm{L_{k_0}}}{\mathrm{L_{k_0}}} \tag{6.1}$$

For a given piece of DNA the value of σ can be determined by using an empirically determined reference plot of σ as a function of the DNA extension at 70 pN (Figure 6.1C, insert, King et al.,

2019).

The degree of supercoiling can be controlled by modulating the time spent in the over-extended position, and through repeated cycles of over-extension and relaxation (Figure 6.1C). Once the required level of supercoiling had been induced the DNA was moved to the protein channel containing labelled Cas9 complex and imaged by confocal microscopy (Figure 6.1A (4./5.) and 6.1D).

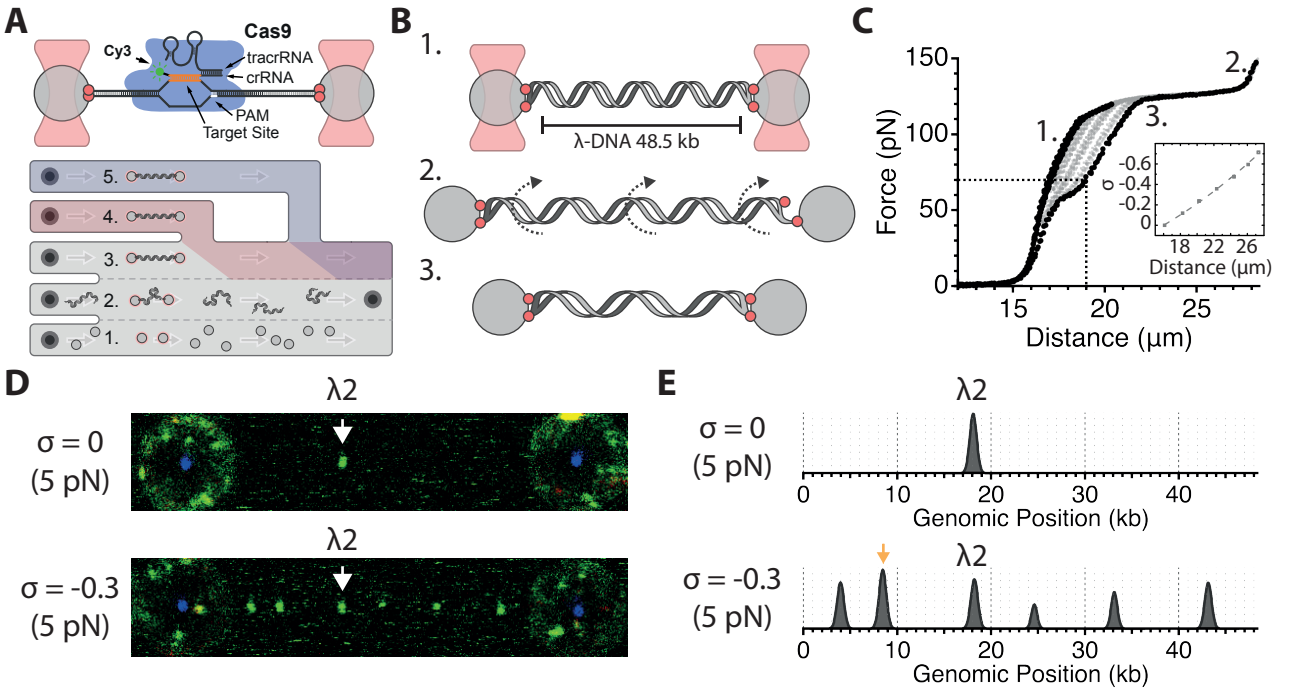


Figure 6.1: Negative supercoiling induces Cas9 off-target binding. (**A**) Experimental design, top. λ-DNA is tethered between two optically trapped beads, Cy3-labelled Cas9 complex is targeted to λ2 site. Schematic of five channel microfluidic flow cell, bottom. Bead Channel (1.), DNA Channel (2.), Buffer Channel (3.) Protein/Imaging Channels (4.) and/or (5.). (**B**) Negatively supercoiling DNA using optical tweezers. Torsionally constrained λ-DNA is held between two optically trapped beads (1.). DNA is held at high extension (>20 μm) at which point, release of biotin-streptavidin interactions can transiently result in unconstrained DNA which unwinds (2.). Biotin-streptavidin interactions are re-formed resulting negatively-supercoiled λ-DNA (3.). (**C**) Example force-distance plot recorded during the generation of negatively super-coiled DNA. Insert, calibration curve of σ value for a given extension at 70 pN, adapted from King et al., 2019. (**D**) 2D confocal scan of λ2-crRNA:tracrRNA Cas9 complex binding to low force stretched, 5 pN, torsionally constrained relaxed DNA (top, $\sigma = 0$) and negatively supercoiled DNA (bottom, $\sigma =$ -0.3). On-target λ2 binding site, white arrows. For negatively supercoiled DNA additional off-target binding sites are observed, bottom. (**E**) Intensity peaks from confocal scans in (D), mapped to λ genome. On-target binding site, λ2.

We first tested on-target binding of dCas9 to torsionally constrained DNA. λ2-crRNA:tracrRNA dCas9 complex was incubated with torsionally constrained but relaxed DNA ($\sigma = 0$) and held at low force (5 pN). A single dCas9 complex was observed bound to the expected λ2 on-target site (Figure 6.1D, top) with no off-target binding events observed across the length of the DNA.

Next, we performed the same experiment but with negatively supercoiled DNA. The DNA was negatively supercoiled to $\sigma \sim -0.3$ (Figure 6.1C), before moving to the dCas9 channel and imaging. As before, dCas9 was observed binding to the expected on-target $\lambda 2$ site; however, in the presence of negatively supercoiled DNA, multiple complexes were also observed bound at off-target sites (Figure 6.1D, bottom). Based on the known sequence of the λ-DNA we mapped the fluorescence intensity peaks of bound Cas9 to the λ genome sequence (Figure 6.1D). As expected, with both the the relaxed and supercoiled DNA, a peak is observed at the $\lambda 2$ on-target site (18.084 kb) (Figure 6.1E, top). In the case of the negatively supercoiled DNA, five additional off-target bindng peaks are also observed (Figure 6.1E, bottom). Interestingly we consistently observed binding at a site at $\sim$ 8.3 kb (Figure 6.2C, yellow arrow). This corresponds to the only off-target site predicted by existing algorithms, at genomic position of 8378 bp, when allowing up to 7 mismatches and the presence of alternative PAM sites (Bae et al., 2014), and was previously observed to be induced by DNA stretching (Newton et al., 2019).

6.2.2 Induced off-target sites correlate with negative supercoiling

We next asked whether the number of induced Cas9 off-target binding sites correlates with the degree of negative supercoiling present in the DNA. Using the optical tweezers assay we generated molecules with different degrees of supercoiling (σ = -0.15, -0.3, -0.4, -0.7), before moving into the dCas9 channel and imaging dCas9 binding events. Even at the lowest level of negative supercoiling tested, σ = -0.15, multiple induced off-target binding events were observed (Figure 6.2A). At higher levels of negative supercoiling (σ = -0.3 to -0.7) a greater number of induced off-target binding events were observed (Figure 6.2A). For each supercoiling state DNA molecules were repeatedly held at 5 pN in the presence of λ-2-crRNA Cas9 and the number of binding events identified and averaged (Figure 6.2B, n $\geq$ 9). A clear increase in the average number of bound Cas9 complexes is observed with increasing levels of supercoiling. This was fit with a simple binding isotherm:

$$\mathrm{N} = \mathrm{N}_0 + (\mathrm{N}_{\max} - \mathrm{N}_0).(\frac{\sigma}{\sigma + \sigma_{50}}) \tag{6.2}$$

Where N is the average number of bound dCas9 complexes, N_0 is the number bound at $\sigma = 0$, N_{max} is the maximum average number bound and σ_{50} is the value of σ at which the average number of bound molecules is half of N_max.

Fitting to this binding isotherm gives $N_{max} = 13 \pm 1$ and $\sigma_{50} = -0.28 \pm 0.06$. This demonstrates that low levels of supercoiling already induce significant amounts of off-target binding. Only a two-fold increase in off-target events is observed from $\sigma = -0.15$ to $\sigma = -0.7$, despite an almost 5-fold increase in the level of supercoiling (Figure 6.2B).

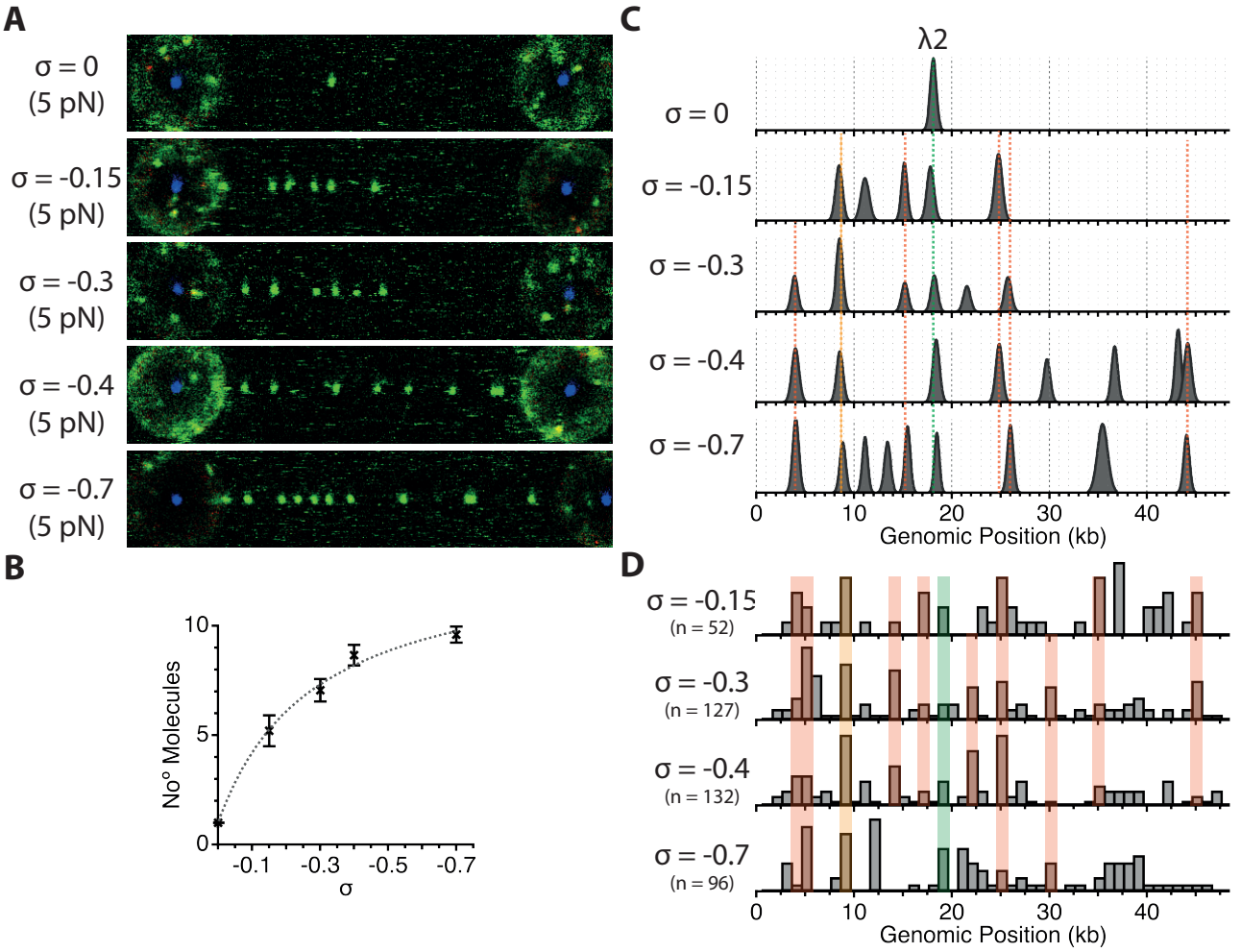


Figure 6.2: Induction of off-target sites correlates with degree of negative supercoiling. (**A**) Representative 2D confocal scans of torsionally constrained λ-DNA with supercoiling states ranging from σ=0 - σ=-0.7 held at low force (5 pN) in the presence of λ2-crRNA:tracrRNA dCas9 complex. (**B**) Plot of number of Cas9 molecules bound at 5 pN as a function of degree of supercoiling. Fit with binding isotherm (See Equation 6.2), grey dashed, $N_{max} = 13 \pm 1$, $\sigma_{50} = -0.28 \pm 0.06$. n geq 9, error bars represent standard error. (**C**) Corresponding intensity profiles from **A** mapped to λ genome. The expected on-target site is present at all sigma values (λ2, black arrow) at increased levels of negative supercoiling additional off-target peaks are observed. Some off-target peaks are observed at the same locations at different supercoiling states (coloured vertical lines). (**D**) Histograms of mapped λ2-crRNA Cas9 binding events at 5 pN at different supercoiling states. Bin size = 1 kb

The observed off-target binding locations were then mapped to the λ genome. It is immediately apparent that the some of the same off-target sites are observed at all levels of supercoiling (Figure 6.2C, dashed lines). However, additional sites are present as the degree of supercoiling increases (Figure 6.2C). We again observed binding at the only predicted off-target site at $\sim$ 8.3 kb at all levels of supercoiling (Figure 6.2C, yellow dashed line). To compare the distribution of off-target sites at different degrees of supercoiling, we built probability density histograms from all the mapped locations of off-target binding observed (Figure 6.2D). There are clear binding hotspots at all levels of supercoiling, many of which occur at the same locations (Figure 6.2D). This further reinforces the idea that Cas9 off-target binding is being induced at specific sites. Interestingly, at higher levels of supercoiling there are fewer clearly defined loci. For example, for σ = -0.7 there are binding events spread over the whole region of 35 - 48 kb (Figure 6.2C). This suggests, that at higher levels of supercoiling more potential off-target sites can be bound, resulting in events which are more evenly distributed across the DNA, rather than at a small number of distinct sites.

6.2.3 Supercoiling induced off-targets correspond to force induced off-target sites

We have previously reported that force stretching λ-DNA induces Cas9 off-target binding (See Chapter 5, Newton et al., 2019). We next compared the locations of supercoiling induced off-target sites to previously observed force induced off-target sites of λ-DNA. Multiple common peaks were observed (Figure 6.3A), suggesting that negative supercoiling of DNA induces off-target binding at low force (5 pN) at the same sites observed only on high force stretched relaxed DNA (40 pN, $\sigma = 0$).

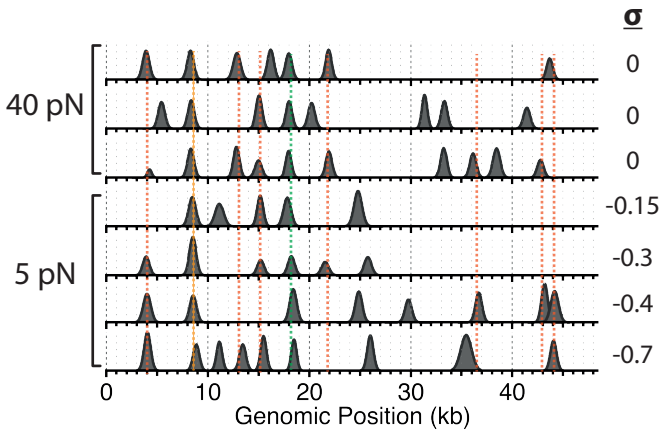


Figure 6.3: Off-target sites induced by force are also induced by supercoiling. Comparison of off-target binding peaks induced at high force (40 pN) without supercoiling ($\sigma = 0$) and at low force (5pN) on negatively supercoiled DNA (σ = -0.15, -0.3, -0.4, -0.7)

In an attempt to identify the specific sequences of force induced off-target binding sites across λ-DNA, we previously developed a simple scoring model (See Section 5.2.6, Newton et al., 2019). Based on the observation that there are common binding locations between the negative supercoiling induced off-targets and previously observed force induced off-targets (Figure 6.3), we hypothesised that the same model may also be valid in this context. Briefly, each potential off-target sequence across the λ genome is assigned a score. This score is based on a number of factors: the number of matches with the guide sequence, with weighting for matches in the PAM proximal seed region; the presence of a canonical or non-canonical PAM sequence; and an estimate of the relative mismatch penalty based on both the ability to wobble base-pair, and the presence or absence of base-paired nearest neighbours (See Section 5.2.6). We took the top 18 hits from across the λ genome (Table 6.1) and generated a predicted profile of binding peaks,

which we compared to the observed binding locations (Figure 6.4).

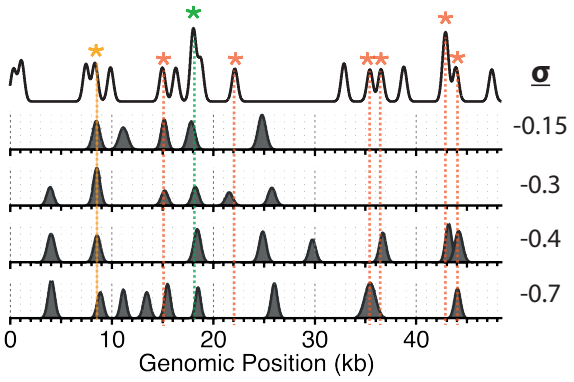


Figure 6.4: A simple model predicts supercoiling induce off-target sites. Comparison of off-target binding peaks predicted by a simple scoring model, top, with off-target peaks observed on low force stretched (5 pN) negatively supercoiled DNA (σ = -0.15, -0.3, -0.4, -0.7).

Supercoiling induced off-target binding was observed at locations overlapping 7/17 of the top predicted sites (Figure 6.4 and Table 6.1). Of particular note is the site at 8.3 kb, which is the most frequently observed off-target site on supercoiled DNA of all σ values, and was previously observed on force stretched DNA (Figure 6.2D, Table 6.1, highlighted with yellow). Furthermore, frequent binding was observed at 15 kb (Figure 6.2D, Table 6.1, highlighted with red). Together this data suggests that at least some of the off-target sites induced by negative supercoiling are the same as those induced by force stretching DNA. This supports the hypothesis of reduction in the energy barrier for DNA melting being the key factor of loss of specificity, whether this is achieved through negative supercoiling or force stretching DNA.

Table 6.1: Predicted Off-target Sequences

Sequence (5' - 3')	Mismatches	Score	Position (nt)	Strand	PAM
TCATAAAATGGTATGCCGAA	10	226	340	+	NGG
GTGCTTTTTGCCATACCACG	8	252	1143	-	NGG
CTGATGCACTTCACGCCACG	10	242	7492	+	NGG
ATGGCCGGTGGCATGTCATG	7	245	8378	+	NAG
CTGCGGATGGTTATGCCACA	10	232	9933	-	NGG
ATATGAATTACAGCGCCATC	10	231	15054	-	NGG
CCCATCCGCGCTACGGCATG	10	230	16359	+	NGG
GTGATAAGTGGAATGCCATG	0	350	18084	+	NGG
AATATCTGCCGAATGCCGTG	8	255	18841	-	NGG
GCATCTCGTTGAAGACCATC	10	228	22180	+	NGG
CGATCATTTGCAACGCCACG	10	244	32979	-	NGG
ATTCAAAGCAGAAGGCTTTG	9	225	35528	-	NGG
GCGATATATTAAAGGTCTTA	9	226	36628	-	NGG
GTTTTACGCAGAATGGCAAG	7	234	38877	-	NAG
CTAATCAGTTCGACGCGATA	9	238	42914	-	NGG
TTTTTACTGAGAACGTCATG	9	238	43071	-	NGG
CTGATGTGATGGCTGCTATG	8	232	43993	+	NGG
ATCATCAACAAAACACAAGG	11	225	47560	-	NGG

6.2.4 Force stretching negatively supercoiled DNA induces further off-target sites

Based on our previous observations of inducing Cas9 off-target activity by force stretching DNA, we next investigated the effect of force in combination with negative supercoiling. A molecule of λ-DNA was tethered between the two beads and negatively supercoiled to approximately σ = -0.15. The DNA was moved to the channel containing Cy3 labelled λ2-crRNA dCas9 complex and held at 5 pN before confocal imaging. As expected Cas9 was observed binding to the λ2 on-target site, as well as multiple off-target sites (Figure 6.5A, 5 pN).

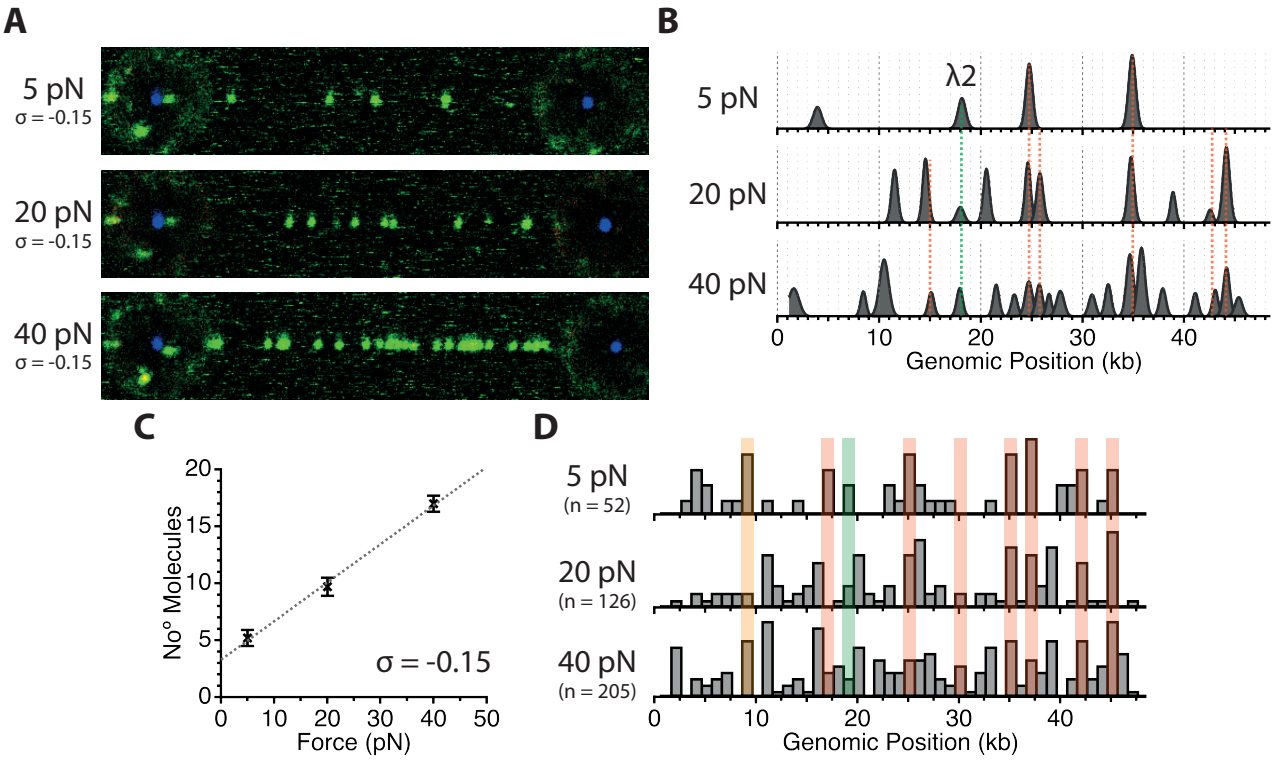


Figure 6.5: Force stretching DNA induces further off-target binding in the presence of supercoiling. (**A**) Representative 2D confocal scans of torsionally constrained λ-DNA with a supercoiling state of σ=-0.15 held at low force (5 pN), moderate force (20 pN) and high force (40 pN) in the presence of λ2-crRNA:tracrRNA dCas9 complex. (**B**) Corresponding intensity profiles from **A** mapped to the λ genome. The expected on-target site is present at all forces (λ2, black arrow). At higher forces an increased number of off-target bound Cas9 molecules are observed. Some off-target peaks are observed at the same locations at different forces (coloured vertical lines). (**C**) Plot of number of Cas9 molecules bound as a function of force for negatively super-coiled DNA, σ = -0.15. Fit with linear fit, grey dashed. n $\geq$ 9, error bars represent standard error. (**D**) Histograms of mapped λ2-crRNA Cas9 binding events at 5 pN, 20 pN and 40 pN. Bin size = 1 kb

Next the DNA was stretched to 20 pN and imaged. Interestingly, binding was observed at additional off-target sites, with further off-target binding induced as the DNA was stretched to 40 pN (Figure 6.5A). As before, we then mapped the intensity peaks of the binding events back

to the λ-DNA sequence (6.5B). A selection of off-target sites were clearly bound at all three forces (Figure 6.5, coloured lines), with additional sites present as the DNA was stretched at higher forces. The average number of binding events shows a strong force dependence (Figure 6.5C). This was fit with a linear fit to obtain an estimate of the number bound molecules at zero force, $N_0 = 3.3 \pm 0.5$, and force dependence of binding, $m_F = 0.34 \pm 0.02$ pN^{-1}. Furthermore, histograms of off-target binding locations reveal a shift from a relatively small number of sites occupied at low force to binding events spread across the length of the DNA at higher force (Figure 6.5D). Together this demonstrates that DNA tension has an additive effect on induction of Cas9 off-target activity in combination with negative supercoiling.

We previously characterised the force induced off-target binding dwell time as a function of force (See Section 5.2.3). For experiments with torsionally constrained DNA it was not possible to record long kymographs for dwell time analysis; prolonged exposure to the excitation laser resulted in DNA nicking and loss of supercoiling.

6.2.5 Negative supercoiling induces off-target cleavage across human genomic DNA

Despite the ability for our simple scoring model to predict a number of the observed peaks from the optical tweezers experiments, it is a biased approach built upon previously described rules for Cas9 off-target site selection such as weighting of the PAM proximal seed sequence and PAM site identities. However, the extent to which these rules are valid in the context of negatively supercoiled DNA is unknown. As such we developed an unbiased next-generation sequencing method to identify off-target sites which are only cut when DNA is negatively supercoiled.

The approach closely follows the CIRCLE-seq protocol for high sensitivity detection of Cas9 off-target sites (Tsai et al., 2017). First genomic DNA was extracted and purified from HEK293T human cells, fragmented and ligated to form DNA circles (Figure 6.6A, See Methods 3.4). At this stage all non-circular DNA was degraded to give a high purity pool of DNA circles of sequences covering the entire human genome. Next, part of the pool of DNA circles was negatively supercoiled with *E. coli* DNA gyrase and both the relaxed DNA circles and negatively

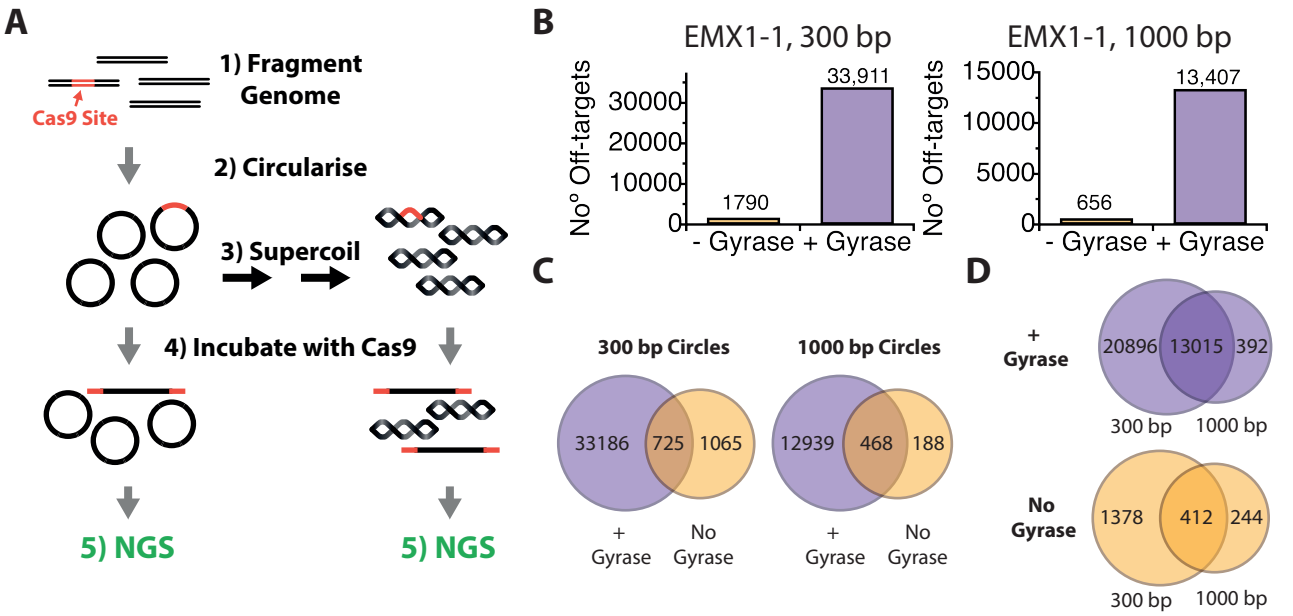


Figure 6.6: Negative supercoiling induces off-target cleavage across human genomic DNA. (**A**) Overview of method for CIRCLE-seq analysis with negative supercoiling (**B**) Number of off-target sites detected for the EMX1-1 crRNA guide sequence across the human genome with and without supercoiling with DNA gyrase for the library prepared with 300 bp fragments, left, and 1000 bp fragments, right. In both instances the number of detected off-target increases by an order of magnitude when target DNA is negatively supercoiled. (**C**) The overlap of sites identified in the unsupercoiled, No Gyrase, and negatively supercoiled, + Gyrase, for the off-targets detected with the EMX1-1 crRNA guide for the 300 bp and 1000 bp fragment libraries. (**D**) Overlap of detected sites between the 300 bp fragment libraries and 1000 bp fragment libraries.

supercoiled circles were incubated with Cas9 complex targeting the EMX1-1 locus (Figure 6.6A, See Methods 3.4). Any circles containing sequences which can be cut by Cas9 are linearised, and all linear fragments were then sequenced (Figure 6.6A, See Methods 3.4). Off-target sites were identified by read alignment to the reference genome, hg38, and and peak calling based on read enrichment between the Cas9 treated sample and no Cas9 control. The threshold maximum number of mismatches was set to 10/20 and the maximum tolerated DNA/RNA gaps set to 3. These parameters are significantly more relaxed than typically used from CIRCLE-seq experiments to ensure the most exhaustive list of off-targets possible is obtained.

The standard CIRCLE-seq protocol uses an average fragment size of 300 bp (Tsai et al., 2017), however as the persistence length of DNA is $\sim$ 150 bp we hypothesised that circles of this size may already be under mechanical strain which could affect Cas9 activity. As a result we prepared two libraries of DNA circles, one with a average fragment size of 300 bp and one with a larger fragment size of 1000 bp, which we then used to test the effect of negative supercoiling on Cas9 specificity. With the 300 bp fragment library, in the absence of supercoiling, 1790 EMX1-1

off-target sites were identified (Figure 6.6). However, when DNA circles were first negatively supercoiled before addition of Cas9, over 30,000 EMX1-1 off-target sites were detected (Figure 6.6B, left). A similar dramatic increase in detected off-targets was observed with the 1000 bp library with a 20-fold increase in detected EMX1-1 off-target sites between relaxed and negatively supercoiled DNA (Figure 6.6B, right). The greater number of detected off-targets in the absence of supercoiling in the 300 bp library, 1790 detected sites, compared to the 1000 bp library, 656 detected sites, could indicate that the small circle size is in itself inducing greater off-target activity. We next asked how much overlap there was between the off-targets sites detected with and without supercoiling (Figure 6.6C); a greater degree of overlapping sites was observed with the 1000 bp fragment library, $> 70\%$, compared to the 300 bp library, $\sim 40\%$ (Figure 6.6C). Furthermore, for the 1000 bp library, a high percentage of sites identified in the absence of supercoiling were present within the sites detected for the 300 bp library, 62 %, and a even greater percentage of the sites identified in the presence of negative supercoiling also detected in the 300 bp library, 97 % (Figure 6.7D). Together this data suggests that the 1000 bp fragment data is in general a subset of 300 bp fragment data, with potentially fewer false positives due to the mechanical strain present in smaller circle sizes. As a result, we will use the larger fragment size for further analysis and experiments.

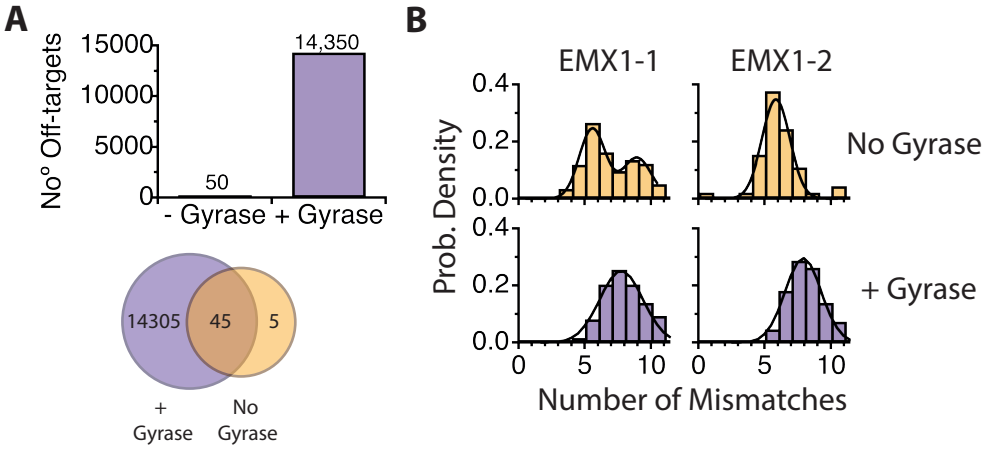


Figure 6.7: Negative supercoiling increases mismatch tolerance. (**A**) Number of off-target sites detected for the EMX1-2 crRNA guide sequence across the human genome with and without supercoiling with DNA gyrase, top. The overlap of sites identified in the unsupercoiled, No Gyrase, and negatively supercoiled, + Gyrase, for the off-targets detected with the EMX1-2 crRNA, bottom. (**B**) Probability distributions of number of mismatches present in detected off-targets for EMX1-1 and EMX1-2 crRNA guides in the presence of relaxed DNA, No Gyrase, and in the presence of negative supercoiling, + Gyrase.

Next, we tested an additional Cas9 guide targeting a different site in the EMX1 locus, EMX1-2. In the absence of supercoiling only 50 EMX1-2 off-target sites were detected, however, when DNA was negatively supercoiled Cas9 cleavage was detected at over 14,000 off-target sites (Figure 6.7A). This further confirms that negative supercoiling generally induces Cas9 off-target activity, independently of the particular on-target site being targeted. Based on the off-target sites detected for EMX1-1 and EMX1-2, we next asked whether the number of tolerated mismatches is altered in the presence of negative supercoiling. On comparison of detected off-targets, a clear increase in the number of mismatches tolerated is observed for both EMX1-1 and EMX1-2 off-targets in the presence of supercoiled DNA (Figure 6.7B). For EMX1-1 and EMX1-2, in the absence of supercoiling, their is a dominant peak of off-targets containing ~ 5 mismatches, with EMX1-1 also displaying some detected off-targets containing higher numbers of mismatches. However, in the presence of supercoiling the average mismatch number increased to 7.8 ± 0.03 for EMX1-1 and 7.9 ± 0.02 for EMX1-2. This data suggests that Cas9 can cleave off-targets with a greater number of mismatches in the presence of negatively supercoiled DNA.

Based on the dramatic increase in off-targets observed in the presence of negatively supercoiled DNA, and the apparent tolerance on an increased number of mismatches, we hypothesised the distribution of mismatches across off-target sites may be altered. We first took all the detected off-target for EMX1-1 guide and grouped them based on the number of mismatches present. For each set of off-targets with the same number of mismatches we then calculated the probability of the mismatches occurring at each position across the 20 bp target site and within the 3 bp PAM sequence (Figure 6.8A).

For the EMX1-1 guide in the presence of relaxed DNA the detected off-targets with ≤ 6 mismatches show a higher probability of mismatches occurring in the PAM distal end, position 14 - 20 (Figure 6.8A, No Gyrase). Despite the dramatic increase in the number of detected off-targets in the presence of negatively supercoiled DNA, the distribution of mismatch probabilities are comparable to those of the relaxed DNA (Figure 6.8A, + Gyrase). We next compared the EMX1-1 mismatch probabilites to those of the EMX1-2 off-targets. In the presence of relaxed DNA only 45 off-target sites were detected with the EMX1-2 guide, as such there are insufficient numbers of off-targets to accurately calculate mismatch probabilities for ever position

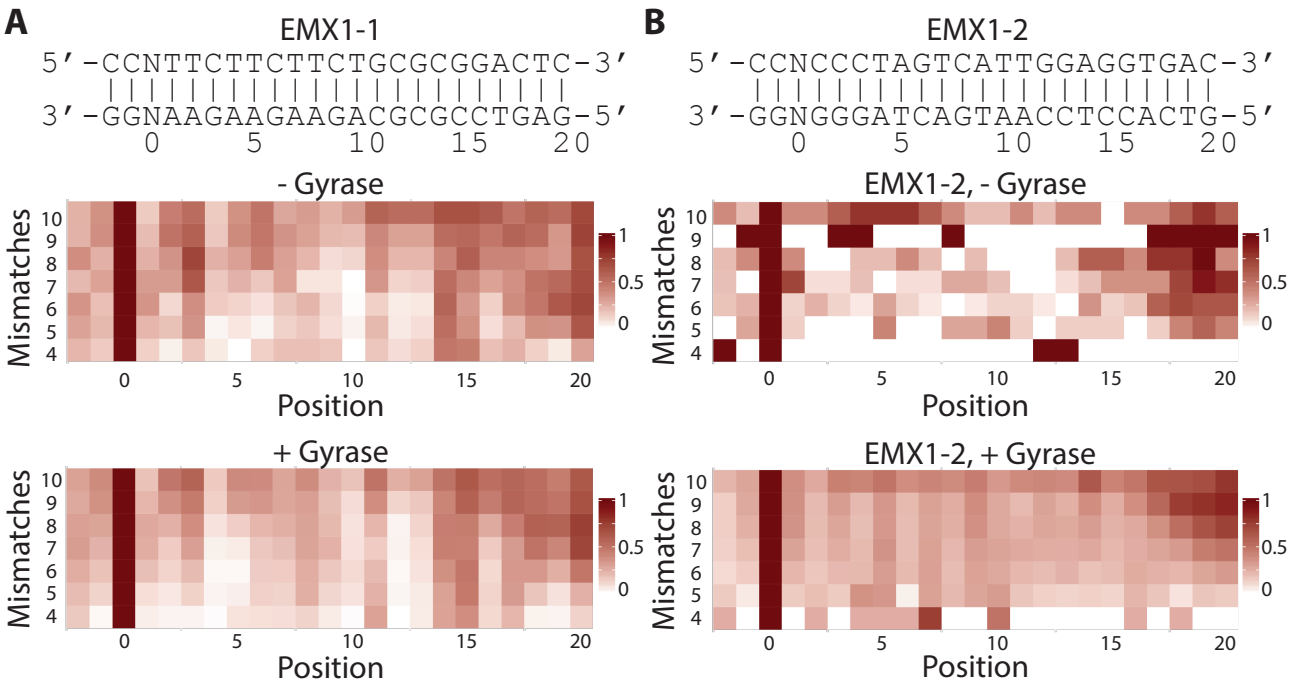


Figure 6.8: Distribution of mismatches across off-target sites. (**A**) EMX1-1 target sequence, top. Below, heatmaps displaying the mismatch probabilities across off-target sites with different numbers of mismatches present, detected for EMX1-1 guide in the presence of relaxed DNA, top, and negatively supercoiled DNA, bottom. (**B**) EMX1-2 target sequence, top. Heatmap displaying the mismatch probabilities across off-target sites with different numbers of mismatches present, detected for EMX1-2 guide in the presence of relaxed DNA, top, and negatively supercoiled DNA, bottom.

(Figure 6.8B, No Gyrase). Again however, there appears to be an overall higher probability of mismatches occurring at the PAM distal end of the target. Interestingly, in the presence of negatively supercoiled DNA the mismatches in the EMX1-2 off-targets appear to be more evenly distributed, with a clear bias towards the PAM distal end only evident with high numbers of mismatches (Figure 6.8B, + Gyrase). Together these data suggest that the distribution of mismatches is not dramatically altered in Cas9 off-targets cleaved in the presence of negatively supercoiled DNA. This again suggests that the effect of negatively supercoiling is instead an increase in the tolerated number of mismatches, resulting in an increase in the potential off-target sites available and as such a dramatic increase in the off-target cleavage sites across the human genome.

6.3 Discussion

In this work we used a recently described method to negatively supercoiling DNA with optical tweezers, we demonstrate that negative supercoiled DNA induces off-target binding even at low forces (Figure 6.1). The number of bound off-target sites correlates with the extent of negative supercoiling with repeat binding observed in the same locations across multiple supercoiling states (Figure 6.2). On comparison of the location of these induced off-target sites with the previously observed force induced off-target sites, we observe that the same sites are being bound (Figure 6.3). Furthermore, the same model which was used to successfully identify the exact sequences of some force induced off-target sites, also predicts may binding sites which overlap with observed supercoiling induced off-targets (Figure 6.4). Together this supports our hypothesis of negative DNA supercoiling reducing Cas9 specificity in an analogous manner to force stretching DNA, due to a reduced energy barrier for DNA melting. This is further demonstrated by the additive effect of applying force to already negatively supercoiled DNA, resulting in and even further inducting of Cas9 off-target binding (Figure 6.5).

In order to understand how this could impact Cas9 specificity in the context of editing the human genome, we developed a novel off-target next-generation sequencing based largely on CIRCLE-seq. This method allows comparison of off-target sites across the human genome in the presence and absence of negative supercoiling (Figure 6.6). Shockingly, tens of thousands of off-target sites are detected on negatively supercoiled DNA. Analysis of the number and distribution of mismatches present in the off-target sites suggests that the presence of negative supercoiling increases the number of tolerated mismatches (Figure 6.7 and 6.8).

An increase in the number of mismatches tolerated by CRISPR/Cas9 massively increases the number of potential sites that could be inadvertantly targeted. Furthermore, as the number of potential sites for off-target activity increases it becomes exponentially harded to detect potential off-target editing events. Any detectable signal of off-target activity get spread across a much larger pool of sites. Although this work does not suggest that Cas9 is cleaving at thousands of sites *in vivo*, is does pose the concern that under certain conditions Cas9 may be active at previously unconsidered sites. Furthermore, the shear number of potential off-target sites

present in a large genome if you allow for and increased mismatch tolerance, makes detection of off-target editing at these sites almost impossible.

In order to fully understand the implications of this it will be vital to test a much wider variety of guide sequences in the supercoiling CIRCLE-seq assay. Such a dataset would allow removal of any guide sequence specific effects to leave a clear picture of overall Cas9 off-targetting rules in the context of negatively supercoiled DNA. We further hypothesise that transcription and DNA replication passing through DNA could expose such 'cryptic' off-target sites. If this is indeed the case it could have significant implications in terms of guide design and predicition of potential off-target sites.

Taken together, the data presented in this chapter, and in combination with Chapter 5, clearly demonstrates that the specificity of the CRISPR/Cas9 complex is not mearly determined by the protein complex and guide sequence but depends on the structural state of the DNA being targeted. Reducing the energy barrier required for target DNA melting increases the number of tolerated mismatches between the guide and target. This could be an important consideration in the development of CRISPR/Cas9 therapeutic applications.

References

Bae, Sangsu, Jeongbin Park, and J.-S. Kim (May 2014). "Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases". *Bioinformatics* 30.10, pp. 1473–1475.

Jacob, R.J., J. Lebowitz, and M.P. Printz (Apr. 1974). "Unpaired bases in superhelical DNA: kinetic evidence". *Nucleic Acids Research* 1.4, pp. 549–558.

King, Graeme A et al. (Dec. 2019). "Supercoiling DNA optically". *Proceedings of the National Academy of Sciences* 116.52, pp. 26534–26539.

Newton, Matthew D. et al. (Mar. 2019). "DNA stretching induces Cas9 off-target activity". *Nature Structural & Molecular Biology* 26.3, pp. 185–192.

Sheinin, Maxim Y et al. (Sept. 2011). "Underwound DNA under Tension: Structure, Elasticity, and Sequence-Dependent Behaviors". *Physical Review Letters* 107.10, p. 108102.

Tsai, Shengdar Q et al. (May 2017). "CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets". *Nature Methods* 14.6, pp. 607–614.

Chapter 7

Conclusion

7.1 Thesis Discussion

In this thesis I have used single-molecule techniques to gain new insights into CRSIPR/Cas9 off-target activity. In Chapter 1 I use single-molecule FRET to monitor binding and R-loop formation of Cas9 complexes bound to hybrid guides, containing DNA bases in the place of the majority of normal RNA bases. I demonstrate that these hybrid guides are capable of directing Cas9 to target sites and formation of the high FRET state associated with full R-loop formation. These guides show reduced stability of this high FRET state consistent with destabilisation of the R-loop and, as a result, increased mismatch discrimination at off-target sites. Although these particular guides where not active *in vivo* this work provides evidence supporting the approach of modulating the hybridisation energy between the guide and target to improve specificity. This work further informs our understanding as to the intrinsic source of Cas9 target discrimination, clearly demonstrating the importance of the guide hybridisation energy in determine Cas9 activity at potential sites.

In Chapter 2 I combined optical tweezers with confocal microscopy in an assay to directly monitor Cas9 target and off-target DNA interrogation. I demonstrate that stretching target DNA induces off-target binding. This is likely through a reduction in the target DNA melting

energy barrier, which leads to an increased in the number of tolerated mismatches. This is supported by the single-molecule FRET experiments, performed with identified force induced off-target site sequences, which demonstrate that reduction in target melting energy through introduction of melting bubbles also induces off-target activity. Furthermore, we hypothesised that other cellular processes which destabilise the DNA could also reduce Cas9 specificity, which I demonstrate to be true in the case of negative supercoiling in Chapter 3.

The induction of off-target binding on negatively supercoiled DNA at low force at the same sites we previously observed at high-forces is perhaps the most important result of this thesis. It is well established that DNA is exposed to dynamic changes in supercoiling state during transcription, DNA replication and through the action of topoisomerases. As such this finding could have significant implications as to the specificity of Cas9 *in vivo*.

Strikingly, the bacterial genomes where CRISPR/Cas9 systems evolved are maintained at a high level of negative supercoiling. This begs the question what the implication of such a dramatic reduction in Cas9 specificity has on its endogenous activity. Early work to understand the mechanism of CRISPR based immunity observed a greater stability of transfected linear DNA compared to circular DNA Garneau et al., 2010 Deveau et al., 2008. First the authors took naive plasmid free *S. thermophilius* and transfected a pNT1 plasmid. Interestingly when looking at the specific sequences acquired in the CRISPR array, and their ability to target the invasive plasmid, they found that even with mismatches in the spacer and in the PAM sequence on the plasmid the plasmid would stil be eliminated when circular, however in some instances a linear form of the invasive plasmid was stable. While the authors offer no explanation at the time, one possible explanation based on these results is that the negative supercoiling present in a plasmid increased the tolerance for mismatches and allows degredation of the plasmid even with imperfect matching between the spacer and the plasmid, or non-canonical PAM sequences. In the linear form there is no supercoiling and so the CRISPR system using mismatched spacers or targeting non-canonical PAM sequences is unable to target the DNA.

Overall these results clearly demonstrate that Cas9 specificity is not an intrinsic property of the Cas9 complex but is dramatically influenced by the state of the DNA being targeted. This

would seem to be directly at odds with the 'conformational checkpoint' described by Chen et al., 2017, whereby the Cas9 complex is described to be capable of detecting the presence of mismatches between the guide and target. Clearly this is not the case as we demonstrate cleavage of sites with as many as 50% of the bases mismatched. Instead our work supports the previously described kinetics models Klein et al., 2018 and Eslami-Mossallam et al., 2020. These models present an energy landscape for binding to on-target sites displaying a downhill slope of progression of R-loop formation (Figure 7.1, pink). The presence of mismatches raises the energy landscape from the position of the mismatch onwards, in the presence of too many mismatches the complex cannot reach the final cleavage competent state.

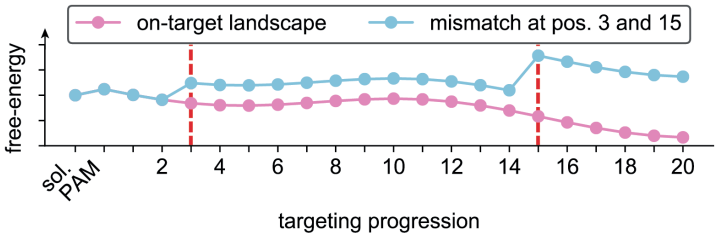


Figure 7.1: Theoretical binding energy landscape. On-target binding, pink, and binding to a site containing two mismatches at positions 3 and 5, purple.

The results from all three chapters of this thesis can be explained in terms of such an energy landscape. Incorporation of DNA bases into the guide can be thought to reduce the slope of the energy landscape, less energy is recovered from each base pairing of a DNA base in the guide to the target. As a result when a mismatch is encountered, the upward shift of the energy landscape is less likely to be overcome by subsequent guide-target base-pairing. This makes Cas9 containing a DNA hybrid guide more sensitive to mismatches than the equivalent RNA guide. In contrast, when the DNA is force stretched or negatively supercoiled, this can be thought of as causing a clockwise rotation of the entire energy landscape around the first point. In this way DNA stretching or negative supercoiling increases the slope of the energy landscape enabling Cas9 to overcome an increased number of mismatches and still reach the final cleavage competent state.

7.2 Future Work

The greater understanding the specificity of CRISPR/Cas9 in this way could be crucial for the development of more specific complexes. While one possible approach of increasing specificity by modulating the hybridisation energy is discussed in Chapter 1, we would also hypothesis that specificity could be improved by increasing the energy cost of unmatched bases being present in the guide binding region of the Cas9 complex. Solving protein structures of the complex bound to the highly mismatched sites demonstrated in Chapter 2 would provide information about the interactions between the protein and the mismatched regions of the guide and DNA target. Reducing the ability for the protein to accommodate such unpaired regions, through targeted mutation of these interacting residues, could be a powerful approach for development of higher specificity complexes.

Furthermore, although this work focuses on the SpCas9 complex, we predict that these findings may be generally applicable to all CRISPR systems which require target site DNA melting and guide hybridisation. It would interesting not only to test novel CRISPR systems in these assays but also to test existing high-fidelity variants. The ability to induce off-target interactions with force or negative supercoiling could make this approach a powerful was of directly comparing the specificity of existing variants.

Although we have show that Cas9 is susceptible to a loss of specificity when targeting force stretched or negatively supercoiled DNA, it is not yet clear to what extend these effect Cas9 targeting in cells. The demonstration in Chapter 2, that previously validated off-target sites can only be cleaved by Cas9 when the double-stranded DNA structure is disrupted, strongly suggests that this does occur in the *in vivo* context. However this could be directly tested by performing experiments to detect off-target sites in cells in the presence of topoisomerase inhibitors or transcription inhibitors.

In regards to the future prospects of therapeutic Cas9 mediated gene therapy, the results presented here do not aim to suggest that safe and efficient Cas9 therapeutics are not possible. In fact, with careful guide selection, and appropriate method of delivery, such approaches have already been shown to be relatively safe. Instead this work highlights our currently

incomplete understanding in the factors that can influence Cas9 specificity. While we have made significant progress in understanding the impact of DNA tension and negative supercoiling on Cas9 specificity *in vitro* much more work is required to fully understand the implications of this *in vivo*.

Perhaps an alternative vision of the future of gene therapy is the use of CRISPR based techniques as a way of screen potential targets. Systems such as Cas9 could be readily designed to target different sites and the most promising targets identified. This could then inform the engineering of specific TALENs to target these sites, which may be able to offer improved specificity. Furthermore, the search is still on for alternative CRISPR systems which could offer advantages over Cas9. Even if it is not clear exactly what flavour of gene therapy will emerge on top it is obvious that routine modification of human DNA is now a reality. Gene therapy is here, and with the right direction it can offer cures to some of the worst diagnoses.

References

Chen, Janice S. et al. (Oct. 2017). "Enhanced proofreading governs CRISPR–Cas9 targeting accuracy". *Nature* 550.7676, pp. 407–410.

Deveau, Hélène et al. (2008). "Phage response to CRISPR-encoded resistance in Streptococcus thermophilus". *Journal of Bacteriology* 190.4, pp. 1390–1400.

Eslami-Mossallam, Behrouz et al. (July 2020). "A kinetic model improves off-target predictions and reveals the physical basis of Sp Cas9 fidelity". *bioRxiv*, p. 2020.05.21.108613.

Garneau, Josiane E et al. (2010). "The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA". *Nature* 468.7320, pp. 67–71.

Klein, Misha et al. (Feb. 2018). "Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules". *Cell Reports* 22.6, pp. 1413–1423.

Appendix A

Appendix

A.1 Off-target Prediction Script

The following script was used to score all the potential off-target sites across the lambda DNA sequence.

A.1.1 createWindows.R

```
library(Biostrings)
library(data.table)
library(dplyr)
library(ggplot2)
library(tidyr)
library(reshape2)

createWindows_std <- function(DNA, query_string, strand = "fwd"){

  if (strand == "fwd") {
    DNA <- DNA
  }

  if (strand == "rev") {
    DNA <- reverseComplement(DNA)
  }

```

```
18  query = DNAString(query_string) #Convert crRNA to DNAstring
19
20  ##Create wobble crRNA
21  temp <- gsub("G", "A", query_string) #Replace G's with A's in crRNA to find
    ↪  G-T wobbles
22  temp2 <- gsub("T", "C", temp) #Replace T's with C's in crRNA to find U-G
    ↪  wobbles
23  query_w <- DNAString(temp2) #Create DNA string of wobble crRNA query
24
25  ##Create sliding windows across DNA sequence
26  le <- width(DNA) #Find length of DNA sequence to scan
27  winSize <- length(query) #Find length of query to define window size
28  windowViews <- NULL # Clear any existing windowViews variable
29  windowViews <- Views(DNA[[1]], start=1:(le-(winSize-1)), end=winSize:(le))
    ↪  #Create windows of length of query as 'views' across DNA sequence
30  windowsDF <- as.data.frame(windowViews)                  # Data frame
    ↪  storage of windows, list of sequence of each window
31  names(windowsDF) <- "seq" #Rename as "seq"
32
33  ##Find perfect matches of query with window
34  sectionSummary <- matrix(as.matrix(windowsDF$seq)) #Convert to martix for
    ↪  easy comparision to the query
35  querySummary <- matrix(query_string, nrow = nrow(windowsDF), ncol =
    ↪  length(query_string), byrow = TRUE) #Create martix of query sequece
    ↪  with same dimensions as matrix of windows
36  perfectMatch <- (sectionSummary == querySummary) * 1 #Logic test two
    ↪  matrices to find locations of all perfect matches
37
38  ##Make Matrix of each base in query and wobble version of query to logic
    ↪  test against each base in windows
39  windowsMatrix <- suppressWarnings(as.matrix(windowViews)) # Matrix storage
    ↪  of windows for logic operations
40  queryMatrix <- matrix(as.matrix(query), nrow = nrow(windowsMatrix), ncol =
    ↪  length(query), byrow = TRUE)  # Matrix of repeated target sequences -
    ↪  same dimentions as windows Matrix
41  query_w_Matrix <- matrix(as.matrix(query_w), nrow = nrow(windowsMatrix),
    ↪  ncol = length(query_w), byrow = TRUE) #Added to handle wobbles
42
43  ##Test base by base matches
44  matchMatrix <- (windowsMatrix == queryMatrix) * 1 #Test bases by bases
    ↪  matches of query with window
45  matchNo <- as.vector(rowSums(matchMatrix)) #Count nuumber of matches
46  mismatchMatrix <- (!matchMatrix) * 1 #Test bases by bases mismatche of
    ↪  query with window
47  mismatchNo <- as.vector(rowSums(mismatchMatrix)) #Count number of
    ↪  mismatches
```

```
48    wobbleMatrix <- ((windowsMatrix == query_w_Matrix) * mismatchMatrix) #Test
      ↪ base by base wobble matches - requires match between wobble query and
      ↪ window but no match between query and window
49    wobbleNo <- as.vector(rowSums(wobbleMatrix)) #Count number of wobble
      ↪ matches
50
51    #Nearest neighbour checks
52
53    # N-1 Matix: Create 0/1 matrix base by base of whether a match is present
      ↪ at base n-1
54    baseBefore <- cbind(matrix( 0, nrow = nrow(matchMatrix)),
      ↪ matchMatrix[,-ncol(matchMatrix)])
55
56    # N+1 Matrix: Create 0/1 matrix base by base of whether a match is present
      ↪ at base n+1
57    baseAfter <- cbind(matchMatrix[,-1], matrix( 0, nrow = nrow(matchMatrix)) )
58
59    # N-1 Matix: Create 0/1 matrix base by base of whether a wobble is present
      ↪ at base n-1
60    wobbleBefore <- cbind(matrix( 0, nrow = nrow(wobbleMatrix)),
      ↪ wobbleMatrix[,-ncol(wobbleMatrix)])
61
62    # N+1 Matrix: Create 0/1 matrix base by base of whether a wobble is
      ↪ present at base n+1
63    wobbleAfter <- cbind(wobbleMatrix[,-1], matrix( 0, nrow =
      ↪ nrow(wobbleMatrix)) )
64
65    ##Apply scores to matricies of matches, wobbles etc.
66    matchMatrix <- matchMatrix * MB
67    wobbleMatrix <- wobbleMatrix * W
68    baseBefore <- baseBefore * NNM
69    baseAfter <- baseAfter * NNM
70    wobbleBefore <- wobbleBefore * NNW
71    wobbleAfter <- wobbleAfter * NNW
72
73    baseScore <- (matchMatrix + baseBefore + baseAfter + wobbleMatrix +
      ↪ wobbleBefore + wobbleAfter) #Calculate individual base scores
74    windowScore <- as.vector(rowSums(baseScore) + rowSums(baseScore[,13:20]))
      ↪ #Calculate windowScore
75
76    windowsDF <- windowsDF %>% mutate(., perfect = as.vector(perfectMatch)) %>%
      ↪ mutate(., matches = matchNo) %>% mutate(., mismatches = mismatchNo) %>%
      ↪ mutate(., score = windowScore) %>% mutate(coord =
      ↪ as.numeric(row.names(.))) %>% mutate(., strand = strand)
77
78    return(windowsDF)
79  }
```

A.1.2 findPAMs.R

```r
library(Biostrings)
library(data.table)
library(dplyr)
library(ggplot2)
library(tidyr)
library(reshape2)

createWindows_std <- function(DNA, query_string, strand = "fwd"){

  if (strand == "fwd") {
    DNA <- DNA
  }

  if (strand == "rev") {
    DNA <- reverseComplement(DNA)
  }

  query = DNAString(query_string) #Convert crRNA to DNAstring

  ##Create sliding windows across DNA sequence
  le <- width(DNA) #Find length of DNA sequence to scan
  winSize <- length(query) #Find length of query to define window size
  windowViews <- NULL # Clear any existing windowViews variable
  windowViews <- Views(DNA[[1]], start=1:(le-(winSize-1)), end=winSize:(le))
  ↪ #Create windows of length of query as 'views' across DNA sequence
  windowsDF <- as.data.frame(windowViews)                    # Data frame
  ↪ storage of windows, list of sequence of each window
  names(windowsDF) <- "seq" #Rename as "seq"

  ##Find perfect matches of query with window
  sectionSummary <- matrix(as.matrix(windowsDF$seq)) #Convert to martix for
  ↪ easy comparision to the query
  querySummary <- matrix(query_string, nrow = nrow(windowsDF), ncol =
  ↪ length(query_string), byrow = TRUE) #Create martix of query sequece
  ↪ with same dimensions as matrix of windows
  perfectMatch <- (sectionSummary == querySummary) * 1 #Logic test two
  ↪ matrices to find locations of all perfect matches

  windowsDF <- windowsDF %>% mutate(., perfect = as.vector(perfectMatch)) %>%
  ↪ mutate(coord = as.numeric(row.names(.))) %>% mutate(., strand = strand)

  return(windowsDF)
}
```

A.1.3 offtargetpredict.Rmd

```
library(Biostrings)
library(data.table)
library(dplyr)
library(ggplot2)
library(tidyr)
library(reshape2)

# Define Scores
MB <- 8      # Score for a Matching Base
MM <- 0     # Penalty for a MisMatched base
W <- 4      # Rescue score for a Wobble mismatch eg. -3 for mismatch +3
 ↪  rescue = 0 for wobble (Substitution of G~A or C~T)
NNM <- 2    # Bonus for a matching neighbour
PAM_score <- 20   # Bonus for presence cononical PAM - NGG
APAM_score <- 0  # Bonus for alt PAM - NAG
NPAM <- 0 # Penalty for no PAM
NNW <- 1    # Bonus for a wobble paired neighbour
S <- 1      # Multipier Bonus given to seed

#Set threshold
threshold <- 225

#Define inputs locally initially
DNA <- readDNAStringSet("lambda.fasta")

# #For Cas12
# crRNA <- ("ATTCAGATGATATGACTATCAAGG") #Guide with on target at ~33.5 kb
# Cas <- "Cas12"
# PAM <- ("TTT")

#For Cas9
crRNA <- ("GTGATAAGTGGAATGCCATG") #crRNA1 a.k.a L2
#crRNA <- ("CAGATATAGCCTGGTGGTTC") #crRNA2 a.k.a L4
#crRNA <- ("GAGTCCGAGCAGAAGAAGAA") #crRNA3 a.k.a N.T.
Cas <- "Cas9"
PAM <- ("GG")
APAM <- ("AG")

#PAM Shift required to align coordinates with crRNA windows
if (Cas == "Cas12") {
  PAM_shift <- 4
}

if (Cas == "Cas9") {
  PAM_shift <- -21
}
```

```
46
47 createWindows_std <- dget("createWindows.R")
48 findPAMs <- dget("findPAMs.R")
49
50 #Forward Strand
51 crRNA_windowMatrix <- createWindows_std(DNA, crRNA, "fwd")
52 PAM_windowMatrix <- inner_join(findPAMs(DNA, PAM, "fwd"), findPAMs(DNA, APAM,
   ↪ "fwd"), by = c("seq","coord", "strand"), suffix = c(".PAM", ".APAM"))
53
54 #Align PAM coordinates with crRNA coordinates
55 PAM_windowMatrix <- PAM_windowMatrix %>% mutate(., coord = coord + PAM_shift)
56
57 #Join crRNA with PAM locations
58 f <- inner_join(crRNA_windowMatrix, PAM_windowMatrix, by =
   ↪ c("coord","strand"), suffix = c(".crRNA",".PAM"))
59
60
61 #Reverse Strand
62 crRNA_rev_windowMatrix <- createWindows_std(DNA, crRNA, "rev")
63 PAM_rev_windowMatrix <- inner_join(findPAMs(DNA, PAM, "rev"), findPAMs(DNA,
   ↪ APAM, "rev"), by = c("seq","coord", "strand"), suffix = c(".PAM",
   ↪ ".APAM"))
64
65 #Align PAM coordinates with crRNA coordinates
66 PAM_rev_windowMatrix <- PAM_rev_windowMatrix %>% mutate(., coord = coord +
   ↪ PAM_shift)
67
68 #Join crRNA with PAM locations
69 r_temp <- inner_join(crRNA_rev_windowMatrix, PAM_rev_windowMatrix, by =
   ↪ c("coord", "strand"), suffix = c(".crRNA",".PAM"))
70
71 #Reverse coordinates for r strand
72 r <- r_temp %>% mutate(., coord = (width(DNA)+1) - coord)
73
74 #Plot scores for + strand
75 plot(f$coord, f$score)
76
77 #Plot scores for - strand
78 plot(r$coord, r$score)
79
80 #Combine f and r strand
81 fr <- rbind(f, r)
82
83 #Filter with PAM only
84 fr_PAM <- fr %>% mutate(., score =
   ↪ (score+(PAM_score*perfect.PAM)+(APAM_score*perfect.APAM))*(perfect.PAM+perfect.APAM)
   ↪ #*(perfect.PAM+perfect.APAM) sets score of any site without a PAM to
   ↪ zero
```

```
fr_PAM_filt <- fr_PAM %>% filter(., score >= threshold)

plot(fr_PAM_filt$coord, fr_PAM_filt$score)

#Plot predicted peaks
fr_toConv <- fr_PAM %>% select(., coord, score, strand) %>% mutate(., score =
↪  (score + 100 - threshold))
fr_toConv$score[fr_toConv$score < 100] <- 0

fr_toConv_combined <- spread(fr_toConv, strand, score, fill = 0) %>% mutate(.,
↪  score = fwd + rev)


x <- seq(-500,500,length.out = 2000)
w = 250
t <- exp(-x^2/(w^2))

convScores <- convolve(fr_toConv_combined$score, t, type = "open")
convScores <- convScores[1000:(length(convScores)-1000)]

frResults <- fr_toConv_combined %>% mutate(., convScore = convScores)

plot(frResults$coord, frResults$convScore)
```