Chemically Modified Guide RNAs to Expand CRISPR Functionality



Lapatrada Taemaitree

Oriel College, University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2018

Declarations

The research presented herein is original and was performed and written by the author, Lapatrada Taemaitree under the supervision of Professor Tom Brown at University of Oxford, United Kingdom. Contributions from colleagues are suitably acknowledged.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to distribute or transmit the thesis on the condition that they attribute it. They are not allowed to use the thesis for commercial purposes and they must not alter, transform or build upon it.

Acknowledgements

I am very thankful to my supervisor Professor Tom Brown for giving me the opportunity to conduct research on a very interesting area of science and for the valuable guidance and support throughout my studies. I am also very grateful to Professor Afaf El-Sagheer and Dr. Arun Shivalingam who contributed to this work, provided advice and helpful suggestions throughout.

For work that was not conducted in the Brown lab, I would like to thank: Professor Grigory Dianov (and his lab members: Siobhan Cunniffe and Samuel Hume) from the Department of Oncology for allowing me to use their tissue cell culture facilities and providing help during my time there; Professor Hagan Bayley (and his lab members: Jianfei Feng, Dr. Sejeong Lee and especially Yujia Qing) from the Department of Chemistry for allowing me to perform the protein work in their lab; Dr. Denis Ptchelkine from the Research Complex at Harwell for the helpful advice on the protein work.

Furthermore, I am very lucky to have worked with the Brown group – Ellia, Phoebe, Jack H, Sarah, Ashley, Mei, Carolyn, Piotr, Yssy, Cameron, Sara, Jack C, Sven, Edward, Josh, Yutong, Anna, Agnes, Pietro, Lynda, Pawan, Joanna, Adam, Jinfang and Yun. They have shared both good and bad times of research with me.

I thank the Royal Thai Government for funding my study and research at Oxford.

Finally, and most importantly, I am very thankful to my mum, dad, sister, nanny and friends for their love and belief in me.

Abstract

CRISPR-Cas, a bacterial adaptive immune system against viruses, has been repurposed into a biological research tool that is invaluable due to its simplicity and programmability. Composed of a Cas endonuclease that is directed to its nucleic acid target by a guide RNA (gRNA), CRISPR has been used for genome editing, live-cell imaging and even diagnostics. In this work, the goal is to chemically modify gRNAs to expand CRISPR functionality. A 'split-and-click' strategy for generation of Cas9 gRNAs *via* CuAAC 'click' ligation of shorter oligonucleotides is developed. The approach provides a simple and cost-effective route to individual or pools of single guide RNAs (sgRNAs), while providing scope for further chemical modification to improve sgRNA activity. The click sgRNAs are functional both *in vitro* and in cells, demonstrating artificial linkages can be tolerated in functionally critical regions of sgRNA. To illustrate the capabilities of the system, work towards an on-field forensic analysis of Short Tandem Repeats (STR) is described using CRISPR, Ramified Rolling Circle Amplification (RAM) and Oxford Nanopore sequencing. sgRNAs were also tagged with small molecule fluorophores to provide greater flexibility in colour choice and expand the colour range for live-cell imaging. Imaging of repetitive elements of telomeres is feasible with these constructs if GFP-tagged dCas9 is used for cross validation; due to high cytoplasmic background fluorescence, methods to turn on fluorescence in the nucleus are also explored. Finally, two CRISPR-AsCas12a gRNAs are chemically ligated to yield a 3'-3' dual gRNA with the aim of altering the DNA cleavage pattern of the Cas12a protein. Chemical ligation in this context impaired gRNA activity *in vitro*, however unexpected single-stranded DNase activity is described for AsCas12a.

Abbreviations

2',4'-cEt	2',4'-constrained ethyl bicyclic nucleic acids
ACE	bis(acetoxyethoxy)-methyl ether
A/C/G/T	Adenine / Cytosine / Guanine / Thymine
ac	Acetyl
AMA	Ammonium hydroxide/methylamine
APS	Ammonium persulfate
ATP	Adenosine triphosphate
BCN	bicyclo[6.1.0]nonyne
BFP	Blue Fluorescent Protein
BNA	Bridged Nucleic Acid
BTT	5-(Benzylmercapto)-1*H*-tetrazole or Benzylthiotetrazole
bp	base pair
BSA	Bovine Serum Albumin
bz	benzoyl
BzH	bis(trimethylsilyloxy)benzhydryloxysilyl
Cas9	CRISPR-associated protein 9
Cas12a	CRISPR-associated protein 12a
CE	Capillary Electrophoresis
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CIRCLE-seq	Circularisation for *in vitro* reporting of cleavage effects by sequencing
CMV	Cytomegalovirus

CODIS	Combined DNA Index System
CPG	Controlled Pore Glass
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
CTD	C-terminal domain
CTP	Cytidine triphosphate
CuAAC	Copper(I)-catalysed Azide Alkyne Cycloaddition
Cy3	Cyanine-3
DAPI	4',6-diamidino-2-phenylindole
DBCO	Dibenzocyclooctyne
dCas9	Catalytically inactivated or deactivated or dead Cas9
DIBO	4-dibenzocyclooctynol
DIFO	Difluorocyclooctyne
DMEM	Dulbecco's Modified Eagle Medium
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
DMT	4,4'-dimethoxytrityl
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	deoxyribonucleoside triphosphate
DPBS	Dulbecco's Phosphate-Buffered Saline
ds	double-stranded
DSB	double-strand break
DTT	Dithiothreitol
EDA	Ethylenediamine
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoretic Mobility Shift Assay

ETT	5-Ethylthio-1*H*-tetrazole or Ethylthiotetrazole
FACS	Fluorescence-Activated Cell Sorting
FAM	Fluorescein amidite
FBS	Fetal Bovine Serum
FISH	Fluorescence In Situ Hybridisation
FP	Fluorescence Protein
FPLC	Fast Protein Liquid Chromatography
FRET	Fluorescence Resonance Energy Transfer
FS	2'-fluoro phosphorothioate
fsgRNA	fluorescent single guide RNA
GFP	Green Fluorescent Protein
gRNA	guide RNA
GTP	Guanosine triphosphate
HDR	Homology-directed Repair
HEK	Human Embryonic Kidney
HEPES	Hydroxyethyl piperazineethanesulfonic acid
HFIP	Hexafluoroisopropanol
His-tag	six-histidine tag
HPLC	High Performance Liquid Chromatography
HPLC-MS	High Performance Liquid Chromatography-Mass Spectrometry
HR	Homologous recombination
ib	isobutyryl
IMAC	Immobilised Metal Affinity Chromatography
IVT	*In vitro* transcription
LB	Luria-Bertani
lcaa	long chain alkylamine
LNA	Locked Nucleic Acid

LS	Large Subunit
M	Mass
MBP	Maltose Binding Protein
MMR	Mismatch repair
MP	2'-*O*-methyl-3'-phosphonoacetate
mRNA	messenger RNA
MS	2'-*O*-methyl phosphorothioate
MSP	2'-*O*-methyl thiophosphonoacetate
NA	Numerical Aperture
NHEJ	Non-homologous End Joining
NHS	*N*-hydroxysuccinimide
NLS	Nuclear Localisation Signal
NMI	*N*-methylimidazole
nt	nucleotide
NUC	Nuclease (domain)
PAGE	Polyacrylamide Gel Electrophoresis
PAM	Protospacer-Adjacent Motif
PCR	Polymerase Chain Reaction
PES	Polyethersulfone (membrane)
PFS	Protospacer-flank site
PI	PAM-interacting (subdomain)
PO	Phosphodiester
Pol	Polymerase
PS	Phosphorothioate
p.t.	proton transfer
RAM	Ramified rolling circle amplification
REC	Recognition (domain)

RFP	Red Fluorescent Protein
RNA	Ribonucleic acid
RNase	Ribonuclease
RNP	Ribonucleoprotein
RP-HPLC	Reversed-phase HPLC
rpm	revolutions per minute
RT	Room temperature
RVDs	Repeat-variable di-residues
SBH	Spacer-blocking hairpin
SDS	Sodium Dodecyl Sulfate
sgRNA	single guide RNA
siRNA	Small (or short) interfering RNA
SNP	Single Nucleotide Polymorphism
snRNA	small nuclear RNA
SPAAC	Strain-promoted Azide Alkyne Cycloaddition
ss	single-stranded
SS	Small Subunit
SSB	single-strand break
STR	Short Tandem Repeat
tac	*tert*-butylphenoxyacetyl
TALE	Transcription Activator-like Effector
TALEN	Transcription Activator-like Effector Nuclease
TAMRA	Carboxytetramethylrhodamine
TBDMS	*tert*-butyldimethylsilyl
TBE	Tris-borate-EDTA
TBTA	Tris(benzyltriazolylmethyl)amine
TC	Thionocarbamate

TCA	Trichloroacetic acid
TCEP	Tris(2-carboxyethyl)phosphine
TEA	Triethylamine
TEAA	Triethylammonium acetate
TEAB	Triethylammonium bicarbonate
TEMED	*N,N,N′,N′*-tetramethylethane-1,2-diamine
THF	Tetrahydrofuran
THPTA	Tris(3-hydroxypropyltriazolylmethyl)amine
TOM	triisopropylsilyloxymethyl
tracrRNA	trans-activating CRISPR RNA
TRIS	tris(hydroxymethyl)amino-methane
tRNA	transfer RNA
UPLC	Ultra Performance Liquid Chromatography
UTP	Uridine triphosphate
v/v	volume by volume
w/v	weight by volume
WED	Wedge (subdomain)
ZFN	Zinc-finger nuclease
ZFP	Zinc-finger protein

Table of Contents

1 General Introduction 18

1.1 Genome Editing 18

1.1.1 Early Attempts 18

1.1.2 Protein Tools 19

1.2 Discovery of CRISPR-Cas9 21

1.2.1 Bacterial Adaptive Immunity 21

1.2.2 Genome Editing Tool 23

1.3 Classification of CRISPR-Cas Systems 24

1.4 CRISPR-Cas9 DNA Targeting and Cleavage 26

1.4.1 Loading of gRNA onto Cas9 27

1.4.2 PAM Recognition 28

1.4.3 Local DNA Melting and RNA Strand Invasion 29

1.4.4 Target DNA Cleavage 30

1.5 sgRNA Interactions with Cas9 31

1.6 Modified Cas9 Applications 34

1.6.1 SpCas9 Variants with Different PAM Specificities 34

1.6.2 High-fidelity Cas9 (SpCas9-HF1) 35

1.6.3 Cas9 Nickases 36

1.6.4 Catalytically Inactive Cas9 (dead Cas9 or dCas9) 36

1.6.5 Delivery Methods 37

1.7 Cas12a: An Alternative Tool for Genome Editing 37

1.7.1 crRNA Loading and Target DNA Recognition 38

1.7.2 PAM Recognition 40

1.7.3 DNA Cleavage 41

1.8 Cas12a crRNA 42

1.9 Basics of Nucleic Acid Structure 44
1.10 Chemical Modifications of the gRNA 49
1.11 Aims and Objectives 53

2 Click sgRNA using CRISPR-Cas9 54

2.1 Background 54
2.1.1 Enzymatic Generation of sgRNAs 54
2.1.2 Oligonucleotide Solid-phase Synthesis 56
2.1.3 Click Chemistry 61
2.1.4 Click Chemical Ligation of Oligonucleotides 62
2.2 Aims and Objectives 64
2.3 Click 37–66 sgRNA Constructs 65
2.3.1 Design Strategy 65
2.3.2 2'-*O*-TBDMS RNA Synthesis and Deprotection 66
2.3.3 2'-*O*-TC RNA Synthesis and Deprotection 71
2.3.4 5'-Iodo Conversion of TC-RNA 74
2.3.5 Click Ligation Optimisation 76
2.3.6 Click Ligation of crRNA-tracrRNA 79
2.3.7 *In vitro* DNA Cleavage 83
2.4 Click ~21–79 sgRNA Constructs 85
2.4.1 Design Strategy 85
2.4.2 5'-Azide Conversion of TC-RNA 86
2.4.3 CuAAC Ligation of ~21–79 sgRNAs 89
2.4.4 *In vitro* DNA Cleavage 90
2.4.5 The Importance of a Biocompatible Triazole Linkage 92
2.4.6 In cell Gene Editing Activity 93
2.4.7 Off-target Effects 95
2.4.8 Split-and-click sgRNA for Library Construction 96
2.5 Conclusions and Future Work 97

3 Fluorescent sgRNAs using CRISPR-dCas9 99

3.1 Background 99

3.1.1 Traditional Approaches 99

3.1.2 Sequence-specific Live-cell Imaging 100

3.2 Aims and Objectives 102

3.3 First-generation Fluorescent sgRNA (fsgRNA I) 102

3.3.1 Design Strategy 102

3.3.2 Construct Synthesis 105

3.4. Second-generation Fluorescent sgRNA (fsgRNA II) 107

3.4.1 Design Strategy and Synthesis 107

3.4.2 *In vitro* Cas9-fsgRNA II Complex Formation 110

3.4.3 Live-cell Imaging 111

3.5 Optimisation of sgRNA Structure 113

3.6 dCas9-fsgRNA Ribonucleoprotein Complex for Imaging 116

3.6.1 dCas9-3xGFP Expression and Purification 116

3.6.2 Evaluating dCas9-3xGFP Activity 119

3.6.3 Ribonucleoprotein Complex Transfection 120

3.7 Fluorescent Click sgRNAs 121

3.8 Cy3-labelled IVT sgRNA 123

3.9 RNase H-activated fgRNA 126

3.9.1 Design Strategy 126

3.9.2 Live-cell Imaging 127

3.10 Conclusions and Future Work 130

4 3'–3' Dual gRNAs using CRISPR-AsCas12a 131

4.1 Background 131

4.1.1 Controlling Gene Editing Outcomes 131

4.2 Hypothesis and Aims 133

4.3 Dual gRNA Generation 134

4.4 AsCas12a Lysate 136

4.4.1 Lysate Preparation 136
4.4.2 Lysate Activity Tests 136
4.5 Bacterial Production of AsCas12a 138
4.5.1 Single-step Purification 138
4.5.2 Single-step Purification Protein Activity Tests 140
4.5.3 Multi-step Purification 141
4.5.4 Multi-step Purification Protein Activity Tests 142
4.6 *In vitro* DNA Cleavage 143
4.6.1 Dual gRNA Activity 143
4.6.2 AsCas12a Non-specific DNase Activity 146
4.7 Conclusions and Future Work 148

5 Nanopore STR Analysis Towards CRISPR Applications 149
5.1 Background 149
5.1.1 Short Tandem Repeats 149
5.1.2 Current Analysis Methods 150
5.1.3 Nanopore Sequencing and Ramified Rolling Circle Amplification 151
5.1.4 STR Extraction using CRISPR 151
5.2 Aims and Objectives 152
5.3 TH01 Ramified Rolling Circle Amplification 154
5.3.1 Generation of Circular DNA Template 154
5.3.2 RAM 154
5.4 Nanopore Sequencing 155
5.4.1 Library Preparation 155
5.4.2 Data Analysis 156
5.5 Conclusions and Future Work 158

6 Conclusions 160
6.1 Summary 160
6.2 Outlook 162

7 Experimental Methods and Materials 166

7.1 Oligonucleotide Synthesis 166

7.1.1 DNA Synthesis 166

7.1.2 RNA Synthesis 167

7.1.3 Modified Monomers 168

7.1.4 Modified Solid Supports 168

7.2 Solid-phase Oligonucleotide Modification 168

7.2.1 5'-Azide Conversion of RNA 168

7.2.2 Selective β-cyanoethyl Removal 169

7.3 Oligonucleotide Deprotection 169

7.3.1 DNA Deprotection 169

7.3.2 2'-*O*-TBDMS RNA Deprotection 169

7.3.3 2'-*O*-TC RNA Deprotection 170

7.4 Oligonucleotide Purification 170

7.4.1 Desalting 170

7.4.2 RP-HPLC 170

7.4.3 Denaturing PAGE 170

7.5 Oligonucleotide Mass Spectrometry 171

7.6 Post-synthetic Oligonucleotide Labelling 171

7.7 Click Oligonucleotide Ligation 172

7.7.1 Non-templated CuAAC 172

7.7.2 Templated CuAAC 172

7.7.3 Non-templated SPAAC 172

7.7.4 Templated SPAAC 173

7.8 *In vitro* Transcription 173

7.9 IVT sgRNA Azido-dA Tailing 174

7.10 Azido-IVT sgRNA Labelling 174

7.11 Plasmid Construction 174

7.11.1 pSpCas9(BB)-2A-Puro-v2-Broccoli 174

7.11.2 pDEST-his-dCas9-3xGFP 174

7.12 Plasmid Production 175

7.12.1 Transformation 175
7.12.2 Small-scale Plasmid DNA Preparation 176
7.12.3 Enzymatic Digestion of Plasmid DNA 176
7.12.4 Large-scale Plasmid DNA Preparation 177
7.13 Bacterial Protein Expression 177
7.14 Protein Purification 178
7.14.1 dCas9-3xGFP 178
7.14.1.1 Nickel-Immobilised Metal Affinity Chromatography 178
7.14.1.2 Size-exclusion Spin-column Buffer Exchange 179
7.14.1.3 Cation Exchange Chromatography 179
7.14.2 AsCas12a 179
7.14.2.1 Single-step Purification 179
7.14.2.2 Three-step Purification 180
7.15 SDS-PAGE 182
7.16 Native PAGE 183
7.17 Electrophoretic Mobility Shift Assay 183
7.17.1 Cas9-fsgRNA 183
7.17.2 dCas9-3xGFP – IVT sgRNA 183
7.18 Agarose Gel Electrophoresis 184
7.19 *In vitro* Cleavage Assays 184
7.19.1 Cas9 184
7.19.2 AsCas12a Lysate 184
7.19.3 Purified AsCas12a 185
7.19.4 Commercial AsCas12a 185
7.20 Real-time Fluorescence Measurements 185
7.21 Cell Culture 186
7.22 Transfection 186
7.22.1 Click sgRNA Gene Editing 186
7.22.2 Live-cell Imaging 186
7.22.3 AsCas12a Protein Expression 187
7.22.4 Lipid Transfection of Ribonucleoprotein Complex 187

7.22.5 Electroporation of Ribonucleoprotein Complex 188
7.23 AsCas12a Cell Lysate Preparation 188
7.24 Confocal Microscopy 188
7.25 Genomic DNA Extraction 189
7.26 T7E1 Assay 189
7.27 Nanopore Sequencing 190
7.27.1 Circularisation of DNA Template 190
7.27.2 Ramified Rolling Circle Amplification 190
7.27.3 Short Tandem Repeat PCR 191
7.27.4 Library Preparation for Sequencing 191
7.27.5 Flowcell Preparation 192
7.27.6 Sample Loading 192
7.27.7 Data Analysis 192

8 Appendix 194

9 References 224

General Introduction

1.1 Genome Editing

DNA stores the genetic information of almost all known organisms. The ability to precisely target DNA and alter its function has been a long-term goal of scientists. Until recently, methods to edit the genome were complex and unintuitive, but the discovery of RNA guided CRISPR-Cas systems has transformed the situation. The focus of this work is to expand the function of the system in order to ask better questions and simplify its use.

1.1.1 Early Attempts

The first genome editing attempts exploited homologous recombination (HR) to integrate foreign DNA into yeast[1] and human genomes[2–4] in order to study function of a particular gene. However, a number of limitations were observed including an extremely low rate of spontaneous DNA incorporation (1 in 10^3–10^9 cells)[4], cell cycle dependence of HR[5] and random integration at undesired sites[6]. This limited the use of this method in many applications.

In a significant step forward, it was demonstrated that the HR events could be massively increased (> 1,000x) when double-strand breaks (DSBs) are introduced at the site of integration[7]. The resulting DSBs are processed by two competing DNA repair pathways: non-homologous end joining (NHEJ) (**Figure 1.1A**) and homology-directed repair (HDR) (**Figure 1.1B**). In NHEJ, the two broken ends of the DNA are ligated without a homologous DNA template. This mechanism is error-prone as substitution and insertion/deletion (indel) mutations of varying length are introduced at the site of the break during ligation[8]. This results in frameshift mutations and potentially loss of gene function (or gene knockout). In contrast, HDR is more precise than NHEJ but it requires a homologous DNA template to function. HDR allows the correct, or desired, genetic information on a DNA template with homologous flanking regions to the DSB to be integrated into the genome through HR. This leads to gene correction or precise gene knock-in.

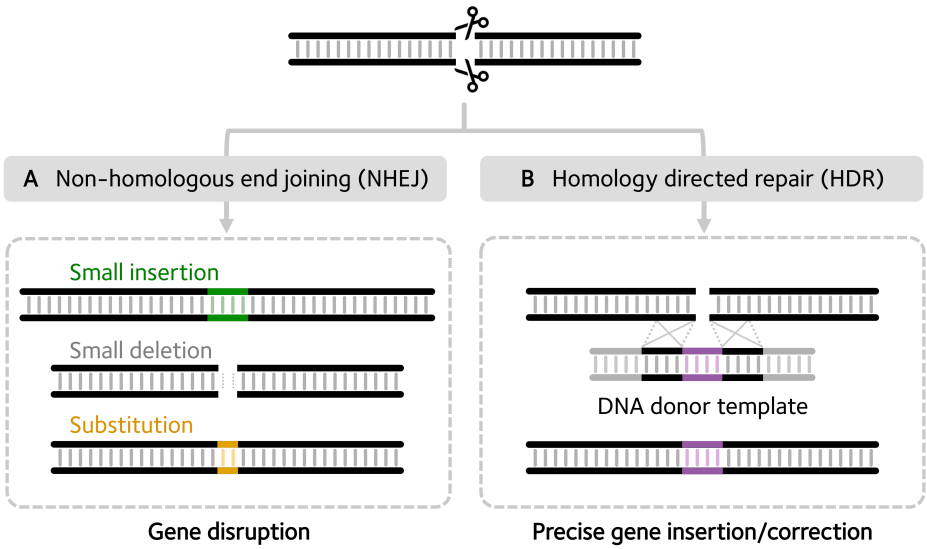


Figure 1.1 | DNA repair of DSBs by (A) error-prone NHEJ and (B) precise HDR.

1.1.2 Protein Tools

Based on this knowledge, multiple genome editing tools have been developed to recognise specific DNA sequences and introduce targeted DSBs. First generation genome editing tools, meganucleases[7] (**Figure 1.2A**), relied on naturally occurring endonucleases that recognise 14–40 bp DNA sequences. Although the enzyme could successfully introduce sequence-specific DSBs,

a large number of targets were inaccessible due to a limited number of known meganucleases. Attempts to expand meganuclease targets have been made[9–11], yet only small gains were possible.

This led to the development of second-generation zinc-finger nucleases (ZFNs, **Figure 1.2B**), which are generated by artificially fusing a zinc-finger DNA recognition unit to a FokI endonuclease that lacks its DNA recognition domain[12]. The zinc-finger DNA recognition unit consists of an array of zinc-finger proteins (ZFPs), small zinc-dependent ~30 amino acid sequences that specifically bind to 3–4 bp of DNA[13]. By mixing and matching different ZFPs, DNA targeting specificity can be easily expanded to sequences 9 to 18 bp in length[14]. Using an inverted pair of ZFNs that bind to opposite strands of double-stranded (ds) DNA, two adjacent FokI domains can dimerise resulting in DSB at a specific location in human cells. This substantially increased targeted gene modification *via* HR[15,16]. However, modular assembly of ZFPs is not straightforward. The spacing and overlapping base specificity of adjacent ZFPs requires careful design and extensive screening to obtain the desired target specificity[17].

To overcome this problem, ZFPs were replaced with transcription activator-like effectors (TALEs)[18,19] to create transcription activator-like effector nucleases (TALENs, **Figure 1.2C**)[20–23]. TALEs are also 33–35 amino acids but contain only two hypervariable residues ("repeat-variable di-residues" or RVDs) which define the specificity of the module for a single base. Their modular assembly by molecular cloning is more straightforward[24], but the nucleic acids encoding TALENs are more repeat rich and far larger than ZFNs (each TALE module specifies one base rather than the three bases specified by a zinc finger of similar size). This makes lenti-viral packaging of TALENs for *in vivo* applications more complex[25]. Furthermore, TALENs (and also ZFNs) require two sets of chimeric proteins per DNA target. As a result, a large DNA cassette (~1500 bp) is needed for the production of proteins to target a single site, making multiplexed gene targeting very challenging. In combination, these design and production limitations restrict the wide-spread of TALENs despite their ability to precisely and efficiently edit the human genome.

Recently a new gene-editing CRISPR-Cas nuclease[26] was discovered (**Figure 1.2D**). In this system, only a single Cas9–gRNA complex is required per target. Moreover, a guide RNA (gRNA) rather than protein motifs confers DNA sequence specificity. It is much simpler to programme RNAs that follow Watson-Crick base-pair rules and introduce multiple gRNAs into cells exogenously, thus massively simplifying multiplexing[27,28]. The simplicity and programmability of CRISPR has made it the method of choice for genome editing applications.

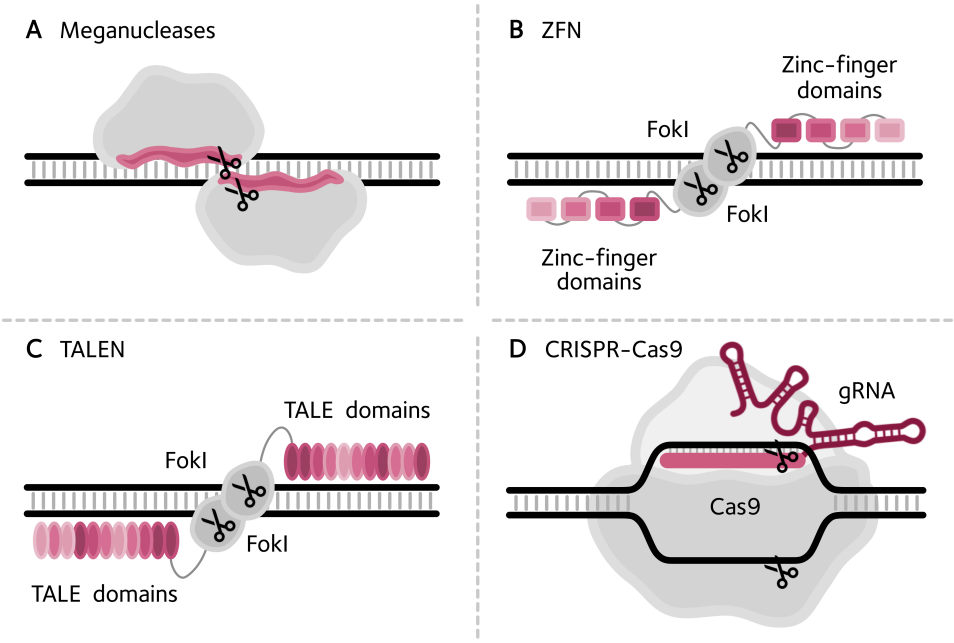


Figure 1.2 | Genome editing tools. (A) Meganucleases, (B) ZFN, (C) TALEN and (D) CRISPR-Cas consist of a DNA recognition domain (pink) and a nuclease domain (grey) that allow site-specific DSBs to be generated in cells, enabling gene editing.

1.2 Discovery of CRISPR-Cas9

1.2.1 Bacterial Adaptive Immunity

In 1987, during the analysis of the *iap* gene in bacteria (*Escherichia coli*), Nakata *et al.* found a strange series of unique sequences, 20–58 bp in length, flanked by short 21–40 bp invariant palindromic repeats downstream of the gene[29]. The biological function of these unusual sequences could not be determined, due to insufficient DNA sequence data at that time. This strange pattern was later found in archaea (*Haloferax mediterranei*)[30] and, with the significant

progress made in Sanger DNA sequencing around 2000, in more than 40% of bacteria and 90% of archaea[31]. This resulted in two key observations.

The first observation was that a group of genes (*Cas*) are well conserved and located adjacent to the strange pattern. Based on sequence homology analysis, Cas genes were presumably involved in DNA replication and gene maintenance as they encoded a helicase, an enzyme that unwinds DNA, and a nuclease, an enzyme that cleaves DNA[32]. Jansen *et al*. later proposed the term CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) to describe this unusual DNA pattern which was widely adopted by the field.

The second observation was that the unique sequences, termed "spacers" were identical to sequences of phage genomes and other mobile genetic elements[33–35]. In addition, bacteria or archaea possessing these spacer sequences were resistant to infection by corresponding viruses[33,34,36], suggesting that CRISPR-Cas could possibly serve as an immunological memory and a defense system in bacteria against phage infection. Barrangou *et al*. provided the first experimental evidence for this[37]. They showed that when phage DNA was integrated into the CRISPR array of *Streptococcus thermophilus* as a spacer sequence, the bacteria became phage resistant and that single-base changes in phage genomes allowed escape from the CRISPR-based immunity. These findings indicate that the acquired resistance of bacteria to phage infection was specific and achieved through precise nucleic-acid complementarity of the spacer sequence and the homologous sequence of the phage genome.

These studies drew attention to the field, which then focused on identifying the components required for CRISPR to function and its biochemical mechanism (**Figure 1.3**). Firstly, the spacer sequences in the CRISPR locus are transcribed to generate short CRISPR RNAs (crRNAs) that guide the activity of Cas enzymes[38]. Secondly, a short motif on the target DNA located next to the spacer, called the protospacer-adjacent motif (PAM), was found to be crucial for CRISPR to function[39,40]. Thirdly, Cas enzymes cleaved the DNA of the invading virus in a double-stranded manner and at a specific site within the spacer region[41]. Finally, specifically for the Cas9-dependent CRISPR system, another RNA called trans-activating crRNA (tracrRNA) was required in

addition to crRNA. During crRNA maturation, tracrRNA hybridises to the pre-crRNA transcript from the CRISPR locus. Following the cleavage by RNase III in the presence of Cas9, a mature crRNA–tracrRNA hybrid is generated and forms a complex with Cas9, which subsequently digests invasive dsDNA inhibiting phage infection[42]. This Cas9-dependent system was classified as a type II CRISPR system. Other types of CRISPR systems were also discovered at a similar time[43,44] but this type was particularly attractive due to its simplicity – it relies on a single Cas9 protein rather than a set of Cas proteins.

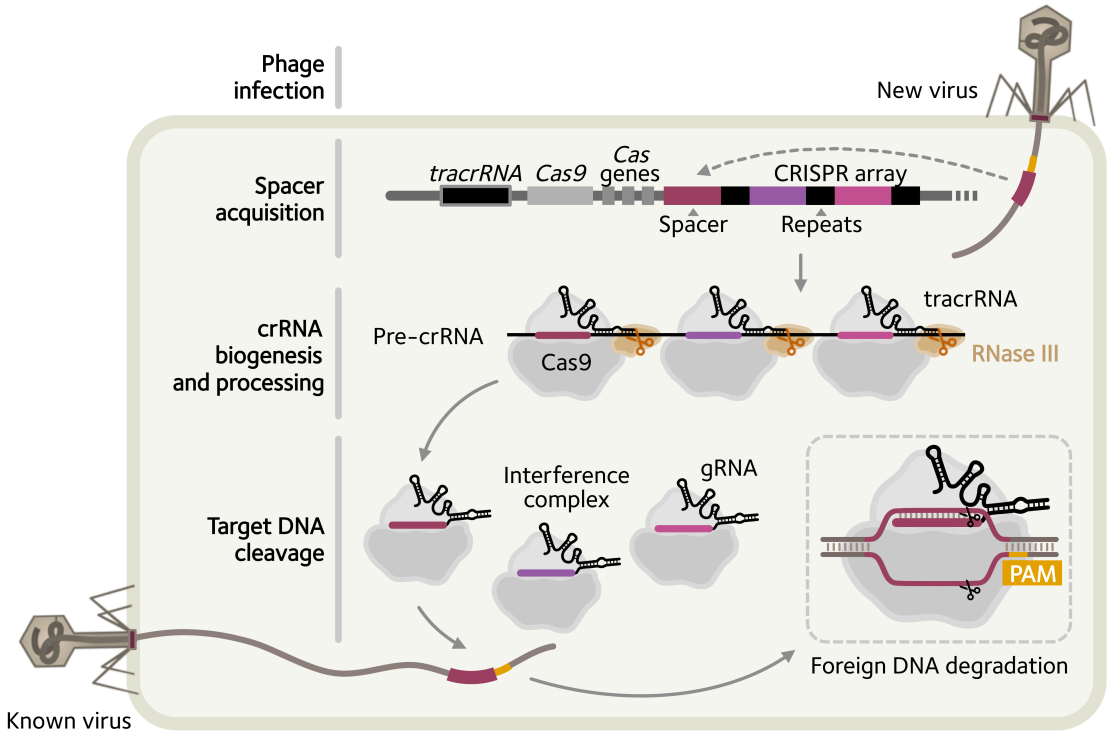


Figure 1.3 | CRISPR-Cas9 adaptive immunity in bacteria. The system consists of three stages: spacer acquisition, crRNA biogenesis and processing and DNA cleavage of phage DNA containing PAM motifs (orange) by the Cas9–gRNA complex.

1.2.2 Genome Editing Tool

The idea of using the CRISPR-Cas9 system as a gene engineering tool then arose from the pivotal findings that the system is transferrable from one organism to another[45] and the system can be programmed to cleave any desired DNA target *in vitro* by customising the crRNA sequence[26,46]. Crucially, the system could be further simplified by fusing crRNA and tracrRNA into a chimeric

single guide RNA (sgRNA)[26] and in 2013, CRISPR-based genome editing in mammalian cells was successfully demonstrated using both crRNA–tracrRNA hybrids and sgRNAs[28,47]. These landmark studies resulted in extensive structural, biophysical and biochemical studies[48–53]. The molecular mechanisms of RNA-guided DNA targeting and DNA cleavage by Cas9 endonuclease in genome editing are now well understood.

1.3 Classification of CRISPR-Cas Systems

To date, CRISPR systems have been divided into two classes[54]: class 1 involves a complex of multiple proteins[38,55] and class 2 relies on a single Cas protein such as Cas9[26,46]. These classes can be subdivided into types (I, III and IV in class 1 and II, V and VI in class 2), depending on the exact proteins required for spacer acquisition, pre-crRNA processing, and most importantly for interference, which can be repurposed for genome editing (**Figure 1.4**).

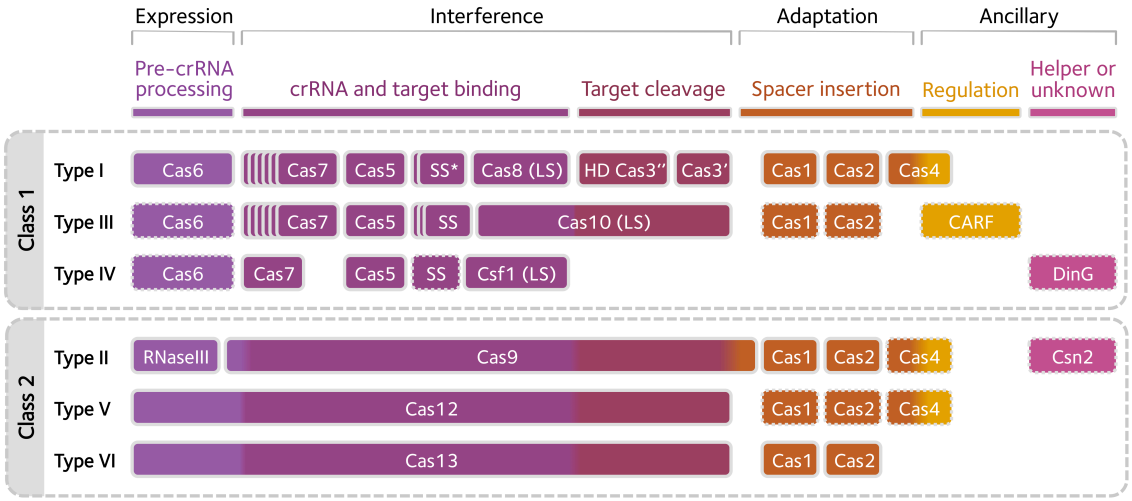


Figure 1.4 | Classification of the CRISPR-Cas systems. SS stands for small subunit. LS stands for large subunit. SS* indicates a putative small subunit that is fused to the Cas8 large subunit for type I. Cas3 is a fusion of the nuclease HD Cas3" and the helicase Cas3' which are encoded by two distinct genes. Dash outlines represent functionally dispensable components. Cas6 with a dashed line for type III indicates that the protein is provided by other CRISPR-Cas loci. The multi-colour proteins reflect their overlapping functions in different stages of the CRISPR-Cas system. This figure was adapted from Makarova *et al.*, 2015[54] and Abudayyeh *et al.*, 2016[56].

The components necessary for the interference step: crRNA binding and cleavage of targets, substrate preferences and PAM requirements are summarised in **Table 1.1**.

Type	Effector nuclease	gRNA (nt)		Target	PAM (5'–3')	Cleavage mechanism
		Spacer	All			
I	Cascade-Cas3[57]	32	61	dsDNA	ATG	Non-target ssDNA cleavage then fully digest dsDNA
III	Csm/Cmr-Cas10 complex [58,59]	39	45	ssRNA/ ssDNA	–	Nontemplate (transcription) ssDNA cleavage coupled to ssRNA (transcript) cleavage
IV	Csf1 complex[54]	N/A	N/A	N/A	N/A	N/A
II	Cas9[41,42]	20	≥100	dsDNA	NGG	Specific blunt-ended DSB 3 bp upstream 5' of PAM
V	Cas12a[60]	20	42	dsDNA	TTTV	Specific DSB with 5' 5-nt overhangs
VI	Cas13a[56]	28	59	ssRNA	3'-A, U, or C	Specific RNA cleavage then non-specific RNA digestion

Table 1.1 | Summary of the key features of different CRISPR-Cas systems.

In terms of targeting, type I, II and V effector nucleases are guided by crRNA to bind and cleave dsDNA following PAM recognition[40,61,62]. In contrast, type III nucleases digest both ssDNA and RNA independently of a PAM motif[59,63]. Since the type VI crRNA–effector protein complex targets ssRNA, it does not recognise the PAM sequence, but instead usually uses a sequence called protospacer-flank site (PFS), which is located on the RNA target adjacent to the spacer sequence[56]. Interestingly, some orthologs (e.g. *Leptotrichia wadei*), however, do not have this requirement[64].

For substrate cleavage, type I, II, III and V nucleases cleave the substrate at specific sites, however the type VI protein (Cas13a) digests the target RNA specifically and completely[56,65] followed by non-specific digestion of other surrounding RNAs[56].

Finally for gRNA requirements, Cas9 in the native CRISPR system forms an effector complex composed of a 40-mer crRNA and a 86-mer tracrRNA with the 20-mer spacer crRNA base-pairing to the target DNA. On the other hand, all other types employ a single crRNA with slight variation in length from 40–58 nt and, similar to type II, about 20 nt are responsible for target hybridisation.

For gene editing, class 2 systems are ideal as this entire interference process could be achieved using a single protein (Cas9 in type II and Cas12a in type V). In this thesis, Cas9 and Cas12a will be focused on as they are widely studied and used in the gene editing field.

1.4 CRISPR-Cas9 DNA Targeting and Cleavage

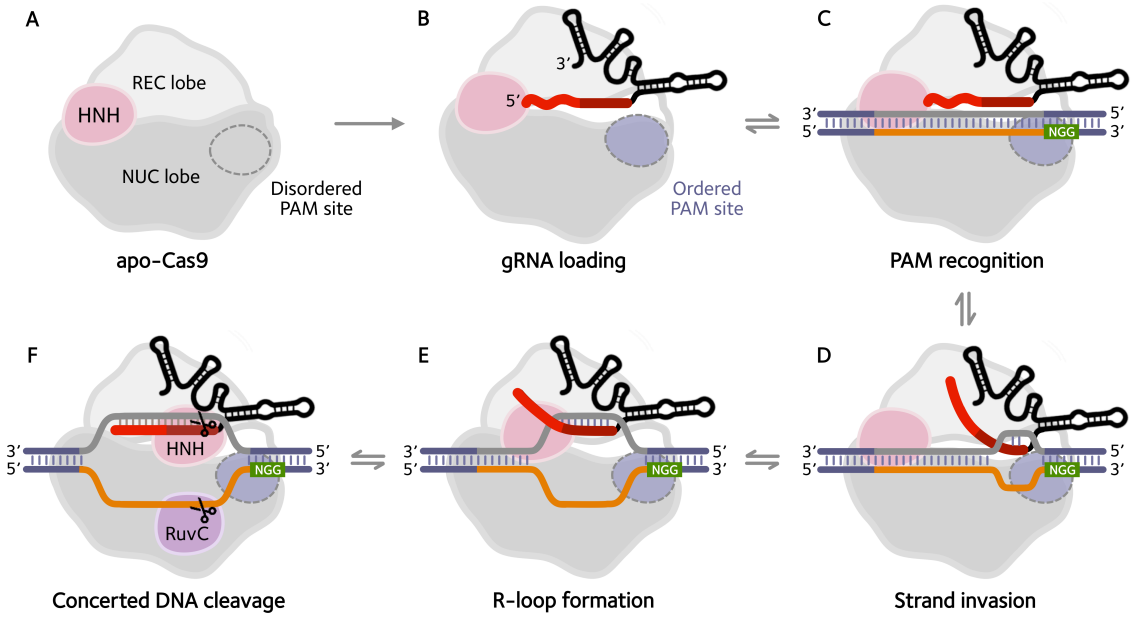


Figure 1.5 | The proposed mechanism of genome editing by CRISPR-Cas9[66]. In the apo state (A), the PAM-interacting domain is disordered. Upon sgRNA loading, Cas9 undergoes a substantial conformational change, causing the PAM-interacting domain to become ordered (B) for PAM recognition (C). The Cas9–sgRNA complex will dissociate from the dsDNA that lacks PAM. If the PAM is recognised, local DNA (PAM-proximal) melting occurs (D), facilitating RNA strand invasion (E). Full hybridisation of the RNA and the target DNA induces a further conformational change of Cas9 to adopt an active conformation, allowing concerted DNA cleavage (F).

Cas9 RNA-guided cleavage of DNA is a multi-step process (**Figure 1.5**) that involves: 1) Cas9 binding to gRNA, 2) ribonucleoprotein (RNP) recognition of PAM on the DNA, 3) strand invasion and gRNA–DNA duplex formation, and 4) cleavage of both DNA strands by Cas9.

1.4.1 Loading of gRNA onto Cas9

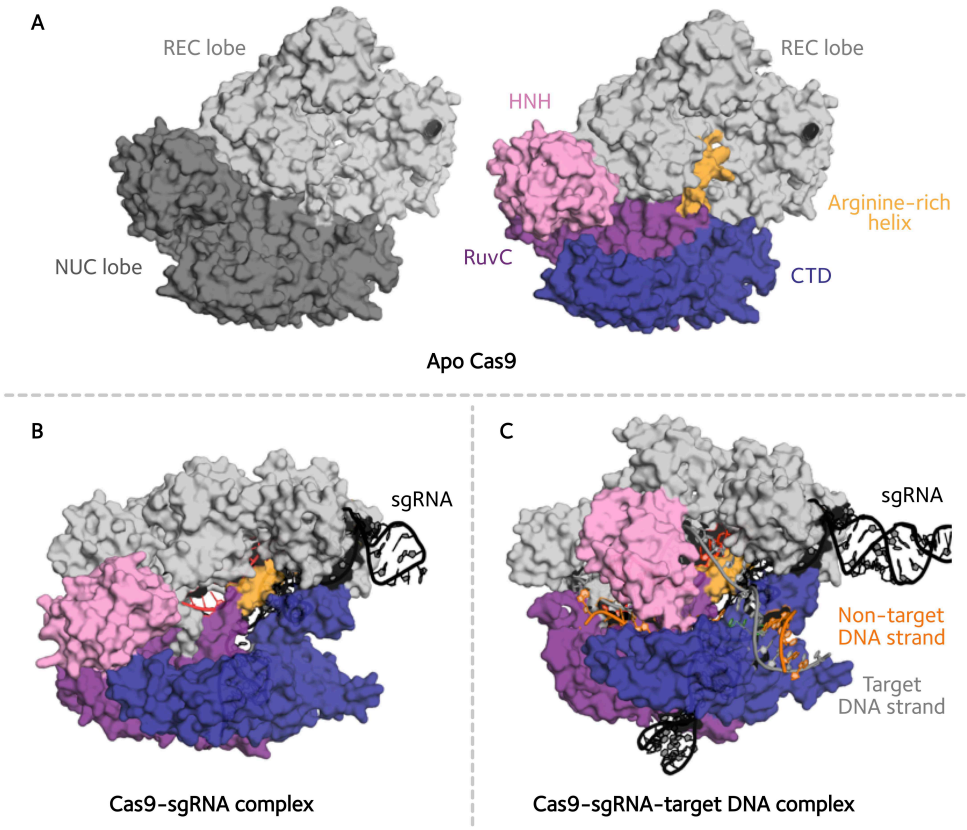


Figure 1.6 | Surface representations of Cas9. Cas9 in the apo state (A) adopts the inactive conformation (PDB: 4CMP[49]). The binding of sgRNA (B) drives Cas9 into a DNA-recognition competent conformation (PDB: 4ZT0[51]). The Cas9–sgRNA complex undergoes a final conformational change upon DNA target binding (C), enabling Cas9 to form an active conformation for DNA target cleavage (PDB: 5F9R[51]).

The crystal structure of *Streptococcus pyogenes* (Sp) Cas9 in the absence of sgRNA ("apo-Cas9", PDB: 4CMP[49]) reveals the protein has a bi-lobe architecture, consisting of a target recognition (REC) lobe and a nuclease (NUC) lobe (**Figure 1.6A**). The NUC lobe contains the HNH and RuvC nuclease domains responsible for cleavage of the target (base-paired with the guide RNA) and non-target DNA strands respectively, as well as a C-terminal domain (CTD) that plays a crucial role in PAM recognition. Comparison of apo-Cas9 and sgRNA-bound Cas9 (PDB: 4ZT0[51]) (**Figure 1.6B**) structures shows that the PAM recognition module in the CTD domain is disordered in the absence of sgRNA, inhibiting non-specific DNA cleavage; if Cas9 binds DNA non-specifically, it departs rapidly[67]. A large conformational change in the REC lobe is observed upon sgRNA binding[49,51]. In addition, the PAM-interacting module of Cas9 that is previously disordered in the

absence of sgRNA is now repositioned for DNA interaction[50]. These observations confirm Cas9 is an RNA-dependent endonuclease. More details of sgRNA interactions with Cas9 will be described in **Section 1.5**.

1.4.2 PAM Recognition

Even if the sequence of the sgRNA is perfectly complementary to that of the target DNA, Cas9 cannot cleave DNA unless the PAM sequence is present. The PAM sequence is crucial for microbes to discriminate between self (in the CRISPR array) and non-self (phage) DNA[68]. If there is a mutation in PAM, the DNA sequence is not recognised *in vitro*[26] and the microbes will not be able to protect themselves from phage infection *in vivo*[69].

Single-molecule studies demonstrated that Cas9 searches for the target DNA *via* three-dimensional collisions. When Cas9 collides with the DNA substrate that lacks PAM, it dissociates immediately. However, the lifetime of Cas9 on target DNA increases when the DNA substrate contains PAM and the sequence is complementary to the sgRNA[67].

The superimposition of the Cas9–sgRNA and the Cas9–sgRNA bound to a PAM duplex (PDB: 4UN3[50]) shows two key residues (R1333 and R1335) are responsible for PAM recognition motif of Cas9 (**Figure 1.7**). These arginine residues form hydrogen bonds with the two guanines in the 5'-NGG-3' PAM sequence of the non-target DNA strand. N can be any bases as it does not make any contact with the PAM recognition site and no interactions are observed between Cas9 and the sequence complementary to PAM (5'-CCN-3'). Thus, 5'-CCN-3' mutations do not affect the target recognition[26,67].

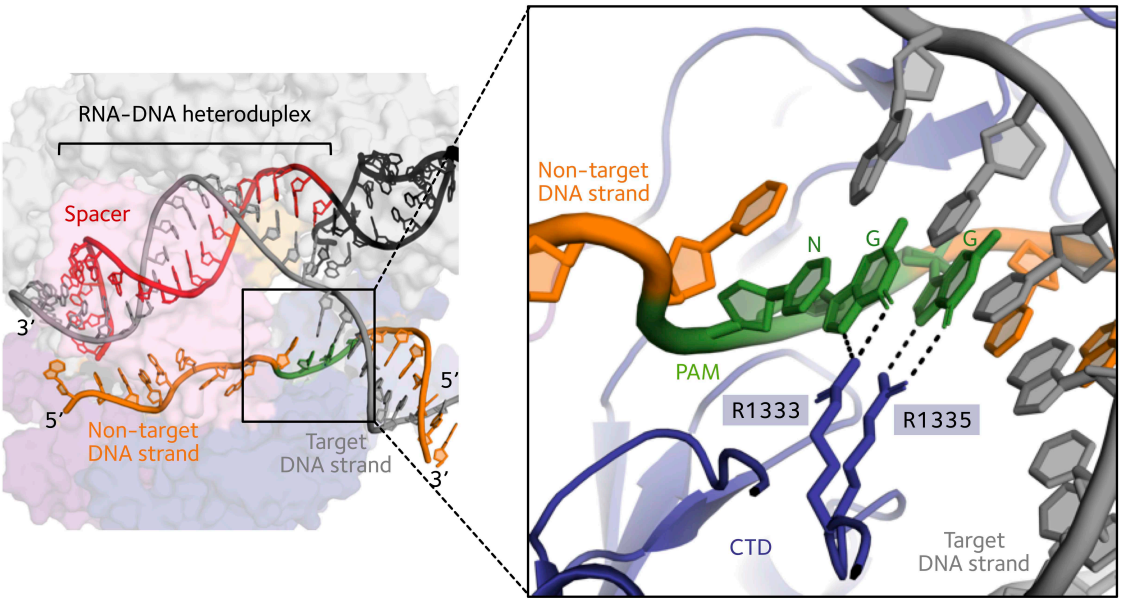


Figure 1.7 | Close-up view of PAM recognition by Cas9. The key residues for PAM recognition are R1333 and R1335 within the CTD domain, which form hydrogen bonds with the two guanines in the 5'-NGG-3' PAM motif (PDB: 5F9R[51]).

1.4.3 Local DNA Melting and RNA Strand Invasion

After PAM recognition, the phosphate group immediately upstream of PAM on the non-target strand (the +1 phosphate position) interacts with K1107 and S1109 residues forming a phosphate lock loop[50] (**Figure 1.8A**). This causes local DNA unwinding, facilitating sgRNA invasion and hybridisation to the target DNA strand to form an RNA–DNA heteroduplex (A-form-like structure)[48]. The non-target DNA strand is also displaced and its PAM-proximal region is stabilised by base stacking onto the PAM duplex[52], resulting in R-loop formation[70] (**Figure 1.8B**). This observation is supported by previous biochemical studies[67]; to mimic the ribonucleoprotein–DNA intermediate before R-loop formation, DNA mismatches were introduced at position 1 and 2 to form a small 2-nt DNA bubble immediately upstream of the PAM site. These mismatches are also not complementary to the sgRNA. Despite this, it was found that the remaining spacer sequence of the sgRNA can still hybridise to target DNA and cause DNA cleavage at a similar level to perfectly complementary sgRNA. The small bubble could bypass the initial RNA–DNA heteroduplex formation that is usually required to drive DNA strand separation and subsequent RNA–DNA hybridisation[67].

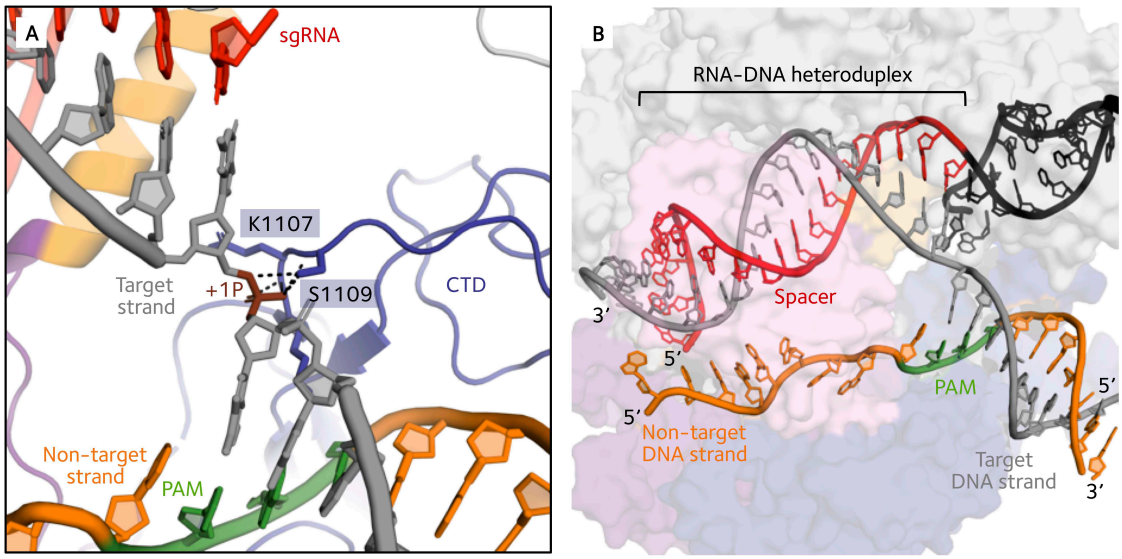


Figure 1.8 | Close-up view of the phosphate lock loop and R-loop formation. Upon PAM recognition, the phosphate upstream the PAM motif (+1 phosphate in brown) forms hydrogen bonds with K1107 and S1109 residues within the CTD domain, facilitating the target DNA melting and rotating upwards to hybridise with sgRNA (A) (PDB: 5F9R[51]). Subsequent RNA strand invasion into the target DNA duplex results in R-loop formation (B) (PDB: 5F9R[51]).

1.4.4 Target DNA Cleavage

Cas9–sgRNA binding to DNA does not always trigger DNA cleavage. From the genome-wide binding analysis by ChIP-seq, it was found that sgRNA with only a 5-nt spacer could cause off-target DNA binding, but these sites are not cut[71]; the extensive base-paring of the full-length spacer to the target DNA was necessary for DNA cleavage. Specifically, the interactions from the 14th to 17th base of the spacer are critical for stabilising the active conformation of Cas9 that enables DNA cleavage and therefore gene editing[72]. sgRNAs with 11 to 15-nt spacers, on the other hand, allow Cas9 to only bind to the target DNA and have been used to regulate gene expression instead[73]. sgRNAs with spacers truncated slightly from 20-nt to 17–18 nt in length greatly reducing off-target effects[74], possibly by reducing non-specific base pairing.

Comparison of Cas9–sgRNA structure with (PDB: 5F9R[51]) and without (PDB: 4ZT0[51]) dsDNA substrate reveals a further conformational change of Cas9 upon target DNA binding (**Figure 1.6B** and **1.6C**). The HNH nuclease domain moves towards the target DNA strand with the one-Mg^{2+}-ion-bound active site of D839, H840, N854 and N863 residues being ~10 Å away from the scissile

phosphate of the DNA (**Figure 1.9A**). This large movement of the HNH domain causes the two hinge linkers that connect the HNH and the RuvC domains to move which in turn results in the displacement of the non-target DNA strand close to the other two-M^{2+}-ion-bound catalytic group of D10, E762, H983 and D986 in the RuvC domain[52] (**Figure 1.9B**, M^{2+} is Mn^{2+} in the crystal structure but Mg^{2+} in real systems). Both strands of the DNA are then cleaved simultaneously between the 3rd and 4th bases proximal to PAM in a metal-ion-dependent manner, generating a blunt-ended double-strand break (DSB)[52] (**Figure 1.10**).

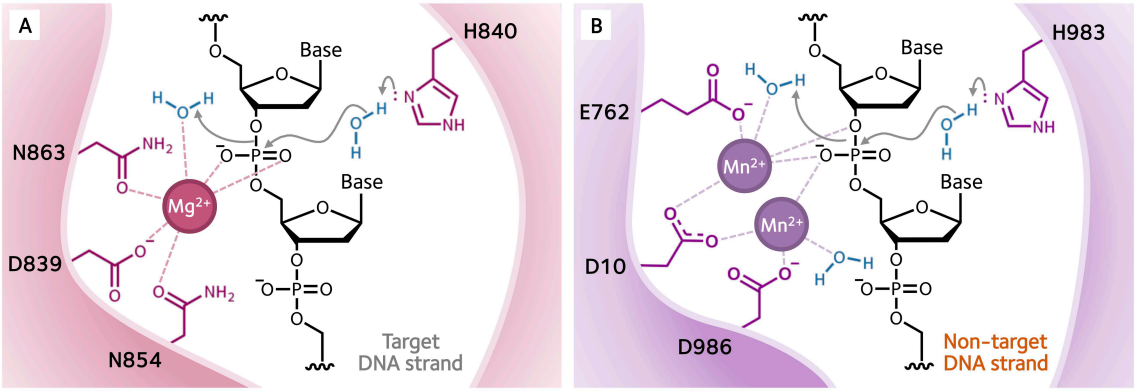


Figure 1.9 | Metal-dependent DNA cleavage. Active site of the HNH domain required a Mg^{2+} ion for target DNA strand cleavage (A) whereas two M^{2+} ions are needed for non-target DNA strand cleavage at the active site of the RuvC domain (B). This figure was modified from Chen and Doudna, 2017[75].

1.5 sgRNA Interactions with Cas9

The sgRNA structure can be divided into six different parts (from 5' to 3' termini): a 20-nt spacer sequence (for target DNA hybridisation), a repeat–antirepeat duplex, a tetraloop (connects crRNA to tracrRNA), stem loop 1, a linker, stem loop 2 and stem loop 3 (**Figure 1.10**).

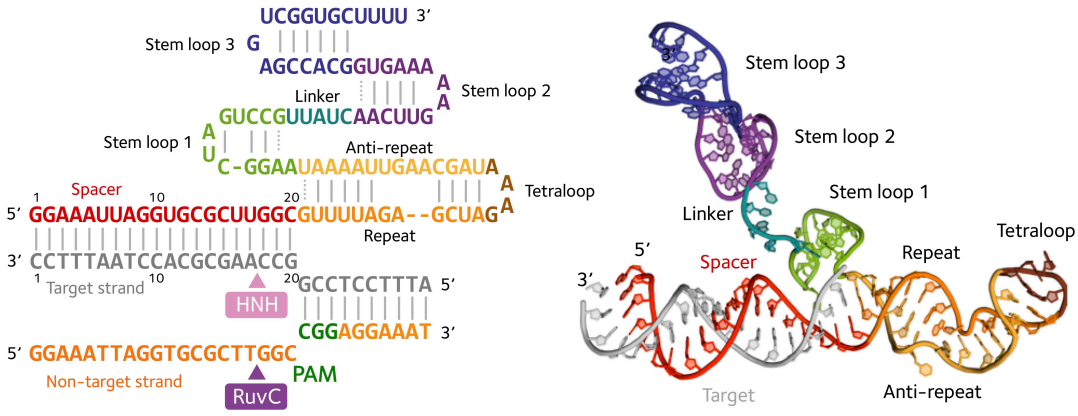


Figure 1.10 | Schematic representation of sgRNA in the CRISPR-Cas9 system. sgRNA consists of a spacer (red), repeat–antirepeat duplex (orange), tetraloop (brown), stem loop 1 (green), linker (teal), stem loop 2 (purple) and stem loop 3 (blue). DNA cleavage occurs between the 17th and 18th position of the spacer region. PDB: 5F9R[51].

Cas9 forms extensive interactions with the PAM-proximal region, repeat–antirepeat duplex, stem loop 1 and the linker (**Figure 1.11**), which are essential for its nuclease activity. Deletion of stem loops 2 and 3 slightly decrease DNA cleavage efficiency *in vitro*[26], but significantly inhibit Cas9 activity in cells[48], possibly due to destabilisation of the ribonucleoprotein complex[76,77]. The tetraloop that protrudes outside of the Cas9–sgRNA complex can be extensively manipulated without affecting Cas9 activity. For example, the replacement of the tetraloop with RNA aptamers enables the recruitment of transcription factors to specific loci, leading to the activation of target genes[78,79].

The 20-nt spacer sequence of sgRNA contains a "seed" region, which is defined as the 10 bases proximal to PAM, and a "non-seed" region, which contains the remaining bases. During target recognition, the A-form seed-region heteroduplex forms many direct contacts with Cas9. The negatively-charged backbone of seed-region heteroduplex is accommodated in a tight positively-charged tunnel formed between the NUC and REC lobes. If there are mismatches in the seed region, the A-form architecture will be perturbed, inhibiting Cas9 binding to the target DNA. In contrast, the heteroduplex at the non-seed region has more conformational flexibility. Thus, mismatches in this region are better tolerated, potentially allowing the sgRNA to bind to off-target sites[51,71,80,81].

Interestingly, since Cas9 recognises the A-form architecture of the RNA–DNA heteroduplex rather than its sequence, it has been programmed to target A-form RNA–RNA duplexes, cleaving the target ssRNA in the presence of a short ssDNA harbouring a PAM sequence ("PAMmer")[82].

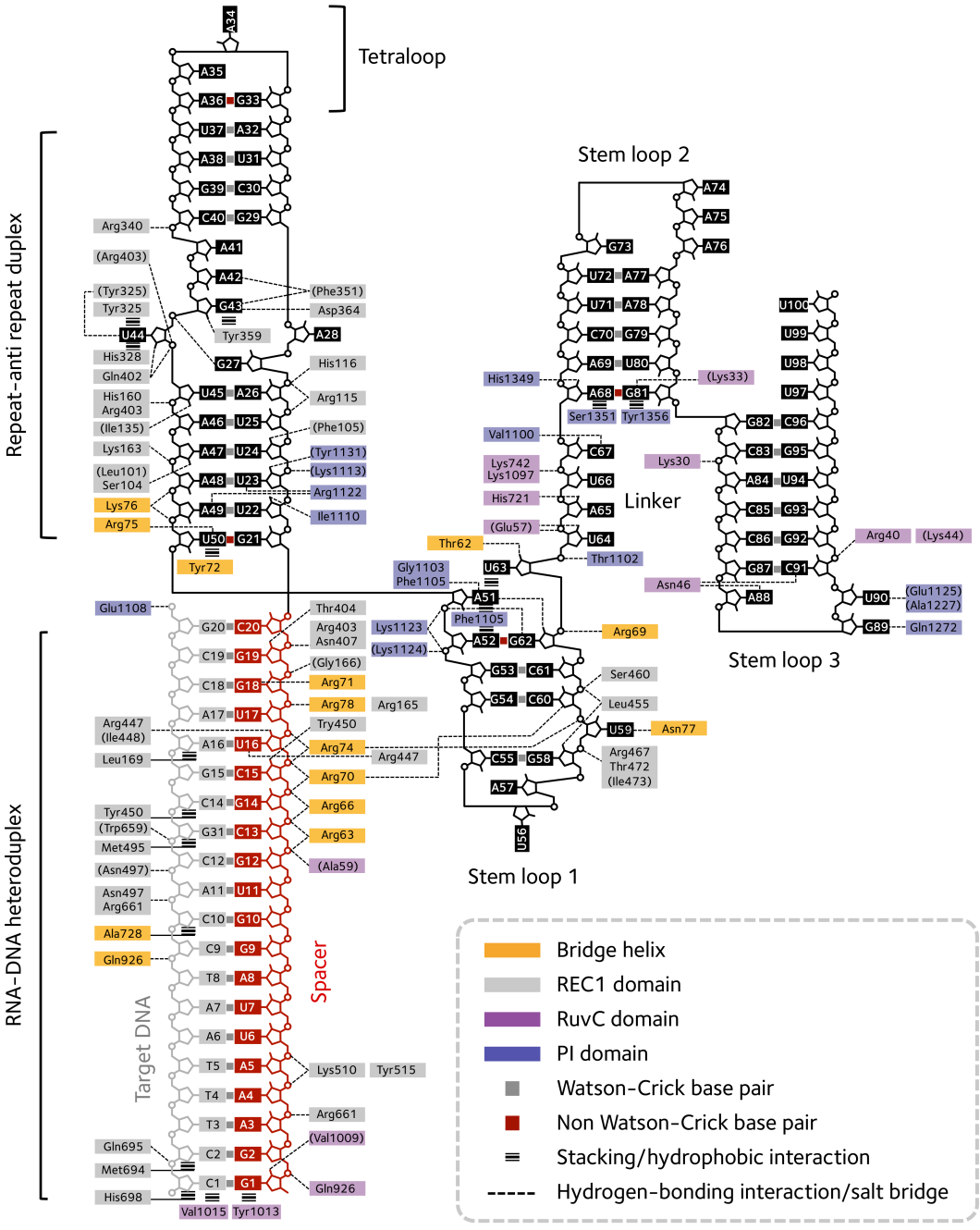


Figure 1.11 | Schematic representation of Cas9 and sgRNA–target DNA interactions. This figure was adapted from Nishimasu *et al.*, 2014[48]. Note that amino acid residues which interact with sgRNA–target DNA though their main chains and side chains are illustrated with and without the bracket, respectively.

1.6 Modified Cas9 Applications

1.6.1 SpCas9 Variants with Different PAM Specificities

To expand the targetable sites to sequences that are not adjacent to a 5'-NGG-3' motif, naturally occurring Cas9 orthologues from other species (e.g. *Francisella novicida* (Fn)[53], *Staphylococcus aureus* (Sa)[40,83], *Neisseria meningitides* (Nm)[84], *Streptococcus thermophilus* CRISPR1 (St1)[46,47], *Streptococcus thermophilus* CRISPR3 (St3)[46,47] and *Campylobacter jejuni* (Cj)[85]) that recognise different PAM sequences (**Table 1.2**) can be used. However, the relatively simple PAM requirement of SpCas9 makes it the most popular orthologue. Thus, engineering SpCas9 to have different PAM specificities has been explored by modifying the PAM-recognition motif that lies in the PAM-interacting (PI) subdomain within the CTD domain[53,86]. Although single mutations in the PAM recognition module (e.g. R1335Q) inhibit DNA cleavage, further mutations in the neighbouring Wedge (WED) subdomain within the same CTD domain restore Cas9 nuclease activity with altered PAM recognition properties. To date, several Cas9 variants have been reported including VQR (D1135V/R1335Q/T1337R), EQR (D1135E/R1335Q/T1337R) or VRER (D1135V/G1218R/R1335E/T1337R) that recognise 5'-NGA-3', 5'-NGAG-3' and 5'-NGCG-3' PAMs respectively[53,86]. Crystal structures of these Cas9 variants indicate that the multiple mutations cooperatively displace the PAM duplex phosphodiester backbone. As a result, the third PAM base can interact with the mutated residue at position 1335 and an additional forth base can be recognised by the mutated amino acid at position 1337[53,86].

Name	Organism	Size (residues)	PAM (5'–3')
SpCas9	*Streptococcus pyogenes*[40,46]	1368	NGG
FnCas9	*Francisella novicida*[53]	1629	NG
SaCas9	*Staphylococcus aureus*[40,83]	1053	NNGRRT
NmCas9	*Neisseria meningitidis*[84]	1082	NNNNGMTT
St1Cas9	*Streptococcus thermophilus* CRISPR1[46,47]	1121	NNAGAAW
St3Cas9	*Streptococcus thermophilus* CRISPR3[46,47]	1409	NGGNG
CjCas9	*Campylobacter jejuni*[85]	984	NNNNACAC

Table 1.2 | Cas9 orthologues. N = A, C, G or T, R = A or G, M = A or C and W = A or T.

1.6.2 High-fidelity Cas9 (SpCas9-HF1)

It has been proposed that off-target DNA cleavage occurs because the Cas9–sgRNA complex has greater affinity for DNA than is necessary for DNA binding and cleavage[74]. This "excess energy" is the result of direct interactions between Cas9 and dsDNA target. Crystal structures of the Cas9–sgRNA–dsDNA complex (PDB: 4OO8[48] and 4UN3[50]) show that several residues of Cas9 make non-specific hydrogen bonds with the phosphodiester backbone of both target and non-target DNA strands. Based on these observations, it was hypothesised that weakening these interactions, for example by substituting the residues with non-hydrogen bonding alanine, would decrease DNA target binding affinity and therefore reduce off-target activity. Two Cas9 variants were designed based on this rational: high-fidelity SpCas9 (SpCas9-HF1)[87], with N497A/R661A/Q695A/Q926A mutations that weaken the interactions with target DNA strand, and enhanced specificity SpCas9 (eSpCas9(1.1))[88], with K848A/K1003A/R1060A mutations that weaken the interactions with non-target DNA strand. These modified proteins significantly reduced off-target activity while having comparable on-target activity to the wild-type version. However, later mechanistic studies suggested these Cas9 variants reduce off-target effects through an alternative mechanism; they inhibit Cas9 adopting the active conformation required for DNA cleavage when binding to targets with mismatches[89]. Based on this finding, an alternative hyper-accurate SpCas9 (HypaCas9) with R692A/M694A/Q695A/H698A mutations was generated with comparable on-target and off-target activity to the SpCas9-HF1 and eSpCas9(1.1)[89]. Concurrently, random mutagenesis revealed a further mutation pattern – M495V/Y515N/K526E/R661Q – that gives even lower off-target activity (evoCas9)[90]. These high-fidelity Cas9 mutants are effective in human cells when delivered as a DNA-encoded vector. However, their on-target activity substantially decreases when delivered as the ribonucleoprotein complex[91] (the advantages of which are discussed in **Section 1.6.5**). As a consequence, through further Cas9 mutant library screening in bacteria, a high-fidelity (HiFi) Cas9 variant was reported containing a single point mutation R691A. Perhaps due to the minimal modification of the wild-type protein, this variant is still effective when delivered as the ribonucleoprotein complex[91].

1.6.3 Cas9 Nickases

Mutation in the catalytic site of one of the two nuclease domains, for example H840A in HNH or D10A in RuvC domains, generates Cas9 nickase (Cas9n) that catalyses the cleavage of one DNA strand while leaving the other stand intact[26,47]. Off-target DNA cleavage by wild-type Cas9 generates DSB, which is often repaired inaccurately by non-homologous end joining (NHEJ) pathway with introduction of indels. If the Cas9 nickase cuts DNA at the off-target site, the resulting single-strand break (SSB) will be repaired precisely without indel formation. The likelihood of two Cas9 nickases cleaving DNA at the same off-target sites and producing DSB is very low. Therefore, a pair of Cas9 nickases are used to cut two proximal DNA target sites to mimic a DSB, minimising off-target DNA cleavage[92,93]. However, this requires the design of two adjacent sgRNAs that recognise two PAMs, limiting the targetable sites.

Interestingly, Cas9 nickase can facilitate efficient single-base editing, for example from C–G to T–A[94] or A–T to G–C[95], without introducing DSBs. By fusing Cas9 nickase with deaminase enzymes, the conversion of a particular base to another base occurs precisely on the non-nicked DNA strand within a 5-nt region (typically positions 13 to 17 of the spacer when counting from the PAM-proximal end). The resulting mismatches undergo DNA mismatch repair (MMR) and subsequent DNA replication yielding the desired base pair mutant. A single base editing can also be achieved using dCas9-deaminase fusion proteins or Cas9-mediated DSBs which undergo HDR in the presence of a donor DNA containing a correct base. However, these approaches are less efficient than the Cas9-nickase-deaminase approach.

1.6.4 Catalytically Inactive Cas9 (dead Cas9 or dCas9)

When the key catalytic residues in both HNH and RuvC domains are mutated (H840A and D10A in the HNH and RuvC domains respectively), Cas9 becomes catalytically inactive (named dCas9) but it is still possible to programme it to bind DNA efficiently and specifically[26]. This dCas9 has been used in many gene regulation applications. For example, when binding to a promoter, dCas9 can sterically block RNA polymerase as well as transcription factor binding to the promoter, inhibiting transcription[96,97]. By fusing dCas9 with specific effector domains such as transcription activators (e.g. VP64) or repressors (e.g. KRAB), specific gene expression can be

efficiently regulated[98–100]. dCas9 fused with epigenetic modifier proteins (e.g. DNA demethylase, histone acetyltransferase and histone demethylase) can also be used to study the effects of chromatin states on gene regulation and the resulting cellular phenotypes[101]. In addition, dCas9 can be fused with fluorescent proteins for probing and live-cell imaging of specific genomic loci[102], allowing their dynamics and conformation to be studied. More details on this specific application will be discussed in **Chapter 3**.

1.6.5 Delivery Methods

Traditional methods for the expression of exogenous proteins in cells involve the use of vectors that encode the protein at the DNA-level. For example, the electroporation or lipid-mediated transfection of DNA vectors into cells is standard, with lenti-viral plasmid transduction offering more prolonged and controlled gene expression levels. However, it has been found that high sustained levels of Cas9 and sgRNA can give rise to more off-target activity[103,104], presumably due to excess complex being active even after the desired target has been edited. As a consequence, expression of Cas9 from faster decaying mRNA[105,106] or even direct delivery of the pre-complexed ribonucleoprotein complex has been explored[105,107,108]. RNPs transfection offers the highest immediate levels of Cas9–sgRNA followed by a rapid decay in levels since the protein cannot be regenerated (cf. expression from mRNA or DNA). This 'fast on, fast off' strategy is reported to generate the fewest off-target effects in literature[91].

1.7 Cas12a: An Alternative Tool for Genome Editing

As an alternative tool for gene editing, Cas12a (previous name: Cpf1) offers both advantages and disadvantages compared to Cas9 (**Figure 1.12**). Firstly, Cas12a only requires a single, short crRNA ~42 nt[60] to function which simplifies the system. Secondly, Cas12a creates a staggered, double-stranded DNA cleavage which allows precise DNA repair or gene insertion *via* HDR to be more efficient over NHEJ. However, Cas9 is still preferred in some applications when gene knockout is the desired outcome of DNA repair due to the preferred NHEJ repair pathway. Finally, Cas12a recognises the 5'-TTTV-3' PAM sequences, allowing the T-rich regions to be targeted in addition

to the G-rich 5'-NGG-3' PAM recognised by Cas9[60]. In effect, these two CRISPR systems are complementary to each other.

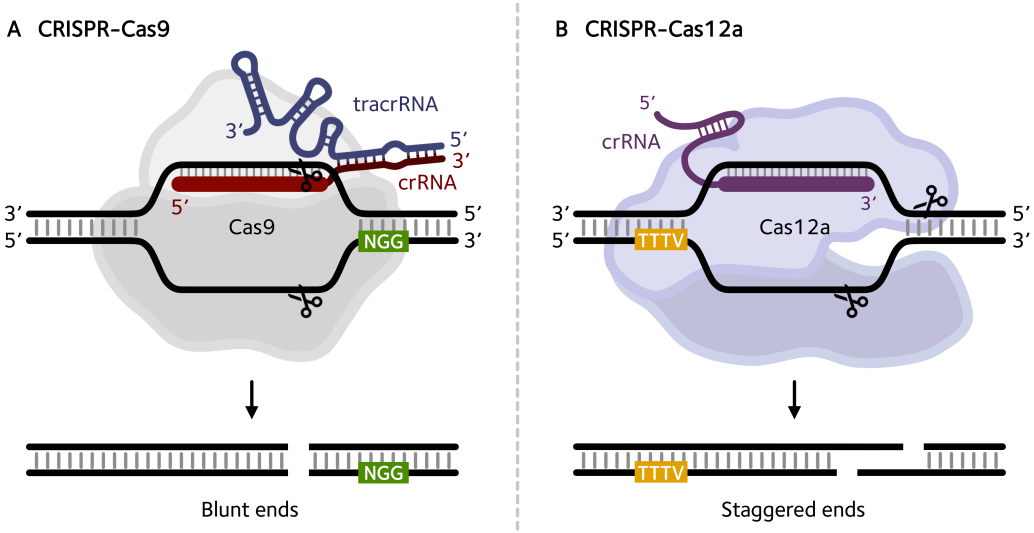


Figure 1.12 | Comparison of the CRISPR-Cas9 (A) and CRISPR-Cas12a systems (B).

Since the discovery of Cas12a is very recent, details of its exact biochemical mechanism are still being determined. In general, Cas12a-mediated DNA cleavage requires crRNA binding, PAM recognition, RNA strand invasion, R-loop formation prior to target cleavage, which is similar to the Cas9-dependent system. Interestingly, Cas12a possesses an intrinsic RNase activity in addition to the DNA cleavage activity. Thus, it can cleave the pre-crRNA transcript into individual mature crRNAs by itself[109]. This has enabled easier multiplexed gene editing as only a single crRNA array is sufficient for gRNA library construction[110].

1.7.1 crRNA Loading and Target DNA Recognition

To date, the available crystal structure of *Acidaminococcus spp. BV3L6* (As) Cas12a–crRNA–DNA complex (PDB: 5B43[111]) shows that Cas12a consists of two lobes: a REC lobe and a NUC lobe which is responsible for DNA cleavage. Comparison of the Cas12a–crRNA complex with and without (PDB: 5ID6[112]) bound dsDNA reveals that Cas12a employs a different mechanism for target DNA recognition compared to Cas9 (**Figure 1.13**). In the Cas9–sgRNA complex, both the PAM-interacting cleft and the PAM-proximal region of the sgRNA (seed region) are preordered before target DNA binding to facilitate PAM recognition and DNA hybridisation, respectively[51].

However, those two parts in the Cas12a–crRNA do not become ordered until binding to DNA. With bound DNA, Cas12a undergoes substantial conformational changes to make contacts with PAM, which in turn allows the PAM-proximal region of the crRNA to adopt A-form conformation for base-pairing with the target DNA. This distinct target recognition mechanism possibly reduces Cas12a binding to off-target sites, which is beneficial to genome editing applications[113].

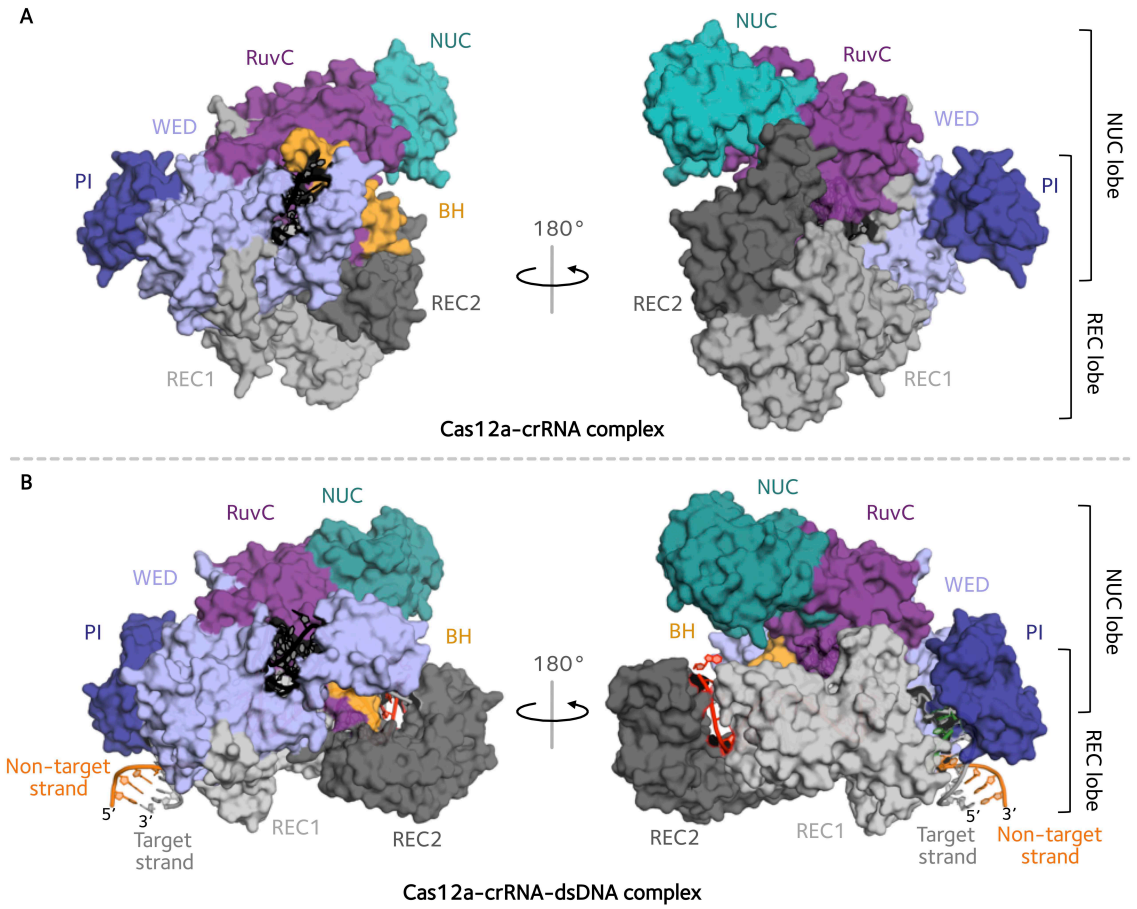


Figure 1.13 | Surface representations of the Cas12a–crRNA complex. A substantial structural rearrangement is observed in the absence of target DNA (A) (LbCas12a, PDB: 5ID6[112]) and after target DNA binding (B) (AsCas12a, PDB: 5B43[111]).

1.7.2 PAM Recognition

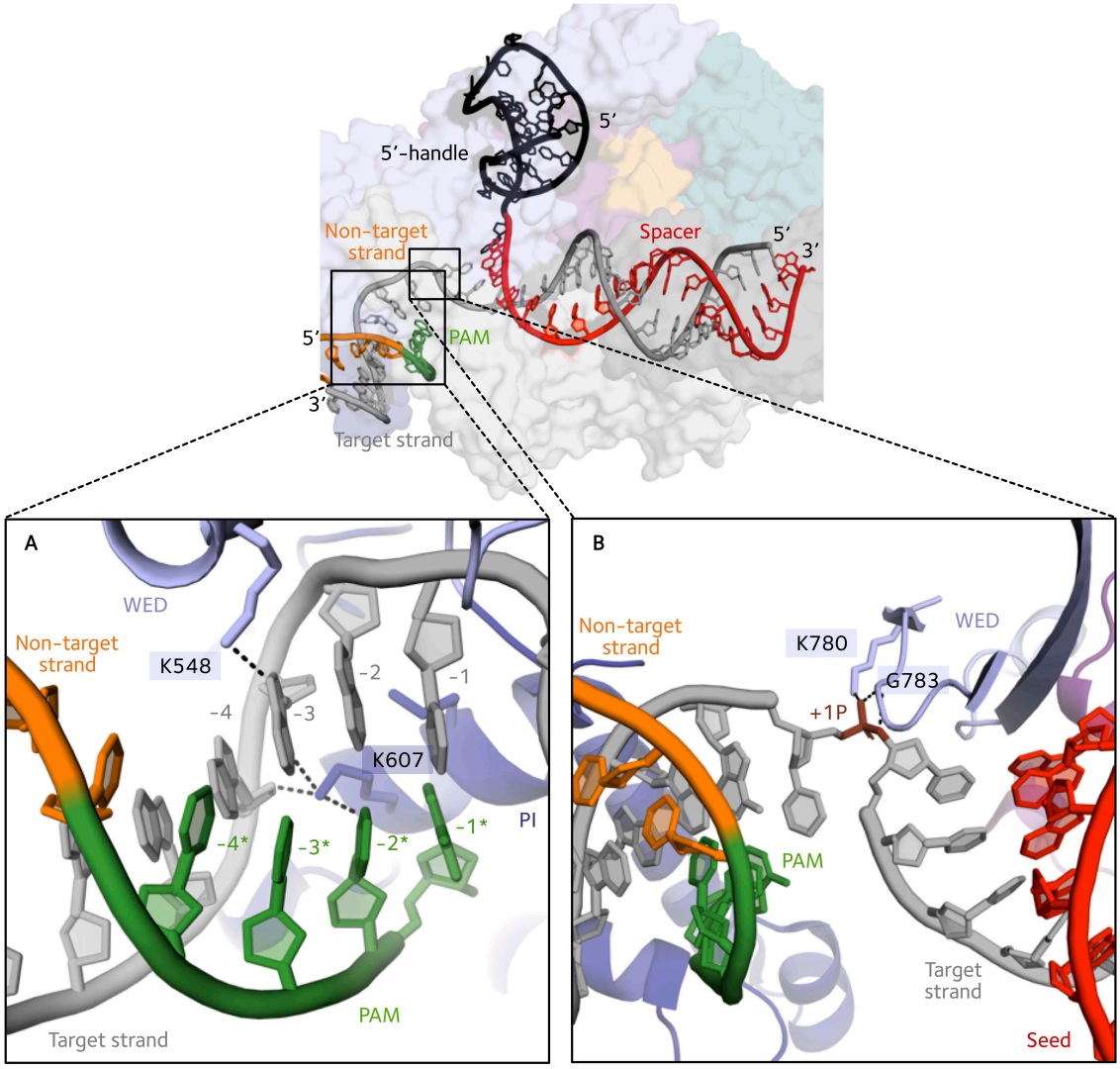


Figure 1.14 | Close-up views of PAM (5'-TTTV-3') recognition and the phosphate lock loop in the CRISPR-Cas12a system. K548 (WED domain) interacts with the complementary strand to PAM and K607 (PI domain) forms extensive hydrogen bonds with many residues within the PAM duplex (A) (PDB: 5B43[111]). Following PAM recognition, the phosphate immediately downstream of PAM (+1 phosphate in brown) makes contacts with K780 and G783 from the WED domain, enabling the target DNA to flip towards the crRNA for RNA–DNA heteroduplex formation (B) (PDB: 5B43[111]).

Similar to the Cas9–sgRNA complex, the Cas12a–crRNA ribonucleoprotein recognises the PAM motif (5'-TTTV-3' where V = A, G or C) on the non-target DNA strand immediately upstream of the spacer[60,114] (**Figure 1.14A**). The key residue critical for PAM recognition is K607 as its side chain forms hydrogen bonds with the third T (-2*), A (-3) and/or A (-4) in the PAM duplex in combination with other residues. Mutation of this residue results in no DNA cleavage activity by

Cas12a[111]. Other important residues, in which their mutations lead to reduced Cas12a activity, are K548, M604 that form a hydrogen bond and van der Waals interactions with A (-3) and A (-2), respectively. Similar to Cas9, following PAM recognition, the first phosphate group next to the PAM duplex on the target DNA strand (+1 phosphate) rotates through interactions with K780 and G783, facilitating crRNA strand invasion and the displacement of the non-target DNA strand and subsequent R-loop formation (**Figure 1.14B**).

1.7.3 DNA Cleavage

dsDNA cleavage occurs in the NUC lobe. Mutational studies of key residues in this lobe reveal that the RuvC domain conserved catalytic residues D908, E993 and D1263 enable cleavage of the non-target DNA strand, similar to Cas9. The HNH domain, which is essential for target DNA strand cleavage in Cas9, is not present in Cas12a. Instead, Cas12a uses residue R1226 located outside the RuvC domain but still within the NUC lobe to cleave target DNA strand outside, rather than inside, the RNA–DNA heteroduplex as observed in Cas9. Such cleavage results in a 5-bp staggered DSB at the 18th (non-target strand) and 23rd (target strand) nucleotides from PAM[60] (**Figure 1.15**), whereas Cas9 cleaves dsDNA in the vicinity of PAM. Mutation of R1226 converts Cas12a into a nickase. However, when catalytic residues of the RuvC domain are mutated, both DNA strands are not cleaved. This suggests that the cleavage of the target strand is dependent on that of the non-target strand by the RuvC domain[111], in contrast to Cas9 where the cleavage of two strands occurs independently. However, it has not been determined how exactly the cleavage of the non-target strand facilitates the cleavage of the target strand.

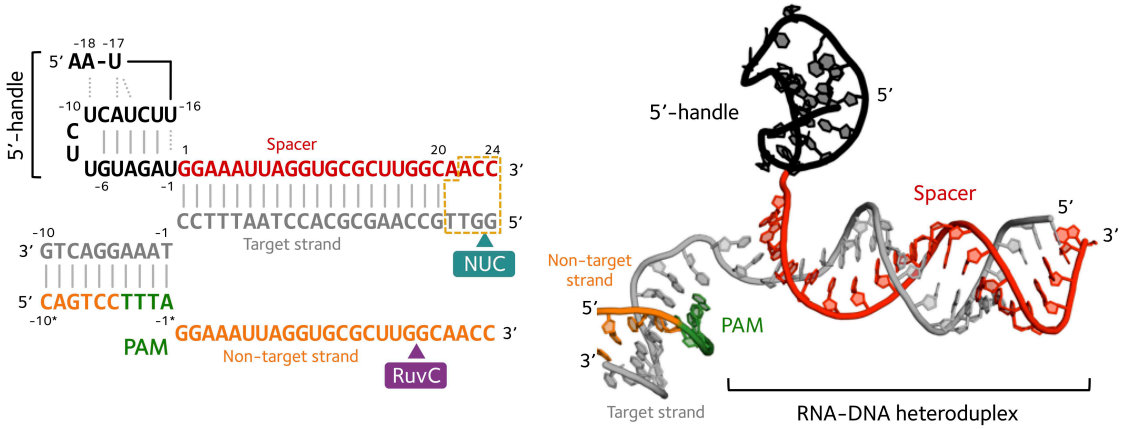


Figure 1.15 | Structure of crRNA in the CRISPR-Cas12a system. crRNA consists of a spacer (red) and a 5'-handle (black). It is also shown that DNA cleavage occurs at 23rd (target DNA strand) and 18th position (non-target DNA strand) of a spacer sequence. Yellow dashed box represents the disordered region in the crystal structure (PDB: 5B43[111]).

1.8 Cas12a crRNA

The crRNA can be divided (from 5' to 3') into a 19-nt scaffold ((-19)–(-1)) required for Cas12a binding and a 24-nt spacer (1–24) needed for target DNA hybridisation. In the Cas12a–crRNA complex, the REC lobe interacts mostly with the ribose 2' OH groups and the phosphate backbone of the spacer whereas the WED and the RuvC domains, within the NUC lobe, form extensive interactions with the scaffold of the crRNA[111] (**Figure 1.16**).

The 19-nt scaffold contains four bases conserved in the Cas12a protein family: U(-1), U(-10), U(-16) and A(-18)[60]. These residues form canonical and non-canonical base pairs, resulting in a pseudoknot structure. Upon binding to Cas12a, the U(-1)–U(-16) base pair forms a further hydrogen bond with H761 and R18/N759 in the WED domain respectively. Disruption of this U–U base pair inhibits Cas12a binding to the crRNA and subsequent DNA cleavage, confirming the importance of the pseudoknot structure[60] (discussed in **Section 1.9**).

The RNA bases from position 1–20 in the spacer hybridises to the 20-nt target DNA, forming a RNA–DNA heteroduplex and the "seed" 5-bp region proximal to PAM is essential for DNA cleavage[60]. The remaining bases from position 21 to 24 do not base pair with the target DNA and are found to be disordered in the crystal structure. Mismatch studies demonstrate the relative

importance of these spacer RNA subdomains; mutation in the PAM-distal 21–24 bases are tolerated but not in the complementary region (especially position 1–18).

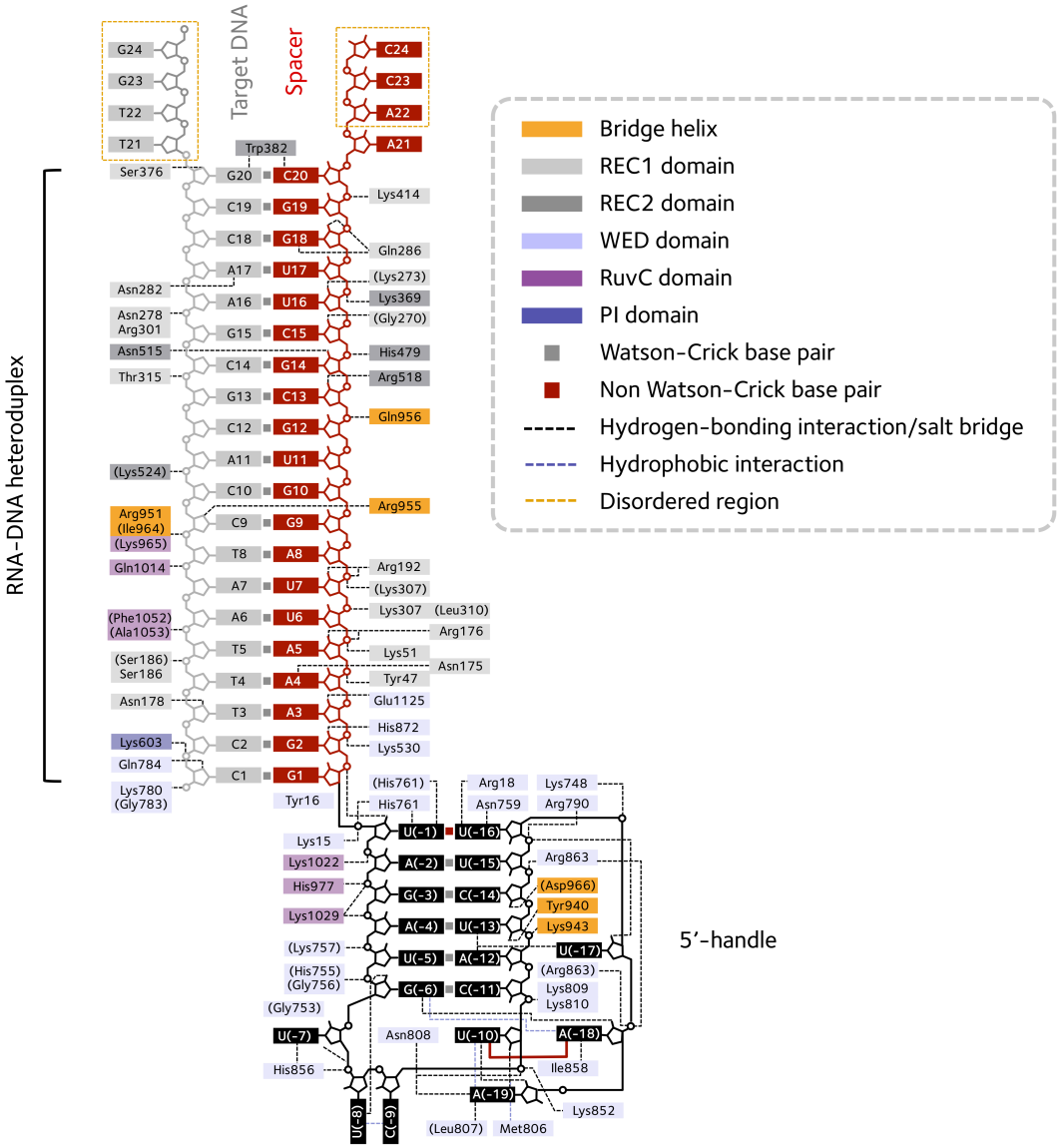


Figure 1.16 | Schematic representation of crRNA–target DNA recognition by Cas12a. This figure was adapted from Yamano *et al.*, 2016[111]. Note that amino acid residues of Cas12a in bracket make contacts with crRNA and target DNA using their main chains while those without bracket use their side chains for interactions.

1.9 Basics of Nucleic Acid Structure

RNA is central to the function of CRISPR systems. Significant advances have been made by modifying the Cas9 protein but modifying the RNA to expand CRISPR functionality has only just begun. The goal of these studies has been to introduce chemical modifications with beneficial effects without compromising the Cas9 activity. In order to do this, a good understanding of nucleic acid structure is necessary.

Nucleic acids have hierarchical levels of complexity from primary to tertiary structure. The primary structure is the linear sequence of molecular building blocks called nucleotides (**Figure 1.17**). A nucleotide contains a phosphate group that is covalently linked to a pentose sugar attached to a nucleobase; the sugar and base alone is termed a nucleoside. The phosphate group of one nucleotide forms a phosphodiester bond with the 3'-OH of the neighbouring nucleotide, resulting in an oligonucleotide. The RNA sugar (a ribonucleotide) has a hydroxyl group at the 2' position whereas the DNA sugar (a deoxynucleotide) has a hydrogen atom at this position instead. The four nucleobases in DNA are the purines: adenine (A) and guanine (G), and the pyrimidines: cytosine (C) and thymine (T). In RNA, T is replaced by uracil (U) which lacks a methyl group at C5' position of the sugar ring.

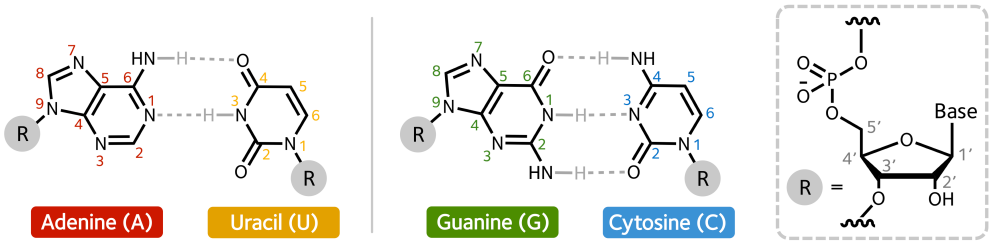


Figure 1.17 | Structures of purine and pyrimidine bases and their base pairing according to the Watson-Crick's base pair rule. Purine bases (A and G) consist of two heterocyclic rings while pyrimidines possess a single ring system. A forms two hydrogen bonds with U (or T) whereas G forms three hydrogen bonds with C.

Secondary structure is when two oligonucleotides are held together by base-stacking interactions with the specificity of one strand for another determined by intermolecular hydrogen bonds. As proposed by Watson and Crick base[115], A pairs with T or U, whereas C pairs with G (**Figure 1.17**). As a result, a secondary regular helical structure is formed that runs in the

opposite direction (anti-parallel, **Table 1.3**). This hides the hydrophobic nucleobases while exposing the negatively charged phosphate backbone to the aqueous environment. The twisting of the helix and the geometry of the nucleobases generate a major groove and a minor groove along the entire duplex with varying conformations termed A-, B- and Z-form (summarised in **Table 1.3**).

Side view			
	A-form	**B-form**	**Z-form**
Helical Turn	11 bp | 25.3 Å	10.4 bp | 35.4 Å	12 bp | 45.6 Å
Rise per bp	2.3 Å	3.4 Å	3.8 Å
Diameter	25.5 Å	23.7 Å	18.4 Å
Glycosidic Conformation	*anti*	*anti*	C: *anti*, G: *syn*
Sugar Puckering	C3'-*endo*	C2'-*endo*	C: C2'-*endo*, G: C3'-*endo*
Screw Sense	Right-handed	Right-handed	Left-handed
Major Groove	Narrow deep	Wide	Flat
Minor Groove	Broad	Narrow	Narrow deep
Top view			

Table 1.3 | Characteristics of A-, B- and Z-form conformations[116,117]. The sgRNA (red)–target DNA (grey) heteroduplex from the CRISPR system (PDB: 5F9R[51]) is an example of an A-form configuration. B- and Z-form are depicted as a black and grey DNA duplex (PDB: 1BNA[118] and 4OCB[119], respectively). This figure was adapted from Shivalingam, 2015[120].

The exact conformation is determined by several factors including sugar pucker (**Figure 1.18A**) and syn-anti conformation of the nucleobase[117] (**Figure 1.18B**). The C2' and C3' of the sugar are displaced from the median plane of C1'-O4'-C4'. The displacement of the C2' above and the C3' below the plane is called C2'-endo while the opposite arrangement is called C3'-endo. The N-

glycosidic bond linking the nucleobase to the C1' of the sugar is also rotatable, giving rise to syn and anti conformation of the bases, where the Watson-Crick base-pairing face of the nucleobase is rotated towards and away from the O4' of the sugar respectively.

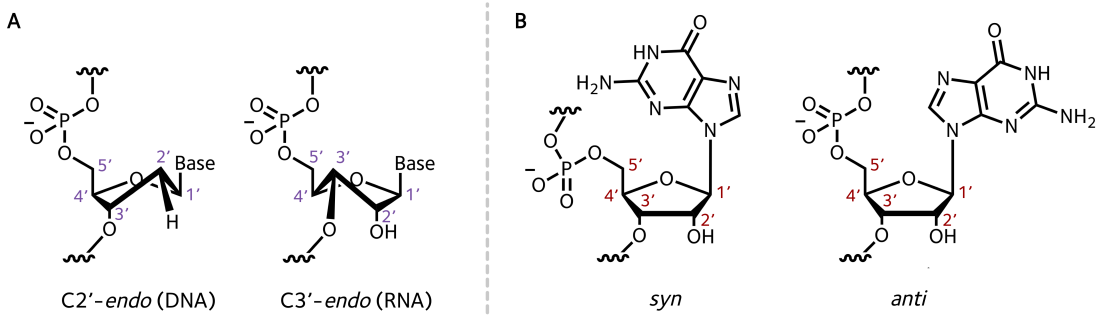


Figure 1.18 | Sugar puckering and the syn-anti conformation of the nucleobase.

dsDNA is predominantly present in the B-form with a right-handed helix, C2'-endo sugar puckering, and base pairs which are perpendicular to the helical axis. Due to the steric bulk of the hydroxyl group at the C2', dsRNA favours C3'-endo pucker, which together with the favourable anti conformation of the bases gives rise to the A-form conformation with a right-handed helix, a very deep major groove and a very shallow minor groove. However, these conformations are not fixed. DNA can adopt the A-form, with the energy barrier between the two forms being low[121]. In addition, when hybridising to RNA, DNA–RNA duplexes take on the A-form configuration, which is a crucial feature observed in CRISPR systems. With distinct differences in the geometry between A- and B-form conformations, gRNA with a spacer solely made of DNA is unable to form a tertiary Cas9–gRNA–target DNA complex, possibly due to the configuration it adopts[122]. Finally, DNA can also adopt the Z-form with left-handed helix in a zigzag pattern and different nucleoside geometries from A- and B-form conformations[123].

Secondary structures also include bulges, hairpins and loops, all of which are present in the biological important RNA molecules[124] as well as the sgRNA in the CRISPR system (**Figure 1.19**). Bulges are produced when a duplex contains unpaired nucleotides. This causes the distortion of base stacking within the duplex, destabilising the helix and providing access to the major groove of base pairs next to the bulge. Hairpins are derived from the folding of one strand onto itself,

creating a duplex stem and a single-stranded loop. The presence of mismatches creates internal loops. Alternative base pairing patterns in addition to normal Watson-Crick base pairs such as G–U wobble base pairs[125] are often used to stabilise these secondary structures.

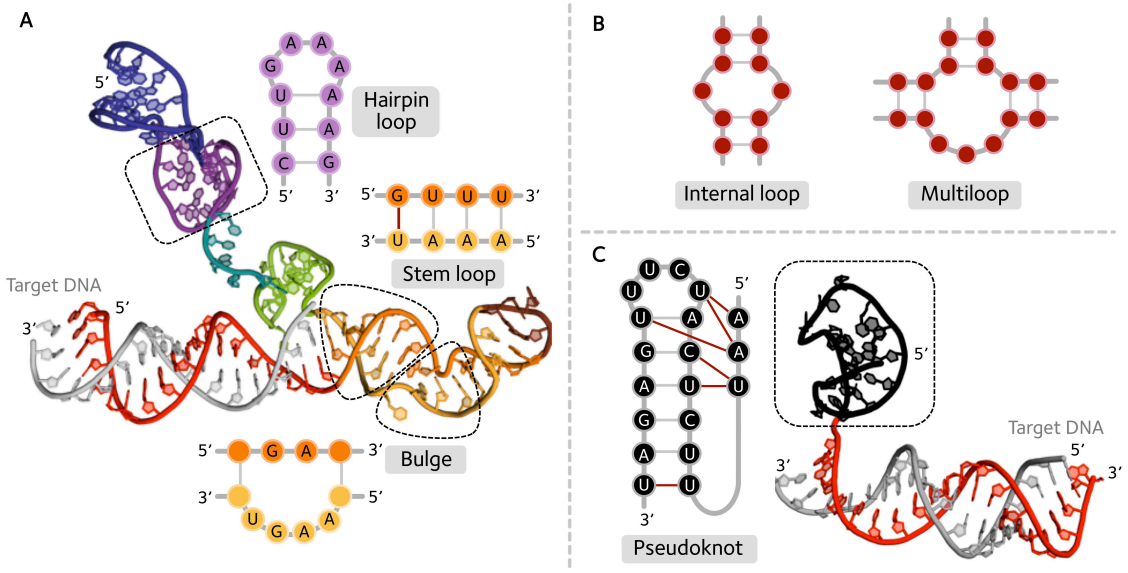


Figure 1.19 | Secondary and tertiary structures of RNA. (A) sgRNA from the CRISPR-Cas9 system (PDB: 5F9R[51]) possesses different secondary structures e.g. hairpin, stem loop and bulge. (B) Other structures such as internal loop and multiloop are also present in the biological system. (C) Tertiary structure e.g. pseudoknot is found in the crRNA structure from the Cas12a (PDB: 5B43[111]).

Base pairing can occur at various positions of the nucleobase as well as sugar (**Figure 1.20A**). Wobble base pairing, despite occurring at the same edge as Watson-Crick base pairs, does not follow the same rules; G can form two hydrogen bonds with U instead of three hydrogen bonds with canonical C (**Figure 1.20B**). Weaker hydrogen bonding is off-set by the improved base stacking of the wobble base pairs, enabling them to nicely fit the normal Watson-Crick helix. This G–U wobble base pairing is observed at the beginning of the repeat–antirepeat region of sgRNA (G21–U50).

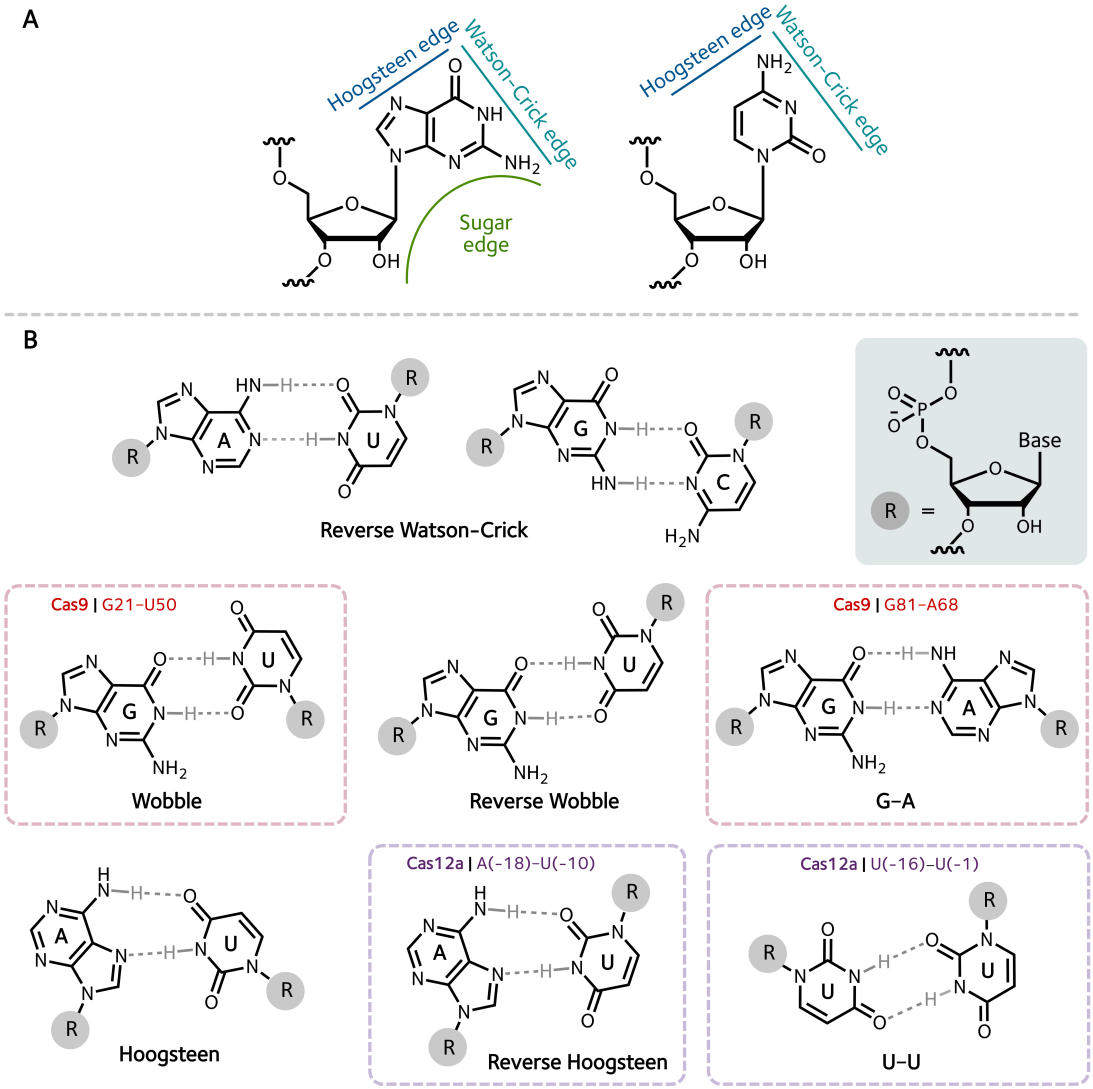


Figure 1.20 | Non-canonical base pairings. In addition to the Watson-Crick edge, the sugar edge and Hoogsteen edge offer alternative locations for base pairing (A). Examples of non-canonical base pairs are shown in (B). Wobble G–U and unusual G–A base pairs are found in the sgRNA of the CRISPR-Cas9 system. Reverse Hoogsteen A–U and unusual U–U base pairs help stabilise the pseudoknot structure of the crRNA in the CRISPR-Cas12a system.

When the secondary structures of nucleic acids interact, a three-dimensional tertiary structure is obtained. Additional non-Watson-Crick base pairings such as Hoogsteen[126] and the reverse version of Watson-Crick, wobble and Hoogsteen base pairings allows the RNA to adopt various functional tertiary structures (e.g. tRNA)[127–129] (**Figure 1.20B**).

Hoogsteen hydrogen bonding occurs at a different edge of the nucleobases, potentially allowing a third nucleotide to hydrogen bond with a Watson-Crick base pair. The reverse versions of the

base pairings described previously are used to facilitate base pairing between parallel strands. A pseudoknot structure within the crRNA of the CRISPR-Cas12a system is an example of RNA tertiary structure (**Figure 1.19C**). The structure is maintained by the reverse Hoogsteen U(-10)–A(-18) base pair (**Figure 1.20** and **1.21A**), a non-canonical U(-1)–U(-16) base pair (**Figure 1.20**), and the U(-13)–U(-17)–A(-12) base triple (**Figure 1.21B**) in addition to standard Watson-Crick base pairs. Disruption of the U(-1)–U(-16) base pair, for example, inhibits Cas12a binding to the crRNA and subsequent DNA cleavage, confirming the importance of the pseudoknot structure[60].

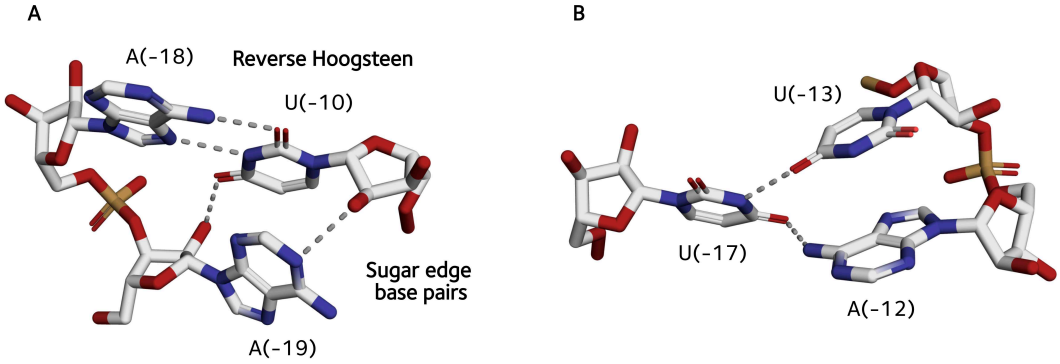


Figure 1.21 | Base triples in the pseudoknot structure of crRNA from the CRISPR-Cas12a system (PDB: 5B43[111]). Hoogsteen base pairing (A–U) allows the third base (U) to interact *via* sugar edge (A). Base triplets can also be generated by other base pairs (B) that are not listed in **Figure 20**.

1.10 Chemical Modifications of the gRNA

The nucleobase, ribose sugar and phosphate backbone, at positions that are not critical for Cas9 interactions are modifiable. So far several sugar and backbone modifications (**Figure 1.22**) have been successfully introduced into the sgRNA at various locations (**Figure 1.23**) to improve CRISPR functionality. Sugar modifications such as 2'-fluoro (2'-F)[130–132] and 2'-*O*-methyl (2'-OMe)[130–132] prevent intra-molecular cleavage by the 2'-OH group and inhibit nuclease degradation, increasing the stability of the sgRNA in cells and hence the gene editing efficiency. Backbone modifications can be combined with 2'-F and 2'-OMe such as 2'-OMe-phosphorothioate (MS)[103,130–133], 2'-F phosphorothioate (FS)[130] and 2'-OMe thiophosphonoacetate (MSP)[103] to further inhibit the hydrolysis of the native phosphodiester bond by cellular nucleases. Interestingly, the sgRNA can carry up to ~70% of 2'-F or 2'-OMe ribose sugars without losing

genome editing activity[130], demonstrating the importance of keeping the C3'-endo pucker[134–136] compared to deoxynucleotides discussed below.

DNA is more chemically stable and cheaper to synthesise than RNA, therefore attempts have been made to replace some of the ribonucleotides in the sgRNA with deoxyribonucleotides[122,137]. For the crRNA–tracrRNA hybrid system, 37 out of 42 RNA nucleotides of the crRNA can be coded as DNA, while maintaining on-target but reducing off-target DNA cleavage activity *in vitro*. In addition, substitution of more than half of tracrRNA (54 out of 87 RNA nucleotides) with DNA does not affect the *in vitro* DNA cleavage efficiency. However, these highly DNA-containing crRNA–tracrRNA hybrid fail to mediate DNA cleavage *in vivo*, possibly due to kinetic destabilisation of the Cas9–gRNA–target DNA complex in cells[122]. Increasing the number of RNA nucleotides (to have 8 DNA bases in the spacer region) begins to restore the DNA cleavage activity in cells (~10–30% relative to an all-RNA construct[130]).

Other sugar modifications have been used to modulate the nucleic acid interactions. 2',4'-constrained ethyl bicyclic nucleic acids (2',4'-cEt) can be incorporated into the crRNA to strengthen its interaction with tracrRNA, improving editing efficiency[131]. Next-generation bridged nucleic acids (2',4'-BNA^{NC}[N-Me]) and locked nucleic acids (LNA) destabilise the crRNA-DNA duplex at off-target sites, which in turn reduces the dwell time of the Cas9–sgRNA complex and hence the off-target DNA is unlikely to be cleaved[138]. 2'-OMe-3'-phosphonoacetate (MP)[139] is found to improve target specificity when placed at specific positions along the spacer region (**Figure 1.23**).

Furthermore, efforts have begun to conjugate chemical handles to 5' and 3' termini of the crRNA. For example, molecules such as synthetic DNA for enhanced HDR or fluorophores for Fluorescence-Activated Cell Sorting (FACS) to enrich crRNA-containing cells have been attached to crRNA[140].

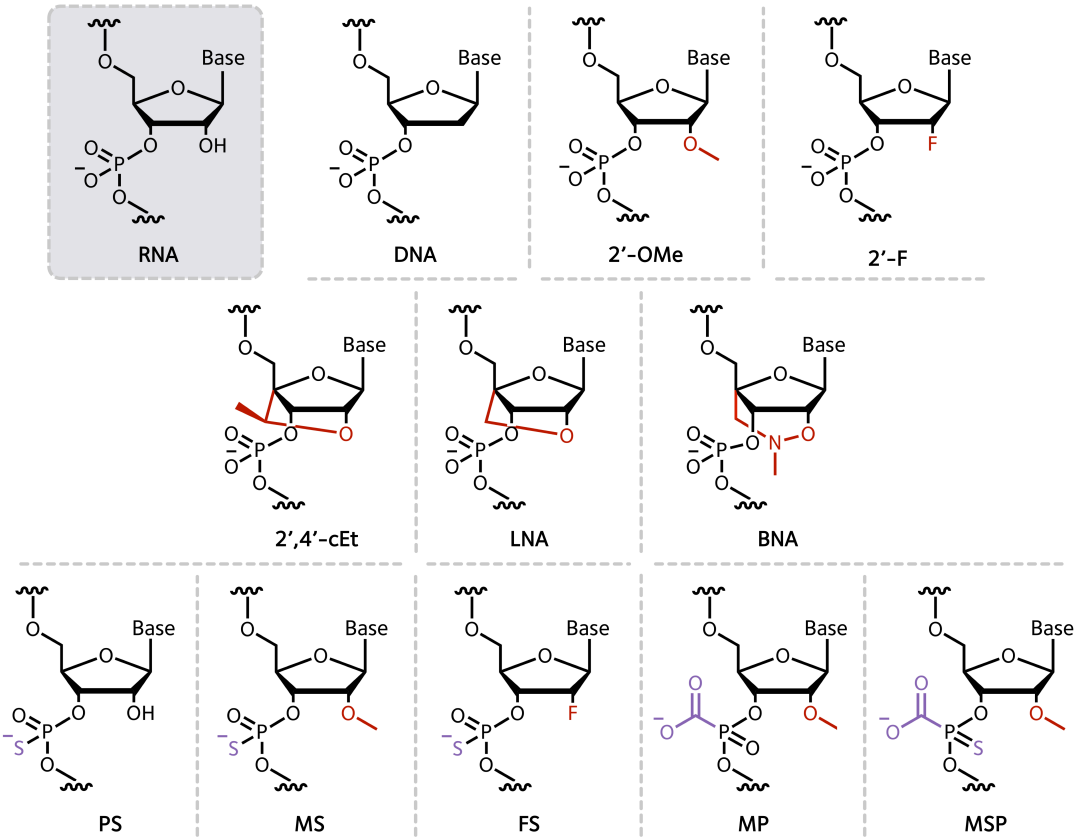


Figure 1.22 | RNA modifications that have been successfully incorporated into the sgRNA of the CRISPR-Cas9 system. Sugar modifications are shown in red while backbone modifications are in purple. Combinations of backbone and sugar modifications are also reported.

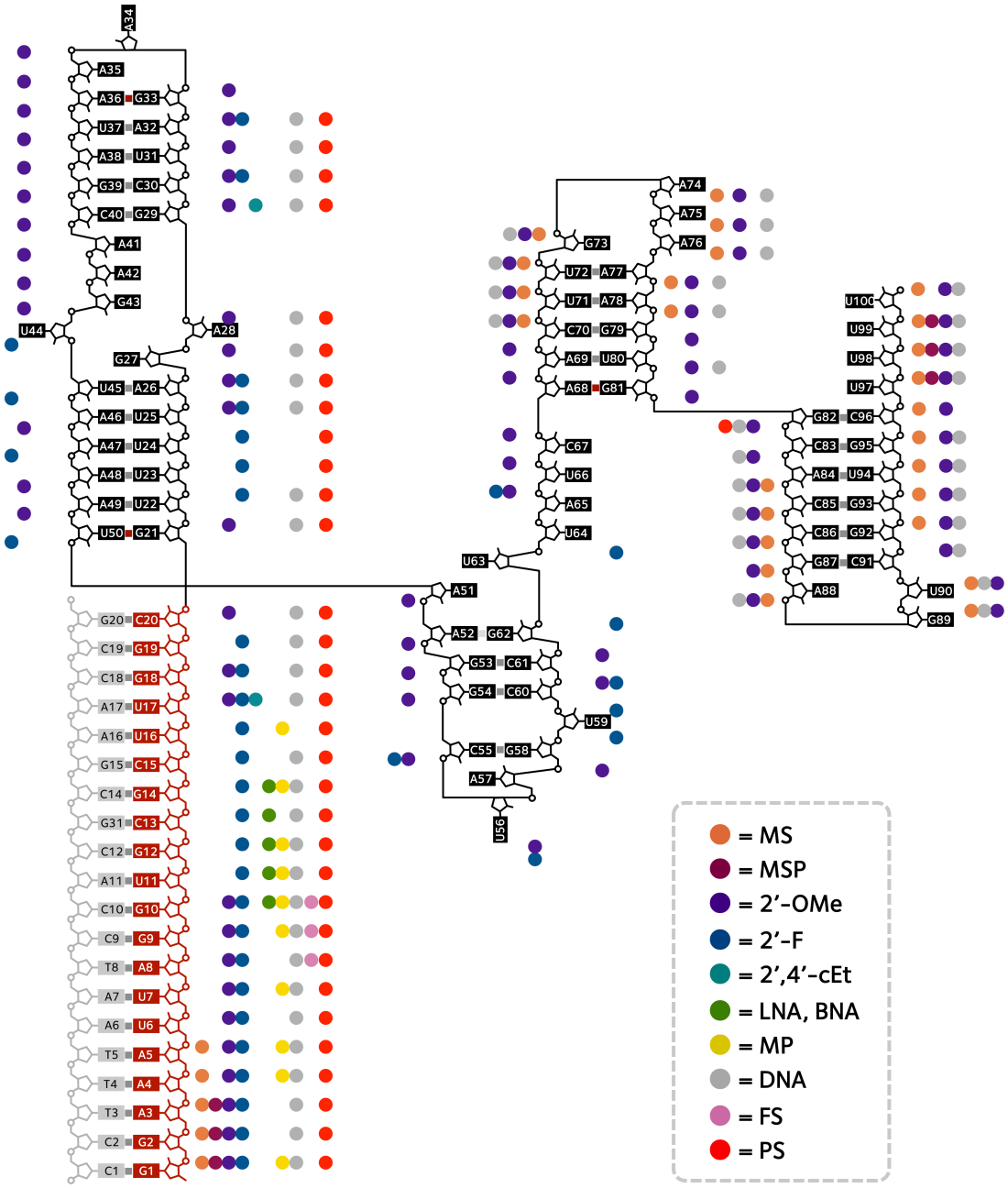


Figure 1.23 | Map of RNA modifications at various positions throughout the sgRNA of the CRISPR-Cas9 system. The positions indicated as modifiable retain approximately 50% activity in comparison to the unmodified sgRNA. The map is constructed from the following references for MS[103,130–133], MSP[103], 2'-OMe[130–132], 2'-F[130–132], 2',4'-cEt[131], LNA/BNA[138], MP[139], DNA[122,137], FS[130] and PS[130–132].

1.11 Aims and Objectives

While biological modifications of the Cas9 protein become mature, chemical modifications of the gRNA have just begun. They have already improved the CRISPR-Cas9 system. However, this work has mainly focused on improving genome editing by enhancing the cellular stability and the target specificity of the sgRNA, intrinsic properties of nucleic acids. Enabling other types of CRISPR functionality using chemically modified gRNAs has not been well explored.

The specific aims of this work are to:

Chapter 2 – improve sgRNA production *via* chemical ligation which is cost-effective, simple, and applicable to library construction.

Chapter 3 – generate fluorescent guide RNAs which are easily customisable for live-cell imaging.

Chapter 4 – alter the genome editing outcomes by chemically ligating pairs of crRNAs.

Chapter 5 – apply the findings from **Chapter 2** for *in vitro* applications such as forensic STR analysis using Oxford Nanopore sequencing technology.

Click sgRNAs using CRISPR-Cas9

2.1 Background

2.1.1 Enzymatic Generation of sgRNAs

RNAs are naturally generated by a process called transcription, where dsDNA acts as a template for complementary RNA production using enzymes called RNA polymerases. These polymerases recognise specific DNA promoter sequences which vary depending on the type of RNA being produced[141]. For example, messenger RNAs (mRNAs) encoding functional proteins are transcribed by RNA polymerase II (Pol II) using an RNA Pol II promoter, while non-coding RNA (e.g. tRNAs and other small nuclear RNAs (snRNAs)) transcription is directed by RNA Pol III-dependent promoters.

To express exogenous sgRNA in human cells, RNA Pol III-driven promoters such as the U3 snRNA[142] , U6 snRNA[28,47,143] and H1[144] have been widely adopted (**Figure 2.1A**) as the gRNA transcript is not subject to extensive post-transcriptional processing such as splicing and end

modifications, which could inhibit gRNA activity. Alternatively, gRNAs can be generated by *in vitro* transcription (IVT) using purified T7, SP6 or T3 RNA polymerases from bacteriophages, and transfected into the cell[145,146] or used in *in vitro* studies[26] (**Figure 2.1B**). IVT is terminated by the enzyme reaching the end of the linear DNA template, a process referred to as "run-off" transcription. Importantly, it has been shown that a partial DNA duplex at the promoter region is sufficient for T7 RNA polymerase to bind and generate a transcript from a single-stranded (ss) synthetic DNA template, greatly simplifying the system[147].

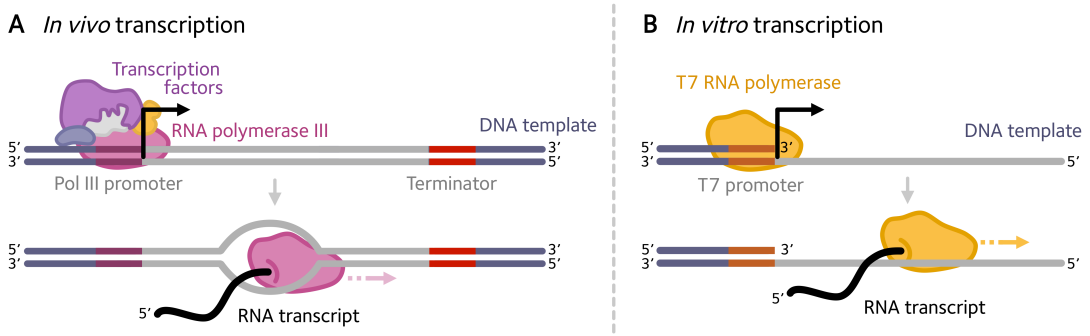


Figure 2.1 | Schematic representation of sgRNA generation approaches. sgRNA can be produced by *in vivo* transcription using the human Pol III promoter which requires several transcription factors to initiate the process (A). Alternatively, sgRNA can be generated by *in vitro* transcription using the bacteriophage promoters without the need of transcription factors (B).

Although effective, *in vivo* and *in vitro* transcription requires specific 5' nucleotides for transcription initiation (e.g. 5'-A, 5'-G, 5'-A/G, 5'-GG, 5'-GA and 5'-G for U3[142], U6[28], H1[144], T7[147,148], SP6[149] and T3[150] promoters, respectively). These requirements constrain the targetable DNA sequences by Cas9. To overcome this limitation, the initiating nucleotides are often appended to the 5' end of a designed 20-nt guiding sequence. The effect of the additional nucleotides on genome editing varies depending on the spacer sequence with reports suggesting little effect on Cas9 activity, improved off-target activity or even reduced on-target activity[26,151–153].

Consequently, many strategies have been developed to generate sgRNAs with a precise 20-nt spacer. In general, they involve flanking the sgRNA with a cleavable RNA motif such as Cys4-cleavable RNA[154], ribozymes[155] or tRNA[156]. Upon processing, these RNAs will be cleaved at a defined position, yielding the gRNA with a defined 20-nt guiding sequence. Moreover,

interspacing multiple gRNAs with these cleavable motifs enables a single RNA transcript to be processed into multiple gRNAs to target multiple locations simultaneously.

Polymerase-based methods are accessible to many research groups, yet there are clear limitations. Although a single sgRNA can be produced from a single synthetic dsDNA, a much larger DNA template, which is beyond the limit of chemical synthesis such as PCR amplicons or DNA plasmids, are required for multiplexed sgRNA generation. This is a time-consuming and labour-intensive process and is even more challenging for arrayed sgRNA libraries. For *in vitro* transcription, extensive purification needs to be performed to remove unincorporated triphosphates, DNA template and pre-mature sgRNA products, however the latter two are often co-purified with the full-length sgRNA. Furthermore, IVT results in errors at the 3' end, and more importantly at the 5' end of the sgRNA within the spacer sequence that is crucial for DNA target specificity[157].

Consequently, ~100-nt synthetic sgRNA offers the ideal solution and allows chemical modifications to be incorporated into the sgRNA, which have been shown to improve genome editing[103,122,130–133,137–139].

2.1.2 Oligonucleotide Solid-phase Synthesis

sgRNA can be prepared using automated solid-phase synthesis[158], a process that allows the four different RNA monomers to be added stepwise from the 3' to 5' direction. In contrast to the generation of the sgRNA by transcription, no template is required. Each RNA monomer is incorporated in a synthesis cycle consisting of 1) activation and coupling, 2) capping, 3) oxidation and 4) detritylation (**Figure 2.2**).

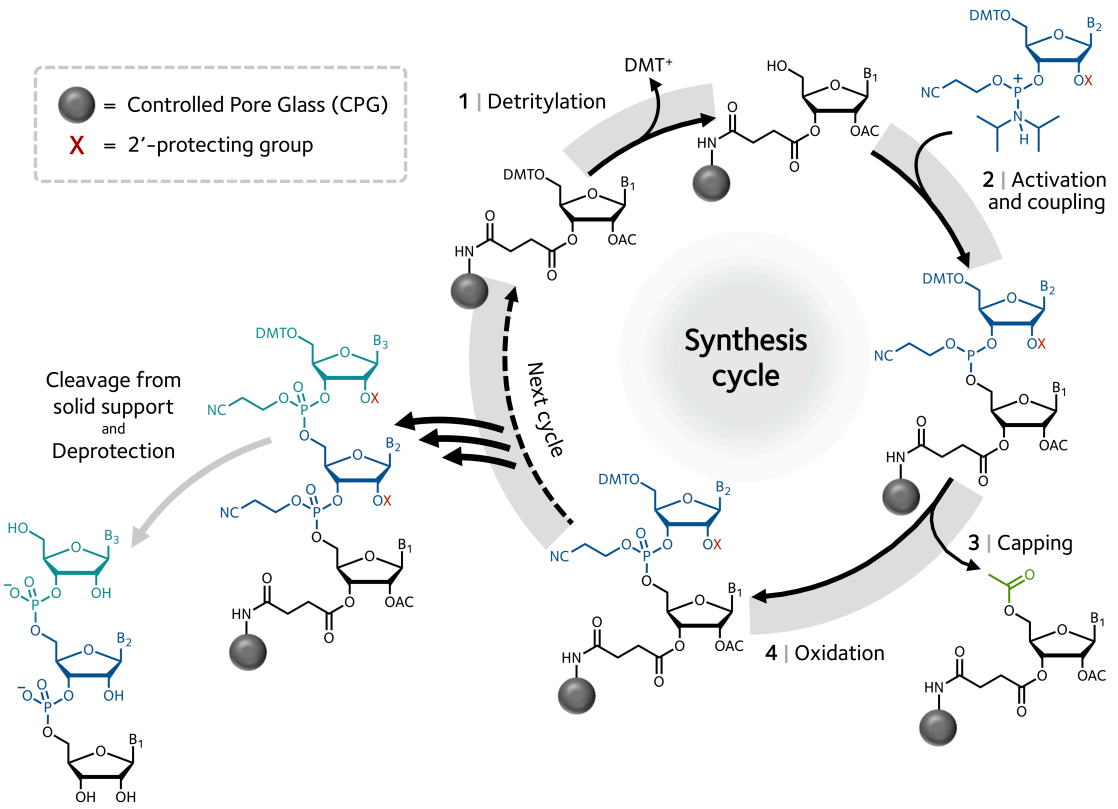


Figure 2.2 | Solid-phase phosphoramidite synthesis cycle of RNA.

The synthesis cycle is initiated by the removal of the 5'-DMT protecting group (detritylation) from an oligonucleotide attached to a controlled pore glass (CPG) solid support by mild acid treatment. This is followed by the coupling of the activated incoming RNA phosphoramidite *via* nucleophilic substitution resulting in a phosphor triester linkage. Since coupling is imperfect, the unreacted 5'-OH is capped by an acetyl group to prevent it from reacting in subsequent cycles, resulting in truncated or "failure sequences". Iodine-mediated oxidation then converts the phosphite triester (P(III)) to phosphate triester (P(V)). This cycle of steps is repeated until the desired oligonucleotide is obtained. At the end of the synthesis, the oligonucleotide is cleaved from the support, deprotected and purified.

Solid-phase RNA synthesis is very similar to DNA synthesis, except the 2'-hydroxyl group on the ribose sugar has to be protected to prevent 3'–2' migration of the phosphate backbone or RNA degradation from base-mediated 2'-OH attack of the phosphate backbone during deprotection (**Figure 2.3**).

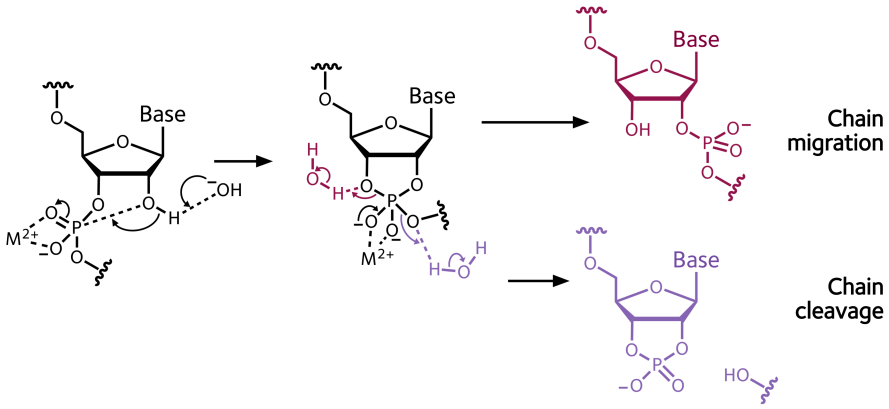


Figure 2.3 | RNA 3'–2' phosphate backbone migration (pink) and chain cleavage (purple) mechanisms. Metal ions (M^{2+}) such as magnesium facilitate these reactions by neutralising the charge of the phosphate backbone.

To date, there are several chemistries available for 2'-OH protection, which include 2'-*O*-*tert*-butyldimethylsilyl (TBDMS)[159], 2'-*O*-triisopropylsilyloxymethyl (TOM)[160,161], bis(acetoxyethoxy)-methyl ether (ACE)[162,163], and 2'-*O*-thionocarbamate (TC)[164] protecting groups (**Figure 2.4**).

TBDMS (**Figure 2.4A**) is the most widely used as the required phosphoramidites are commercially available at a reasonable cost. The TBDMS group can be easily and selectively removed by treatment with triethylamine trihydrofluoride (**Figure 2.5**)[165,166], however its steric bulk inhibits monomer coupling resulting in much longer coupling times compared to DNA synthesis. These factors combined limit the length of RNA that can be synthesised in good yields and purity to ~70 nucleotides[167]. This problem can be reduced by the use of more powerful activators such as 5-ethylthio-1*H*-tetrazole (ETT)[168] or 5-(benzylmercapto)-1*H*-tetrazole (BTT)[169] rather than the standard 1*H*-tetrazole.

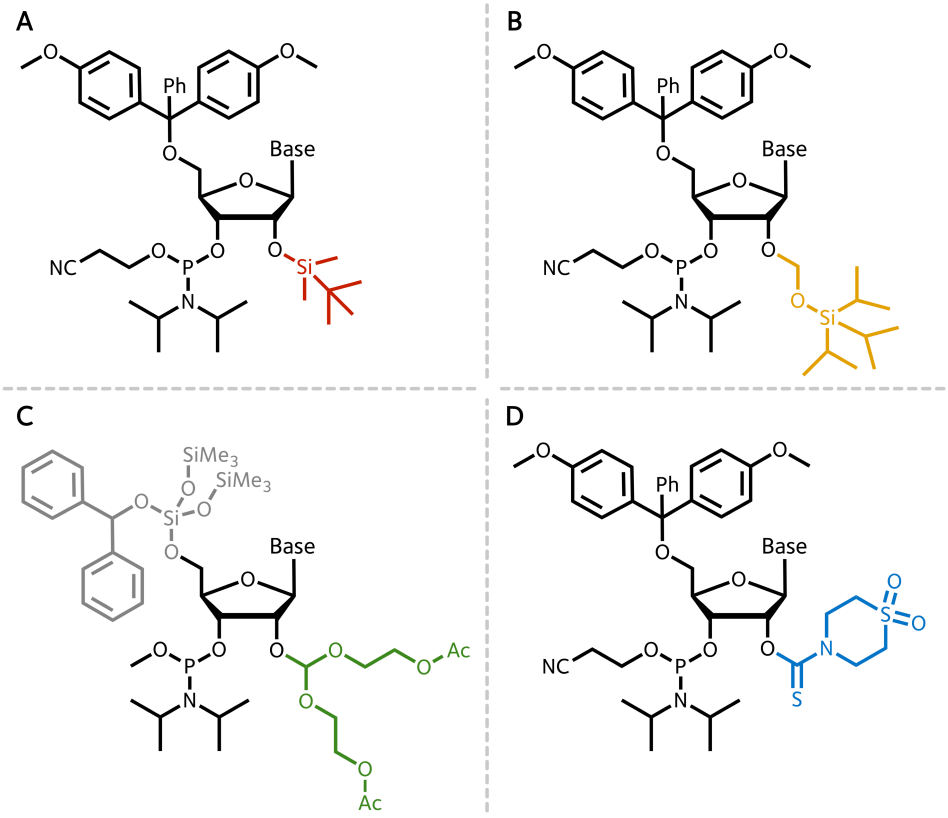


Figure 2.4 | 2'-protecting groups of RNA. (A) TBDMS. (B) TOM. (C) ACE. (D) TC.

The TOM protecting group (**Figure 2.4B**) reduces the steric bulk around the phosphoramidite coupling site and reduces 2'–3' protecting group migration during the monomer production, making monomer synthesis easier. It is still compatible with the TBDMS-based RNA synthesis and importantly requires a far shorter coupling time. As a result, long RNA up to 84 nucleotides have been successfully synthesised using the TOM-protected monomers[161].

The acid-labile ACE protecting group (**Figure 2.4C**) is a very different approach aimed at improving RNA synthesis[162,163]. In contrast to TBDMS and TOM strategies, the ACE method employs completely different chemistry to protect the RNA phosphoramidite. The fluoride-labile bis(trimethylsilyloxy)benzhydryloxysilyl (BzH) group is used instead of the acid-labile DMT to protect 5'-OH of ribose sugar. The cyanoethyl group, which is cleavable in the presence of fluoride, is replaced by a methyl group for phosphate protection. In addition to improvements in yields, purity and length, the 2' ACE-protected RNA is soluble in water, enabling the RNA to be purified and handled while being protected. As a result, RNA degradation is greatly minimised.

Despite showing great promise, the TOM and ACE strategies are not broadly used by the community as the required monomers are not commercially available.

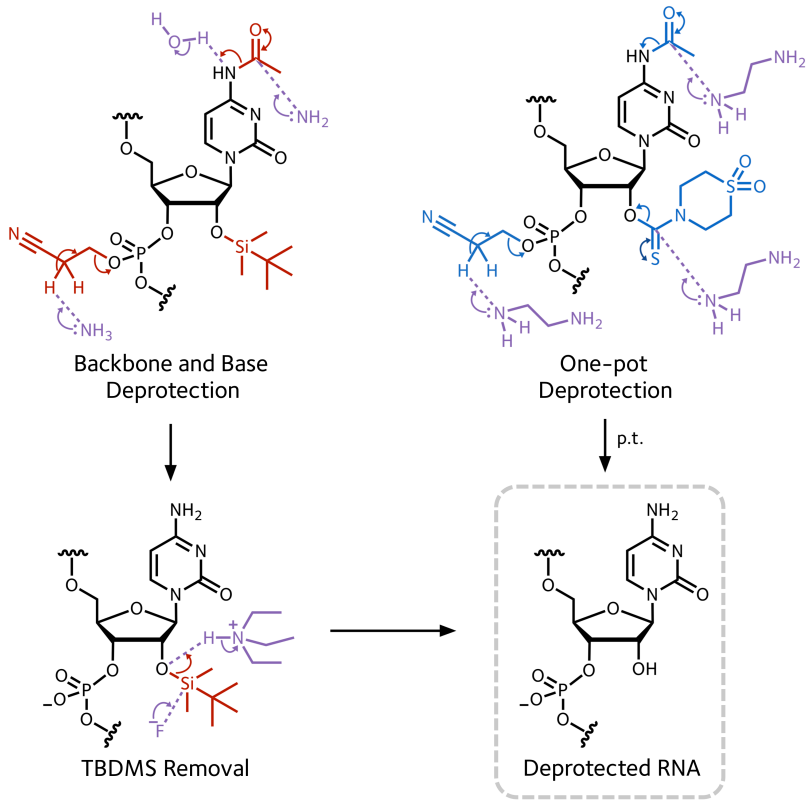


Figure 2.5 | Step-wise deprotection of 2'-*O*-TBDMS (red) oligonucleotides compared to one-pot deprotection of 2'-*O*-TC (blue) oligonucleotides. One nucleotide is shown to illustrate these mechanisms.

Recently, the TC protecting group (**Figure 2.4D**) has been introduced as another candidate for long RNA synthesis (~100 nucleotides)[164]. Usually the deprotection of nucleobases and phosphate groups is performed prior to that of the 2'-OH group in all the methods described above (**Figure 2.5A**). TC, however, can be removed in the same conditions as the protecting groups on the nucleobases (**Figure 2.5B**). Thus, the removal of all protecting groups can be achieved in a single step using ethylenediamine, which greatly simplifies the workflow. Importantly, like TBDMS chemistry, the TC phosphoramidite monomers of A, C, G and U are commercially available.

2.1.3 Click Chemistry

Due to imperfect coupling (~98.5–99.5%), solid phase synthesis using the phosphoramidite approach has an upper limit in terms of achievable oligonucleotide length. Beyond this length, yield and purity significantly decrease. For example, the maximum yield of a 100-mer is 36.6% if coupling efficiency is 99%[170] (E^N where E = coupling efficiency and N = number of cycles). This is even lower for RNA synthesis due to the steric hindrance of the chosen 2'-OH protecting group, which reduces coupling efficiency. Furthermore, undesired side-reactions such as depurination[171–173] can occur from prolonged exposure to chemicals in different synthesis steps in both DNA and RNA synthesis.

To overcome these problems, shorter oligonucleotides can be synthesised and covalently linked to form larger constructs. Enzymes such as T4 DNA ligase can be used to connect the oligonucleotides bearing a 5'-phosphate and 3'-OH groups. However, ligase-based approaches are not efficient, require very specific buffer conditions, are not tolerant of structured oligonucleotides and are restricted to a particular substrate (e.g. T4 DNA ligase only joins dsDNA or dsRNA but not ssDNA or ssRNA). As a result, chemical ligation methods are attractive. For this purpose, Copper(I)-catalysed Azide Alkyne Cycloaddition (CuAAC) chemical ligation is well-developed in our group.

Azide-alkyne cycloaddition initially discovered by Huisgen relies on extremely high temperature and gives a mixture of two isomeric products depending on the orientation of the alkyne and azide[174] (**Figure 2.6A**). However, in the presence of a Cu(I) catalyst, the specific formation of a 1,4-triazole can be accelerated by 10^7–10^8 fold[175] (**Figure 2.6B**). The CuAAC reaction is the first and the most popular example of click chemistry, which is defined as follows:

"The reaction must be modular, wide in scope, give very high yields, generate only inoffensive byproducts that can be removed by nonchromatographic methods, and be stereospecific (but not necessarily enantio-selective). The required process characteristics include simple reaction conditions (ideally, the process should be insensitive to oxygen and water), readily available

starting materials and reagents, the use of no solvent or a solvent that is benign (such as water) or easily removed, and simple product isolation." [175]

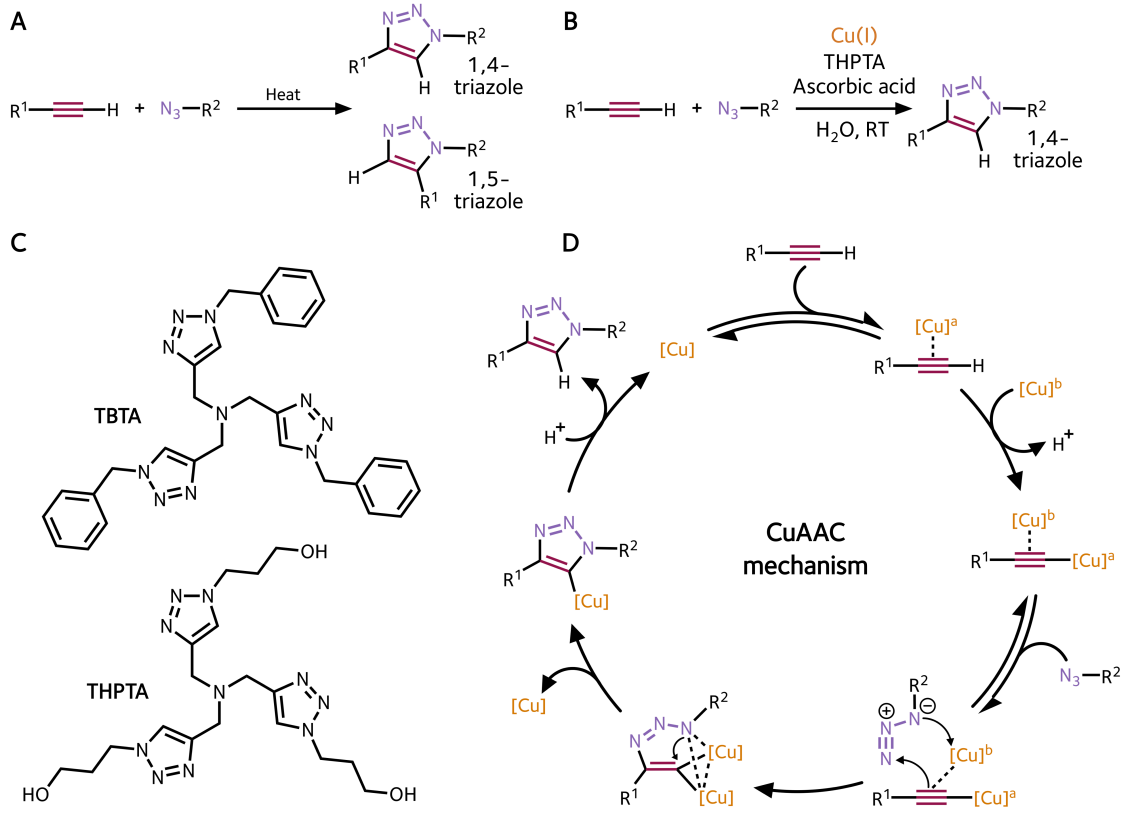


Figure 2.6 | Azide–alkyne cycloaddition and its proposed mechanism. (A) shows the un-catalysed Huisgen reaction. (B) shows the aqueous copper-catalysed reaction and the ligands (C) used for this. (D) is the proposed mechanism for CuAAC.

To perform the reaction in water, two additional components are often added. Ascorbate is used as the reducing agent to convert Cu(II) to the active Cu(I)[176] and ligands such as tris(benzyltriazolylmethyl)amine (TBTA)[177] or tris(3-hydroxypropyltriazolylmethyl)amine (THPTA)[178] (**Figure 2.6C**) serve as a Cu(I)-stabilising agent, enhancing click reaction efficiency. The proposed mechanism for the CuAAC[179] reaction is illustrated in **Figure 2.6D**.

2.1.4 Click Chemical Ligation of Oligonucleotides

The initial triazole linkage in DNA (**Figure 2.7B**) pioneered by our group is formed between the oligonucleotides containing 3'-azide and 5'-propargyl amino groups[180]. However, its excessive length and the rigid amide restricts its uses in biological applications; it prevents DNA

polymerases from reading through the bases around the triazole accurately, and leads to the enzyme skipping the base adjacent to the triazole linkage[180]. As a consequence, a new triazole linkage with improved properties was developed (**Figure 2.7C**).

Structural studies[181] show that this new triazole does not greatly alter the characteristics of canonical DNA; it is biocompatible, allowing DNA polymerases as well as T7 RNA polymerases[182,183] to accurately read through the bases flanking the triazole *in vitro* and *in vivo*[182,184]. CuAAC has also been used to successfully ligate many pieces of RNAs together to construct a functional large hairpin ribozyme[185]. Even though a long alkyne chain containing non-biocompatible triazole linkage was used in this work, it was not a problem as the triazole linkages were inserted between nucleobases across the major groove of the RNA helix. In the same study the biocompatible triazole was used in a hammerhead ribozyme. Both ribozymes were functional.

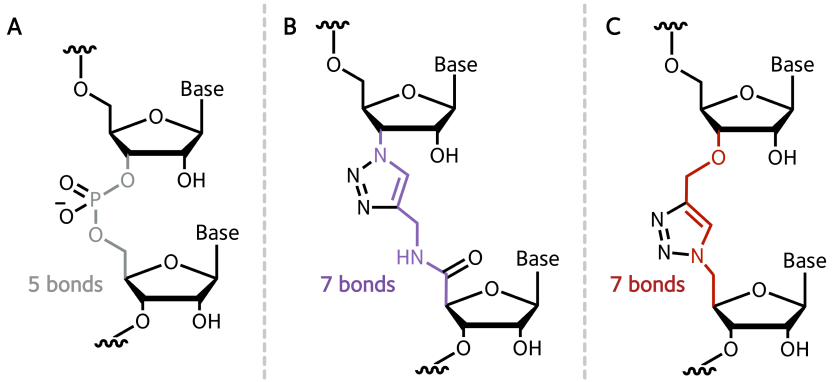


Figure 2.7 | Click linkages pioneered by our group. (A) shows the native phosphodiester bond with a five-bond spacing between two adjacent nucleotides. (B) shows the initial triazole linkage developed in our group with seven-bond spacing, and (C) shows the biocompatible triazole linkage with seven-bond spacing.

Despite the excellent efficiency of the CuAAC reaction there is a potential problem. Excess copper ions can inhibit enzymatic processes and cause cellular toxicity. An alternative click reaction has been developed that proceeds rapidly without a copper catalyst. This spontaneous reaction is called a Strain-Promoted Azide Alkyne Cycloaddition (SPACC)[186] (**Figure 2.8A**). It requires the alkyne functional group to be incorporated into an 8-membered ring structure such

as bicyclo[6.1.0]nonyne (BCN), dibenzocyclooctyne (DBCO), 4-dibenzocyclooctynol (DIBO) and difluorocyclooctyne (DIFO) (**Figure 2.8B**). The release of the ring strain upon reacting with azide (caused by sp-hybridisation changing to sp^2) accelerates the reaction allowing it to occur at ambient conditions. The reaction is also compatible with RNA[187] and surprisingly the resulting bulky linkage can be read through by reverse transcriptase with one nucleotide omitted around the modified linkage[187].

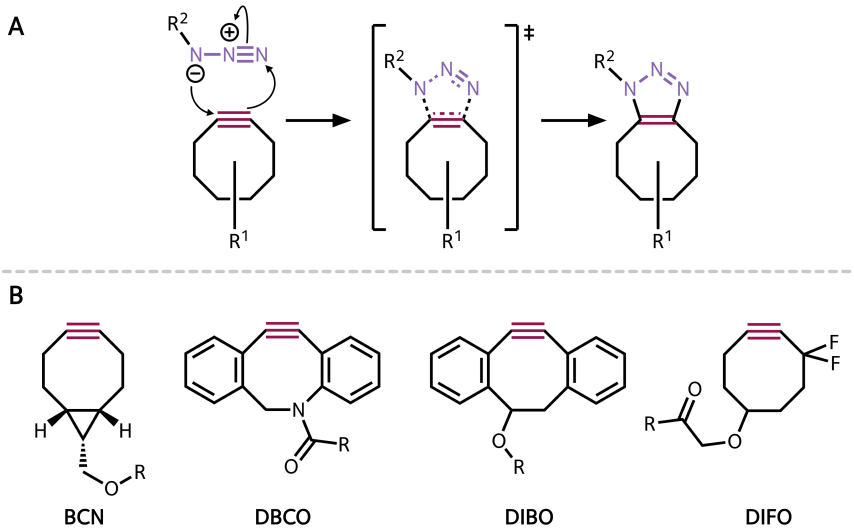


Figure 2.8 | The SPAAC mechanism (A) and examples of cyclic alkyne derivatives (B) required for it.

2.2 Aims and Objectives

In this chapter, CuAAC and SPAAC click ligation will be explored in the CRISPR system, particularly in the generation of synthetic sgRNA. The sgRNA will be split into two parts: a variable part containing a DNA-targeting sequence and a fixed part bearing the tracrRNA sequence. These two components can be produced in higher purity and can be chemically modified to improve activity, for example through the introduction of bridged nucleic acids[138] that improve on-target activity or phosphorothioates[130–132] and 2'-OMe groups that enhance nuclease stability (see **Chapter 1**, **Section 1.10**). They will then be ligated using click chemistry to generate sgRNAs that are functional *in vitro* and in cells. The goal is to create a simple system where many different DNA-targeting RNA components can be mixed and matched with a fixed RNA component to create a library of sgRNAs.

2.3 Click 37–66 sgRNA Constructs

2.3.1 Design Strategy

The initial goal was to explore the viability of different RNA click ligation strategies. Therefore, the sgRNA was broken into two parts at the centre of the artificial tetraloop (**Figure 2.9**). This position is known to accommodate large modifications[188,189] and gives a DNA-targeting 37-mer crRNA and a fixed 66-mer tracrRNA. These lengths are a reasonable starting point for good quality RNA synthesis in acceptable yields. Upon click ligation, it was envisaged that a range of sgRNAs with different linkages would be obtained (**Figure 2.9**), allowing us to evaluate the effects of chemical synthesis and click chemistry on the function of sgRNAs. Moreover, at this position, the artificial linkages should form no contacts with Cas9 isolating the effects of synthesis from those of the artificial linkage on Cas9 activity.

The standard length of the crRNA is 34 nt (20-mer segment for DNA targeting plus 14-mer for hybridising to the tracrRNA), however GGG was appended to the 5' end of crRNA for it to be directly comparable to the unmodified *in vitro*-transcribed (IVT) sgRNA. The 37-mer crRNA was synthesised with a terminal 3'-*O*-propargyl nucleotide (crRNA I, **Figure 2.9**) or a 3'-C7-amino group (crRNA IIa, **Figure 2.9**) that was subsequently coupled with a DBCO-NHS ester. The 66-mer tracrRNA was synthesised with a terminal 5'-OH, which was converted to azide post-synthetically (tracrRNA IV, **Figure 2.9**), or a long linker amine (tracrRNA IIIa, **Figure 2.9**) that can be labelled with an azide-NHS ester.

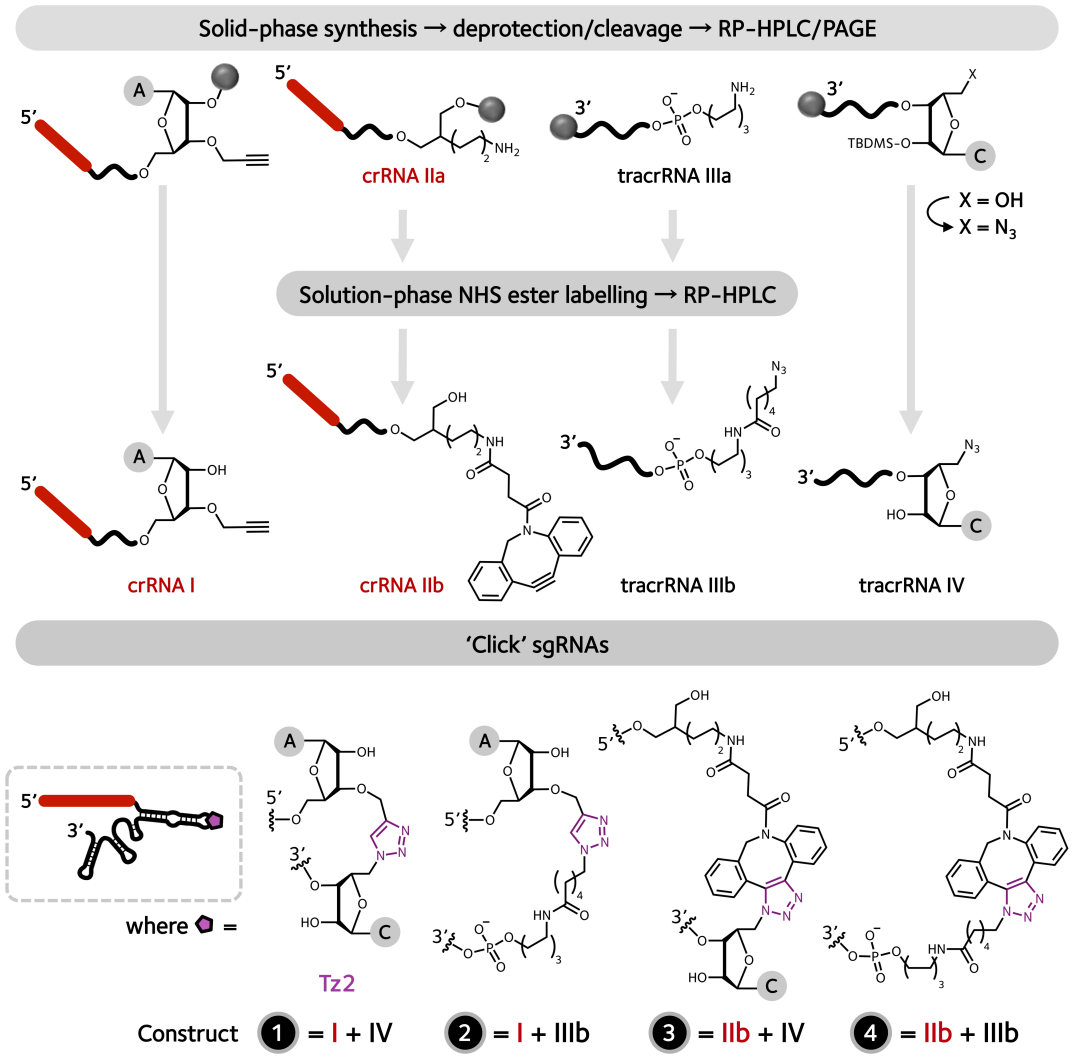


Figure 2.9 | Generation of click 37–66 sgRNA constructs. crRNA I was obtained directly from solid phase synthesis. tracrRNA IV underwent conversion to azide on the solid support yielding 5'-azido tracrRNA. crRNA IIa and tracrRNA IIIa were synthesised with terminal amines which were coupled with DBCO-NHS and C6-azide-NHS esters respectively. Click sgRNAs bearing different linkages were generated upon mixing and matching the different crRNAs and tracrRNAs. Full sequences can be found in **Appendix Table 3** and **4**.

2.3.2 2'-*O*-TBDMS RNA Synthesis and Deprotection

As discussed in **Section 2.1.2**, there are multiple options for protecting the 2'-OH groups for RNA synthesis. TBDMS and TC protecting groups were used for this work due to their commercial availability. Initially, 2'-TBDMS RNA monomers with a *tert*-butylphenoxyacetyl (tac) protecting group on A and G and an acetyl protecting group on C (**Figure 2.10**) were used to generate 3'-propargyl crRNA I, and a 66-mer tracrRNA which was converted from a 5'-OH group to a 5'-azide

(tracrRNA IV). The azide conversion step (**Figure 2.11**) occurs before deprotection to ensure that only the terminal 5'-OH is converted to a 5'-iodo intermediate and then to a 5'-azide. Based on mass spectrometry analysis (**Figure 2.11**), the 5'-azide conversion was successful using standard conditions[185].

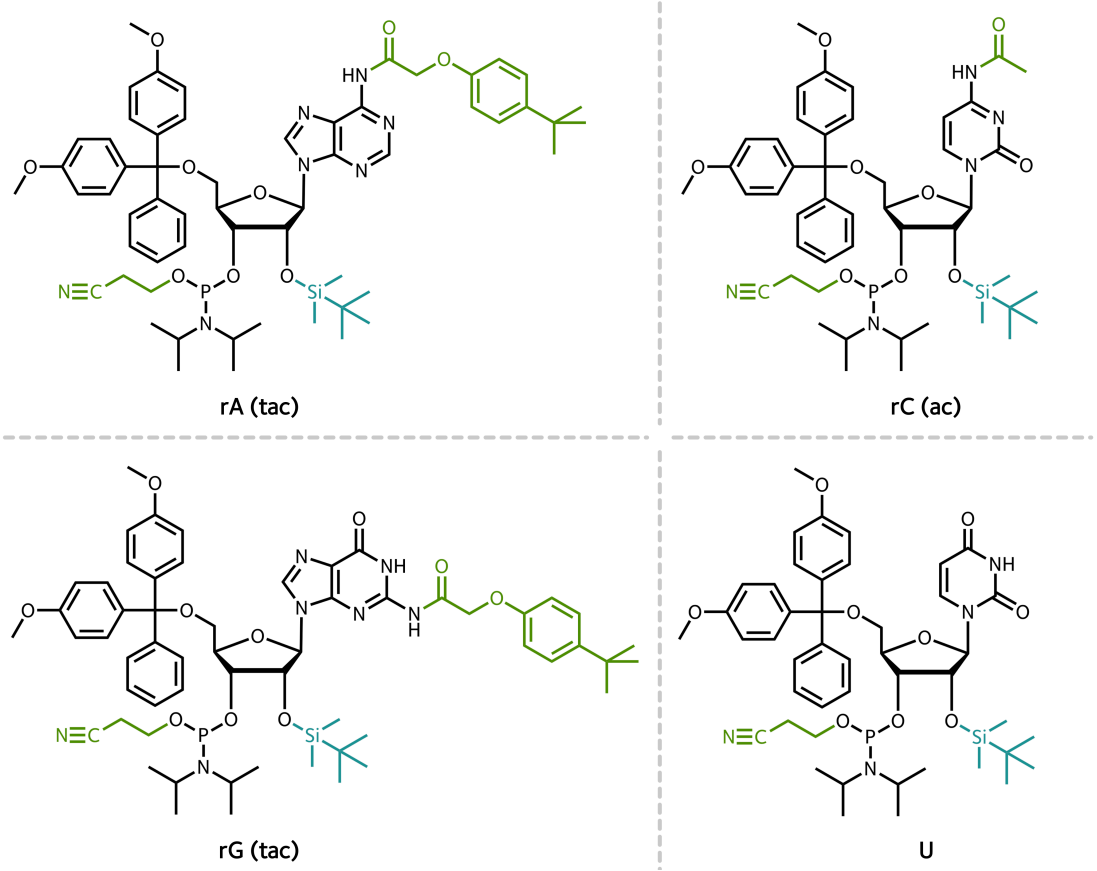


Figure 2.10 | TBDMS-RNA monomers. Cyanoethyl and nucleobase protecting groups (green) along with *tert*-butylphenoxyacetyl (tac) on A / G and acetyl (ac) on C are removed after RNA assembly by treatment with aqueous/ethanolic ammonia or metylamine. Following base deprotection, the 2'-TBDMS group (blue) is removed using triethylamine trihydrofluoride. Mechanisms for deprotection are shown in **Figure 2.5**.

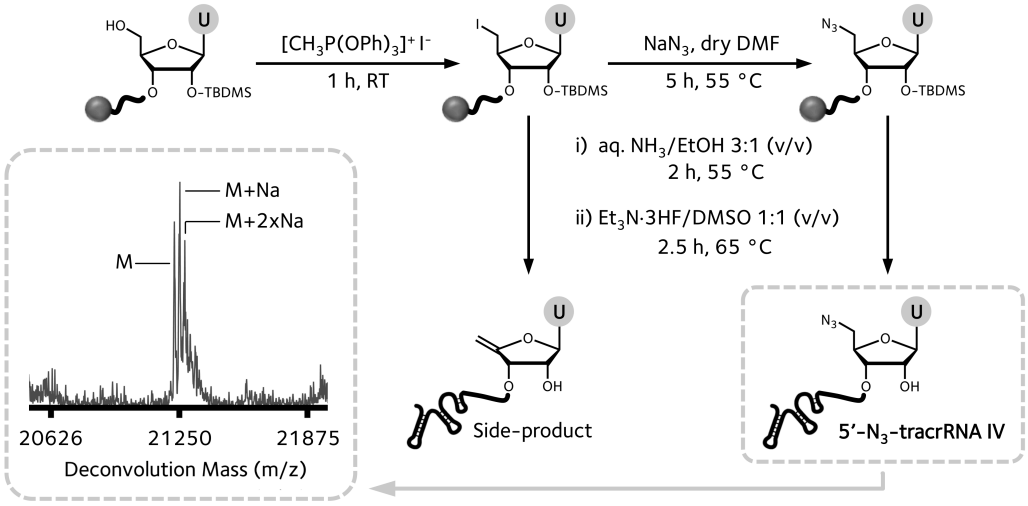


Figure 2.11 | Azide conversion scheme. The 5'-OH-tracrRNA was converted to a 5'-iodo-tracrRNA intermediate prior to azide conversion yielding the tracrRNA IV (sequence in **Appendix Table 3**). Incomplete iodo conversion can lead to the formation of an alkene side-product *via* elimination of the iodo group reaction during deprotection in ammonia, which is observed by mass spectrometry. M = 21225 (expected); 21227 (found). Note that Na^+ adducts were also detected.

After solid-phase synthesis is complete, the RNA resin was exposed to ethanolic ammonium hydroxide (1:3)[190,191], a standard deprotection condition, prior to the 2'-TBDMS removal using triethylamine trihydrofluoride[165,166,168]. This cleaves RNA from the solid support and removes tac and cyanoethyl protecting groups. Ethanol was added to enhance 2'-TBDMS protected RNA solubility and inhibit the hydrolysis of the 2'-TBDMS group[191–193]. Premature removal of TBDMS allows the attack of 2'-OH on the adjacent phosphodiester group, leading to 3'–2' backbone migration or chain cleavage (see **Figure 2.3**). HPLC-MS traces of both 37-mer and 66-mer RNAs showed incomplete nucleobase deprotection. The major HPLC peak corresponds to the fully deprotected RNA, however a minor peak representing the desired RNA with a single tac protecting group (+191) was also observed (**Figure 2.12**). The 'plus-tac' peak of the 66-mer tracrRNA IV (**Figure 2.12B** and **2.12C**) was larger than that of the 37-mer crRNA I, reflecting the difficulty of deprotection after long RNA synthesis.

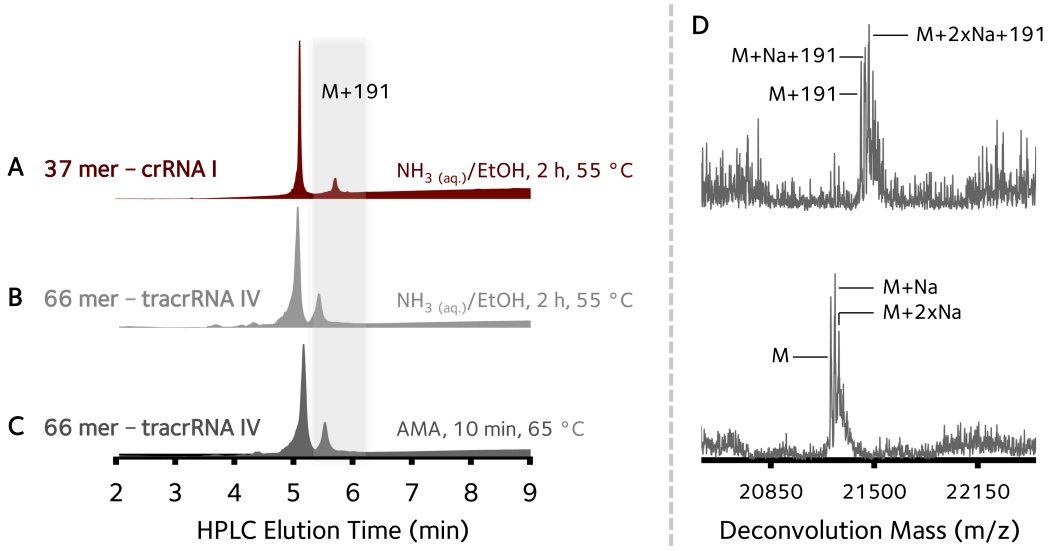


Figure 2.12 | HPLC profiles of crude TBDMS-synthesised RNAs after deprotection by standard conditions. (A) 37-mer crRNA I and (B) 66-mer tracrRNA IV were deprotected using $NH_{3\ (aq.)}$/EtOH (3:1) for 2 h, 55 °C. (C) 66-mer tracrRNA IV was also deprotected using Ammonium hydroxide/MethylAmine (AMA, $NH_{3\ (aq.)}$:40% $MeNH_{2\ (aq.)}$, 1:1 v/v) for 10 min, 65 °C. The crude mixture consisted of fully deprotected RNA (main peak) and a single tac-containing RNA peak (small peak) due to incomplete nucleobase deprotection. The 37-mer crRNA I contained a relatively smaller amount of 'plus-tac' contaminant than the 66-mer tracrRNA IV (mass +191). Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. (D) illustrates 66-mer tracrRNA IV deconvolution from (C) with the fully deprotected RNA in bottom panel and the 'plus-tac' contaminant in top panel. M (37-mer crRNA I) = 11999 (expected); 11999 (found). M (66-mer tracrRNA IV) = 21225 (expected); 21227 (found). The oligonucleotides used can be found in **Appendix Table 3**.

In order to reduce the plus-tac contaminant, a more nucleophilic Ammonium hydroxide/Methylamine (AMA) solution was used to deprotect 66-mer tracrRNA IV that enables ultra-fast deprotection in 5–10 minutes[194,195], with an acetyl protecting group on C preventing transamination side reaction by methylamine[194]. However, the HPLC-MS trace (**Figure 2.12B** and **2.12C**) still showed similar levels of incomplete deprotection.

Therefore, attempts were made to further optimise the deprotection step. Firstly, for ammonia/ethanol mixtures, the incubation time was increased to allow full deprotection, and it was also performed at room temperature to avoid RNA degradation (**Figure 2.13C** and **2.13D**). Unfortunately, no significant improvement was observed using the different conditions. The biggest change was the large amount of truncated fragments seen as a broad curve immediately

in front of the main HPLC peak, which illustrates the variability of RNA synthesis from batch to batch (**Figure 2.13A**, **Figure 2.12A**, the same condition $NH_{3\ (aq.)}$/EtOH (3:1), 2 h, 55 °C).

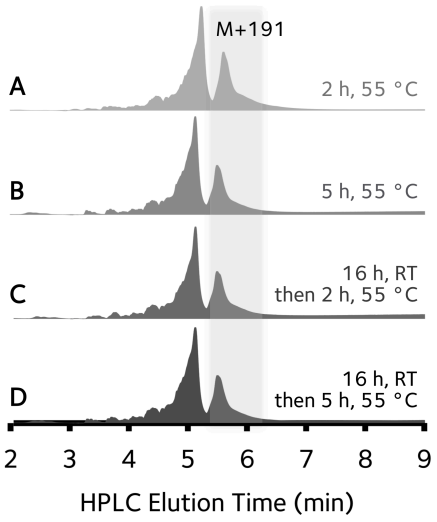


Figure 2.13 | HPLC chromatograms of the crude tracrRNA IV deprotected using $NH_{3\ (aq.)}$/EtOH (3:1) with varied temperature and incubation times. Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. The trace shows fully deprotected RNA (main peak) and a tac-containing RNA impurity (small peak, +191). Deprotection was carried out for (A) 2 h, 55 °C, (B) 5 h, 55 °C, (C) 16 h, RT (25 °C) then 2 h, 55°C and (D) 16 h, RT (25 °C) then 5 h, 55 °C. M = 21225 (expected); 21227 (found). The oligonucleotide used can be found in **Appendix Table 3**.

Next, AMA conditions were optimised. Temperature and time were varied to increase deprotection with ethanol added for the reasons described earlier. DMSO was also included to help solubilise the partially deprotected RNA[196] (**Figure 2.14**). Yet, similar to ethanol/ammonia deprotection conditions, the 'plus-tac' by-product still could not be eliminated. Since this by-product was unavoidable for tracrRNA IV, the standard $NH_{3\ (aq.)}$/EtOH (3:1, 2 h, 55 °C) condition was used to deprotect all subsequent RNAs synthesised by the TBDMS route for convenience.

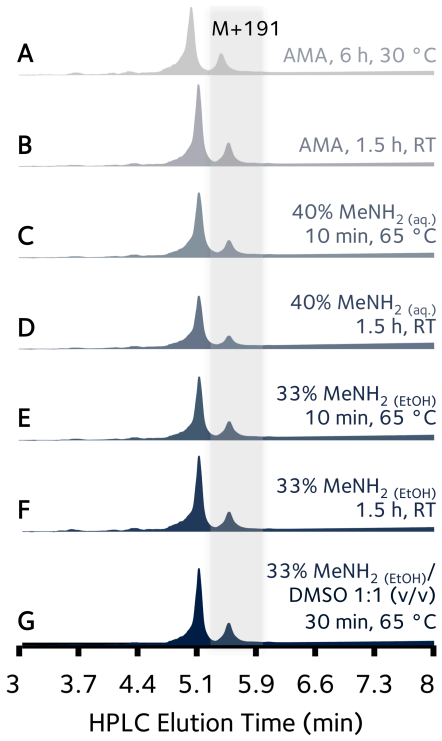


Figure 2.14 | HPLC traces of the crude tracrRNA IV deprotected using various $MeNH_2$-based conditions. Deprotection was carried out under (A) AMA, 6 h, 30 °C, (B) AMA, 1.5 h, RT (25 °C), (C) 40% $MeNH_{2\ (aq.)}$, 10 min, 65 °C, (D) 40% $MeNH_{2\ (aq.)}$, 1.5 h, RT (25 °C), (E) 33% $MeNH_{2\ (EtOH)}$, 10 min, 65 °C, (F) 33% $MeNH_{2\ (EtOH)}$, 1.5 h, RT (25 °C), and (G) $MeNH_{2\ (EtOH)}$/DMSO 1:1 (v/v), 30 min, 65 °C. Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. The trace shows fully deprotected RNA (main peak) and a tac-containing RNA impurity (small peak, +191). M = 21225 (expected); 21227 (found). The oligonucleotide used can be found in **Appendix Table 3**.

Imperfect monomer coupling results in a large amount of truncated tracrRNAs along with the full-length 66-mer oligonucleotide. These were removed by denaturing PAGE purification, a technique that separates RNAs based on size. However, RNA strands that retained a hydrophobic tac protecting group were not removed by PAGE purification and required further reversed-phase (RP) HPLC purification, a technique that separates oligonucleotides based on their hydrophobicity, to obtain the pure oligonucleotide.

2.3.3 2'-*O*-TC RNA Synthesis and Deprotection

2'-TC chemistry was also explored for RNA synthesis. It was used to synthesise a 37-mer crRNA with a 3'-C7-amine group and the 66-mer tracrRNA with a 5'-C6-amine for subsequent coupling to DBCO- and azide-NHS esters respectively for click ligation reactions (crRNA IIb and tracrRNA IIIb, **Figure 2.9**).

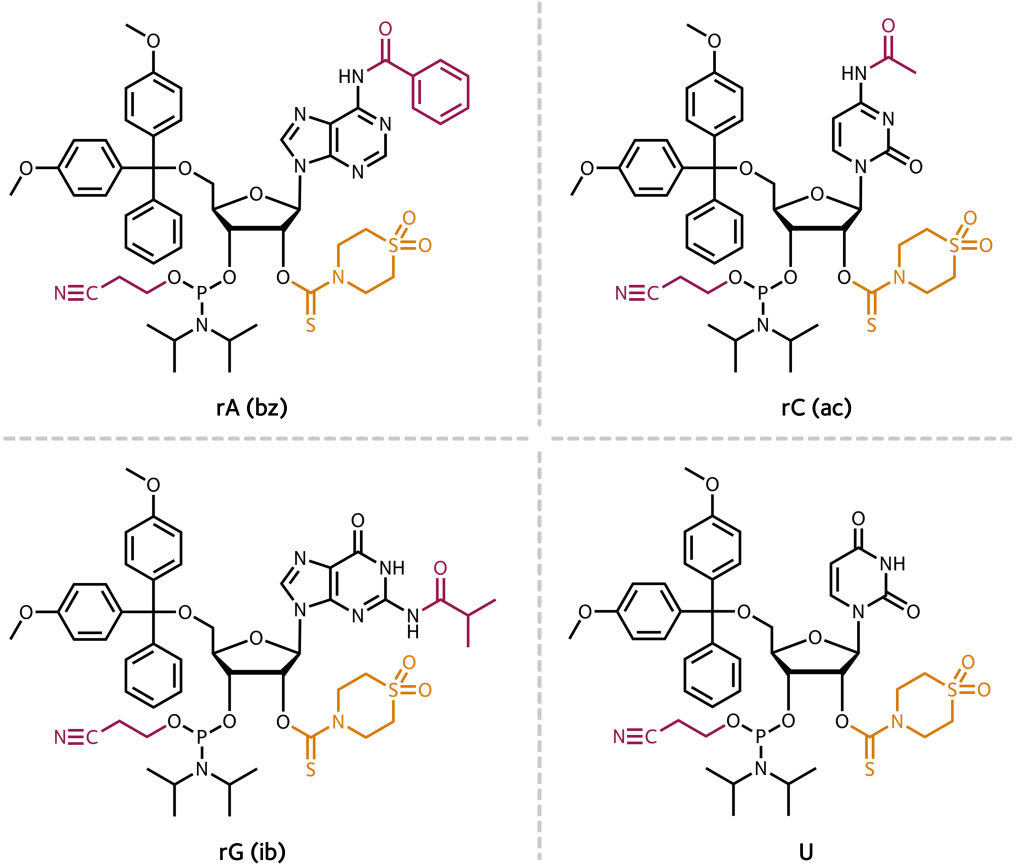


Figure 2.15 | TC-RNA monomers. Cyanoethyl and nucleobase protecting groups (pink): benzoyl (bz) on A, acetyl (ac) on C and isobutyryl (ib) on G are removed together with 2'-TC group (orange) using only ethylenediamine (EDA) in single-step deprotection.

TC-RNA monomers contain the same nucleobase protecting groups as standard DNA monomers (**Figure 2.15**). In contrast to TBDMS, the removal of the TC group occurs simultaneously with nucleobase deprotection using strongly nucleophilic ethylenediamine (EDA, **Figure 2.5**). This one-step deprotection minimises the handling time of RNA, and therefore reduces RNA degradation. The HPLC-MS profiles of the crude RNAs contain only a major peak for fully deprotected RNA with no detectable incompletely-deprotected RNA contaminants (**Figure 2.16**). Due to cleaner synthesis, RP-HPLC was sufficient to purify the full-length RNAs. Moreover, the 5'-alkyl-amino group improved purification. Whereas the full-length 37-mer RNA and all failure sequences contain a 3'-amino group, only the full-length 66-mer RNA should contain a 5'-C6-amino moiety (if no degradation occurs during deprotection). This increases the hydrophobicity of the desired RNA, increasing its retention and allowing it to be separated from the failure sequences.

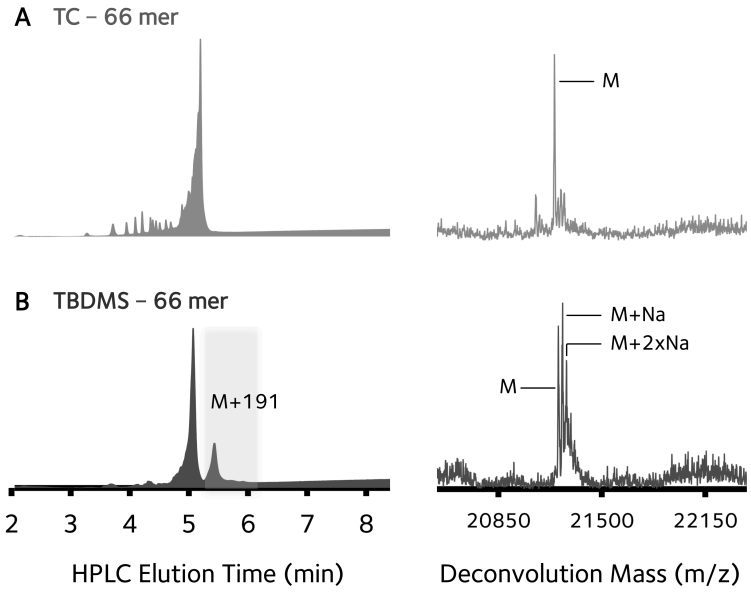


Figure 2.16 | HPLC of crude 66-mer tracrRNA IV synthesised by TC (A) and TBDMS chemistry (B) respectively. Note that the 5'-OH of TBDMS-tracrRNA IV underwent conversion to azide whereas the 5'-OH of TC-tracrRNA was not converted. The plus-tac (+191) contaminant was only generated from deprotection of TBDMS-RNA. The area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. M (TBDMS-tracrRNA with 5'-N_3) = 21225 (expected); 21227 (found). M (TC-tracrRNA with 5'-OH) = 21200 (expected); 21202 (found). The oligonucleotide sequence can be found in **Appendix Table 2** and **3**.

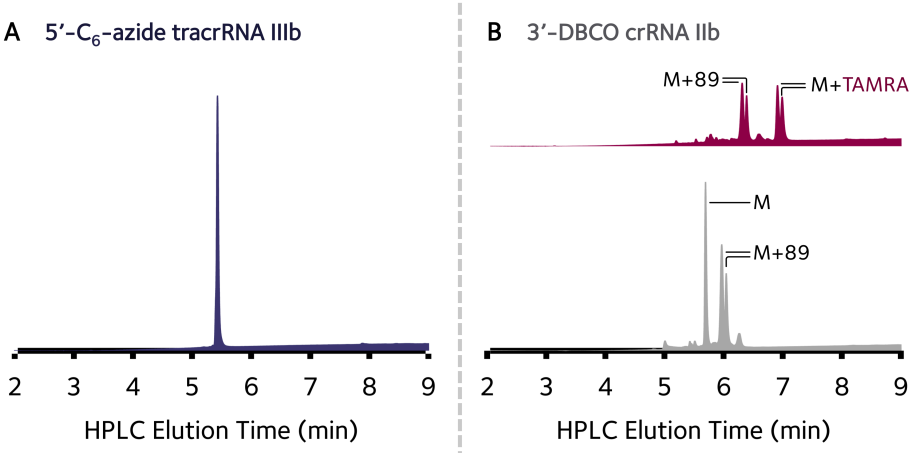


Figure 2.17 | HPLC chromatograms of RP-HPLC-purified (A) 66-mer tracrRNA IIIb and (B) 37-mer crRNA IIb. The contaminant (+89 adduct) was detected along with the DBCO-crRNA IIb. The activity of the DBCO-crRNA IIb was confirmed by reacting it with TAMRA-N_3 *via* the SPAAC reaction. Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. M (tracrRNA IIIb) = 21521 (expected); 21523 (found). M (crRNA IIb) = 12456 (expected); 12458 (found). The oligonucleotides used can be found in **Appendix Table 3**.

The purified crRNA IIa and tracrRNA IIIa were then labelled with DBCO- and azide-NHS esters respectively, passed through size exclusion column (a NAP™ column) to remove excess NHS ester, and purified by RP-HPLC to isolate the labelled RNA from the unlabelled RNA.

The 66-mer long linker 5'-C6-azide tracrRNA IIIb was observed as a single peak by mass spectrometry (**Figure 2.17A**). However, for 37-mer 3'-DBCO crRNA IIb, a +89 adduct in addition to the expected mass peak was also detected (**Figure 2.17B**). Since the amino precursor crRNA IIa was HPLC-purified, the expected adducts should be the salt counter-ions: copper (+63), sodium (+23), potassium (+39), triethylamine (+101) and hexafluoro-2-propanol (+168) adducts (the last two are from the mass spectrometry buffer), but the +89 mass does not correspond to any of these values. One possible explanation could be that an impurity in the mass spectrometer HPLC column reacted with the cyclic alkyne. To test this hypothesis, the HPLC-purified 37-mer crRNA IIb was reacted with large excess of TAMRA-azide. Provided that the contaminants are generated during mass spectrometry, all the 37-mer RNA should be successfully labelled with TAMRA and none of the +89 mass 37-mer RNA should be observed. However, a similar amount of +89 contaminant was still observed, indicating its presence prior to mass spectrometry (**Figure 2.17B**). The origin of this impurity is unknown but the 37-mer RNA with a terminal 3'-DBCO was still used.

2.3.4 5'-iodo Conversion of TC-RNA

Since the TC approach was very efficient in the synthesis of 66-mer tracrRNA IIIa, it was also used to produce a 66-mer 5'-OH tracrRNA for conversion to a 5'-azide. However, it has been reported by Sigma-Aldrich[197] that the 2'-thiocarbamate protecting group is susceptible to oxidation and generates a 2'-carbamate adduct that is not removed during ethylenediamine deprotection. This side-reaction is dependent on the iodine concentration, with the TC group remaining intact under 0.02 M iodine concentrations used during synthesis but not with 0.1 M iodine.

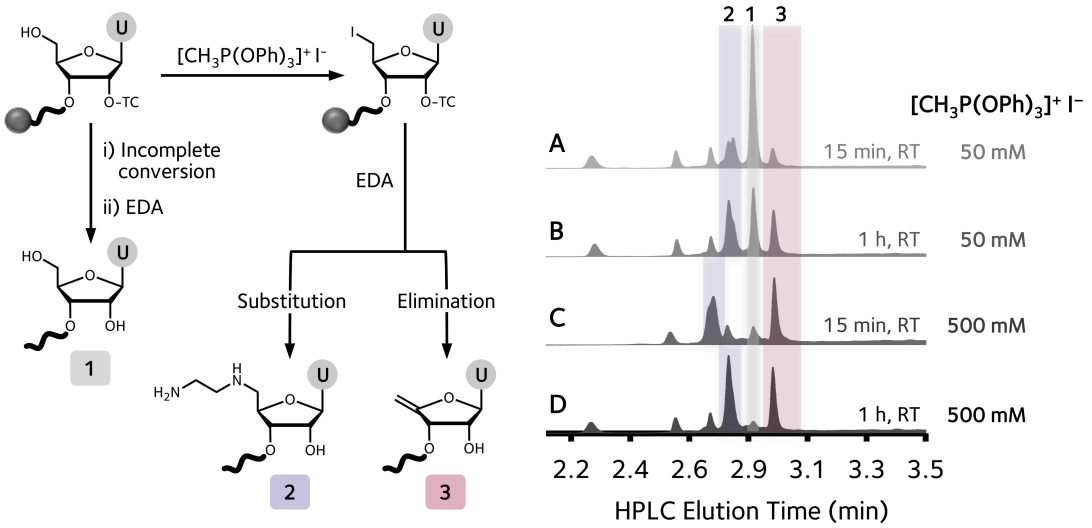


Figure 2.18 | Iodo conversion of short (10 mer) TC-RNA. Various concentrations of methyltriphenoxyphosphonium idodide ($[CH_3P(OPh)_3]^+ I^-$) solution and time: (A) 50 mM, 15 min, (B) 50 mM, 1 h, (C) 500 mM, 15 min and (D) 500 mM, 1 h were used to convert to the 5'-OH of the TC-RNA to 5'-iodo at RT. Deprotection by EDA/toluene 1:1, 6 h, RT were performed immediately after iodo conversion. Unsuccessful iodo conversion generated product 1 after deprotection with EDA. 5'-iodo-RNA reacted with EDA during deprotection and was transformed into product 2 or product 3 *via* two possible mechanisms: nucleophilic substitution or elimination. Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. The oligonucleotide sequence can be found in **Appendix Table 2** (Code: IodoOpt TC-RNA).

To form the 5'-iodo intermediate required for 5'-azide generation, a 0.5 M iodide solution is used. Thus, for this step, the effect of iodide concentration on a short TC-protected RNA (10 nt) was evaluated, with the aim of achieving efficient conversion to the iodo intermediate without oxidation of the TC group. Surprisingly, iodo conversion proceeded efficiently at 0.5 M iodide concentrations with no detectable TC oxidation (**Figure 2.18D**), with reduced iodide concentration impairing the reaction (**Figure 2.18A** and **2.18B**). Thus, 0.5 M iodide was used for iodo conversion of the 66-mer tracrRNA. Unfortunately, a broad peak containing a large number of species was observed, with deconvolution to give the exact mass of these products not possible (**Figure 2.19**). This suggests oligonucleotide degradation or incomplete deprotection. Thus, it appears that the TC method cannot be used to synthesise long RNAs that require conversion of the 5'-OH to iodo/azide groups. All RNA oligonucleotides requiring a 5'-azide group were prepared using TBDMS chemistry, and other oligonucleotides prepared with either TC or TBDMS chemistry.

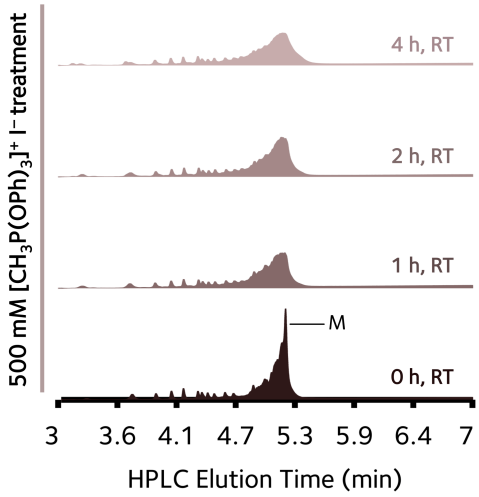


Figure 2.19 | Time-course conversion of 5'-hydroxyl to 5'-iodo group on 66-mer TC-RNA. The RNA was treated with 500 mM $[CH_3P(OPh)_3]^+ I^-$ solution and the reaction was stopped at various time points. Full length 66-mer RNA was observed without the treatment (0 h, RT). In the presence of $[CH_3P(OPh)_3]^+ I^-$, incomplete deprotection of the 2'-TC group was observed even after 1 h treatment. Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. M = 21200 (expected); 21202 (found). The oligonucleotide used can be found in **Appendix Table 2** (Code: TC-tracrRNA).

2.3.5 Click Ligation Optimisation

Prior to ligation of 3'-alkyne and 5'-azide modified RNAs, reaction conditions were first optimised using DNA which is cheaper. Using a standard protocol[198], untemplated ligation of DNAs proceeded efficiently. However, using the same conditions for RNAs, no significant ligation was observed (**Figure 2.20A**). Despite their different sequences, neither the DNA or RNA oligonucleotides used are expected to form strong secondary structures (calculated mfold[199] ΔG = –2 to –3 kcal/mol, 37 °C, 1 M mono-valent ions) and the final products are of similar length (~80 bases). This suggests RNA–RNA click ligation is inherently more difficult. To make the reaction more favourable, the reactive ends of the two strands can be brought in close proximity through the addition of an oligonucleotide ('splint') that hybridises to both strands. This significantly improved click ligation (**Figure 2.20A**).

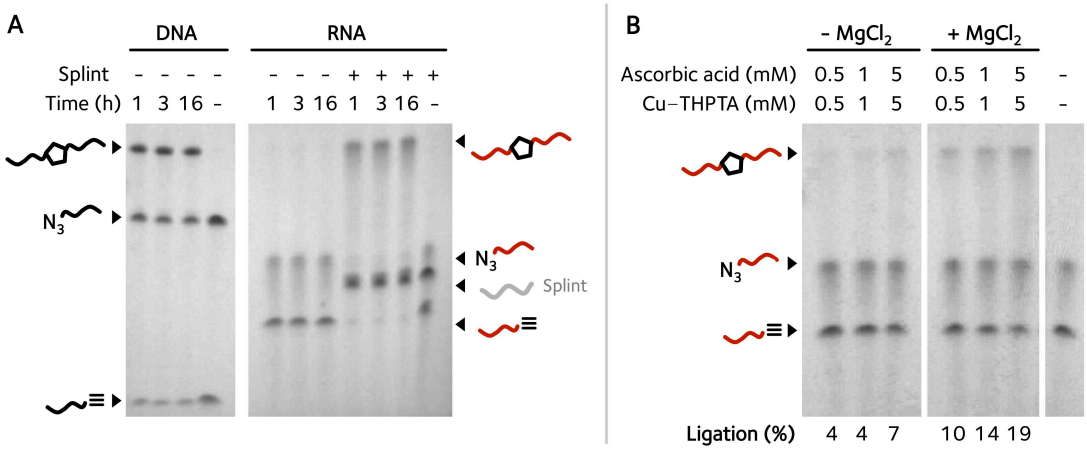


Figure 2.20 | Initial CuAAC optimisation on RNA and the effects of magnesium chloride. (A) Time-course CuAAC ligation of DNA and RNA. CuAAC ligations of DNA (left) and RNA (right) were performed with and without a templating DNA strand. Reactions contained 20 µM of each oligonucleotide, 0.2 M TEAA buffer (pH 7.0), 0.5 mM ascorbic acid, 0.5 mM Cu–THPTA complex and 10% DMSO. Templated reactions also contained 200 mM NaCl. The ligation reaction proceeded for 1 to 16 h at RT and was loaded onto a 10% denaturing PAGE gel. The oligonucleotide sequences can be found in **Appendix Table 1**. (B) The effect of $MgCl_2$ on CuAAC ligation efficiency. Non-templated CuAAC ligations of RNA contained 0.2 M TEAA buffer (pH 7.0), 0.5–5 mM ascorbic acid, 0.5–5 mM Cu–THPTA complex and 10% DMSO. $MgCl_2$ was added to the final concentration of 5 mM. The ligation reaction proceeded for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. Ligation efficiency was calculated using the equation $f_{product}/(f_{tracrRNA} + f_{crRNA} + f_{product})$ x100 and ImageJ software. Note that f stands for band intensity divided by oligonucleotide molar extinction coefficient at 260 nm. The oligonucleotides used can be found in **Appendix Table 1**.

However, untemplated ligation is still attractive as it provides greater flexibility – it does not rely on the generation of a splint oligonucleotide and therefore does not require its removal after ligation. One factor that is essential in click ligation is a copper catalyst. Since RNA contains a large number of hydroxyl groups at 2' position, they could potentially cause the oligonucleotide to tightly coordinate copper ions, sequestering it away from the ligation point. Therefore, $MgCl_2$ was added to outcompete copper binding to the 2'-OH groups and phosphate backbone, increasing the availability of copper for the cycloaddition reaction. Promisingly, CuAAC of RNAs, without a DNA splint, significantly improved (19% yield with 5 mM $MgCl_2$ and 5 mM Cu–THPTA complex, **Figure 2.20B**).

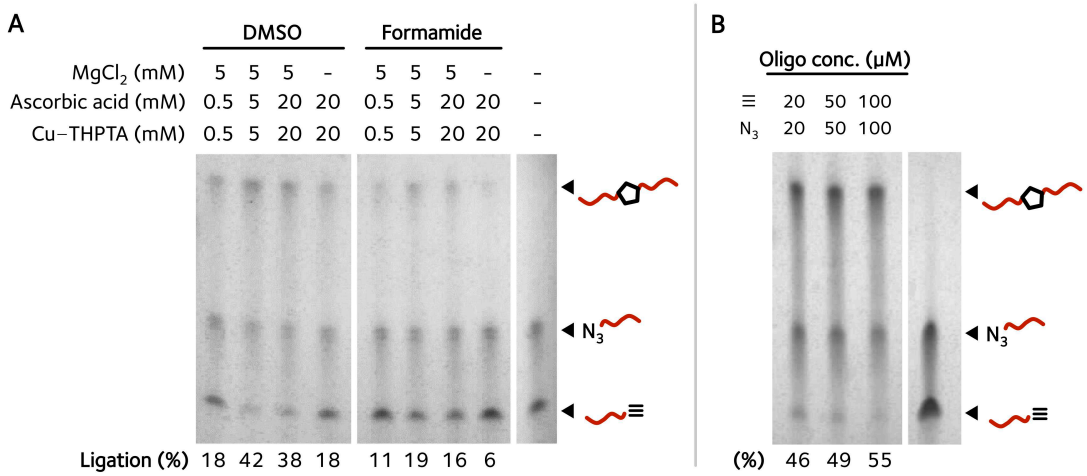


Figure 2.21 | The effect of solvents and RNA concentration on non-templated CuAAC optimisation. (A) The effect of organic solvents and copper concentration on CuAAC ligation efficiency. Non-templated CuAAC ligations were performed in 0.2 M TEAA buffer (pH 7.0) containing 0.5–20 mM ascorbic acid, 0.5–20 mM Cu–THPTA complex and 50% organic solvents: DMSO (left) and formamide (right). $MgCl_2$ was added to the final concentration of 5 mM. The ligation reaction proceeded for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. (B) The effect of RNA concentration on CuAAC ligation efficiency. Non-templated CuAAC ligations of various RNA concentrations were performed in 0.2 M TEAA buffer (pH 7.0), containing 5 mM ascorbic acid, 5 mM Cu–THPTA complex, 50% DMSO and 5 mM $MgCl_2$. The ligation reaction proceeded for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. Ligation efficiency was calculated using the equation $f_{product}/(f_{tracrRNA} + f_{crRNA} + f_{product})$ x100 and ImageJ software. Note that f stands for band intensity divided by oligonucleotide molar extinction coefficient at 260 nm. The oligonucleotides used can be found in **Appendix Table 1**.

Next, the effect of organic solvents on the CuAAC efficiency was investigated. RNAs, and in particular the tracrRNA, are expected to form secondary structures. This could result in the alkyne and azide functional groups being "hidden" within the biomolecule, making them unable to react with each other. Therefore, polar organic solvents DMSO and formamide were added to disrupt the RNA hydrogen bonding interactions, exposing the alkyne and azide groups for the reaction. Comparison of ligation in **Figure 2.20B** with **Figure 2.21A** (lanes with 5 mM $MgCl_2$ and 5 mM Cu–THPTA complex) suggests that increasing the amount of DMSO to 50% of the total reaction volume enhances CuAAC efficiency from 19% to 42%. DMSO also appears to improve CuAAC efficiency more than formamide (**Figure 2.21A**). In parallel, increasing the copper concentration by 10-fold from 0.5 to 5 mM appears to increase the amount of click product generated. Thus, both DMSO and higher copper concentrations were used for further untemplated CuAAC.

The other factor that affects non-templated CuAAC is the concentration of the oligonucleotides. The higher the oligonucleotide concentrations, the greater the chance of the 3'-alkyne modified RNA and the 5'-azide RNA interacting and reacting. It was found that 50 μM was a good compromise between sgRNA production scale and product yields (**Figure 2.21B**).

In summary, the best non-templated CuAAC conditions for RNA ligation contain 5 mM $MgCl_2$, 50% DMSO and 50 μM RNAs. The reaction proceeds efficiently within 1–2 h at room temperature and yields are sufficient for further assays (**Figure 2.21B**).

2.3.6 Click Ligation of crRNA–tracrRNA

Initially, alkyne-modified crRNA I was ligated to long linker 5'-C6-azide-containing tracrRNA IIIb using the above optimised non-templated CuAAC conditions. Templated CuAAC was also carried out as a control. Surprisingly, the non-templated reaction performed better than the templated reaction (61% vs 52% ligation efficiency for the non-templated vs templated reaction, both with 5 mM Cu-THPTA and 5 mM ascorbic acid, **Figure 2.22**). Furthermore, increasing the amount of copper(I) beyond 5 mM does not improve the CuAAC reaction (61% vs 58% for the non-templated reaction using 5 mM and 12.5 mM Cu–THPTA respectively, further reactions used 12.5 mM Cu–THPTA to ensure there is sufficient copper). Mass spectrometry analysis (**Figure 2.24B**) confirmed that the full-length sgRNA with linkage 2 was obtained in high purity.

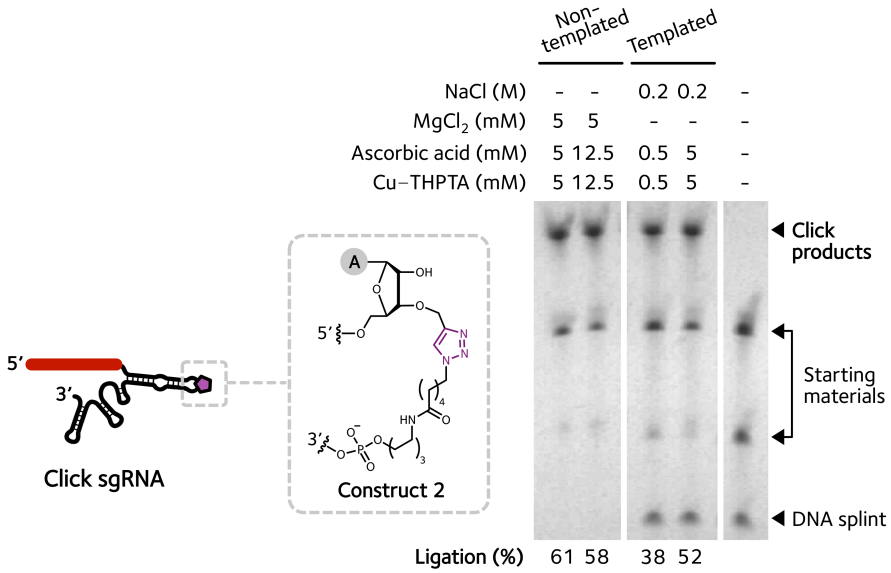


Figure 2.22 | Templated and non-templated CuAAC ligations of sgRNA with linkage 2. For the templated reaction, 37-mer crRNA I, 66-mer tracrRNA IIIb were annealed in the presence of a DNA splint (in 200 mM NaCl, at 1:1:1 ratio and 20 µM final concentration) prior to CuAAC consisting of 0.5 or 5 mM ascorbic acid, 0.5 or 5 mM Cu–THPTA complex and 10% DMSO. The non-templated CuAAC contained 50 µM of 37-mer crRNA I, 66-mer tracrRNA IIIb, 5 or 12.5 mM ascorbic acid, 5 or 12.5 mM Cu–THPTA complex, 50% DMSO and 5 mM $MgCl_2$. All reactions proceeded for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. Ligation efficiency was calculated using the equation $f_{product}/(f_{tracrRNA} + f_{crRNA} + f_{product})$ x100 and ImageJ software. Note that f stands for band intensity divided by oligonucleotide molar extinction coefficient at 260 nm. The oligonucleotide sequences can be found in **Appendix Table 3** and **4**.

Next, the focus was to optimise copper-free SPAAC ligation to generate sgRNAs. The 37-mer crRNA IIb bearing DBCO at the 3' end was ligated to a 66-mer tracrRNA IIIb carrying a long linker 5'-C6-azide in a templated and non-templated manner. Since DBCO can be readily inactivated by copper binding, and our HPLC systems have previously been used to purify CuAAC ligation reaction, EDTA was added to chelate any residual copper in the sample to create a copper-free environment for SPAAC. However, all the SPAAC conditions tested gave comparable amounts of sgRNAs (~15%, **Figure 2.23A**) with linkage 4, which was significantly less than the CuAAC reaction (~50%, **Figure 2.22**).

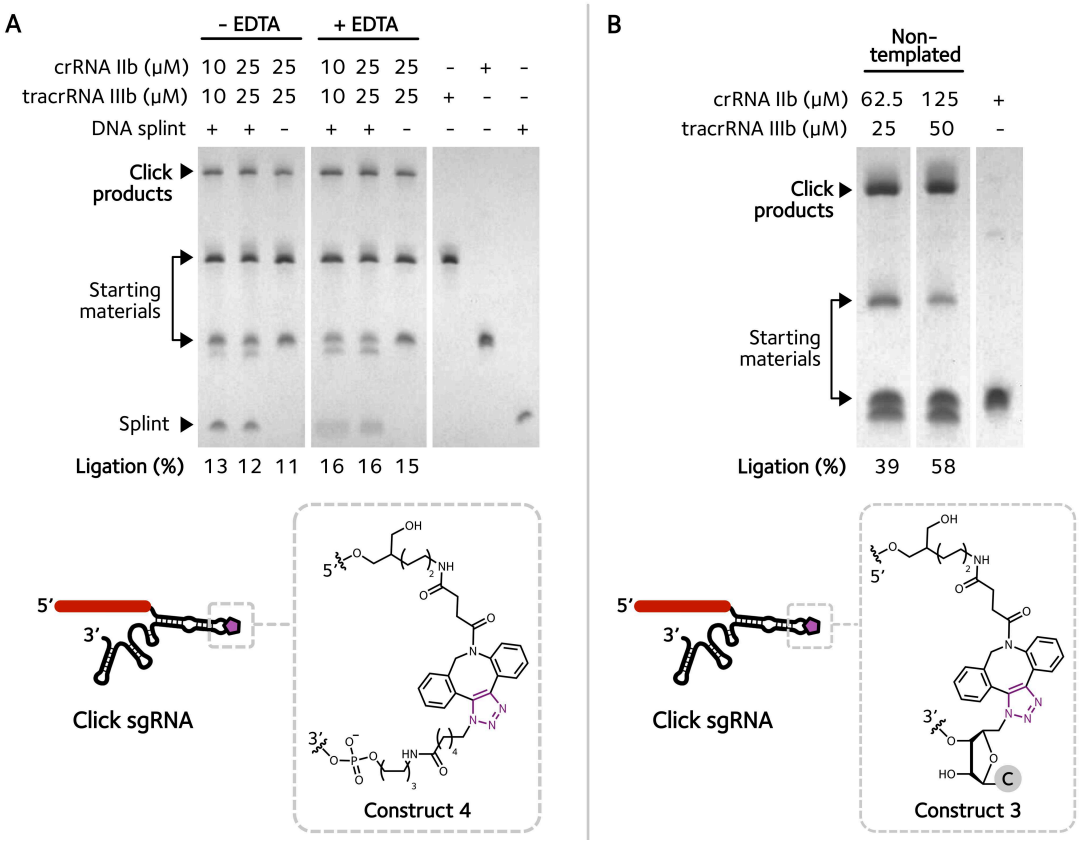


Figure 2.23 | SPAAC optimisation. (A) Templated and non-templated SPAAC ligations of sgRNA. For a templated reaction, 37-mer crRNA IIb, 66-mer tracrRNA IIIb at various concentrations were annealed in the absence and presence of a DNA splint (at 1:1:1 ratio), 200 mM NaCl, 20 mM HEPES and with and without 5 mM EDTA (pH 8.0). All reactions proceeded for 2 h at RT before loading onto a 10% denaturing PAGE gel. (B) The effect of RNA concentration on non-templated SPAAC ligation efficiency. Reactions contained 25 and 50 μM of RNAs, 200 mM NaCl, 20 mM HEPES and 5 mM EDTA (pH 8.0). The ligation reaction proceeded for 2 h at RT before loading onto a 10% denaturing PAGE gel. Note that heating the sample (for annealing) caused the formation of an unknown +33 adduct for DBCO oligonucleotides (confirmed by mass spectrometry), therefore the DBCO starting material appears as two bands on the gel. Ligation efficiency was calculated using the equation $f_{product}/(f_{tracrRNA} + f_{crRNA} + f_{product})$ x100 for (A) and $f_{product}/(2xf_{tracrRNA} + f_{product})$ x100 for (B) and ImageJ software. The different equation for (B) assumes an equal amount of crRNA is consumed as tracrRNA in ligation and is necessary as the shorter crRNA is used in excess. Note that f stands for band intensity divided by oligonucleotide molar extinction coefficient at 260 nm. The oligonucleotide sequences can be found in **Appendix Table 3**.

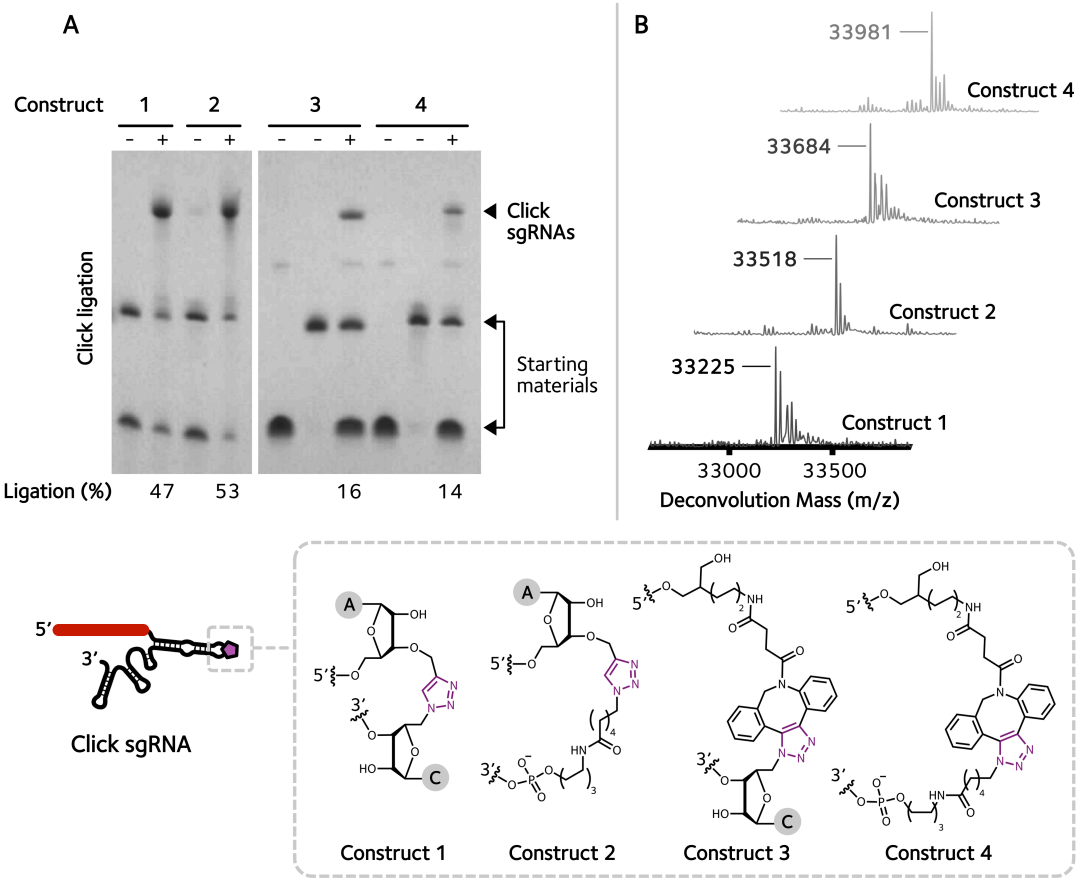


Figure 2.24 | Ligation efficiency and mass spectrometry of all clicked 37–66 sgRNAs. (A) CuAAC and SPAAC ligations gave click sgRNAs, which contain the following linkages (C): (1) biocompatible triazole (Tz2), (2) long triazole, (3) long DBCO-triazole, and (4) very long DBCO-triazole. (B) Mass spectrometry of all clicked sgRNAs. The deconvolution mass showed the high purity of all clicked sgRNAs with correct observed mass. M (construct 1) = 33224 (expected); 33225 (found). M (construct 2) = 33518 (expected); 33518 (found). M (construct 3) = 33681 (expected); 33684 (found). M (construct 4) = 33977 (expected); 33980 (found). Sodium adducts were also detected. Reaction conditions are described in **Experimental Chapter 7**, **Section 7.7**. Ligation efficiency was calculated using the equation $f_{product}/(2xf_{tracrRNA} + f_{product})$ x100 and ImageJ software. This equation assumes an equal amount of crRNA is consumed as tracrRNA in ligation and is necessary as the shorter crRNA is used in excess. Note that f stands for band intensity divided by oligonucleotide molar extinction coefficient at 260 nm. The oligonucleotide sequences can be found in **Appendix Table 3** and **4**.

The low efficiency of the SPAAC reaction is probably due to the 37-mer crRNA IIb not being pure after HPLC purification as discussed previously, or the formation of an unknown +33 adduct (confirmed by mass spectrometry) upon heating the sample during annealing that causes the crRNA IIb to migrate as two bands on denaturing PAGE (**Figure 2.23**). Therefore, 37-mer crRNA IIb (with the contaminant) was used in greater excess (2.5 equivalent) in order to push the reaction to completion. However, this requires the reaction to be performed in a non-templated manner

as a DNA splint would bring the inactive DBCO-containing strand in close proximity to 66-mer tracrRNA IIIb, preventing the active DBCO strand from being able to react. SPAAC ligation was substantially improved when the concentration of the 37-mer crRNA IIb was increased to 125 μM (58%, **Figure 2.23B**).

Using optimised CuAAC and SPAAC ligation conditions, clicked 37–66 sgRNAs with linkages 1–4 were generated (**Figure 2.24A**), purified and characterised by mass spectrometry (**Figure 2.24B**), ready for functional testing. Among these four linkages, construct 1 is the easiest to make. CuAAC gave the most consistent yields compared to SPAAC. Construct 1 also does not require NHS ester labelling, a step which requires a second round of HPLC purification. The additional handling of the sample in solution could cause greater RNA degradation (see RNA chain cleavage and backbone migration in **Figure 2.3**). Even though construct 1 requires a 5'-OH to azide conversion step, this occurs on the solid support while the protecting groups are still present on RNA bases, sugar and phosphate, minimising RNA degradation.

2.3.7 *In vitro* DNA Cleavage

To evaluate the activity of the synthetic sgRNA generated from CuAAC and SPAAC ligation, *in vitro* DNA cleavage was performed using Cas9. The unmodified sgRNA produced by enzymatic *in vitro* transcription was used as a control. The reaction contained sgRNA, Cas9 and a DNA template (plasmid) with a ratio of 50:10:1. As Cas9 is a single-turnover protein[67], sgRNA was used in greater excess to ensure that every Cas9 protein molecule forms a complex with the sgRNA and cleaves a target DNA. It was found that all the clicked sgRNAs could mediate DNA cleavage as efficiently as IVT sgRNA and the artificial linkages introduced did not inhibit the sgRNA activity over longer periods of time (1 h and 16 h, **Figure 2.25**). These results are consistent with reports that the tetraloop of the sgRNAs can be extensively modified without loss of DNA cleavage activity[188,189]. It should be noted that a construct similar to one example of our sgRNA constructs (construct 2) was published during the course of the work[188]. However, this study did not show CuAAC ligation efficiency or *in vitro* activity of the sgRNA construct.

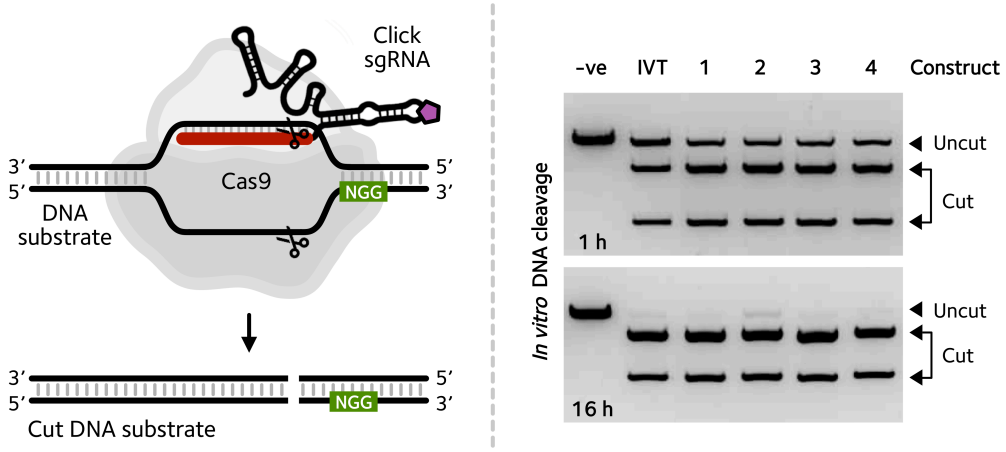


Figure 2.25 | *In vitro* DNA cleavage by clicked 37–66 sgRNAs. The DNA cleavage reaction contained 30 nM Cas9, 150 nM sgRNA and 3 nM linearised plasmid DNA, and proceeded for 1 h or 16 h at 37 °C. Samples were loaded onto a 1% SYBR gold-stained agarose gel. Constructs are shown in **Figure 2.24** and the oligonucleotide sequences can be found in **Appendix Table 4**.

The structure of sgRNA suggests that the region beyond the DNA-targeting 20-mer of the 37-mer crRNA can hybridise to the 66-mer tracrRNA and allow Cas9 to cleave the target DNA. To confirm that ligation of these two components together is necessary, *in vitro* DNA cleavage mediated by clicked sgRNA (linkage 1) and unclicked crRNA I hybridised to tracrRNA IV was performed in a concentration-dependent manner. The gRNA:Cas9 ratio was varied from 1:1 to 10:1. It was found that DNA cleavage only remained efficient at higher concentrations of unclicked gRNA while the DNA cleavage by the clicked sgRNA was concentration-independent (**Figure 2.26**). Despite being feasible *in vitro*, delivery of excessive amount of unclicked gRNA into cells would not be favourable, potentially giving rise to cell death. Therefore, click ligation is essential as it brings the reaction down to the bi-molecular (sgRNA and Cas9) rather than tri-molecular (crRNA, tracrRNA and Cas9) level, potentially reducing the amount of sgRNA required for gene editing.

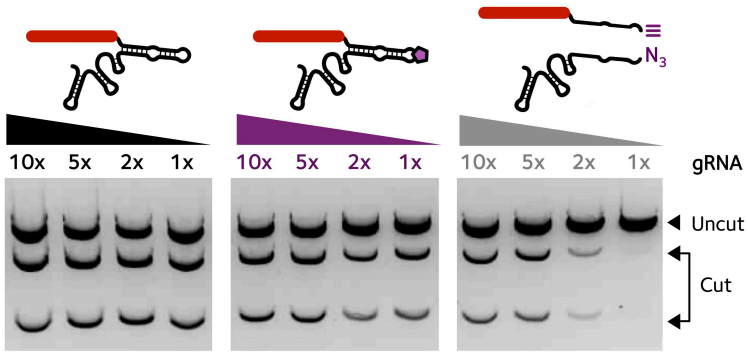


Figure 2.26 | *In vitro* DNA cleavage mediated by clicked and unclicked sgRNA construct 1 (Tz2 linkage). The reaction with the lowest amount of sgRNA contained 30 nM sgRNA (1x), 30 nM Cas9 and 3 nM target DNA. The concentration of sgRNA was increased to 60 nM (2x), 150 nM (5x) and 300 nM (10x) while the concentrations of Cas9 and target DNA were kept constant. The oligonucleotide sequences can be found in **Appendix Table 3** and **4**. Note the clicked linkage is Tz2 (Construct 1 in **Appendix Table 4**).

2.4 Click ~21–79 sgRNA Constructs

2.4.1 Design Strategy

From these encouraging results, the next goal was to move the click linkage in order to improve the split-and-click sgRNA system. The shorter the RNA synthesis the purer the final RNA, therefore moving the split point towards the variable DNA-targeting region to obtain a ~21-mer DNA-targeting RNA and a fixed 79-mer Cas9 binding RNA has potential benefits because the ~21-mer DNA-targeting RNA will be much purer, which is essential for optimum target specificity. In addition, synthesising ~21-mer RNA is much more cost effective than 37-mer RNA when a new gene-targeting sgRNA is required. The fixed 79-mer RNA can be synthesised in a large scale with a small portion used for ligation to different ~21-mer RNAs, providing a simple cost-effective way of constructing a library of sgRNAs. In this design, the ligation site was chosen to allow azide conversion to be performed on a terminal U base, which is more efficient than on other bases (especially relative to purines)[200]. Based on the crystal structure of sgRNA in complex with Cas9 (PDB: 4OO8[48]), placing the biocompatible triazole between bases G1 and U2 (**Figure 2.27**) should minimise the interference with the Cas9 binding.

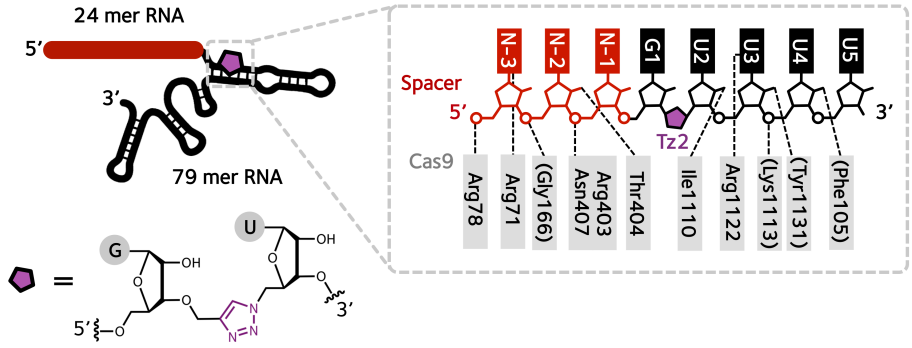


Figure 2.27 | New ~21–79 sgRNA construct design. sgRNA, with an identical sequence identical to the click 37–66 construct, was split into a variable 24-mer DNA-targeting RNA and a constant Cas9-interacting tracrRNA. Upon ligation, a biocompatible triazole (Tz2, purple) is generated at a position where the interactions with Cas9 are minimal. Note that amino acids of Cas9 without the bracket interact with sgRNA through their side chains, while those within the bracket make contacts *via* their main chains.

2.4.2 5'-Azide Conversion of TC-RNA

However, shortening the crRNA means the length of tracrRNA has to increase. Thus, ideally the 79-mer tracrRNA should be produced *via* TC chemistry due to its higher purity and yields compared to TBDMS chemistry. As shown previously, the iodo conversion step failed for RNA with a 2'-TC protecting group. Therefore, to by-pass this issue, the tracrRNA was synthesised with a terminal 5'-iodo-dT. Surprisingly, after NaN_3 treatment, the 79-mer RNA could not be completely deprotected using EDA/toluene (1:1) for 6 h at room temperature, a standard deprotection condition for TC-RNA. A mixture of the desired RNA plus various protecting groups (present as shoulder of peaks higher than the fully deprotected RNA) was observed by HPLC-MS (**Figure 2.28A**). It appears that the azide removed the TC group but not the nucleobase protecting groups, reducing the solubility of the partially deprotected RNA in subsequent deprotection solution (EDA/toluene), resulting in further inefficient cleavage of nucleobase protecting groups. To test this hypothesis, short TC-RNA (13 mer) was treated with NaN_3 and subsequently deprotected by EDA. Even though the majority of the RNA was fully deprotected, a +71 contaminant was seen indicating unsuccessful removal of isobutyryl (ib) on the guanine base (**Figure 2.28B**). The incomplete deprotection issue was more pronounced when DNA containing identical nucleobase protecting groups as TC-RNA, but no TC group, was deprotected by EDA (**Figure 2.28C**). A series of optimisation was carried out to solve this

problem. This included 'one-pot' azide conversion and deprotection by mixing NaN_3 with neat EDA. However, the RNA was severely degraded. Another attempt involved performing azide conversion in mixed solvent system (water:DMF, 1:9 v/v), which improves sodium azide solubility and allows higher azide concentrations to be used. Surprisingly, azide conversion was successful and fully deprotected 5'-azido tracrRNA was observed. However, a significant amount of RNA degradation as well as incomplete deprotection (as shown by higher retention time peaks on HPLC) prevented this method from being viable (**Figure 2.28D**).

As a consequence, the 79-mer tracrRNA was synthesised using more stable TBDMS protecting groups and then converted to azide at the 5' terminus. Although a significant amount of truncated products were observed (as expected from the TBDMS strategy), the full-length RNA could be efficiently isolated from those impurities using RP-HPLC with a very shallow gradient (**Figure 2.29B**). Mass spectrometry indicated that the tracrRNA was sufficiently pure for subsequent click ligation (**Figure 2.29B**). DNA nucleotides were also incorporated at various positions along tracrRNA that do not inhibit Cas9 activity as reported by Rueda *et al*[122]. The aim was to reduce the cost and increase the efficiency of synthesis, increase the stability of the sgRNA, simplify purification and thereby improve the purity of the RNA. As expected, tracrRNA-DNA was produced in higher yield than the all-RNA tracrRNA (for 1 µmol synthesis scale – tracrRNA: 65.2 nmol (crude) and 5.2 nmol (RP-HPLC); tracrRNA-DNA: 164.6 nmol (crude) and 35.6 nmol (RP-HPLC)).

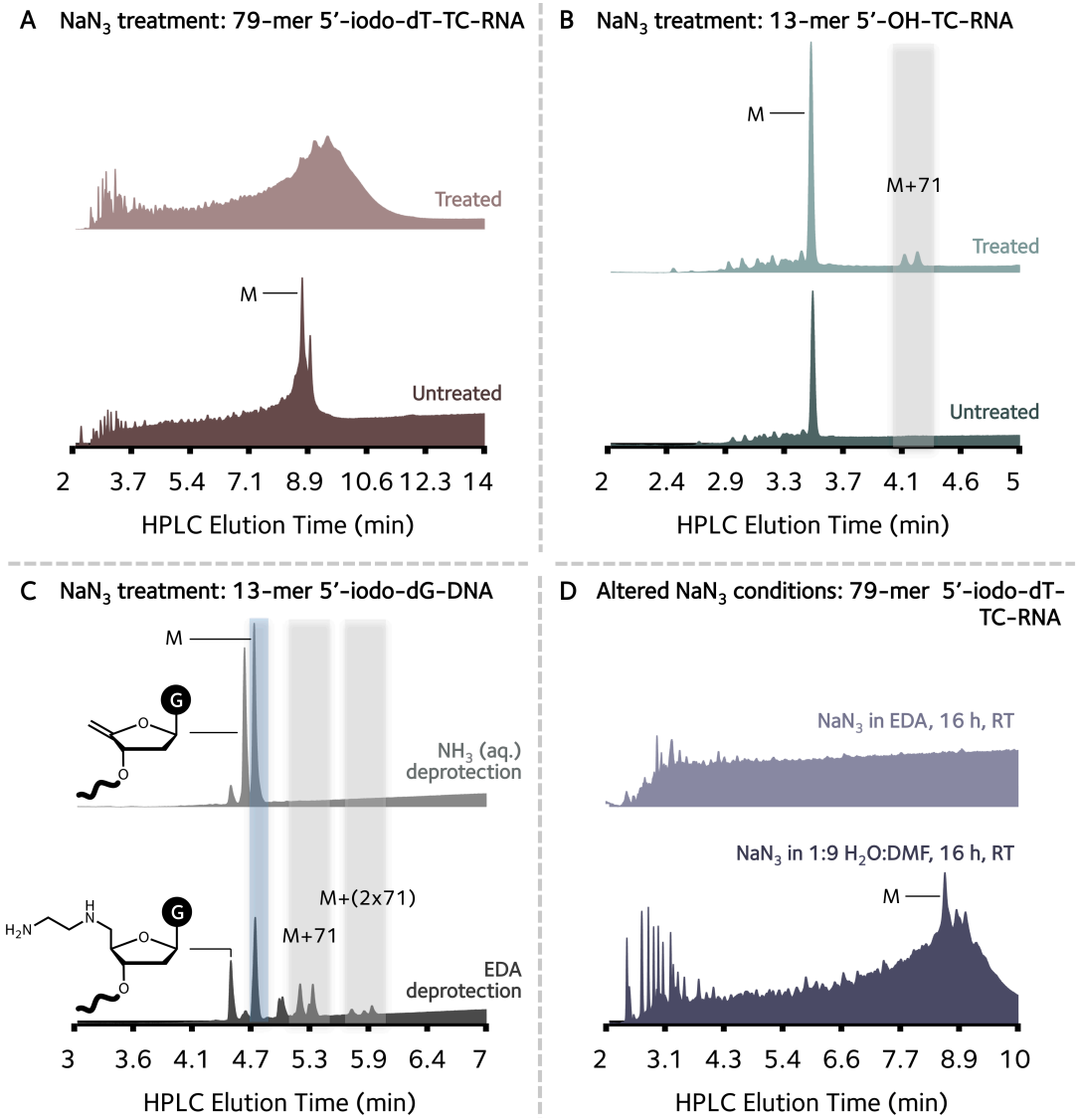


Figure 2.28 | Azide conversion tests on TC-RNAs and DNA. (A) EDA deprotection of sodium azide treated 5'-iodo-dT 79-mer TC-RNA. M = 25520 (expected); 25517 (found). (B) Azide treatment of an unmodified 13-mer TC-RNA followed by EDA deprotection. M = 3816 (expected); 3814 (found) and +71 is an isobutyryl group on G. (C) Azide treatment of 13-mer DNA deprotected by $NH_{3\ (aq.)}$ or EDA. M = 3055 (expected); 3055 (found) and +71 is an isobutyryl group on G. Conversion of 5'-iodo to azide did not go to completion, resulting in elimination of the iodo group and the formation of an alkene during $NH_{3\ (aq.)}$ deprotection (top panel). During EDA deprotection, the 5'-iodo intermediate was substituted by EDA (by-product in the bottom panel). (D) NaN_3 conversion conditions were altered in an attempt to overcome incomplete deprotection. One-pot azide conversion and deprotection (mixture of NaN_3 and EDA) severely degraded the TC-RNA (top panel). Addition of water to increase the solubility and hence the concentration of NaN_3 improved conversion to azide and RNA deprotection, yet significant RNA degradation and incomplete deprotection was also observed (bottom panel). M = 25435 (expected); 25438 (found). Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. The oligonucleotide sequences can be found in **Appendix Table 2**.

Furthermore, the 24-mer DNA-targeting RNA was synthesised with a terminal 5'-C6-amine that can be post-synthetically labelled with different molecules such as fluorophores or ssDNA donor[140] to expand the sgRNA functionality. Here, the 24-mer RNA was labelled with Cy3-NHS ester. Overall, this ~21–79 mer split sgRNA is feasible synthetically.

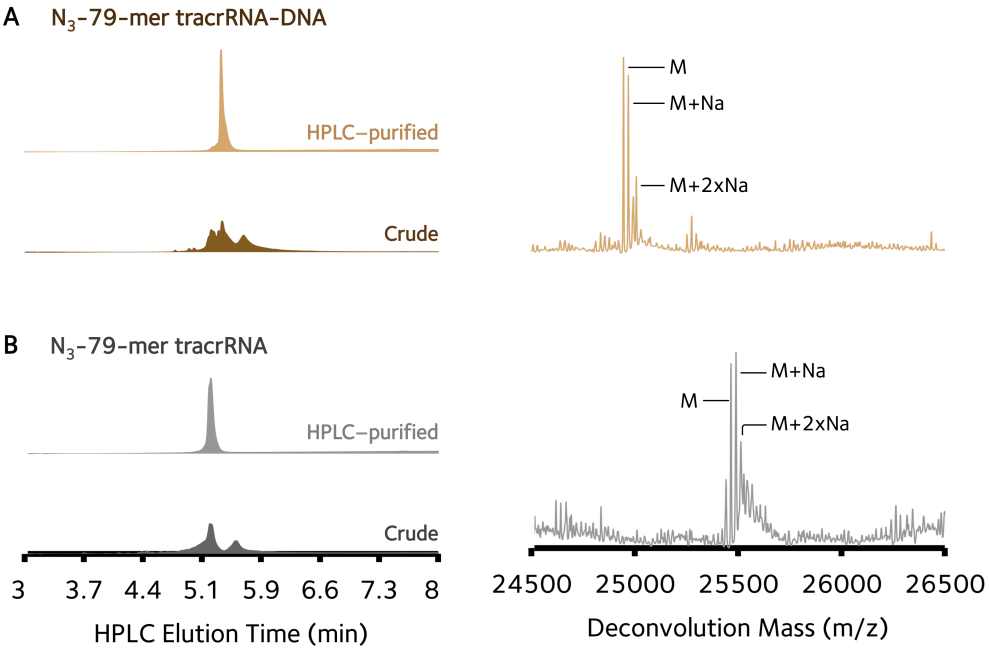


Figure 2.29 | HPLC purification and mass spectrometry of (A) 5'-azido 79-mer tracrRNA-DNA and (B) 5'-azido 79-mer tracrRNA. Note that the area under the curve of all HPLC traces is normalised to 1, allowing direct comparison of peak shape and height. The expected masses of fully deprotected species plus sodium (+23) adducts were observed. M (N_3-tracrRNA-DNA) = 25436 (expected); 25437 (found). M (N_3-tracrRNA) = 25018 (expected); 25021 (found). The oligonucleotide sequences can be found in **Appendix Table 5**.

2.4.3 CuAAC Ligation of ~21–79 sgRNAs

As the 24-mer crRNA is variable, it is not practical to synthesise a new DNA splint every time a click ligation needs to be performed. Thus, non-templated CuAAC is preferred for this design strategy. With the previously optimised non-templated CuAAC condition, the 24-mer 3'-alkyne crRNA with and without C6-amine or Cy3 at 5' terminus was efficiently ligated to the 79-mer 5'-azide tracrRNA and tracrRNA-DNA (**Figure 2.30**), yielding full-length sgRNAs (confirmed by mass spectrometry, **Figure 2.30B** and **2.30C**).

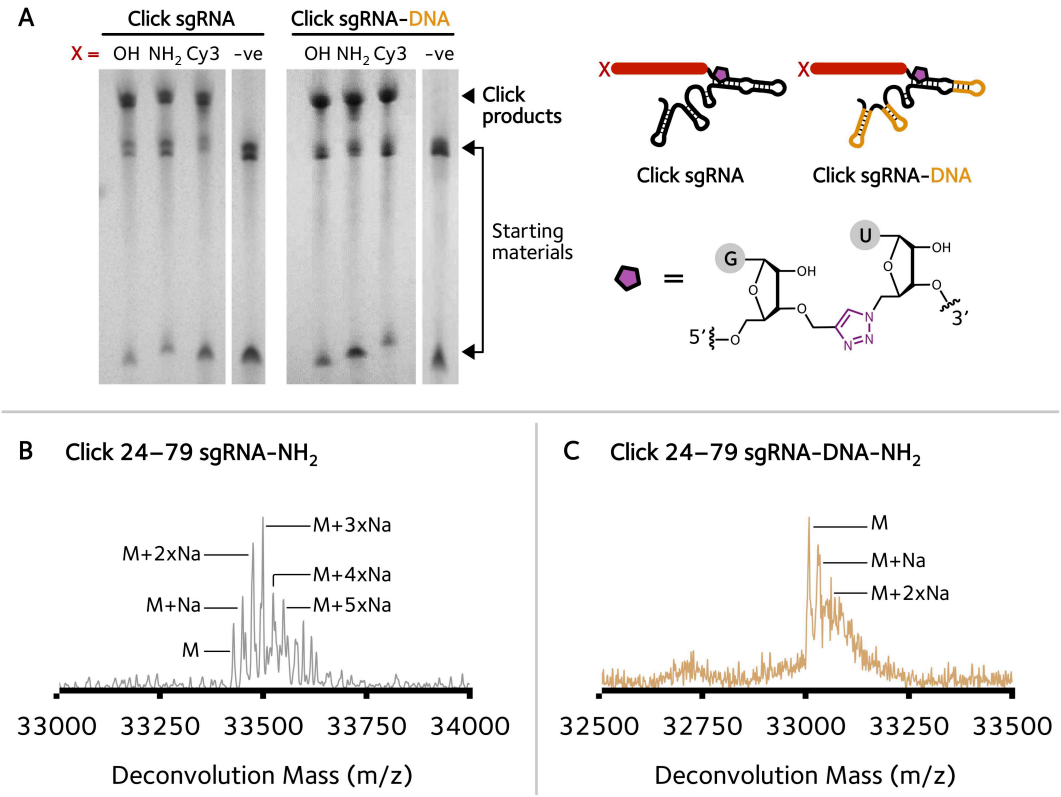


Figure 2.30 | Non-templated CuAAC ligation to generate sgRNA and sgRNA-DNA for *in vitro* DNA cleavage. (A) The CuAAC reactions contained 50 µM of 24-mer crRNA with 3'-alkyne and 5'-OH / C6-NH_2 / Cy3, 79-mer 5'-N_3-tracrRNA or tracrRNA-DNA, 0.2 M TEAA buffer (pH 7.0), 12.5 mM ascorbic acid, 12.5 mM Cu–THPTA complex, 50% DMSO and 5 mM $MgCl_2$. All reactions proceeded for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. The oligonucleotide sequences can be found in **Appendix Table 5** and **6**. Note that "NH_2" refers to a 5'-C6-NH_2 group and "OH" is the sgRNA with an unmodified 5' end. (B) and (C) are examples of mass deconvolution of clicked 24–79 sgRNA and sgRNA-DNA with 5'-C6-NH_2 respectively. M (sgRNA-NH_2) = 33425 (expected); 33426 (found). M (sgRNA-DNA-NH_2) = 33007 (expected); 33009 (found). Note that Na^+ adducts were also detected.

2.4.4 *In vitro* DNA Cleavage

All clicked sgRNAs were used in the *in vitro* DNA cleavage assay and compared to the control IVT sgRNA. Unclicked 24–79 gRNA and gRNA-DNA were also tested as a negative control. The clicked 24–79 sgRNAs and sgRNA-DNAs with a 5'-OH group mediated DNA cleavage slightly less efficiently (~30–50%) than IVT sgRNA after 1 h, however the cleavage reaction went to completion after 16 h (**Figure 2.31**). This indicates that the presence of DNA nucleotides did not impair Cas9 activity. Absolutely no DNA cleavage was observed even after 16 h when using unclicked gRNA and gRNA-DNA showing click ligation is crucial for this ~21–79 splitting design.

Surprisingly, the clicked sgRNAs with a 5'-C6-amino group performed slightly better (45–52% after 1 h, **Figure 2.31**) than those with a 5'-OH group (36–46% after 1 h, **Figure 2.31**) for unknown reasons that need further investigation. However, when a Cy3 dye was attached to this amino group, Cas9 activity was reduced significantly even after 16 h incubation (46–52%, **Figure 2.31**). This contradicts the results from Lee *et al*[140], who showed that Cas9 can tolerate large modifications at the 5' end of crRNA with only a slight drop in activity (~10%) when a rhodamine dye is appended at this position. This could probably be due to the different dye properties – Cy3 is more hydrophobic than rhodamine[201].

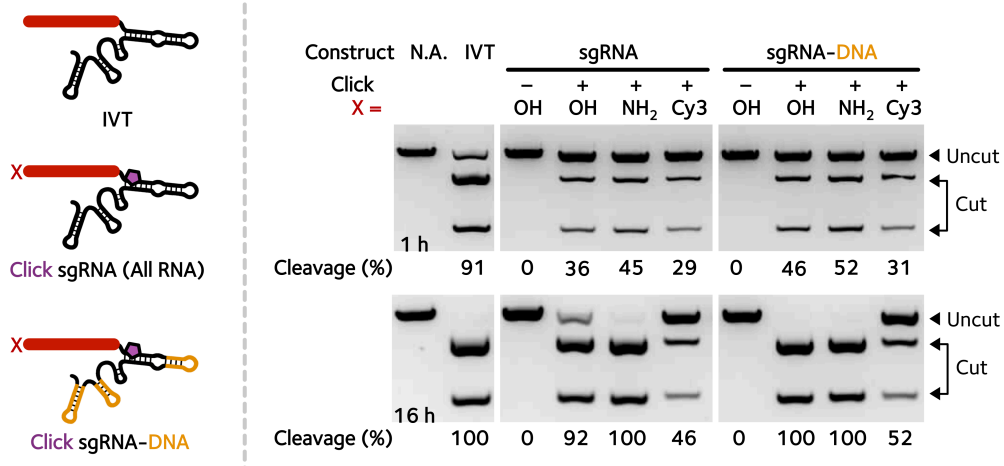


Figure 2.31 | *In vitro* DNA cleavage of clicked 24–79 sgRNAs and sgRNA-DNAs. The DNA cleavage reaction contained 30 nM Cas9, 150 nM sgRNA and 3 nM linearised plasmid DNA. The reaction was left for 1 h or 16 h at 37 °C and loaded onto a 1% SYBR gold-stained agarose gel. The cleavage efficiency was quantified using an Agilent Bioanalyzer and the equation f_{cut}/f_{total} x100, where f stands for fraction. The oligonucleotide sequences can be found in **Appendix Table 6**. Note that "NH_2" refers to a 5'-C6-NH_2 group and "OH" is the sgRNA with an unmodified 5' end.

Following CuAAC ligation, denaturing PAGE was performed to purify the clicked product from the starting materials as well as copper ions, which could inhibit downstream applications. Nevertheless, PAGE purification might be avoidable provided that residual starting materials do not inhibit Cas9 activity. To mimic the crude CuAAC reaction, the PAGE-purified sgRNA was independently titrated with a 24-mer crRNA and a 79-mer tracrRNA-DNA. As expected, the 24-mer crRNA did not affect Cas9 cleavage activity. In contrast, the tracrRNA-DNA acted as a competitive inhibitor for Cas9 as its concentration increased (**Figure 2.32A**). Therefore, it is important to keep a tracrRNA as a limiting component in CuAAC reaction. Next, the actual crude

clicked sgRNA-DNA (with copper ions removed simply by using a gel-filtration column and no RP-HPLC purification) was tested and it appeared to mediate DNA cleavage with reduced efficiency by ~50% compared to PAGE-purified clicked sgRNA-DNA (**Figure 2.32B**). The ability to use the crude sample greatly simplifies the preparation of sgRNA, allowing researchers to stock 24-mer crRNAs and 79-mer tracrRNA and perform CuAAC ligation when needed.

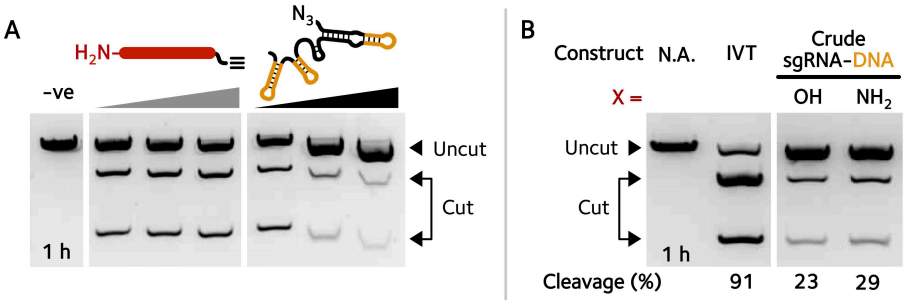


Figure 2.32 | The inhibitory effects of click sgRNA-DNA starting materials on *in vitro* DNA cleavage and the potential use of desalted clicked 24–79 sgRNA-DNA for *in vitro* DNA cleavage. (A) The DNA cleavage reaction consisting of 30 nM Cas9, 150 nM sgRNA and 3 nM linearised plasmid DNA was titrated with 0, 2 and 4 equivalents of 24-mer crRNA and 79-mer sgRNA-DNA (relative to the click sgRNA-DNA). The reactions proceeded for 1 h at 37 °C and were loaded onto a 1% SYBR gold-stained agarose gel. (B) Crude clicked 24–79 sgRNA-DNAs with and without a 5'-C6-NH_2 group were used for *in vitro* DNA cleavage. Note that "NH_2" refers to a 5'-C6-NH_2 group and "OH" is the sgRNA with an unmodified 5' end. The cleavage reactions contained 30 nM Cas9, 150 nM sgRNA and 3 nM linearised plasmid DNA, proceeded for 1 h at 37 °C and was loaded onto a 1% SYBR gold-stained agarose gel. The cleavage efficiency was quantified using an Agilent Bioanalyzer and the equation f_{cut}/f_{total} x100, where f stands for fraction. The oligonucleotides used can be found in **Appendix Table 5** and **6**.

2.4.5 The Importance of a Biocompatible Triazole Linkage

To determine whether other linkages could also be used to generate the clicked sgRNA, the 24-mer crRNA was synthesised with a long serinol-based 3'-alkyne linker (**Appendix Table 5**) and then clicked to the 79-mer tracrRNA or tracrRNA-DNA with 5'-azide using the optimised non-templated CuAAC condition. The resulting clicked sgRNAs with a very long linker reduced Cas9 activity in comparison with the short triazole-containing sgRNAs (**Figure 2.33**), indicating that short biocompatible triazole linkage is important in the design.

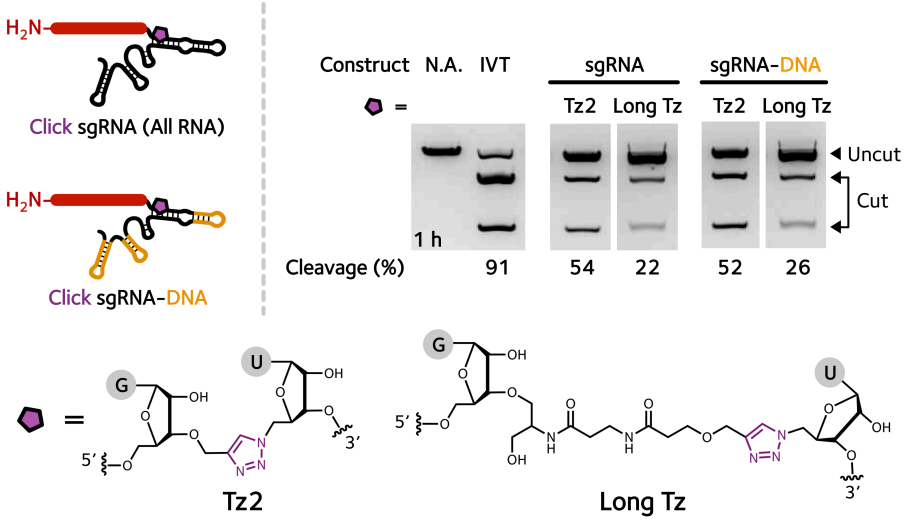


Figure 2.33 | *In vitro* DNA cleavage by clicked 24–79 sgRNA and sgRNA-DNA containing a long triazole linkage. The cleavage reactions contained 30 nM Cas9, 150 nM sgRNA and 3 nM linearised plasmid DNA, proceeded for 1 h at 37 °C and were loaded onto a 1% SYBR gold-stained agarose gel. The cleavage efficiency was quantified using an Agilent Bioanalyzer and the equation f_{cut}/f_{total} x100, where f stands for fraction. The oligonucleotide sequences can be found in **Appendix Table 6**.

2.4.6 In cell Gene Editing Activity

Next, the activity of clicked ~21–79 sgRNAs was examined in cells, which is the ultimate aim of most CRISPR-Cas applications. A 21-mer crRNA targeting the *EMX1* gene was synthesised with a terminal 5'-C6-NH_2 that has shown to improve DNA cleavage *in vitro*. This crRNA was ligated to either a 79-mer 5'-azido tracrRNA or tracrRNA-DNA and purified by denaturing PAGE (**Figure 2.34A**). U2OS cells already transfected with a lipid-plasmid complex encoding Cas9 were subsequently transfected with the clicked or control IVT sgRNAs using lipid reagents. After several days in which DSBs are generated and subsequently repaired, NHEJ-mediated indel mutations within the *EMX1* gene were evaluated. Genomic DNA was extracted and the region of interest amplified by PCR. Amplicons of mixed length are generated which after re-hybridisation yields imperfectly hybridised duplexes. These hetero-duplexes are then recognised and cleaved by T7E1 endonuclease, a mismatch endonuclease. The more DNA cleavage, the greater the editing efficiency represented as % indels (where indels = 100 x $(1 - (1 - (b + c)/(a + b + c))^{0.5})$, where a is the integrated intensity of the uncleaved PCR amplicon, and b and c are the integrated intensities of each digestion product)[93].

The clicked sgRNA could mediate gene editing with slightly less efficiency (20%), than IVT sgRNA (37%) (**Figure 2.34B**). Unexpectedly, the clicked sgRNA-DNA did not cause detectable indels, contradicting the results from *in vitro* DNA cleavage which was efficient. This finding is in agreement with Rueda *et al*[122], who demonstrated that the incorporation of DNA nucleotides into the guide RNA does not affect the Cas9 activity *in vitro* but dramatically reduce cellular activity of Cas9. Biophysical studies showed that a Cas9–DNA-containing gRNA complex did not tightly bind to target DNA or adopt the active conformation for DNA cleavage[122]. The low binding affinity *in vitro* could possibly be compensated for the high concentrations of the gRNA and target DNA, but not in a complex cellular environment where the concentrations of those components cannot be easily controlled. It is possible this effect is worse for clicked sgRNA-DNA due to a triazole linkage also being introduced. It would also be interesting to study the binding kinetics of clicked sgRNA (all RNA) to Cas9 to improve the future design.

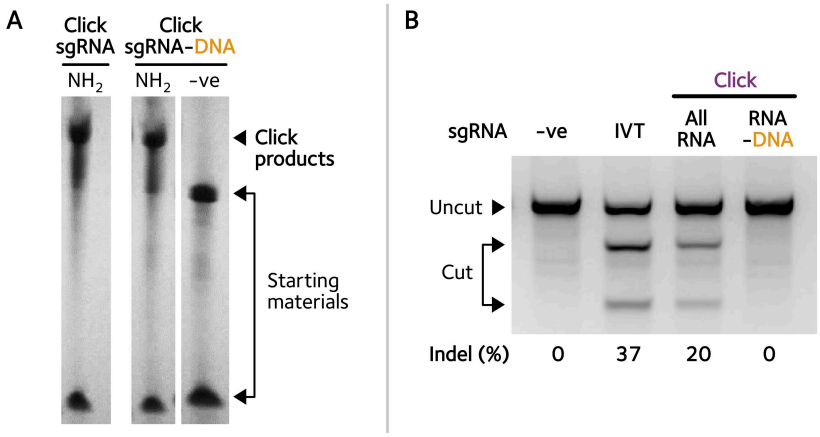


Figure 2.34 | Non-templated CuAAC ligation of *EMX1*-targeting sgRNA and sgRNA-DNA and their in cell DNA cleavage activities. (A) The CuAAC reaction contained 50 µM of 21-mer crRNA with 3'-alkyne and 5'-C6-NH_2, 79-mer 5'-N_3-tracrRNA or tracrRNA-DNA, 0.2 M TEAA buffer (pH 7.0), 12.5 mM ascorbic acid, 12.5 mM Cu–THPTA complex, 50% DMSO and 5 mM $MgCl_2$. All reactions proceeded for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. The oligonucleotide sequences can be found in **Appendix Table 5** and **7**. Note that "NH_2" refers to a 5'-C6-NH_2 group on the final construct. (B) The clicked sgRNAs with and without deoxyribonucleotides were transfected into U2OS cells and their ability to mediate indel formation was determined by T7E1 assay. Cleavage was quantified using an Agilent Bioanalyzer (biological repeats, n = 3). Indel = $(1 - (1-f_{cut}/f_{total})^{0.5}) \times 100$.

2.4.7 Off-target Effects

Off-target activity is one of the major concerns in the CRISPR-Cas field for therapeutic applications as it may lead to the adverse effect – editing the genome at unwanted locations can potentially cause the inactivation of the tumour suppressor genes or the activation of the oncogenes[80,152,202]. Thus, the target specificity of the clicked ~21–79 sgRNA with a 5'-C6-NH_2 group was tested using CIRCLE-seq[203], a read-efficient Next-Generation Sequencing approach that uses U2OS genomic DNA *in vitro* to estimate biologically relevant off-target effects. This work was performed by Dr. Arun Shivalingam in the Brown group. The detected targets were aligned and the observed frequency of each base at each position of the DNA-targeting and PAM region was determined. To simplify the comparison of IVT and clicked sgRNAs, the frequencies for IVT sgRNA were subtracted from the clicked sgRNA. The frequencies were then plotted as a heat map (**Figure 2.35A**) where: 1) colourless squares indicate no significant difference in specificity, 2) red squares show a preference for this base using IVT sgRNA, 3) blue squares show a preference for this base using clicked sgRNA. Black boxes indicate the intended on-target base. The intensity of the colours, in general, are weak and close to colourless indicating a minimal difference in specificity. For on-target sites (black boxes), a comparable number of positions are slightly favoured (light blue) or dis-favoured (light red) for clicked sgRNA compared to IVT sgRNA.

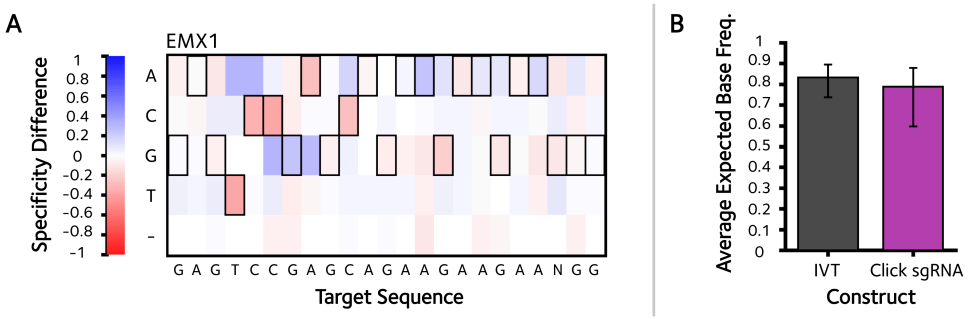


Figure 2.35 | Off-target profiles of genome editing by clicked 24–79 sgRNA. (A) shows the difference in specificity of IVT and clicked sgRNA using CIRCLE-seq. Statistically significant cleavage sites were changed into a heat-map of the observed base frequency at each position of the target sequence. Values for IVT sgRNA were subtracted from clicked sgRNA. The red intensity shows lower specificity and blue higher specificity for that specific base of clicked sgRNA. Black boxes show the expected on-target base and '–' indicates deletions. (B) is a bar graph of the median expected base frequency across the DNA-targeting and PAM region. Error bars are the Q1 and Q3 quantiles. Note that the clicked sgRNA used have a 5'-C6-NH_2 modification and the oligonucleotide sequences can be found in **Appendix Table 7**.

Taking the median on-target base frequency across the DNA-targeting and PAM region suggests there is minimal difference in base specificity (**Figure 2.35B**).

2.4.8 Split-and-click sgRNA for Library Construction

Lastly, another beneficial feature of this click ~21–79 sgRNA production system is that it potentially allows simple unbiased library construction (the constructs are formed in denaturing conditions). To evaluate this application, a 24-mer crRNA with 5'-C6-amine was independently labelled with Cy3- and ATTO 647N-NHS esters. Since these two dyes are spectrally far apart from each other, the emission of one dye does not overlap with the other making interpretation of the signal straightforward. A ratio of 0.75:0.75:1 of ATTO 647N-crRNA:Cy3-crRNA:79-mer tracrRNA-DNA was clicked *via* non-templated CuAAC and the ligation efficiency was analysed on denaturing PAGE. It was found that Cy3-sgRNA and ATTO-sgRNA produced were similar (42% of 0.75 equivalents Cy3-crRNA and 65% of 0.75 equivalents of ATTO-crRNA, note 79-mer tracrRNA-DNA is 1 equivalent in the reaction, **Figure 2.36**), confirming the unbiased performance of the reaction. This could be particularly useful for live-cell imaging or gene regulation, where a sgRNA library with roughly equal amounts of many different sgRNAs is needed.

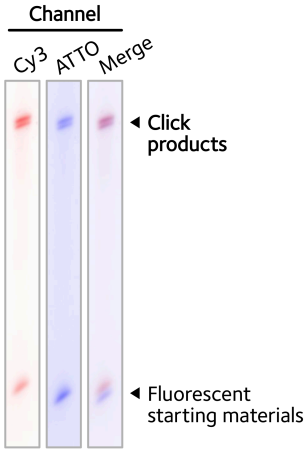


Figure 2.36 | Split-and-click ligation bias. Cy3 and ATTO 647N-labelled crRNA concentration was 37.5 µM with the 79-mer tracrRNA-DNA concentration 50 µM. The reactions contained 0.2 M TEAA buffer (pH 7.0), 12.5 mM ascorbic acid, 12.5 mM Cu–THPTA complex, 50% DMSO and 5 mM $MgCl_2$. The ligations were performed for 2 h at RT before desalting by a NAP™ column and loading onto a 10% denaturing PAGE gel. The coupling efficiency as determined by $f_{product}/f_{total}$ x100, where f is fraction. Cy3 24-mer coupling = 42% and ATTO 647N 24-mer = 65% using ImageJ gel analysis. The oligonucleotide sequences can be found in **Appendix Table 5** and **6**.

2.5 Conclusions and Future Work

The sgRNA can be split into parts and chemically ligated to form a functional sgRNA. During this work, it was found 5'-azide conversion is feasible for TBDMS-RNA but not TC-RNA. SPAAC and, in particular, CuAAC reactions were optimised to give efficient non-templated ligation. These strategies have been applied to the production of sgRNA from 37-mer and 66-mer RNA fragments. The design was improved further by splitting a sgRNA into 24-mer and 79-mer RNAs and using a biocompatible triazole linkage for more efficient and cost-effective synthesis. The performance of the clicked ~21–79 sgRNA construct *in vitro* and in cells confirms that this simple sgRNA production strategy provides pure RNA with comparable functionality to RNA generated *via* enzymatic reactions. Moreover, the chemical ligation approach should allow different functional moieties such as LNA[138], PS[103,130,131], fluorophores[140] or amine[140] groups to be easily incorporated into the synthetic sgRNA, potentially improving its activity and expanding its functionality. It was demonstrated that the click ligation products only require simple purification by desalting, increasing their accessibility for researchers who would like to make sgRNA libraries.

Since the activity of the clicked sgRNA was slightly lower than that of IVT sgRNA, the triazole linkage introduced appears to have a destabilising effect. Future work should focus on the effects of the biocompatible triazole linkage on sgRNA binding to Cas9 and its DNA target as well as moving the linkage to different positions in the sgRNA (few nucleotides around the current position). To further simplify the preparation step, the 5'-azido tracrRNA could be made directly by solid-phase synthesis, incorporating an azide-containing monomer, avoiding the need of azide conversion. Currently the addition of an RNA monomer containing an azide at the C5' position *via* a solid-phase phosphoramidite approach is not possible because P(III) in the phosphoramidite could reduce the azide *via* a Staudinger reaction[204]. Attempts should therefore be made to make an azide-containing H-phosphonate or triester P(V) monomer that can be coupled using alternative solid-phase methods to prevent reduction of the azide[205–207]. Other chemical linkages that possess biocompatible properties (i.e. can be tolerated by enzymes) additional to a triazole such as a phosphoramidate[208] or amide[209] linkage (**Figure 2.37**) should also be explored as alternatives for the generation of sgRNA.

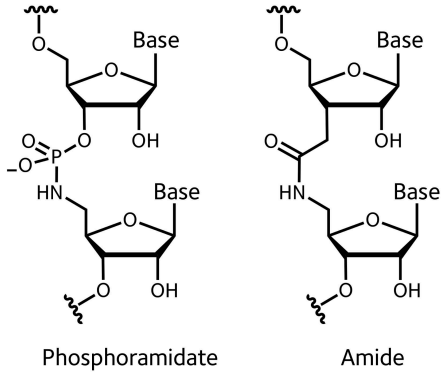


Figure 2.37 | Alternative artificial backbones that can be generated by chemical ligation: phosphoramidate and amide.

Fluorescent gRNAs using CRISPR-dCas9

3.1 Background

3.1.1 Traditional Approaches

Imaging has historically provided the most direct way of studying the effect of genome architecture on biological function. To this end, small molecules such as DAPI and Hoechst 33342[210,211] have been used to non-specifically stain DNA and generate fluorescence, creating a global image of DNA structure in cells. However, there are many other brighter fluorophores that do not bind to DNA. They have been attached to oligonucleotides that hybridise to DNA[212] or RNA[213] targets *via* Watson-Crick base pairing, giving rise to site-specific nucleic acid fluorescence. These oligonucleotides are called Fluorescence In Situ Hybridisation (FISH) probes.

In practice, these fluorescent probes require cells to be 'fixed' – cellular activity is stopped but DNA structure and protein interactions to some extent remain intact. The DNA is then partially denatured allowing the probes to hybridise and any non-specific probe binding is removed by

washing the cells. This provides excellent spatial and temporal information on genome architecture. For example, FISH has enabled enhancer-promoter interactions in transcription activation[214] to be studied and allowed the 24-colour karyotyping of human chromosomes[215,216] for detection of chromosome aberrations. However, fixing cells prevents the study of real-time dynamics and perturbs the natural environment[217]. In principle, imaging in live cells would be ideal.

3.1.2 Sequence-specific Live-cell Imaging

One of the key factors that determine the success of live-cell imaging is a high signal to background ratio since excess probe cannot be efficiently removed. As a result, imaging highly abundant (repetitive) genomic loci such as the centromere or telomere is relatively simple; a single DNA-binding protein can be labelled with a fluorescent protein (FP) to bind to the same region, enriching the fluorescence signal. Customisable DNA-recognition modules such as zinc fingers[218] and TALEs[219] fused to FPs have been successfully used in this context. Imaging of non-repetitive elements, such as genes on the other hand is more challenging as multiple DNA sequence-specific proteins are required. Engineering multiple zinc fingers and TALEs is not straightforward. Similar to genome editing, CRISPR-Cas9 offers a solution; multiple DNA sites can be targeted by complementary sgRNA using a single protein. The Cas9 nuclease can be catalytically deactivated (dCas9, D10A/H840A mutations), fused to a green fluorescent protein (GFP) and programmed by sgRNA to label specific loci either repetitive or non-repetitive regions of the genome in living human cells[102] (**Figure 3.1A**). Alternatively, sgRNA can be modified to carry the "Broccoli" aptamer that generates fluorescence upon binding to a small molecule fluorophore[220] (**Figure 3.1B**).

For labelling of a non-repetitive genomic locus, it was demonstrated that 26–36 different sgRNAs are required per target site to yield detectable fluorescence signal[102,221]. Ideally the number of unique sgRNAs used per target site should be kept as low as possible. Therefore, more than one GFP (3xGFP) have been fused to dCas9 protein to enhance the signal intensity[222]. However, further addition of FPs (~25 kDa) can inhibit imaging due to the massive size of the dCas9–sgRNA complex and the difficulty of its production. Consequently, efforts have been made to

engineer dCas9 to carry a repeating small peptide array called Suntag, which recruits multiple copies of GFP-tagged antibodies that reversibly bind and amplify the fluorescent signal per dCas9–sgRNA complex[223] (**Figure 3.1C**). Alternatively, sgRNAs can be fused to a large number of aptamers that bind to fluorescent proteins. Using this strategy and a highly specialised microscopy setup, currently as few as 4 sgRNAs are required to image one non-repetitive genomic locus[224] (**Figure 3.1D**).

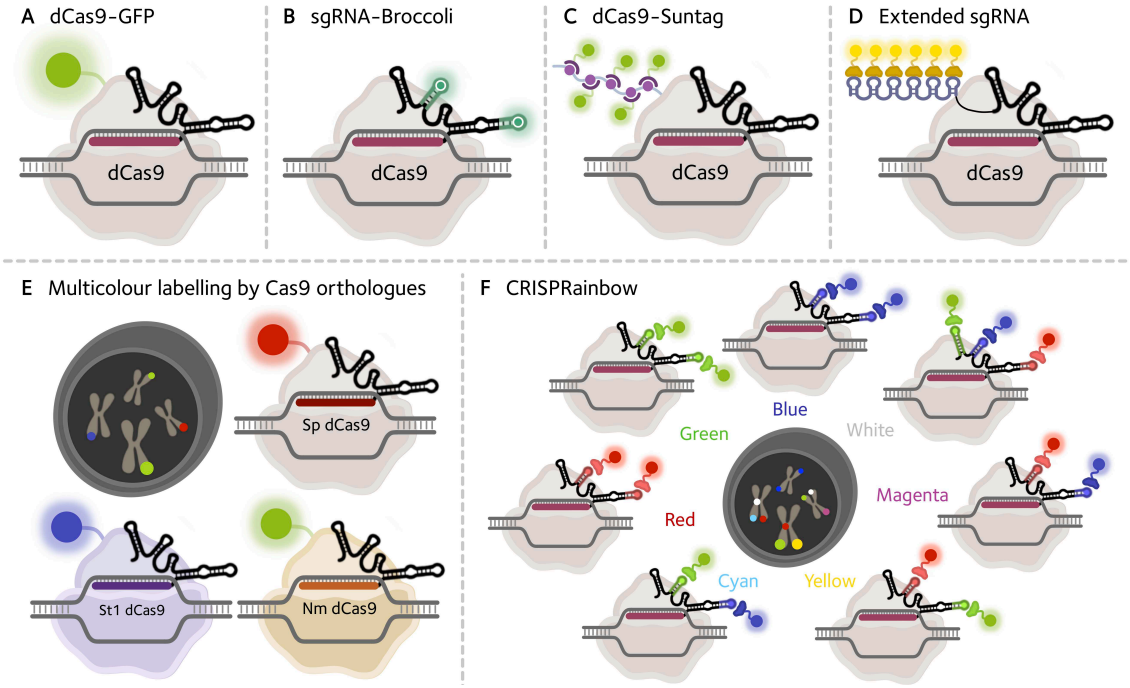


Figure 3.1 | CRISPR-Cas9 based approaches for live-cell imaging. (A) shows the first attempt using a dCas9-GFP fusion protein. (B) shows the incorporation of Broccoli aptamers into sgRNA and their binding to fluorophores. (C) demonstrates dCas9 fused to a Suntag array that binds multiple fluorescently-labelled antibodies, resulting in fluorescence signal amplification. (D) illustrates fluorescence signal is enhanced when a large number of aptamers are appended to sgRNA and recruit multiple fluorescent proteins to the dCas9–sgRNA complex. (E) depicts dCas9 from different species can be linked to different fluorescent proteins for multicolour labelling. (F) demonstrates the colour palette can be expanded to seven colours by mixing and matching two aptamers on the sgRNA.

Simultaneously imaging different genomic loci with different colours offers an opportunity to measure the distance between inter- and intrachromosomal loci, revealing information on chromosome structure and dynamics[222,225]. Initial attempts involve the use of orthogonal dCas9 proteins from three different species (Sp, Nm, and St1) that are fused to different fluorescent proteins (RFP, GFP, and BFP, respectively)[222] (**Figure 3.1E**). However, the targetable sites are

limited by PAM specificities and the maximum number of colours is three. To minimise this problem, different aptamer hairpins recognising different fluorescent proteins can be included into sgRNA in various combinations. Spectral overlapping of individual fluorescence signals generates new colours (secondary colours: cyan = blue + green, yellow = green + red, magenta = red + blue, white = red + blue + green) in addition to the primary colours (red, green and blue). This strategy is called CRISPRainbow[225] (**Figure 3.1F**).

While these strategies are promising, they still have limitations. dCas9 fusion constructs often require the delivery of many large plasmids (~10 kb) into the same cell. This is not simple and may lead to cell toxicity due to the large size of the final complex. Similarly sgRNA with multiple appended aptamers can become very large (~400 nt) and highly structured, making their *in vitro* production and/or transcription in cells difficult[226,227]. This becomes more problematic when many of these modified sgRNAs have to be produced.

3.2 Aims and Objectives

We propose a simple chemical solution to the above problems. In this chapter, synthetic sgRNAs will be sequence-specifically tagged with very bright, highly stable small molecule fluorophore(s) to create fluorescent sgRNAs (fsgRNAs), which should greatly reduce the size of the sgRNA and significantly increase the number of colours available, potentially enabling multiple loci to be imaged. Only a dCas9 expression plasmid is required, simplifying access to live-cell imaging. So far, only few attempts have been made to incorporate the fluorophore into gRNAs but they have been used for FACS enrichment of transfected cells[140] and for fixed-cell imaging (CASFISH[228]). Here our fsgRNAs will be used for live-cell imaging. In addition, we also aim to reduce the background fluorescence from unbound fsgRNAs.

3.3 First-generation Fluorescent sgRNA (fsgRNA I)

3.3.1 Design Strategy

This project was carried out in parallel with the click sgRNA project (**Chapter 2**). As mentioned in **Chapter 2**, previous reports had shown that the tetraloop of the sgRNA could be extended[102] or,

for imaging purposes, replaced by an RNA aptamer[229] without disrupting Cas9 binding. Furthermore, it was shown Cas9 still recognises sgRNA without stem loop 2[230]. Therefore, for our imaging design, the sgRNA was split into two fragments: a DNA-targeting crRNA (36 mer) and Cas9-binding tracrRNA (53 mer) lacking stem loop 2, which would be connected *via* a fluorophore-containing DNA loop (85 mer) using chemical ligation (**Figure 3.2**).

Similar to **Chapter 2**, the sgRNA was split into shorter fragments in order to improve synthesis quality and yields. In addition, only the 36-mer crRNA needs to be synthesised when a new DNA-targeting fsgRNA is required, simplifying the preparation. As a proof of concept, the crRNA was synthesised with a sequence targeting telomeres[102], a region of the genome that contains multiple TTAGGG repeats. Therefore, only one fsgRNA should be required to generate a detectable signal. Moreover, the cells that would be used (human bone osteosarcoma U2OS cells), contain ~65 choromosomes[231] which should facilitate visualisation. 2'-OMe modifications were introduced at the 5' terminus of crRNA to increase stability of the fsgRNA when transfected into cells (to prevent enzymatic degradation). For the tracrRNA, stem loop 2 was deleted in order to improve synthesis and was also produced with a 2'-OMe ribose sugar at the 3' terminus for the same reason as the crRNA. Alexa 488 was chosen as a fluorophore due to its good photostability and hydrophilicity compared to other dyes. The hydrophilicity should reduce localisation of the dye to the cellular/nuclear membrane, improving entry of the fsgRNA into the nucleus[232]. Alexa 488 dyes would be conjugated to amino-C6-dT modifications on a DNA loop. The loop was chosen to be DNA as DNA synthesis is more efficient and cost-effective than RNA. Moreover, the sequence of the DNA loop was chosen to enable it to form a hairpin, a rigid structure that allows each fluorophore to be well separated, minimising dye–dye quenching.

For the chemical ligation strategy, the aim of the design was to control the direction of ligation without the use of DNA splints for each ligation point, simplifying the system. Therefore, 2'-ethoxyazido dT was added to the 3' end of crRNA (f-crRNA I, **Figure 3.2**). A hexynyl was appended to the 5' end of the DNA loop as well as a 2'-ethoxyazido dT at the 3' end (f-DNA loop Ia, **Figure 3.2**). Finally, a C6-amino group was placed at the 5' terminus of tracrRNA (f-tracrRNA Ia, **Figure 3.2**) for subsequent labelling with DBCO-NHS ester to yield f-tracrRNA Ib (**Figure 3.2**). This

order of functional groups allows chemical ligation *via* SPAAC and CuAAC in a stepwise manner (Figure 3.2).

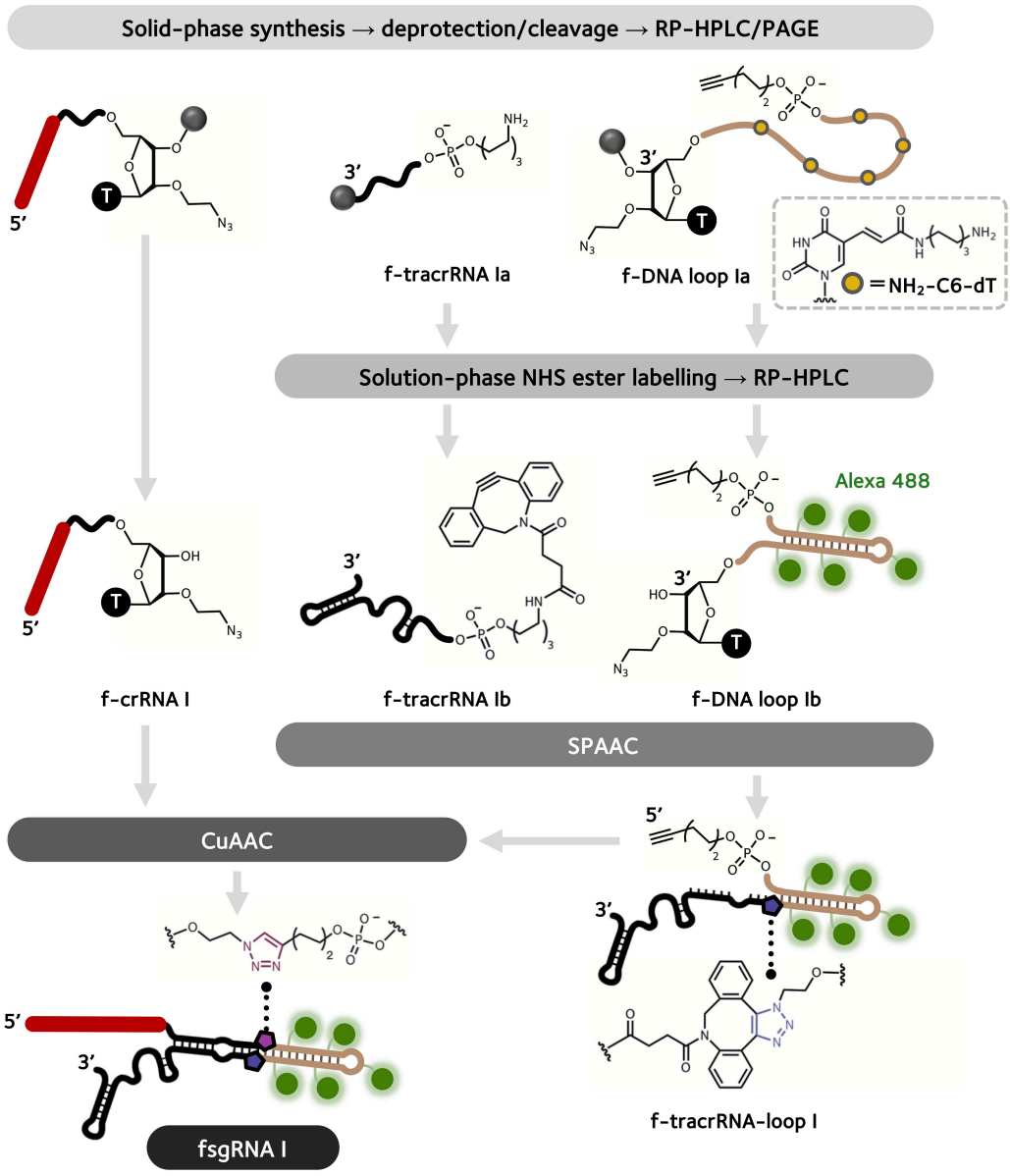


Figure 3.2 | Schematic representation of the first generation fsgRNA (fsgRNA I) design. A DNA loop with terminal 3'-azide and 5'-hexynyl modifications, and five internal NH_2-C6-dT modifications (f-DNA loop Ia) is labelled with Alexa 488-NHS ester to generate f-DNA loop Ib. A 5'-NH_2-C6 tracrRNA (f-tracrRNA Ia) is labelled with DBCO-NHS ester to generate f-tracrRNA Ib. SPAAC ligation of f-tracrRNA Ib and f-DNA loop Ib gives intermediate structure f-tracrRNA-loop I, which is ligated to DNA-targeting f-crRNA I using the CuAAC reaction. This stepwise process gives the final construct for imaging fsgRNA I. The oligonucleotide sequences can be found in Appendix Table 9.

The SPAAC reaction needs to be performed first, otherwise both ends of the DNA loop could be self-ligated in the subsequent CuAAC. Furthermore, the presence of copper, even small

quantities, could bind to the alkyne of DBCO and inhibit its reaction with azide-containing oligonucleotides. After non-templated SPAAC, a template will be generated for the crRNA to hydridise to, resulting in templated CuAAC ligation with the other end of the DNA loop giving fsgRNA I (**Figure 3.2**).

3.3.2 Construct Synthesis

With this strategy in place, the long f-DNA loop Ia (**Figure 3.2**) was first purified by denaturing PAGE due to its length. The strong stability of the DNA loop resulted in a smeared band as the structure could not be fully denatured even in a 7 M urea gel (see **Figure 3.3**, lane 1 for example of DNA loop smearing). However, after purification, mass spectrometry analysis of the PAGE-purified sample showed the correct oligonucleotide mass with a small quantity of N-acetylated DNA loop (+42) generated during deprotection in NH_3. Further purification using RP-HPLC still could not remove the impurity due to the length of the oligonucleotide and the minor difference in hydrophobicity caused by the acetyl group. Therefore, the doubly purified f-DNA loop Ia was coupled with Alexa 488-NHS ester, accepting the impurity could give some molecules with the DNA loop containing four dyes incorporated instead of five. It was felt this should not significantly reduce the fluorescent signal for initial studies. The products from NHS ester-coupling reaction consisted of the fully labelled DNA loop with five (~60%) and four dyes (~40%, **Figure 3.3A**).

Next, the SPAAC ligation of Alexa 488-labelled f-DNA loop Ib and f-tracrRNA Ib was performed for 2 h at room temperature. However, this was not successful when using an excess of the DNA loop relative to the f-tracrRNA Ib (3:1) with the final concentration of the DBCO-labelled f-tracrRNA Ib being 200 μM (**Figure 3.3B**). Increasing concentration of salt in order to bring two components in closer proximity and increasing the temperature to increase the reaction rate only slightly improved the reaction, but was insufficient for the subsequent CuAAC ligation (**Figure 3.3B**).

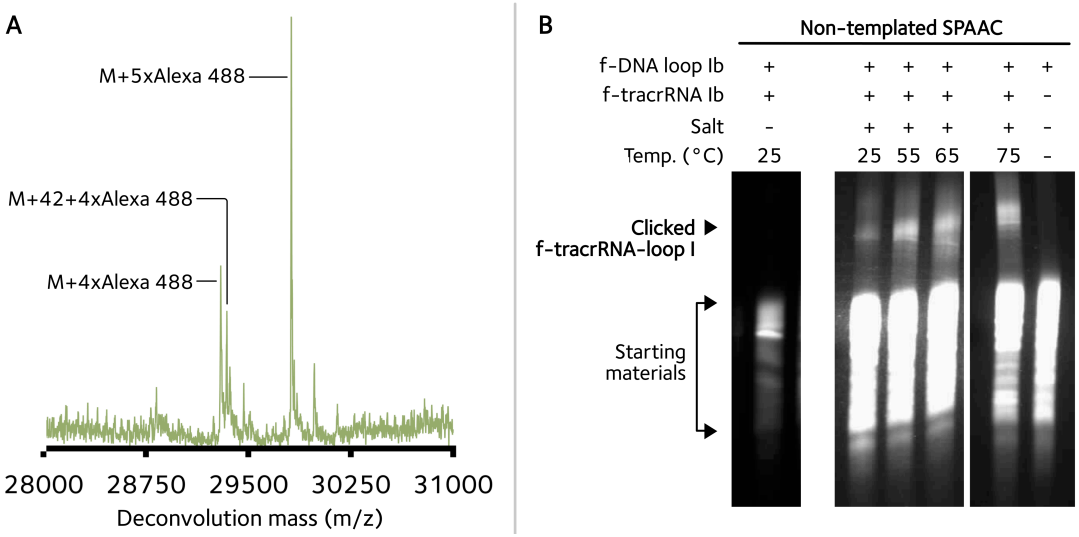


Figure 3.3 | Generation of f-DNA loop Ib and SPAAC ligation of f-DNA loop Ib and f-tracrRNA Ib. (A) shows the deconvolution mass of the Alexa 488-labelled f-DNA loop Ib. Both 4, 4+acetyl (+42) and 5 additions of the dye were observed. M (f-DNA loop I)+5xAleaxa 488 = 29818 (expected); 29818 (found). (B) shows SPAAC ligation. The reaction contained f-DNA loop Ib and f-tracrRNA Ib (3:1) with the final concentration of the f-tracrRNA Ib being 200 µM. Salt was added to a final concentration of 10 mM NaH_2PO4/Na_2HPO4 and 100 mM NaCl. The ligation was performed for 2 h at various temperatures (25, 55, 65 or 75 °C), and loaded onto an 8% denaturing PAGE gel. Note the no-salt reaction was loaded on a different gel with different exposure time for imaging. Fluorescence was then visualised to determine the f-DNA loop Ib starting material (as a streaky band due to the partly denaturation of the stable DNA duplex) and the desired product f-tracrRNA-loop I (slower mobility band). The oligonucleotide sequences can be found in **Appendix Table 9**.

Mass spectrometry of the DBCO-labelled f-tracrRNA Ib showed a +63 adduct, possibly due to binding of copper ions which could inhibit the SPAAC reaction. To confirm whether the source of copper was from the HPLC column (as it had been used to purify CuAAC products) or the mass spectrometry column, a test DNA oligonucleotide with a 5'-amino-dT modification was labelled with DBCO-NHS ester (DBCO-DNA I), purified or not purified by RP-HPLC, and reacted with a 3'-amino-dT DNA oligonucleotide labelled with C6-azide-NHS ester (N_3-DNA I) *via* the SPAAC reaction in a templated manner. As shown in **Figure 3.4**, HPLC-purified DBCO-DNA I could efficiently react with N_3-DNA I, suggesting that the copper contaminant was definitely present in the mass spectrometry column (different from +89 mass obtained from the HPLC column in **Chapter 2**, **Section 2.3.3**). Thus, the problem of unsuccessful ligation between the f-DNA loop Ib and the f-tracrRNA Ib is likely due to the azide group on the 3' end of the DNA loop

being unavailable for click ligation for unknown reasons. The fsgRNA chemical ligation strategy was therefore re-designed.

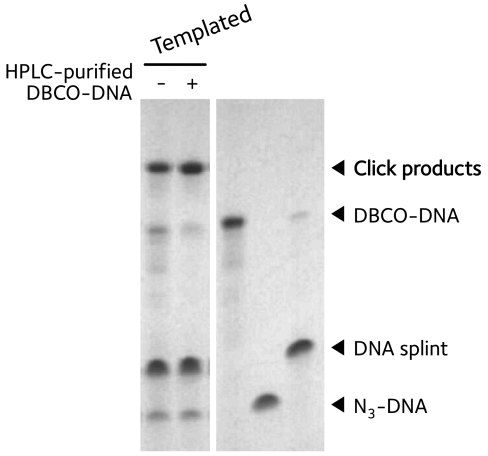


Figure 3.4 | SPAAC ligation of DNA oligonucleotides to determine source of copper ion adduct peaks in mass spectrometry (+63). DBCO-labelled DNA was either purified by RP-HPLC ('+') or used after desalting to remove excess NHS ester ('–') and reacted with N_3-labelled DNA. The templated SPAAC was performed without annealing in the presence of 0.2 M NaCl and consisted of 7:1:1 DBCO-DNA:N_3-DNA:DNA splint with the final concentration of N_3-DNA being 50 μM. For HPLC-purified DBCO-DNA (+), the greater accuracy of the oligonucleotide concentration allowed a 1:1:1 ratio of DBCO-DNA:N_3-DNA:DNA splint in the final reaction with N_3-DNA I being 50 μM. All the reactions proceeded for 2 h at RT and were loaded on a 10% denaturing PAGE gel. The oligonucleotide sequences can be found in **Appendix Table 9**.

3.4 Second-generation Fluorescent sgRNA (fsgRNA II)

3.4.1 Design Strategy and Synthesis

The fsgRNA was still divided into three parts: a crRNA, a tracrRNA and a DNA loop bearing five fluorophores. However, these oligonucleotides were re-designed such that the crRNA could hybridise to the overhang of the DNA loop and tracrRNA could hybridise to crRNA more effectively (**Figure 3.5**). This ensures the correct orientation of the oligonucleotides for a single step CuAAC ligation of the three components to generate the fsgRNA II.

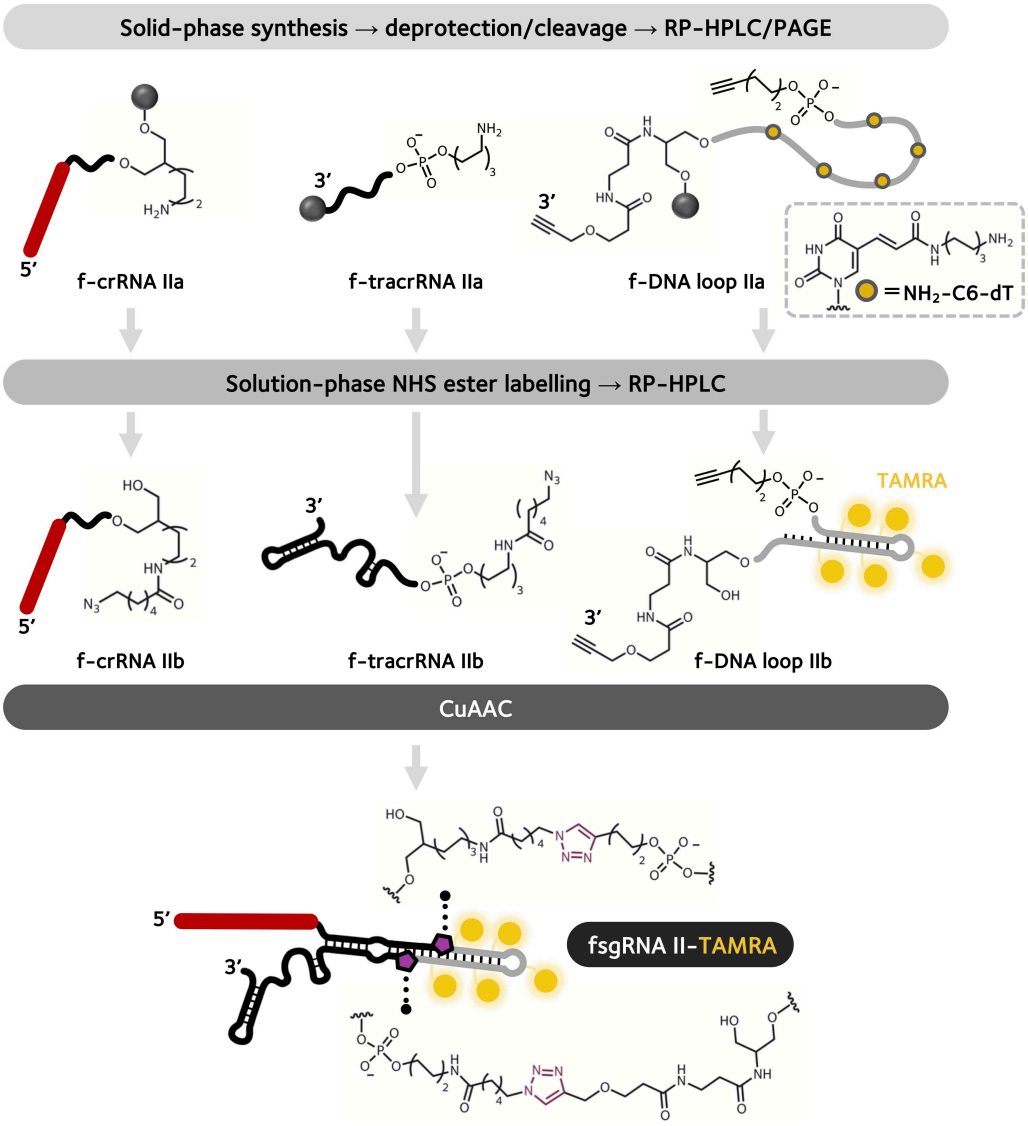


Figure 3.5 | Schematic representation of the second generation fsgRNA (fsgRNA II) design. A DNA loop with terminal 3'-serinol-alkyne and 5'-hexynl modifications, and five internal NH_2-C6-dT modifications (f-DNA loop IIa) is labelled with TAMRA- (or ATTO 647N-) NHS ester to generate f-DNA loop IIb. This loop has an extended 3'-overhang that allows templated hybridisation of f-crRNA IIb and f-tracrRNA IIb, both of which are generated from azide-C6-NHS ester labelling of the f-crRNA IIa and f-tracrRNA IIa respectively. CuAAC ligation of the three components gives fsgRNA II-TAMRA (or II-ATTO). All the oligonucleotide sequences can be found in **Appendix Table 10**.

The 2'-OMe groups on the RNAs and amino-C6-dT bases on the DNA were kept, as with the previous design, except the length of the oligonucleotides and functional groups for chemical ligation were altered. While the length of the DNA loop was fixed (85 mer) to enable incorporation of five fluorophores with optimal spacing, the length of crRNA and tracrRNA was increased (49 mer and 58 mer, respectively) to allow for efficient hybridisation. In this design, the

DNA loop contained terminal 5'-hexynyl and 3'-serinol-alkyne groups (f-DNA loop IIa) to prevent self-ligation during the CuAAC reaction. The crRNA was synthesised with a 3'-C7-amine (f-crRNA IIa), while the tracrRNA was produced with a 5'-C6-amine (f-tracrRNA IIa). Both were later coupled with an azide-C6-NHS ester to yield f-crRNA IIb and f-tracrRNA IIb respectively.

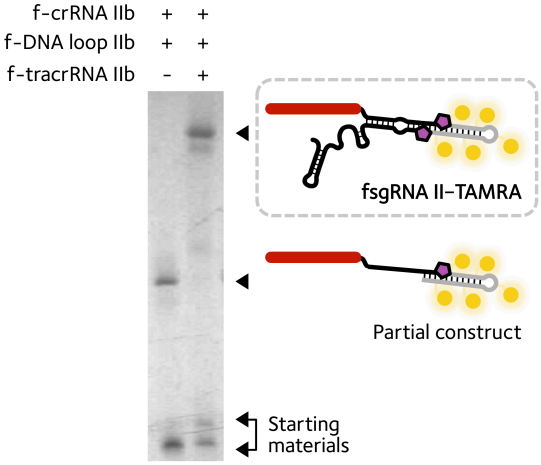


Figure 3.6 | Self-templated CuAAC of second generation fsgRNA II-TAMRA. Following the annealing of 1:1:1 f-crRNA IIb : TAMRA-labelled f-DNA loop IIb : f-tracrRNA IIb (10 µM each final concentration) in the presence of 0.2 M NaCl, $CuSO_4$ (2 mM final concentration), sodium ascorbate (20 mM final concentration) and THPTA ligand (14 mM final concentration) were added and the reaction was left for 2 h at RT. Note that this CuAAC ligation condition was different to those in **Chapter 2** and **4**, which were optimised later. After desalting using a NAP™ column, the reaction was loaded onto an 8% denaturing PAGE gel. The oligonucleotide sequences can be found in **Appendix Table 10**.

As expected, templated CuAAC to give fsgRNA II proceeded well (**Figure 3.6**). Since unmodified IVT sgRNA would be used as a control, the cells needed to express dCas9 with 3xGFP to allow visualisation by confocal microscopy. Thus, fsgRNA II was prepared with TAMRA instead of Alexa 488 dyes, which has the same excitation/emission wavelength as GFP, to allow cross validation of live-cell imaging foci through the colocalisation of GFP and TAMRA signals. To test if the fsgRNA design was limited to a particular dye, fsgRNA II was also prepared with ATTO 647N dyes, which are more hydrophobic than TAMRA but still have emission well separated from GFP. Note that mass spectrometry characterisation was not possible due to the large size of the constructs.

3.4.2 *In vitro* Cas9–fsgRNA II Complex Formation

After purification of the constructs, the binding activity of the fsgRNA II constructs was examined *in vitro* in the presence of commercially available Cas9 protein using the Electrophoretic Mobility Shift Assay (EMSA). Target DNA was not included in this assay as Cas9 would cleave it after forming a complex with the fsgRNA. Without Cas9, the unbound fsgRNA II-TAMRA was present as a single band at the bottom of a 5% native PAGE gel (**Figure 3.7**). However, upon the addition of Cas9, a strong new band was observed at the top of the gel that corresponds to the larger Cas9–fsgRNA II-TAMRA complex. Concurrently the intensity of the unbound fsgRNA II-TAMRA decreased significantly. This indicates that Cas9 binds to the fsgRNA II-TAMRA. To determine if the binding was specific, IVT sgRNA-Competitor was used as a competitive inhibitor to bind to Cas9. If the binding of Cas9 and fsgRNA II-TAMRA was non-specific or weak, only a few equivalents IVT sgRNA-Competitor should be required to disrupt the pre-formed Cas9–fsgRNA II-TAMRA complex. However, it was found that a large excess of IVT sgRNA-Competitor (20–200x greater than fsgRNA II-TAMRA) was required to disrupt the pre-formed complex of Cas9 and fsgRNA II-TAMRA. Once disrupted, the unbound fsgRNA II-TAMRA band at the bottom of the gel was regenerated and the Cas9–fsgRNA II-TAMRA complex band at the top of the gel decreased in intensity, demonstrating that the Cas9 binding of fsgRNA II-TAMRA was specific. The trends with the fsgRNA II-ATTO construct were similar but less clear (**Figure 3.7**). Concentration-dependent Cas9 binding and displacement of the fsgRNA by a competitor caused weaker changes in band intensity, possibly due to the ATTO dye preventing the fsgRNA construct from entering the gel.

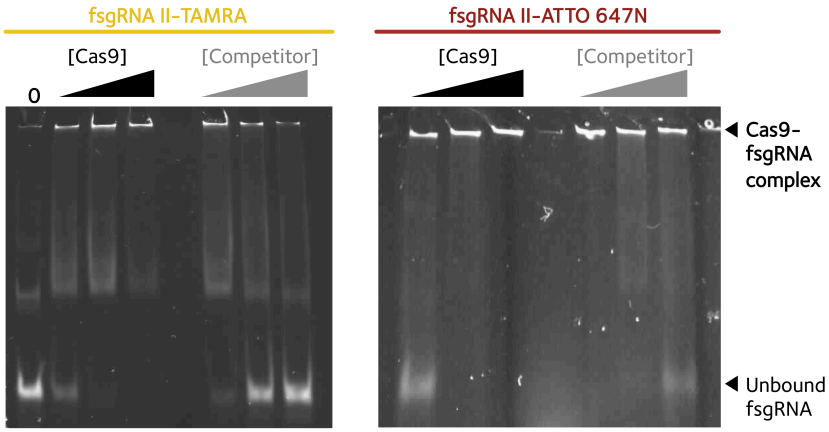


Figure 3.7 | *In vitro* binding of fsgRNA II to Cas9 *via* EMSA. fsgRNA II-TAMRA (0.1 μM, left) or fsgRNA II-ATTO 647N (0.1 μM, right) were mixed with increasing concentrations of Cas9 (0, 0.1, 0.2 and 0.4 μM) in a buffer containing 20 mM HEPES, 100 mM NaCl, 5 mM $MgCl_2$ and 0.1 mM EDTA. IVT sgRNA-Competitor (2, 20 and 200 equivalents relative to fsgRNA) was also added in the indicated lanes. The complex formation proceeded for 15 min at 37 °C before loading onto a 5% native PAGE gel. The oligonucleotide sequences of fsgRNA II precursors and IVT sgRNA-Competitor can be found in **Appendix Table 10**.

3.4.3 Live-cell Imaging

Next, fsgRNA II carrying both dyes were transfected into cells using Lipofectamine reagents, cationic lipids that have widely been used to mediate DNA and siRNA delivery[233–235]. The cationic lipids consist of positively charged heads that interact with the negatively charged oligonucleotides. The lipid-fsgRNA complex is taken into cells *via* endocytosis and is located within vesicles called endosomes. Next, the cationic lipids interact with the anionic lipid from the endosomal membrane to form ion pairs, which subsequently destabilises the lipid bilayer of the endosomal membrane and allows the release of nucleic acids into cytoplasm[236–238]. For successful transfection, this process must take place before the endosomes containing the lipid-fsgRNA complex reach and fuse with lysosomes where fsgRNA is degraded[239]. The fsgRNA in the cytoplasm should form a complex with dCas9-3xGFP, which has a nuclear localisation signal (NLS), expressed within the cells. The complex should then enter the nucleus for telomere targeting. Since telomeres contain repetitive sequences, a large number of dCas9–fsgRNA complexes could be recruited to the same region, enriching the fluorescent signals and resulting in a fluorescent spot. As both dCas9 and fsgRNA were fluorescently tagged, the same spot

should be seen as two colours when using different lasers, confirming the colocalisation of dCas9 and fsgRNA.

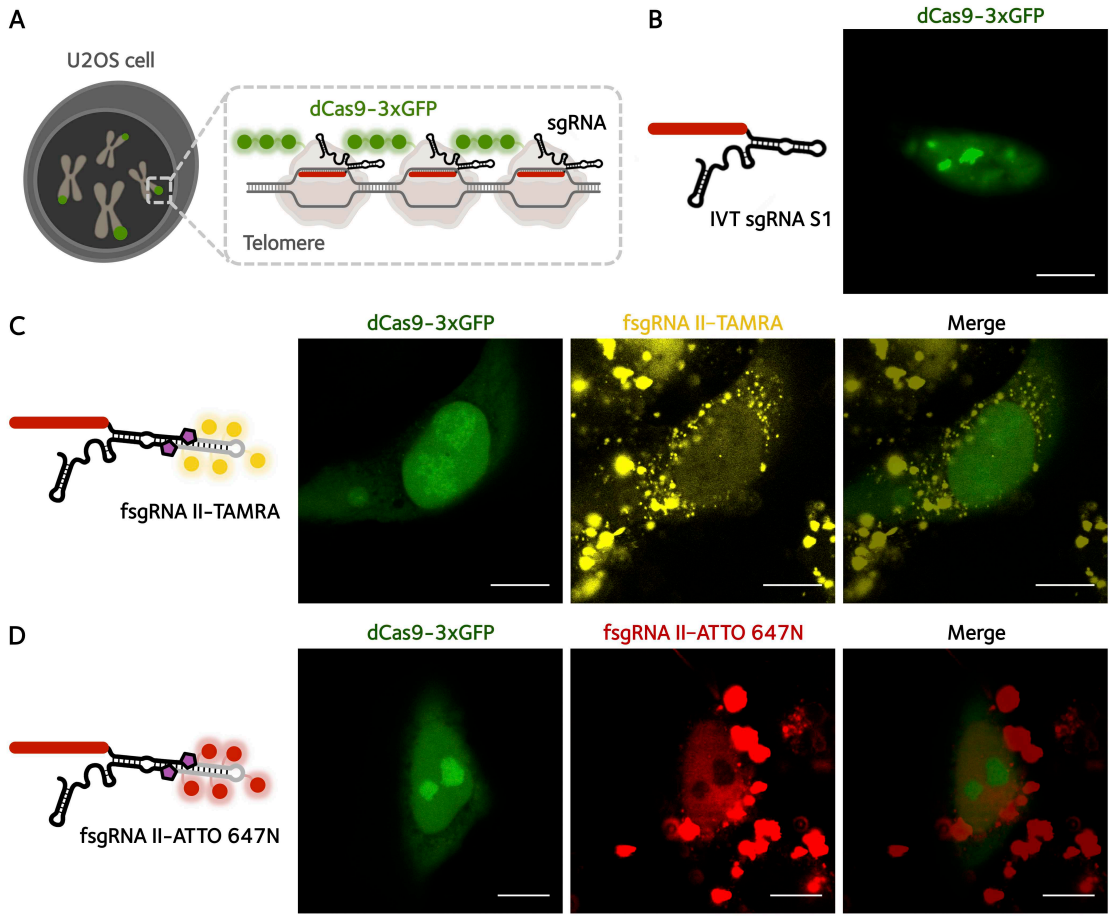


Figure 3.8 | Live-cell imaging using IVT sgRNA S1, fsgRNA II-TAMRA and fsgRNA II-ATTO. (A) schematic diagram of telomeric foci. (B) a control non-fluorescent sgRNA S1 (obtained from *in vitro* transcription), (C) fsgRNA II-TAMRA or (D) fsgRNA II-ATTO 647N were transfected into dCas9-3xGFP-expressing U2OS cells using Lipofectamine reagents. After 24 h, live cells were imaged by confocal fluorescence microscopy using different lasers to visualise GFP (green), TAMRA (yellow) and ATTO 647N (red). Scale bar = 10 µm.

Firstly, unmodified non-fluorescent IVT sgRNA S1 was introduced into dCas9-3xGFP expressing cells. A diffuse GFP staining of the nucleus was observed with larger brighter spots representing nucleoli[102,240] where small noncoding RNA species are enriched[241] and non-specifically interact with unbound dCas9-3xGFP (**Figure 3.8B**). No small bright foci indicative of telomeres were observed. Similarly, no small nuclear foci were seen when fsgRNA II-TAMRA was transfected into cells in the place of IVT sgRNA (**Figure 3.8C**). Instead, a diffuse nuclear staining was seen suggesting the dCas9–fsgRNA–DNA complex was not formed. Bright large granules outside the

nucleus were also observed indicating endosomal escape of the lipid-fsgRNA II complex[242,243] was not complete or fsgRNA II was digested within endosomes/lysosomes. The same pattern was also detected in the cells transfected with fsgRNA II-ATTO (**Figure 3.8D**), showing the result is dye-independent. Most importantly, it should be noted that the control IVT sgRNA S1 failed to label the telomeres (**Figure 3.8B**) compared to litrature[102]. This suggests that the introduction of fluorophores might not be the main reason for unsuccessful imaging. As a result, the structure of the sgRNA was optimised.

3.5 Optimisation of sgRNA Structure

To determine the optimal fsgRNA structure for live-cell imaging, many different modified sgRNA structures were chosen from the literature[74,102,230], which are reported to be functional in cells, and were generated *via in vitro* transcription. Their expected secondary structure in the absence of the variable DNA-targeting 20-mer was then confirmed using mfold[199]. Stem loop 2 was re-introduced (IVT sgRNA S2 / S3 / S4, **Figure 3.9**). The DNA-targeting sequence was reduced to 18 mer to increase specificity[74] (IVT sgRNA S3 / S4 / S5 / S6, **Figure 3.9**). The repeat–antirepeat hairpin was extended by 5 bp to further enhance dCas9 binding[102] (IVT sgRNA S4 / S5, **Figure 3.9**) and the extension sequence was also varied (IVT sgRNA S6, **Figure 3.9**). These IVT sgRNAs were then transfected into cells and the strength of the nuclear foci were compared to background.

As shown in **Figure 3.9B**, when the stem loop 2 was re-introduced, a large number of fluorescent foci emerged with significantly reduced nuclear background. This indicates the importance of the stem loop 2 for stable dCas9–fsgRNA binding to target DNA, an important feature for live-cell imaging. Interestingly, sgRNA lacking stem loop 2 (IVT sgRNA S1) initiates DNA cleavage[230], suggesting the sgRNA requirements for live-cell imaging and gene editing are different. Shortening the DNA-targeting sequence did not improve GFP spot formation (**Figure 3.9C**). Appending a 5-bp extension to the original sgRNA structure further improved dCas9 binding as shown by the increased fluorescent foci formed and minimal nuclear background observed (**Figure 3.9D**). Surprisingly, when comparing fluorescent foci, the removal of stem loop 2 was partially compensated for by the presence of the 5-bp extension (**Figure 3.9E**) suggesting that

the 5-bp extension stabilised the dCas9–sgRNA complex. Moreover, changing the sequence of the 5-bp extension still gave rise to the stabilising effect (**Figure 3.9F** compared to **Figure 3.9E**).

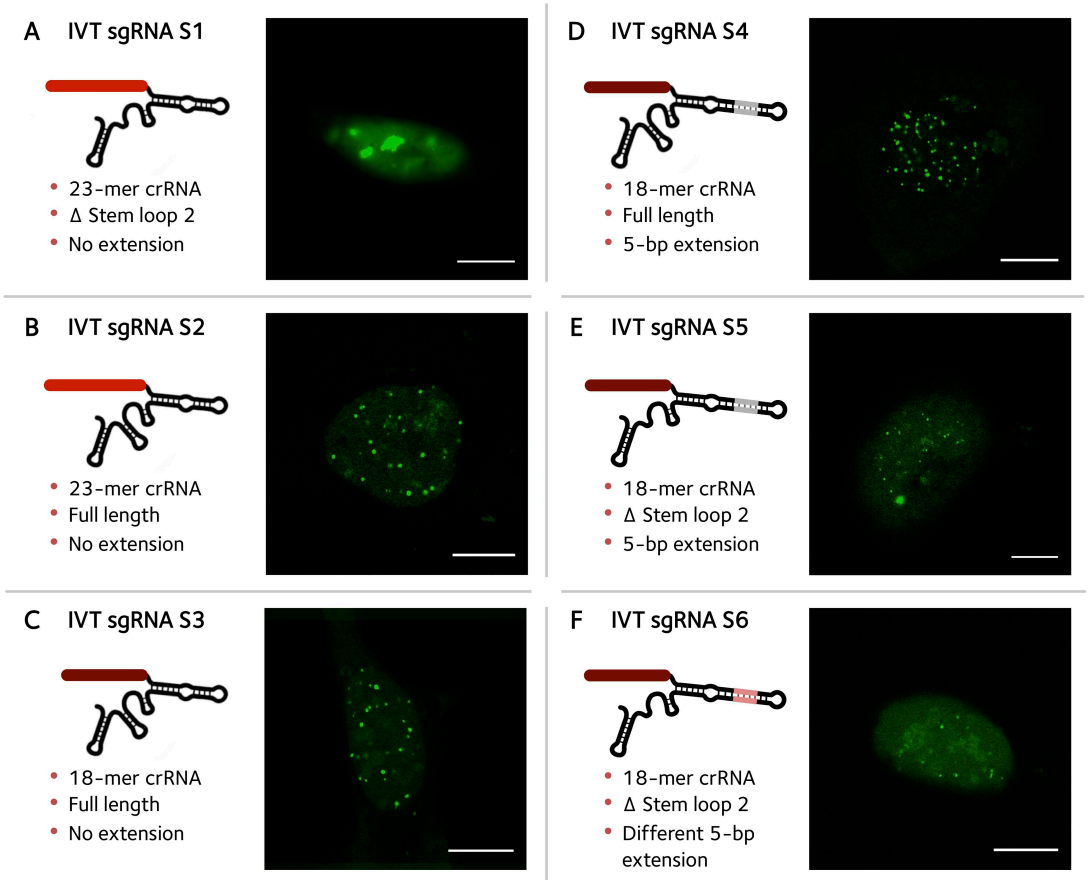


Figure 3.9 | Live-cell imaging using various different IVT sgRNA structures. The following IVT sgRNAs were transfected into U2OS cells expressing dCas9-3xGFP (green) and live-cell imaging was carried out after 24 h. (A) IVT sgRNA S1 with a 23-mer RNA for DNA targeting, stem loop 2 deleted and no extension at the tetraloop (the control IVT sgRNA used in previous live-cell imaging, **Figure 3.8A**). (B) IVT sgRNA S2 with a 23-mer RNA for DNA targeting, stem loop 2 included and no extension at the tetraloop. (C) IVT sgRNA S3 with shortened DNA-targeting RNA (18 mer), stem loop 2 included and no extension at the tetraloop. (D) IVT sgRNA S4 with shortened DNA-targeting RNA (18 mer), stem loop 2 included and 5-bp extension at the tetraloop. (E) IVT sgRNA S5 with shortened DNA-targeting RNA (18 mer), stem loop 2 deleted and 5-bp extension at the tetraloop. (F) IVT sgRNA S6 with shortened DNA-targeting RNA (18 mer), stem loop 2 deleted and different sequence of 5-bp extension at the tetraloop. Green fluorescent spots indicate the localisation of the dCas9-3xGFP–IVT sgRNA complexes at telomeres. Scale bar = 10 μm. The oligonucleotide sequences can be found in **Appendix Table 12**.

When comparing the images from fsgRNA II-ATTO (**Figure 3.8C**) and IVT sgRNA S6, both of which contain the same 5-bp extension but lack stem loop 2 (**Figure 3.9F**), there could be two factors that inhibited live-cell imaging using fsgRNA II-ATTO. One could be that the presence of two

bulky triazole linkages destabilises the 5-bp extension, and negatively affects the dCas9–fsgRNA or dCas9–fsgRNA–target DNA complex. The second could be that the presence of fluorophores prevents the fsgRNA from entering the nucleus. Therefore, further optimisation was performed to address these hypotheses.

One of the quickest ways to study the destabilising effects caused by the triazole linkages is to introduce RNA mismatches (U–U mismatch ΔG = 0.5–1.5 kcal/mol depending on the local bases[244]) adjacent to the 5-bp extension region of IVT sgRNA to partly mimic triazole destabilisation (IVT sgRNA M1, **Figure 3.10A**). When this mismatch construct was transfected into cells, the number of fluorescent foci was significantly reduced compared to perfectly complementary IVT sgRNA S4 (**Figure 3.9D**). This indicates that RNA mismatches disrupted the 5-bp duplex extension and destabilised the dCas9–sgRNA or dCas9–sgRNA–target DNA complex. The mismatch-containing sgRNA M1 in fact generated fewer spots than the native IVT sgRNA S2 lacking the 5-bp extension (**Figure 3.9B**). It should be noted that the hairpin containing the mismatches still forms as expected (confirmed by mfold[199]), but the local hairpin stability cannot be accurately predicted. These findings suggest the destabilising effects from triazole linkages might be more profound than initially expected resulting in no fluorescent spots in cells (**Figure 3.8B** and **3.8C**).

In order to refine the position of the triazole linkages in future design, IVT sgRNAs were generated where the RNA mismatches are moved further away from Cas9 by increasing the extension length. As expected, the formation of fluorescent foci was restored when the mismatches were further away by 6 bp (IVT sgRNA M2, **Figure 3.10B**). However, moving the mismatches even further (by 10 bp relative to the initial location) did not improve fluorescent spot formation (IVT sgRNA M3, **Figure 3.10C**). Since synthesis of longer RNAs is required for constructing IVT sgRNA M3, for future designs, the location of triazole linkages should be in the same position as the RNA mismatches in IVT sgRNA M2.

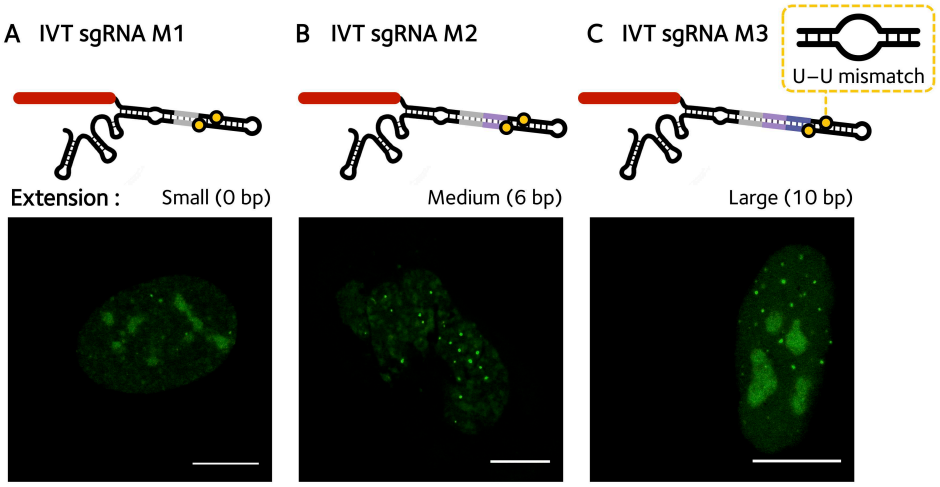


Figure 3.10 | Live-cell imaging using IVT sgRNAs containing RNA mismatches. (A) IVT sgRNA M1 contained two U–U RNA mismatches at the exact same place as triazole linkages in fsgRNA II constructs. (B) IVT sgRNA M2 contained two U–U RNA mismatches, which were moved away from the original position by 6 bp. (C) IVT sgRNA M3 contained two U–U RNA mismatches, which were moved even further away (10 bp) from the original location. All of these IVT sgRNAs were transfected into dCas9-3xGFP-expressing U2OS cells and live-cell imaging was performed 24 h later. Green fluorescent spots indicate the localisation of the dCas9-3xGFP–IVT sgRNA complexes at telomeres. Scale bar = 10 μm. The oligonucleotide sequences can be found in **Appendix Table 12**.

3.6 dCas9–fsgRNA Ribonucleoprotein Complex for Imaging

Another factor that possibly contributed to unsuccessful live-cell imaging is the inefficient entry of fluorophore-containing fsgRNA into the nucleus. Since dCas9-3xGFP contains a Nuclear Localisation Signal (NLS), it was hypothesised that it could escort the fsgRNA into the nucleus if the complex was formed *in vitro* and transfected into cells with fsgRNA. To test this, dCas9-3xGFP was expressed and purified *in vitro*.

3.6.1 dCas9-3xGFP Expression and Purification

To date, dCas9-3xGFP had been expressed in human cells under the control of the human promoter (CMV). Therefore, cloning was required to construct a plasmid that could express dCas9-3xGFP in bacterial cells under the control of the bacterial promoter (tac promoter) (**Figure 3.11A**). Since the plasmid expressing AsCas12a protein in bacterial cells was available (from work in **Chapter 4**), two PCR reactions were performed to copy the dCas9-3xGFP gene from the human vector, and the AsCas12a vector without the AsCas12a gene (i.e. the plasmid backbone

with the tac promoter and a six-histidine tag). These PCR amplicons were generated with tailed primers that provide sequence homology at the 5' and 3' ends of the two amplicons. Next, the two overlapping PCR amplicons were joined *via* Gibson Assembly that uses 5' exonuclease – to create overhangs for efficient ligation, DNA polymerase – to fill the gap after two overhangs annealed, and DNA ligase – to seal the nicks. The resulting plasmid was then transformed into bacterial cells, isolated and purified.

Expression of dCas9-3xGFP was then performed in bacterial cells using the protein induction condition as of the original vector for Cas12a[245]. Since the six-histidine tag was fused to the N-terminus of the protein, purification by Immobilised Metal Affinity Chromatography (IMAC) that uses nickel (Ni^{2+}-bound resin) to selectively bind to His-tagged dCas9-3xGFP was performed according to the published Cas9 purification protocol[246]. SDS-PAGE showed that fairly pure dCas9-3xGFP was obtained using IMAC, however there were some impurities remaining as a few bands below the main dCas9-3xGFP band (**Figure 3.11B**). Therefore, the protein was subject to another round of purification using ion exchange chromatography[246]. Since the net charge of dCas9-3xGFP is positive (ca. +22 for Cas9[247], +4 for Lysine-rich NLS and -7 for each superfolder GFP[248]), the HiTrap SP HP column packed with a cation exchange resin was used to allow only the positively charged dCas9-3xGFP to be immobilised and later eluted as a pure protein using a linear KCl gradient from 0.1 M to 1 M. As result, the purity of the protein was improved as shown in SDS-PAGE (**Figure 3.11B**) and it was sufficient for subsequent *in vitro* activity assays and live-cell imaging.

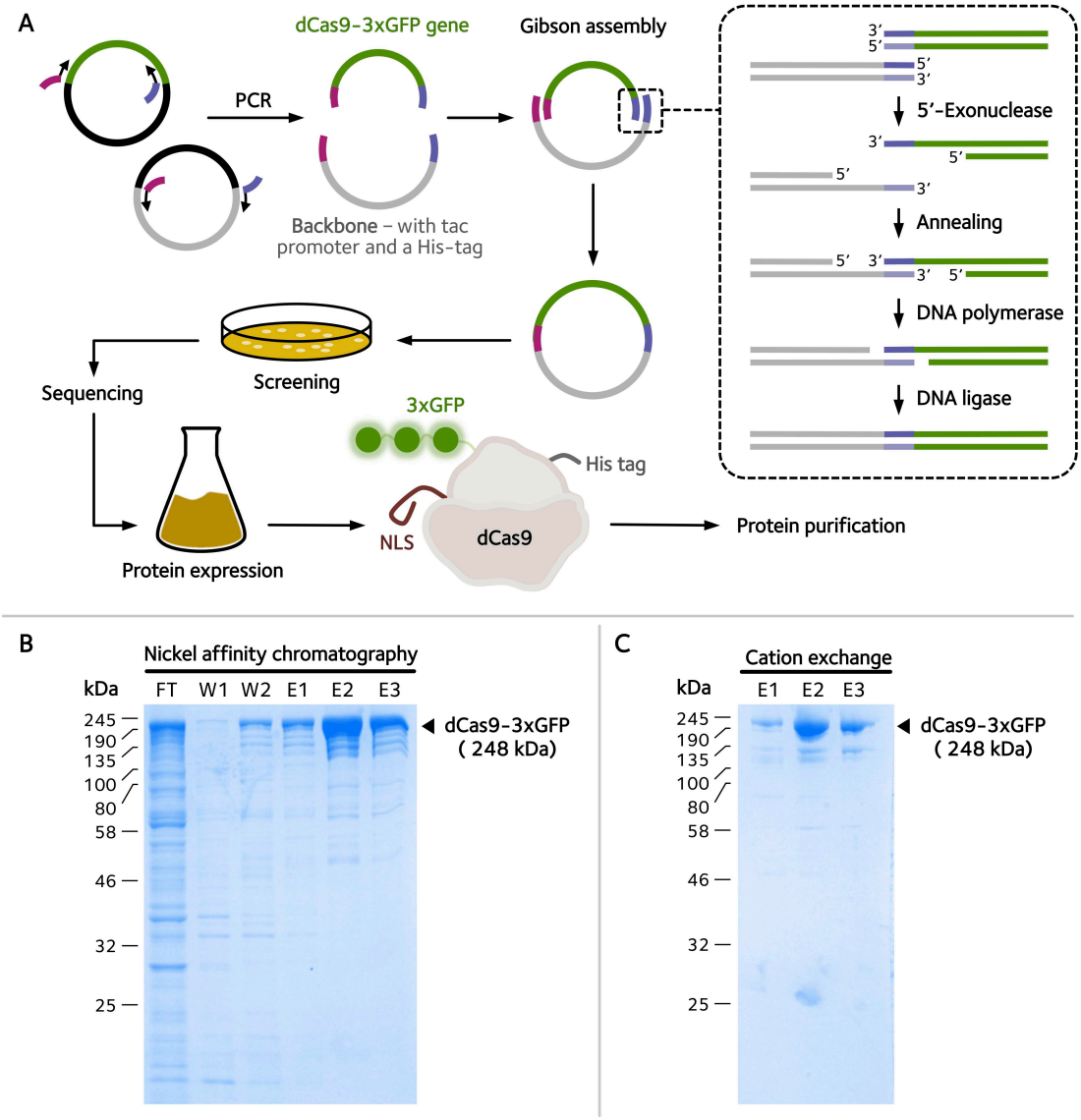


Figure 3.11 | Expression and purification of dCas9-3xGFP. (A) Cloning strategy to create bacterial dCas9-3xGFP expression vector. The dCas9-3xGFP gene and a backbone containing a tac promoter and his tag were PCR-amplified from different vectors. The primer can be found in **Appendix Table 24**. Upon Gibson assembly of the dCas9-3xGFP vector, transformation and screening the desired plasmid was obtained and sequenced. Transformation of the plasmid into bacterial cells allowed protein expression. After ~18 h, cells were then harvested and lysed to release the protein for subsequent purification. (B) The his-tagged dCas9-3xGFP was purified *via* nickel affinity chromatography and examined by SDS-PAGE. (C) Elution fractions 2 and 3 from nickel affinity chromatography were combined and subject to further purification *via* cation exchange chromatography. The purity of dCas9-3xGFP was determined by SDS-PAGE. Note that FT = Flow Through (the protein of interest and contaminants that did not bind to resin), W = Wash (the protein of interest and contaminants that weakly or non-specifically bound to resin) and E = Elution (the protein of interest that specifically bound to the resin).

3.6.2 Evaluating dCas9-3xGFP Activity

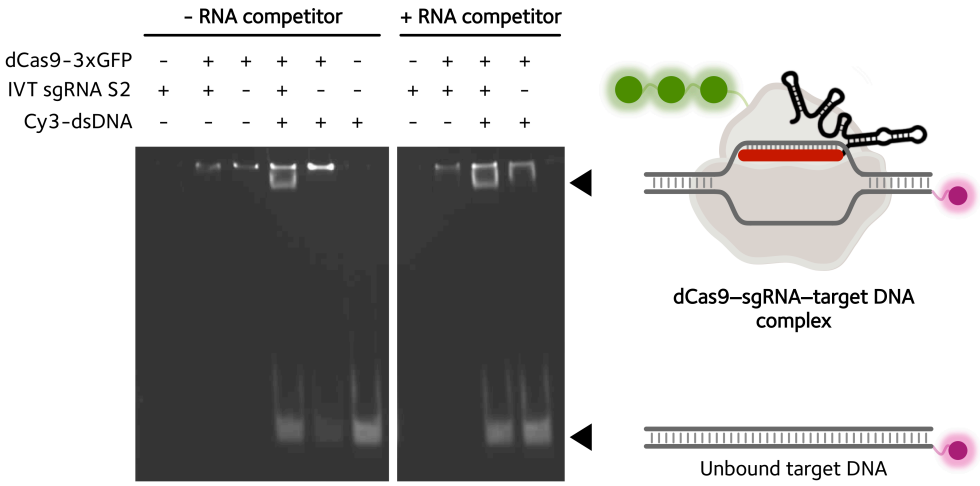


Figure 3.12 | *In vitro* binding of purified dCas9-3xGFP to IVT sgRNA S2 and Cy3-labelled target DNA using EMSA. The binding reaction contained 0.6 µM Cas9, 0.6 µM IVT sgRNA S2, 0.6 µM Cy3-labelled dsDNA, 20 mM HEPES, 100 mM NaCl, 5 mM $MgCl_2$, 0.1 mM EDTA and 100 µg/mL BSA. The RNA competitor (crRNA46-alk) was added to a final concentration of 30 µM as specified. After complex formation for 30 min at 37 °C, all the reactions were loaded onto an 8% native PAGE gel. The oligonucleotide sequences can be found in **Appendix Table 12**, **14** and **18**.

Next, the activity of the protein was tested using EMSA in the presence of IVT sgRNA S2 and Cy3-labelled target DNA. Without dCas9-3xGFP, the target DNA moved as a single band at the bottom of a native PAGE gel (**Figure 3.12**). In the presence of dCas9-3xGFP, the intensity of the unbound target DNA decreased while the intensity of the protein at the well increased, indicating there is non-specific binding between dCas9 and target DNA. However, in the presence of IVT sgRNA S2, a new band was formed which represents the complex of dCas9–sgRNA–target DNA. This demonstrates that the purified dCas9-3xGFP was active. To further confirm the identity of the purified protein, random RNA (crRNA from the CRISPR-Cas12a system in **Chapter 4**) was added to the protein–DNA mixture instead of IVT sgRNA S2, and a far fainter band at a similar position the dCas9–sgRNA–target DNA complex was observed. This band is possibly the result of Cas9 non-specifically interacting with the RNA and DNA to make a complex that is sufficiently negatively charged to enter the gel. However, during the run time of the gel, these weaker interactions are disrupted and the final observed band is weaker than the correct dCas9–sgRNA–target DNA complex. Furthermore, the random RNA was also added in the presence of dCas9, IVT sgRNA S2 and target DNA and the intensity of the dCas9–sgRNA–target

DNA complex band was evaluated. As expected, the intensity of the dCas9–sgRNA–target DNA complex was not reduced. These observations in combination confirmed that the purified protein is active dCas9-3xGFP.

3.6.3 Ribonucleoprotein Complex Transfection

As an initial test, dCas9-3xGFP was complexed to IVT sgRNA S2 targeting the telomere *in vitro*. The complex then was transfected into cells using Lipofectamine reagents. Unfortunately, no fluorescent foci were observed inside the nucleus. Instead diffuse staining was observed, possibly due to localisation of complex to the cell membrane (**Figure 3.13**). This suggests that lipid-based transfection would not be a suitable method for dCas9–fsgRNA complex delivery. This may be due to the reduced Cas9–sgRNA complex negative charge, reducing its interactions with the cationic lipids, resulting in less efficient lipid-RNP complex formation and hence poor cell uptake. Previous studies had illustrated the use of lipid-based transfection of RNP for successful genome editing[247], suggesting only a small amount of the complex needs to enter the nucleus for genome editing. In contrast, live-cell imaging needs a significantly larger amount of dCas9–sgRNA complex for a detectable fluorescent signal.

As a consequence, transfection was attempted using electroporation (Amaxa® Nucleofector® II, the Cell Line Nucleofector® Kit V, Programme X-001), a lipid-free technique that uses an electrical pulse to generate temporary pores in cell membranes to allow the RNP to enter the cells. As shown in **Figure 3.13**, the dCas9-3xGFP–IVT sgRNA S2 complex was located in cytoplasm, indicating that it could enter the cell but not the nucleus for unknown reasons. Once more this shows the different requirements of live-cell imaging and genome editing, where electroporation is effective for gene editing[105,146]. Thus, it appears ribonucleoprotein transfection cannot be used for fluorophore-containing fsgRNA, however more studies exploring protein concentration and the exact transfection method are required to support this conclusion.

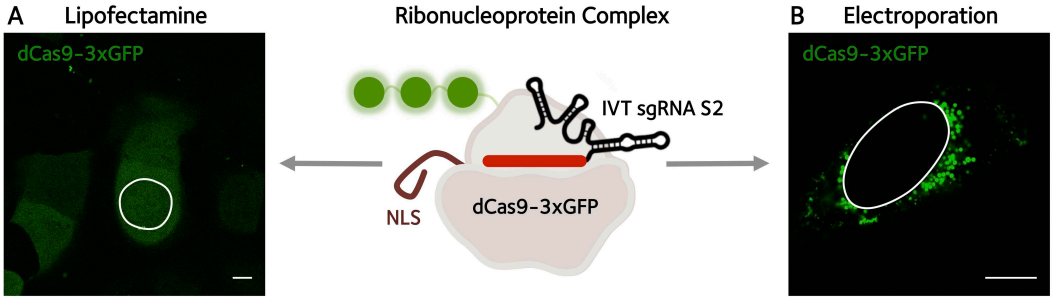


Figure 3.13 | Live-cell imaging by transfection of the ribonucleoprotein complex of dCas9-3xGFP and IVT sgRNA S2 using (A) Lipofectamine reagents and (B) electroporation. With Lipofectamine-based approach, the complex of dCas9-3xGFP and IVT sgRNA S2 was formed *in vitro* in Opti-MEM prior to transfection. On the other hand, Nucleofector solution was used for dCas9-3xGFP–IVT sgRNA S2 complex formation *in vitro* prior to electroporation. The GFP signal is shown in green. Scale bar = 10 μm. The IVT sgRNA S2 sequence can be found in **Appendix Table 12**.

3.7 Fluorescent Click sgRNAs

In parallel, the click ~21–79 sgRNA strategy from **Chapter 2** was used for live-cell imaging. The 21-mer DNA-targeting RNA with a 5'-C6-amine group and a 3'-propargyl was labelled with Cy3 or ATTO 647N. Upon ligation with the 79-mer 5'-azido tracrRNA *via* non-templated CuAAC, a fsgRNA containing a biocompatible triazole linkage was generated (codes: click sgRNA-Cy3, click sgRNA-ATTO, click sgRNA-DNA-Cy3 and click sgRNA-DNA-ATTO), characterised by mass spectrometry (**Appendix Table 16**) and transfected into cells for telomere labelling. The click sgRNA and sgRNA-DNA without a fluorophore attached at the 5' end was used as a control (codes: click sgRNA-NH_2 and click sgRNA-DNA-NH_2), and unexpectedly, no fluorescent foci emerged (**Figure 3.14**); similar results were observed for Cy3- and ATTO-constructs. This could be due to the instability of the dCas9–click sgRNA or dCas9–click sgRNA–target DNA complex. This is consistent with the in cell gene editing results in **Chapter 2**, **Section 2.4.6**, which showed that the click sgRNA could mediate indel formation but with reduced efficiency compared to IVT sgRNA. The destabilising effect of a triazole linkage may therefore be more detrimental in live-cell imaging, where a stable dCas9–sgRNA–target DNA complex is strongly required.

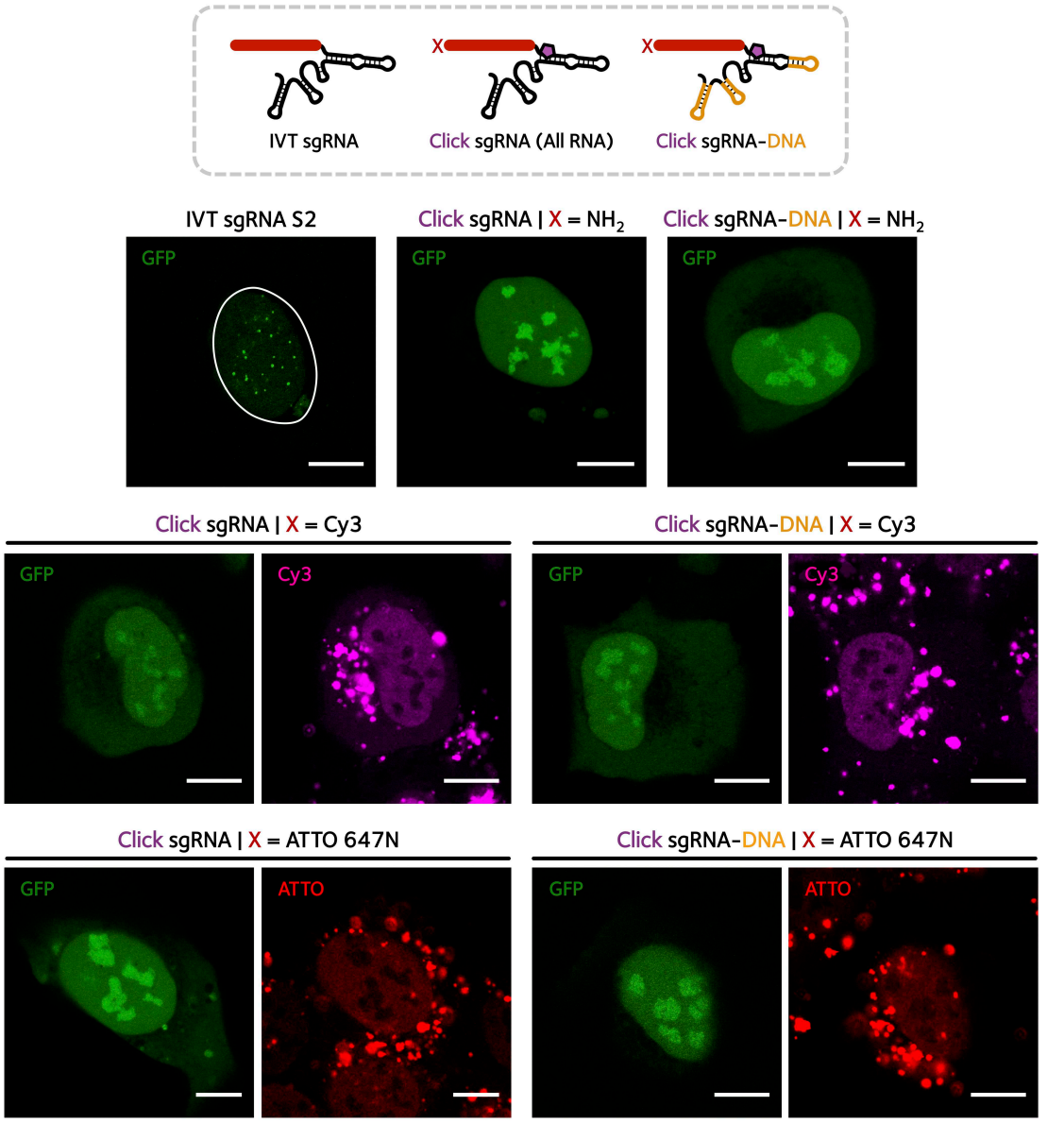


Figure 3.14 | Live-cell imaging using fluorescent click 21–79 sgRNA constructs. The 21-mer DNA targeting RNA was synthesised with a C6-amine at the 5' end which was labelled with Cy3 or ATTO 647N-NHS esters to obtain a fluorescent DNA-targeting RNA. Upon clicking with 79-mer tracrRNA or tracrRNA-DNA, the fluorescent sgRNA and sgRNA-DNA were generated and transfected into dCas9-3xGFP-expressing U2OS cells. After 24 h, live-cell imaging was performed with different lasers used to visualise GFP (green), Cy3 (pink) and ATTO 647N (red). Scale bar = 10 µm. The oligonucleotide sequences can be found in **Appendix Table 16**.

3.8 Cy3-labelled IVT sgRNA

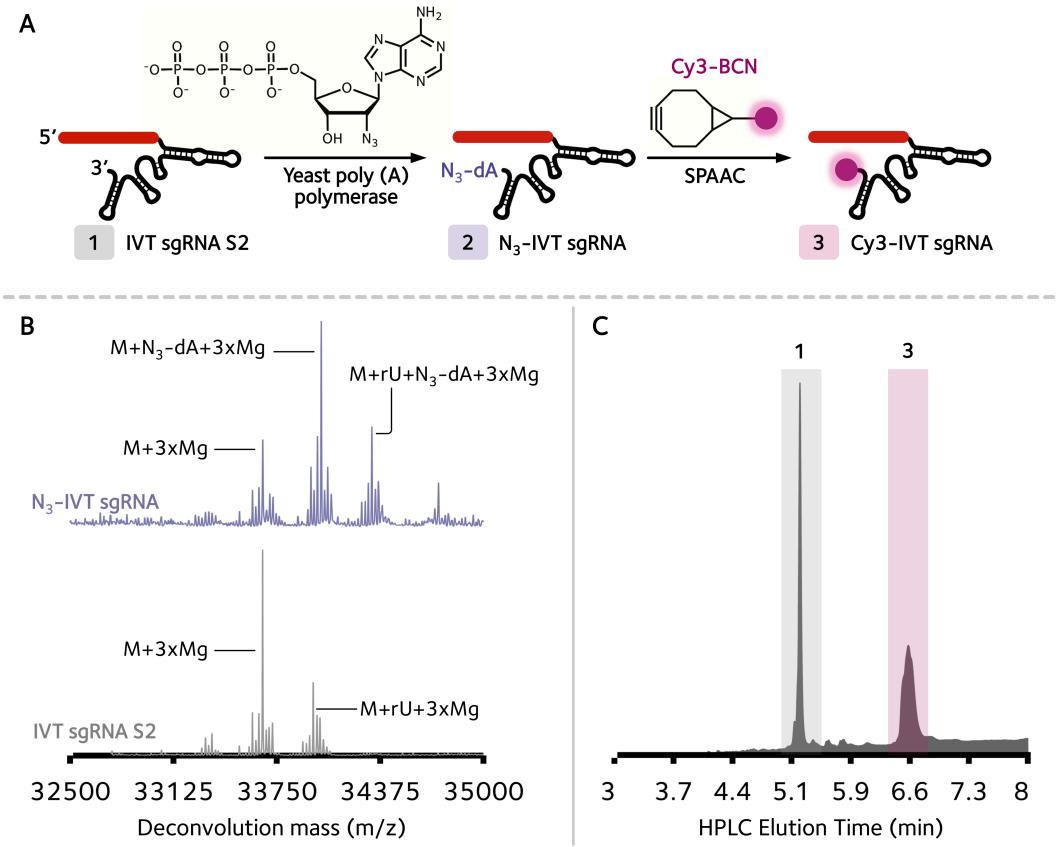


Figure 3.15 | Generation of Cy3-labelled IVT sgRNA. (A) An IVT sgRNA S2 was appended with a 2'-N_3-dA by yeast poly (A) polymerase at the 3' terminus, giving an N_3-IVT sgRNA (2) which was then reacted with Cy3-BCN *via* SPAAC to yield a Cy3-IVT sgRNA (3). (B) Deconvolution mass of an N_3-IVT sgRNA (top panel) showed unincorporated IVT sgRNA starting material and the desired N_3-IVT sgRNA. For comparison the mass of IVT sgRNA S2 is shown in the bottom panel. Note that an additional rU was added during *in vitro* transcription, thus the N_3-dA incorporation of this species was detected. Mg^{2+}/Na^{+} adducts were also observed. M (IVT sgRNA S2)+3xMg = 33677 (expected); 33663 (found). M (N_3-IVT sgRNA)+3xMg = 34013 (expected); 34019 (found). (C) HPLC profiles of an IVT sgRNA that was successfully incorporated with Cy3 (pink) and an azide-free IVT sgRNA (grey). Due to incomplete labelling with 2'-N_3-dA, residual IVT sgRNA S2 (grey) was present in the HPLC profile. M (Cy3-IVT sgRNA+3xMg) = 34839 (expected); 34845 (found). The oligonucleotide sequences can be found in **Appendix Table 12** and **13**.

To investigate the feasibility of using fluorophore-containing sgRNA in live cells, sgRNA was generated *via in vitro* transcription, appended with a 2'-azido dA at the 3' terminus using poly A (yeast) polymerase and then labelled with Cy3-BCN *via* the SPAAC reaction (Cy3-IVT sgRNA, **Figure 3.15A**). Cy3 was used as it is available as a BCN conjugate. Mass spectrometry showed that incorporation of 2'-azido dA into IVT sgRNA gave three types of products: no incorporation,

one incorporation relative to the expected IVT sgRNA S2 mass and one incorporation relative to the expected IVT sgRNA S2 mass plus a terminal uracil (the final product is an artefact from IVT, **Figure 3.15B**). After labelling with Cy3-BCN, RP-HPLC was required to separate the Cy3-IVT sgRNA from unlabelled species (**Figure 3.15C**) and was characterised by mass spectrometry (**Appendix Table 13**). Even though fsgRNA could be produced enzymatically, it should be noted that the efficiency of the process was poor and gave low yields compared to chemically synthesised fsgRNA.

Next, the Cy3-IVT sgRNA was transfected into cells using Lipofectamine reagents. As shown in **Figure 3.16A**, the Cy3-IVT sgRNA could enter the cells and mediate fluorescent spot formation in both GFP and Cy3 channels. However, a major concern was the strong background signal from a number of bright granules around the nucleus that correspond to fsgRNA trapped inside endosomes. An attempt was made to reduce this background signal by performing FRET such that only fsgRNA that forms a complex with dCas9 would show GFP–Cy3 energy transfer and emission. This does reduce unbound fsgRNA signal from the endosomes, however due to the high density of Cy3-labelled sgRNA in the endosomes, some non-FRET Cy3 emission was still observed (**Figure 3.16A**). Reducing the concentration of Cy3-IVT sgRNA by 10 fold with the aim of reducing background also resulted in the significantly weaker signal of telomeric foci (**Figure 3.16B**). Therefore, a new strategy was required to reduce the background fluorescence.

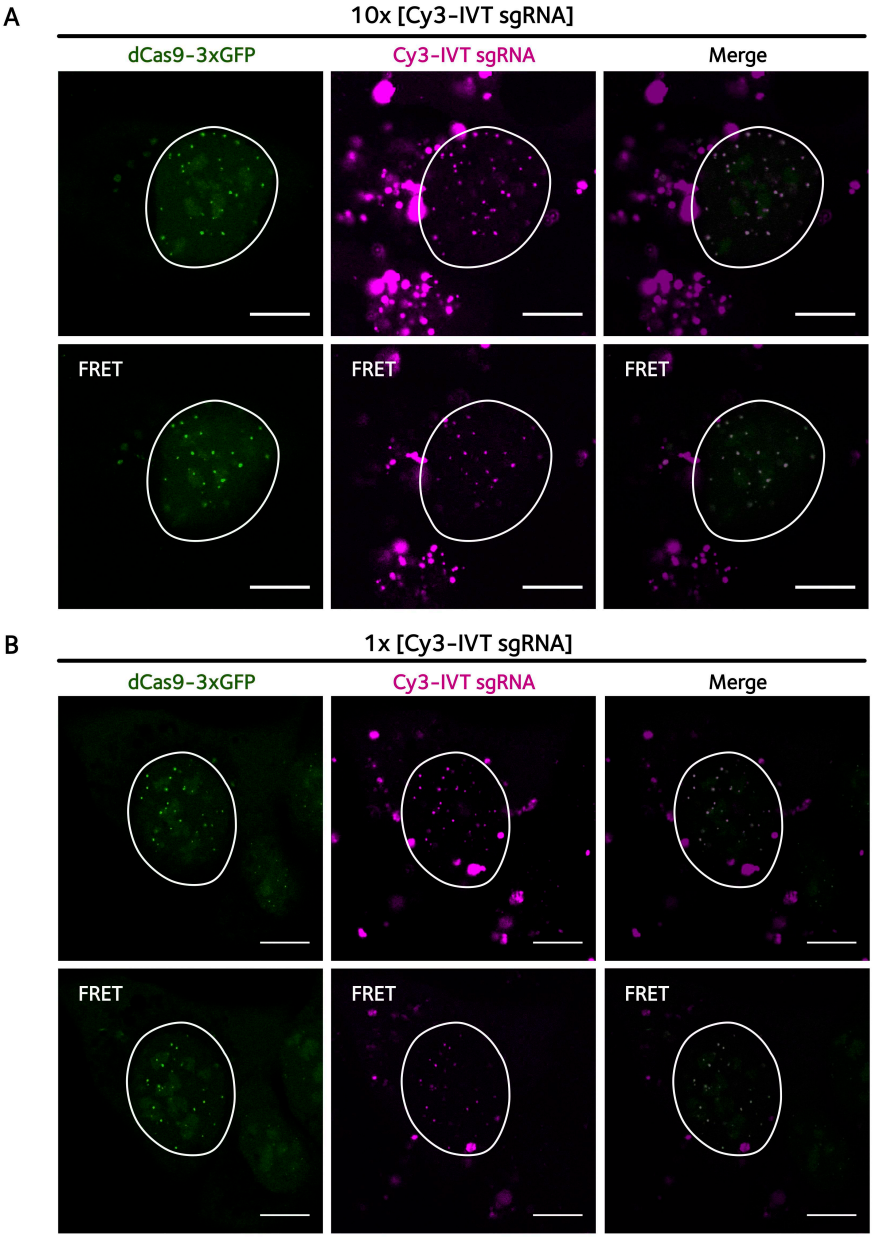


Figure 3.16 | Live-cell imaging using different concentrations of Cy3-labelled IVT sgRNA with and without FRET. (A) High and (B) lower concentrations of Cy3-IVT sgRNA were transfected into dCas9-3xGFP-expressing U2OS cells. Live-cell imaging was performed without FRET (top panels) and with FRET (bottom panels). Note that the position of fluorescent spots was different between non-FRET and FRET due to the telomeric dynamic motion in live cells and drift of the microscope objective. After 24 h, live-cell imaging was performed with different lasers used to visualise GFP (green) and Cy3 (pink). Scale bar = 10 µm. The sequence of Cy3-IVT sgRNA can be found in **Appendix Table 13**.

3.9 RNase H-activated fgRNA

3.9.1 Design Strategy

Before optimising the fsgRNA design any further (i.e. increasing the number of fluorophores or performing multi-colour multi-foci imaging), a big concern was the strong background fluorescence arising from endosomal trapping. It was felt this problem needed to be addressed. It had been previously demonstrated that sgRNA could be inactivated by the introduction of the spacer-blocking hairpin (SBH) with a cleavable RNA unit and subsequently activated upon cleavage by externally expressed CTS1 nuclease or internally expressed RNase H endonuclease[249]. A similar concept was therefore applied to develop a new fsgRNA design. For the initial study, the RNase H was chosen as an activator as it is predominantly located in the nucleus[250], thus it should enable a nuclear 'turn on' system. Moreover, the use of RNase H avoids the expression of exogenous proteins, which would require an additional plasmid transfection, thus complicating the study. In this design, the fluorescent guide RNA (fgRNA) was split into the crRNA–tracrRNA hybrid system with the key modifications being the addition of an external Cy3-linked 11-mer DNA extension on the 3' end of the crRNA and a quencher-linked 11-mer RNA extension on the 5' end of the tracrRNA (**Figure 3.17**). Upon hybridisation of the modified crRNA and tracrRNA, a fgRNA with a stable DNA–RNA heteroduplex should be generated which would be susceptible to nuclear RNase H cleavage, resulting in the release of the quencher thereby removing static quenching of the fluorophore. Thus, only the DNA–RNA heteroduplex of the fgRNA that is successfully delivered into the nucleus should be activated and become fluorescent whereas those outside the nucleus would remain inactive and invisible (dimmer).

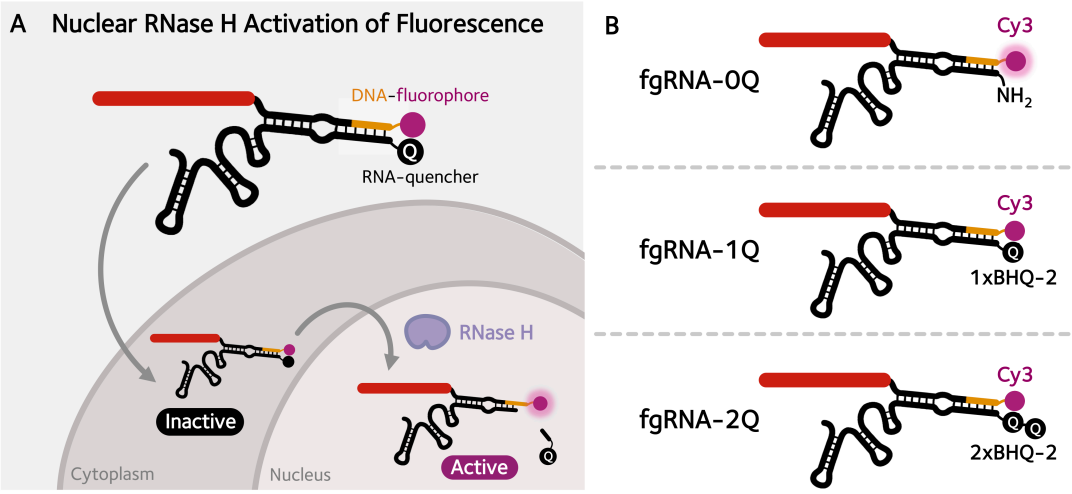


Figure 3.17 | RNase H-activated fgRNA design. (A) The RNase H-activated fgRNA contains a DNA sequence linked to a fluorophore on the crRNA and an RNA sequence linked to a quencher on the tracrRNA. Upon crRNA–tracrRNA hybridisation, the fluorophore is quenched and the fgRNA remains dim in the cytoplasm. Once it enters the nucleus where RNase H is present, the quencher is cleaved, enhancing fluorescent signal. (B) The fgRNA constructs used for this study consist of a control fgRNA with and without BHQ-2 quenchers (5'-C6-amino linker only = fgRNA-0Q; one BHQ-2 quencher = fgRNA-1Q; two BHQ-2 quenchers = fgRNA-2Q). The oligonucleotide sequences can be found in **Appendix Table 17**.

3.9.2 Live-cell Imaging

The crRNA appended with fluorophore-DNA and the tracrRNA containing zero (fgRNA-0Q), one (fgRNA-1Q) and two BHQ-2 quenchers (fgRNA-2Q, **Figure 3.17B**) were generated directly from the synthesizer and were characterised by mass spectrometry (**Appendix Table 17**). The fgRNA was produced *in vitro* by hybridising the crRNA and the tracrRNA in 1:1.25 ratio to ensure that all fluorophores were completely quenched. Then the fgRNA was transfected into cells using Lipofectamine reagents. As illustrated in **Figure 3.18A**, the wide-field images showed that the background fluorescence was substantially decreased when increasing the number of quenchers and using identical microscope excitation and emission settings. However, the background fluorescence outside the nucleus could not be eliminated completely, which could possibly be derived from the degradation of the fgRNA within the endosome/lysosome organelle, freeing the fluorophore from the quencher. In the nucleus, the Cy3 signal of telomeres became weaker when the number of quenchers increased (**Figure 3.18B**). Concurrently, the higher excitation power was required to observe the GFP nuclear fluorescent foci relative to the

IVT sgRNA, resulting in higher nuclear GFP background signal. This suggests the BHQ-2, which can also quench GFP as well as Cy3 emission, is not fully cleaved and released in the fgRNA constructs.

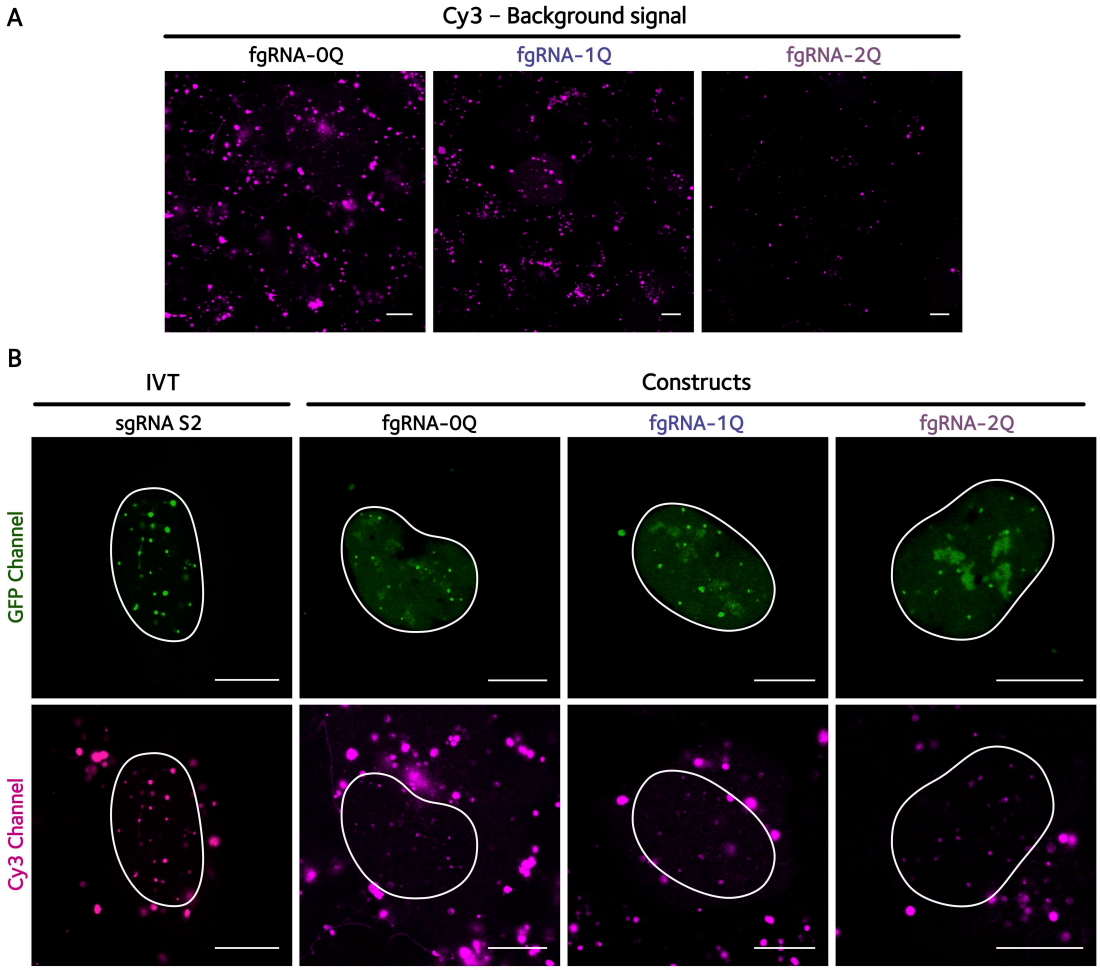


Figure 3.18 | Live-cell imaging using RNase H-activated fgRNA OQ / 1Q / 2Q. (A) shows that background fluorescence signal was significantly reduced when the BHQ-2 quencher was introduced (identical excitation and emission settings were used). (B) illustrates live-cell imaging by the fgRNA that contains no (0Q), one (1Q) or two BHQ-2 (2Q) quenchers. Transfection was performed in exactly the same way as the previous fsgRNA constructs using Lipofectamine reagents. After 24 h, live-cell imaging was performed with different lasers used to visualise GFP (green) and Cy3 (pink). Scale bar = 10 μm. The oligonucleotide sequences can be found in **Appendix Table 17**.

Weaker Cy3 signal might be caused by the unsuccessful RNase H-mediated release of the quencher within the nucleus. It could be that the presence of dCas9-3xGFP inhibited the cleavage of DNA–RNA heteroduplex by RNase H. To test this assumption, the effect of Cas9 on the cleavage of DNA–RNA heteroduplex by RNase H was examined *in vitro*.

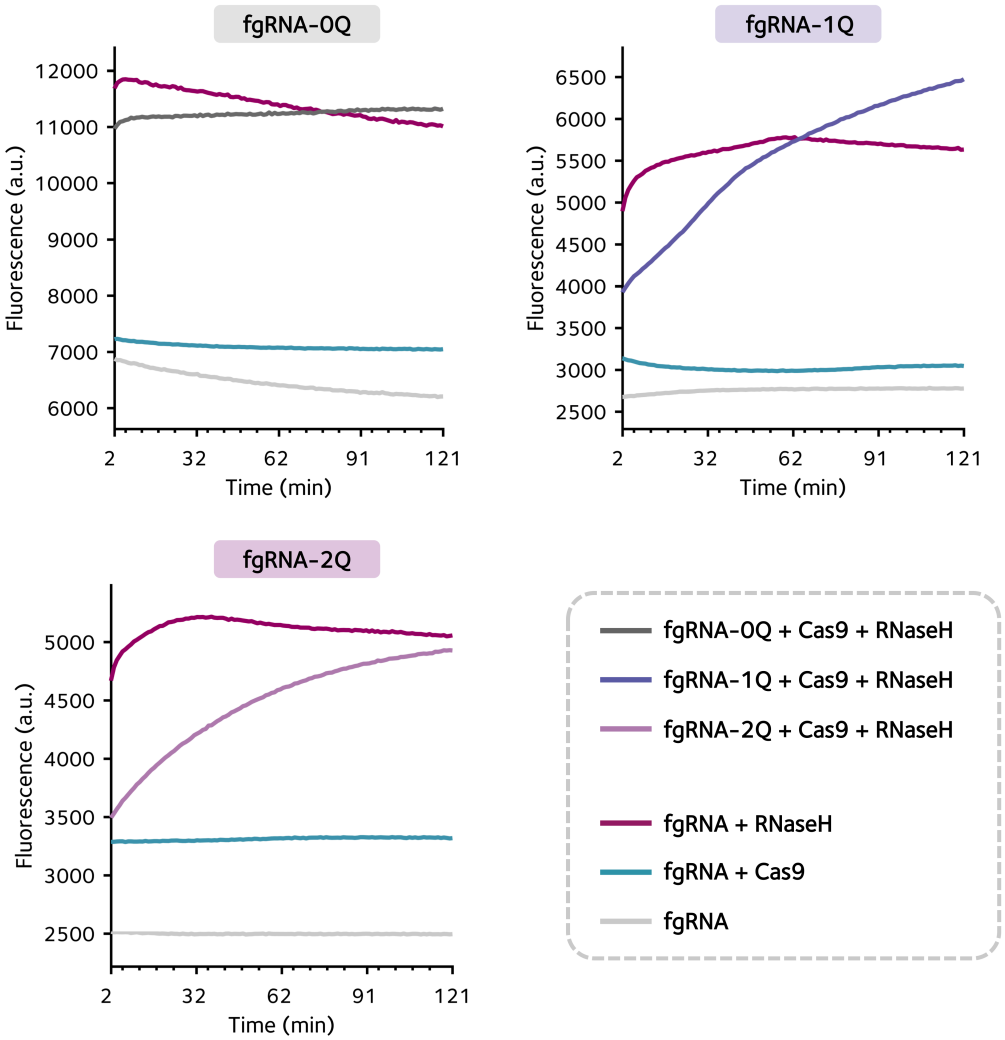


Figure 3.19 | Real-time fluorescence measurement of RNase H-activation of fgRNA *in vitro*. RNase H-activated fgRNAs (0Q, 1Q and 2Q) were placed in buffer containing 50 mM Tris-HCl, 7.5 mM KCl, 3 mM $MgCl_2$ and 10 mM DTT. Cas9 was then added and incubated before the addition of RNase H. RNase H cleavage of the RNA-DNA heteroduplex resulted in enhanced fluorescence that was monitored over time from t = 2 min. The oligonucleotides used can be found in **Appendix Table 17**.

Cas9 instead of dCas9-3xGFP was used due to the lower background fluorescence, which makes data interpretation simpler. Real-time fluorescence detection showed that without Cas9 the fluorescent signal was generated rapidly upon the addition of RNase H. However, when Cas9 was present, the fluorescent signal rose more slowly (**Figure 3.19**), indicating that Cas9 inhibited RNase H cleavage. Full cleavage was observed but at a far slower rate, which may translate into worse activity in cells where RNase H concentration is more difficult to determine. Thus, in future design, the DNA-RNA heteroduplex should be moved further away so that it is more accessible for RNase H.

3.10 Conclusions and Future Work

In summary, alternative CRISPR-based approaches for live-cell imaging have been explored with the focus on chemical modifications of sgRNA. It has been shown that stable dCas9–sgRNA or dCas9–sgRNA–target DNA complex formation is a key factor that contributes to successful live-cell imaging in contrast to the genome editing applications. Small-molecule fluorophores such as Cy3 could be used to specifically label genomic DNA regions with almost undetectable nuclear background. However, the strong granular cytoplasmic background due to the endosomally trapped fsgRNA is a major concern as it could give rise to false positive artefacts. Attempts have been made to reduce such background noise including the use of RNase H-activated fgRNA. Although the quencher in the design could significantly reduce the background fluorescence, spots corresponding to the unbound fgRNA within the endosomes were still observed and the desired nuclear Cy3 foci were dimmer. More optimisation of this design is required as the approach is only feasible when fluorescently tagged dCas9 is used for cross-validation. The click sgRNA strategy from **Chapter 2** was also explored, but live-cell imaging was not possible due to the destabilising effect caused by the triazole linkage, a factor that does not appear to be as important for gene editing. Finally, the design of a sgRNA was optimised to include a DNA loop by chemical ligation, which could be beneficial for other applications.

Future work should include the development of approaches that more efficiently reduce background signal outside the nucleus (e.g. using fluorophores in DNA-targeting region of the sgRNA whose fluorescence only activates upon DNA binding). Alternative delivery agents (e.g. nanoparticles or cell penetrating peptides) should also be explored that increase the entry of fsgRNA into the nucleus and/or avoid the endosomal trapping issue. Provided the problem of granular background signal is solved, the types and the number of fluorophores will need to be studied to improve the imaging of non-repetitive genomic regions and offer an alternative for multi-colour live-cell imaging. The click sgRNA approach is also worth considering (see **Chapter 2**) since it will provide a simple and efficient way of constructing a sgRNA library for imaging of non-repetitive DNA regions. In addition, the length of the linker for dye attachment at the 5' end of crRNA could be altered to increase the spacing between the dye and the dCas9 protein.

3'–3' Dual gRNAs using CRISPR-AsCas12a

4.1 Background

4.1.1 Controlling Gene Editing Outcomes

CRISPR-Cas9 allows efficient gene disruption, but precise gene insertions remain a challenge to introduce. This is partly due to the 'blunt' dsDNA breaks induced by Cas9 that favour DNA repair by error-prone NHEJ over more accurate HDR (see **Chapter 1**, **Figure 1.1**). NHEJ occurs throughout the cell cycle, while HDR occurs only during late S and G2 phase when DNA replication is complete and a homologous DNA template is available[5]. This results in a ~4:1 ratio of NHEJ:HDR in proliferating mammalian cells,[251] and genome editing efficiencies of 20–60% through NHEJ and 0.5–20% by HDR[27,28,145]. Several attempts have been made to enhance HDR: 1) cells can be chemically synchronised to remain in G2 phase by blocking the M phase prior to Cas9–gRNA and donor DNA template delivery[252]; 2) proteins such as Rad51, a ssDNA-binding protein that facilitates strand invasion in HDR, can be upregulated using the HDR enhancer[253]; 3) competing NHEJ pathways can be suppressed[253–255]. The last route is achieved by gene silencing,

small-molecule inhibition or proteolytic degradation of NHEJ key components such as Ku70/Ku80, proteins that bind and align the two broken ends to promote end joining, and/or DNA ligase IV, an enzyme that joins both ends[253–255]. These methods have increased HDR rates, however they disturb the natural DNA repair that occurs at other locations in the genome, potentially altering cell growth and even phenotype.

Therefore, new strategies that do not significantly perturb the native DNA damage response have been developed. It was observed that after cleavage, Cas9 does not release both strands of the DNA substrate simultaneously[256]; the 3' end of the non-target strand is released first. Designing ssDNA templates complementary to this strand can help increasing the local concentration of the donor template at DSB, thus increasing the HDR events by ~50 fold[256]. Similarly, the availability at DSB can be increased by covalently linking the ssDNA template to Cas9[257,258] or onto sgRNA[140,259]. In some cases, the HDR-repaired DNA can be re-edited as the PAM sequence may remain intact, which could ultimately result in a NHEJ-mediated product. To overcome this problem, the PAM sequence in the ssDNA template is mutated such that it is not recognised by Cas9 when integrated into the genome[260].

Finally, a pair of 'nickase' Cas9 variants (RuvC mutant D10A or HNH mutant H480A), proteins which cut only one of the two strands (see **Chapter 1**, **Section 1.6.3**), can be used to target two adjacent sites in order to generate 'staggered' cuts. These cuts can give either 5'- or 3'-overhangs, with 5'-overhangs generally favouring HDR repair[261]. Interestingly, the new CRISPR-Cas12a system naturally generates these 5'-overhangs using only a single PAM site (**Figure 4.1A**).

However, there remain limitations. The 5'-overhangs generated are complementary to each other, therefore the likelihood of ligation of the nicks to yield the original genomic sequence is high. To overcome this issue, the use of four 'nickase' Cas9 recognition sites or two Cas12a proteins to excise small duplexes, which can dissociate and thereby generate non-complementary staggered cuts, is possible. However, this approach is severely restricted by the tri-/tetra-nucleotide (PAM) recognition requirement of the systems.

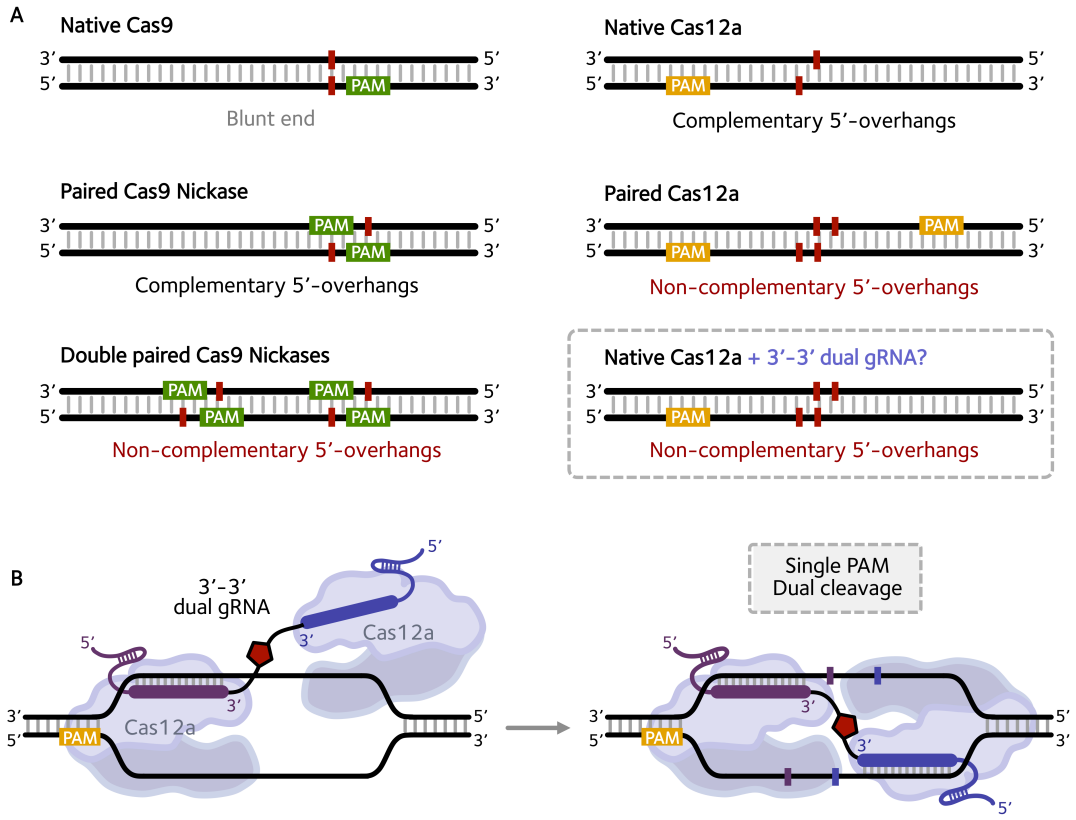


Figure 4.1 | Overview of CRISPR-Cas9/Cas12a cleavage patterns. (A) highlights the growing PAM motif dependence for generating non-complementary 5'-overhangs that favours DNA repair by HDR (B) illustrates the proposed 3'–3' backbone modified guide RNA (crRNA) for Cas12a that reduces PAM dependence and facilitates the removal of short dsDNA to generate non-complementary 'sticky' ends.

4.2 Hypothesis and Aims

Closer inspection of the CRISPR-Cas9 mechanism of action (see **Chapter 1**, **Section 1.4**) reveals the PAM site is required for initial duplex interrogation and strand invasion as opposed to DNA cutting; Cas9 can cleave ssDNA without a PAM site[26]. Given that strand invasion at one CRISPR-Cas9 or Cas12a site is likely to generate a local pocket of unwound DNA next to the protein, we hypothesise that a second CRISPR-Cas9/Cas12a–gRNA complex would not require a PAM site, if it could be colocalised with the first target site. In cells, this is likely to be a rare event with low efficiency. Yet, chemically, this could be made highly favourable by covalently linking the two guide RNAs (3'–3') required in CRISPR proteins (**Figure 4.1B**). In the crystal structure of crRNA–dsDNA complex of the CRISPR-Cas12a system (PDB: 5B43[111], **Chapter 1**, **Figure 1.15**), it was observed that only 20 bases of the 23-mer guiding sequence hybridise to target DNA; the 3' end

of the crRNA and the 5' end of the target DNA are both disordered. Therefore, it may be possible to modify these ends without interfering with Cas12a binding. The crystal structure of Cas12a–sgRNA, reported after this project had begun, suggests the seed-region is not pre-organised like Cas9[113]. However, we hypothesise that RNA–DNA hybridisation could still force the active conformation of Cas12a. A bi-functional guide RNA would only require a single PAM site, with the unnatural 3'–3' linkage facilitating base-pairing to the second cleavage site.

4.3 Dual gRNA Generation

The dual gRNA is composed of two individual crRNAs: a PAM-dependent crRNA that generates the first DNA cleavage product, and a PAM-independent crRNA that will hypothetically generate a second DNA cleavage product (**Figure 4.2A**). To optimise the spacing between the two components, the crRNAs were appended with an additional 2 to 4 nucleotides at the 3' end of the DNA-binding sequence, which already has three bases that are not involved in DNA-binding. Therefore, one set of 42, 44 and 46-mer crRNAs were synthesised with a 2'-ethoxyazido-dT at the 3' terminus (termed crRNA42-N_3, crRNA44-N_3 and crRNA46-N_3 respectively), and another set was prepared with a 3'-*O*-propargyl nucleotide at the 3' end (termed crRNA42-alk, crRNA44-alk and crRNA46-alk). Following RP-HPLC purification, their purities were confirmed using mass spectrometry.

Next, dual gRNAs 42, 44 and 46 were generated *via* non-templated CuAAC (**Figure 4.2B**) as described in **Chapter 2**. The click dual gRNAs were purified and successfully characterised by mass spectrometry (**Figure 4.2C**). Next, they were used for the DNA cleavage assay.

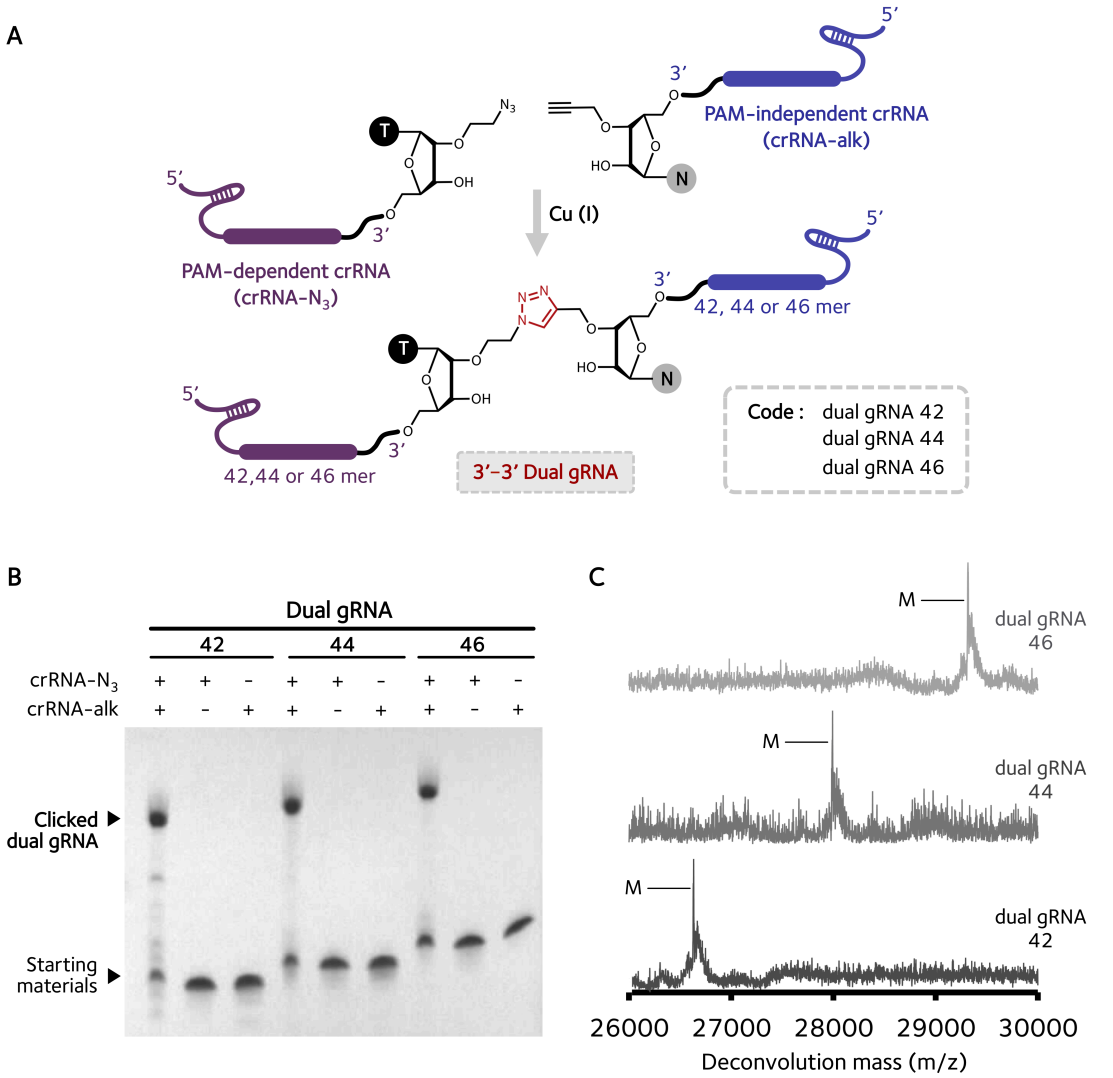


Figure 4.2 | Generation of dual gRNAs *via* CuAAC ligation. (A) Schematic representation of dual gRNA production. 42, 44 or 46 mer of the PAM-dependent crRNA was synthesised with 2'-ethoxyazido-dT at the 3' terminus (termed crRNA42-N_3, crRNA44-N_3 and crRNA46-N_3 respectively). 42, 44 or 46 mer of the PAM-independent crRNA was generated with a 3'-propargyl ribonucleotide at the 3' terminus (N = rA, rG, rC or rU; termed crRNA42-alk, crRNA44-alk and crRNA46-alk respectively). (B) With the optimised non-templated CuAAC reaction from **Chapter 2**, 3'–3' dual gRNA 42, 44 and 46 were obtained in good yields as shown by a 10% denaturing PAGE gel. (C) Deconvolution mass of all different dual gRNA constructs. M (dual gRNA 42) = 26631 (expected); 26632 (observed). M (dual gRNA 44) = 27991 (expected); 27992 (observed). M (dual gRNA 46) = 29316 (expected); 29311 (observed). The oligonucleotide sequences can be found in **Appendix Table 18** and **19**.

4.4 AsCas12a Lysate

4.4.1 Lysate Preparation

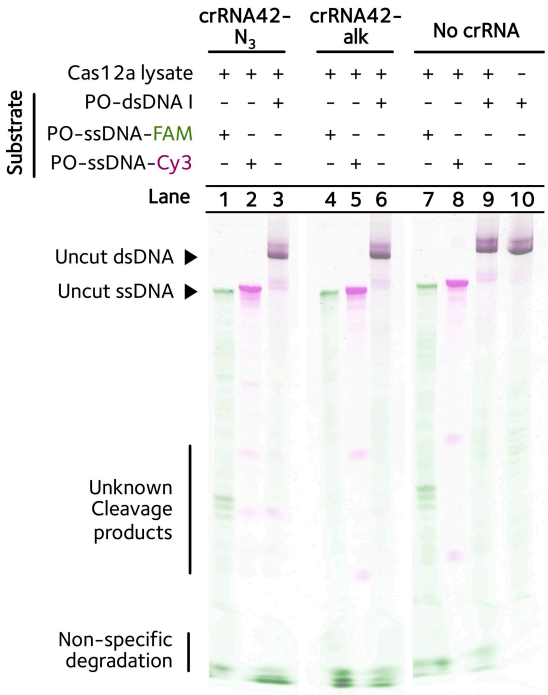
Since AsCas12a was not commercially available at the start of this project, to quickly establish the viability of the hypothesis in the aims, HEK293T cells were transfected with a vector that expresses AsCas12a and the protein was extracted as a crude cell lysate. A GFP-expressing plasmid was co-transfected with the Cas12a-encoding plasmid (1:1) to confirm successful transfection and protein expression. The cells were then harvested 24 h post-transfection and sheared *via* sonication to disrupt the cell membrane and allow the release of the protein. The resulting Cas12a-containing lysate was used without further purification for *in vitro* testing as reported in literature[60].

4.4.2 Lysate Activity Tests

To test activity of the lysate, the PAM-dependent crRNA (crRNA42-N_3, positive control) was designed to cleave a 90-bp dsDNA (PO-dsDNA I) that was labelled with FAM at the 5' end of the non-target strand (PO-ssDNA-FAM) and Cy3 at the 5' end of the target strand (PO-ssDNA-Cy3). Upon Cas12a-crRNA complex assembly, the dsDNA should be digested asymmetrically yielding 37-mer FAM and 48-mer Cy3-labelled fragments. This should enable simple visualisation of activity of the lysate on denaturing PAGE.

Firstly, the dsDNA was not fully denatured even in the presence of 7 M urea causing it to migrate more slowly than the ssDNA (**Figure 4.3**). Without crRNA but in the presence of lysate, some non-specific degradation was observed for dsDNA and ssDNA as shown by the bands at the very bottom of the gel (Lane 7–9, **Figure 4.3**). This suggests the presence of the nucleases in the cell lysate. After the addition of the crRNA42-N_3 (PAM-dependent, positive control) to the DNA-lysate mixture, only non-specific degradation was seen and the majority of the dsDNA remained uncleaved (Lane 1–3 compared to Lane 7–9, **Figure 4.3**). For the FAM-labelled ssDNA, although a cleaved product was observed in the presence of the crRNA, it was also present in the lysate-only lane suggesting it is a non-specific cleavage product. The crRNA42-alk (PAM-independent, negative control) also gave the same pattern as the crRNA42-N_3 (Lane 4-6, **Figure 4.3**).

Collectively, these observations suggested that none or a very small amount of Cas12a was present in the lysate which mainly contained contaminating nucleases that caused non-specific DNA degradation.



crRNA	Substrate	Expected product
42-N_3	PO-dsDNA I	37-nt FAM/48-nt Cy3
	PO-ssDNA-FAM	Full length
	PO-ssDNA-Cy3	48-nt Cy3
42-alk	PO-dsDNA I	Full length
	PO-ssDNA-FAM	42-nt FAM
	PO-ssDNA-Cy3	Full length
N/A	All substrates	Full length

Figure 4.3 | *In vitro* DNA cleavage using HEK293T lysate containing AsCas12a. The reaction contained 5 µL Cas12a lysate (note that the amount of Cas12a could not be determined), 1 µM crRNA42-N_3 or crRNA42-alk, 1 µM PO-dsDNA I produced from hybridisation of PO-ssDNA-FAM and PO-ssDNA-Cy3, in a reaction buffer composed of 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/mL BSA and 5 mM DTT. The reaction proceeded for 3.5 h at 37 °C and was loaded onto a 12.5% denaturing PAGE gel. The expected outcomes of the reactions are listed in the table to the right. The oligonucleotide sequences can be found in **Appendix Table 18** and **20**.

To optimise this assay, the DNA templates were first re-synthesised with 4–5 phosphorothioates incorporated at both ends of the ssDNA to minimise non-specific degradation (PS-DNA I). In addition, since transfection seemed successful, new Cas12a lysate was prepared with a varying number of sonication cycles; increasing sonication cycles should disrupt the cells more and release more Cas12a into the lysate. When testing the lysate, if the number of sonication cycles was too high (10 cycles), even PS-containing dsDNA could not resist the non-specific degradation (Lane 6–7, **Figure 4.4**). This indicates that the greater the number of sonication cycles, the greater the amount of non-specific nucleases released. However, it was found that a 37-mer FAM-labelled DNA fragment band was generated in the presence of the positive control

PAM-dependent crRNA42-N_3 (Lane 1–5, **Figure 4.4**), indicating the presence of Cas12a. Increasing the number of sonication cycles increased the amount of the 37-mer FAM-labelled DNA produced but no further improvement was observed after 3 cycles. Surprisingly, the other expected 48-mer Cy3-labelled DNA fragment could not be observed for unknown reasons. It is possible that the 48-mer Cy3-labelled DNA was further digested non-specifically by contaminating nucleases in the lysate. Thus, the protein was purified to improve the DNA cleavage assay.

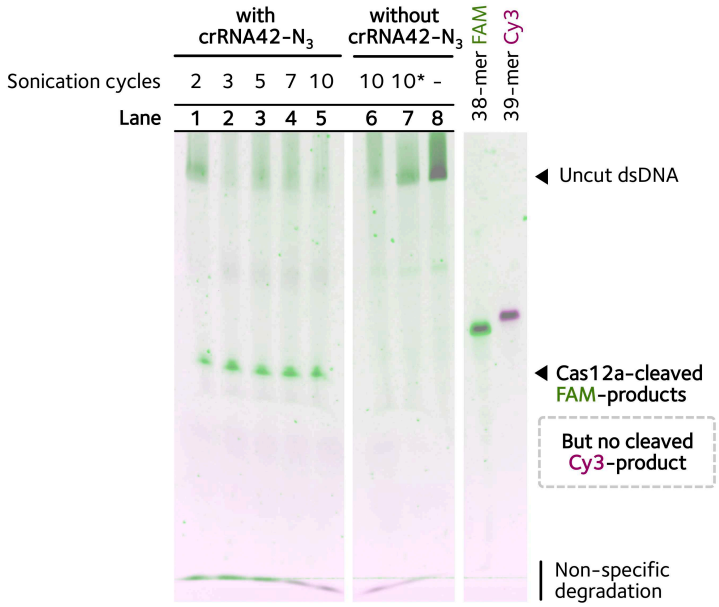


Figure 4.4 | *In vitro* DNA cleavage using HEK293T lysate containing AsCas12a and PS-containing dsDNA. The reaction contained 30 µL Cas12a lysate with 2, 3, 5, 7 or 10 sonication cycles (note that the amount of Cas12a could not be determined) or the Cas12a-free lysate obtained from 10 sonication cycles (10*), 3 µM crRNA42-N_3, 0.3 µM PS-dsDNA I and a buffer containing 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/mL BSA and 5 mM DTT. The reaction proceeded for 20 min (relatively short reaction time was used here to minimise non-specific degradation) at 37 °C and was loaded onto a 12.5% denaturing PAGE gel. The 38-mer FAM-labelled ssDNA and the 39-mer Cy3-labelled ssDNA were also loaded as control size markers. The oligonucleotide sequences can be found in **Appendix Table 18** and **20**.

4.5 Bacterial Production of AsCas12a

4.5.1 Single-step Purification

The vector for bacterial expression of Cas12a was available from Addgene and used directly after preparing the plasmid on large scale. The vector was transfected into Rosetta cells and the

protein was expressed as fusion with a six-histidine tag to the N-terminus to facilitate purification by IMAC. After sonication to release the protein, Nickel-bound resin was used to selectively capture to the His-tagged Cas12a, which was subsequently eluted as a pure protein using imidazole. SDS-PAGE showed that a large quantity of Cas12a was produced (**Figure 4.5**) with some impurities, as shown as a series of bands below the main Cas12a band, which were co-purified. The eluted protein was applied to an ultrafiltration centrifugal device containing a polyethersulfone (PES) membrane in order to remove imidazole, buffer exchange and concentrate the protein. With a molecular-weight cut-off of 100 kDa, the device should be able to remove any impurities that have a smaller molecular weight than Cas12a (156 kDa). Unexpectedly, some impurities were not removed even after five successive centrifugation steps with slight protein loss during the process, suggesting the impurities may aggregate in solution and become larger than the cut-off limit. However, Cas12a (fraction E4) was used for DNA cleavage as this level of purity might be sufficient.

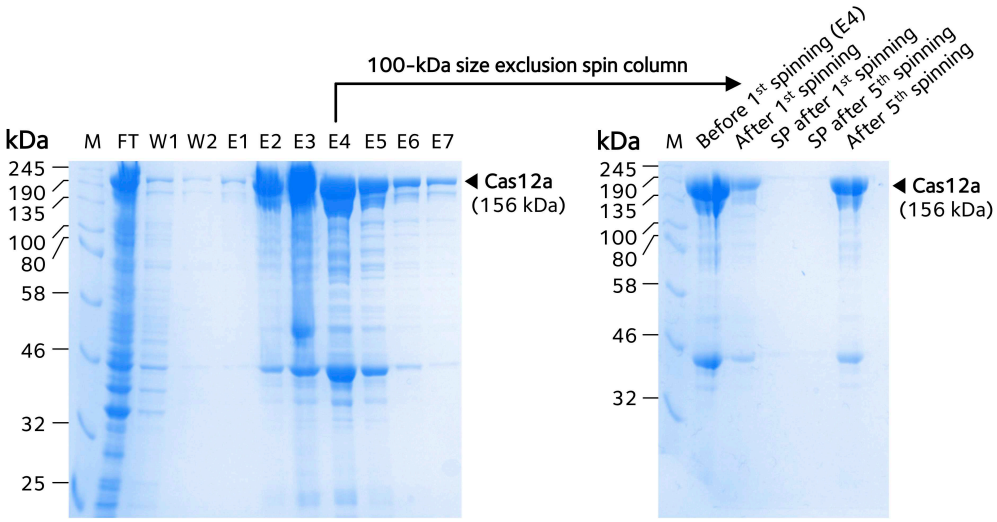


Figure 4.5 | Single-step purification of His-tagged AsCas12a *via* nickel-IMAC. Coomassie Blue-stained 10% SDS-PAGE gel (left) showed different fractions obtained throughout the purification of Cas12a. M = Colour Prestained Protein Ladder, broad range (11–245 kDa). Marker band sizes are indicated in kDa. FT = Flow-through fraction (the Cas12a and contaminants that did not bind to resin). W1–W2 = Wash fractions (Cas12a and contaminants that weakly or non-specifically bound to resin). E1–E7 = Elution fractions (Cas12a that specifically bound to the resin was eluted and collected in a series of fractions). E4 was subject to further de-salting and buffer exchange using a 100-kDa size exclusion spin column. Coomassie Blue-stained 10% SDS-PAGE gel (right) showed different fractions from the spin column. After the 1st and the final 5th centrifugation step, the concentrated protein fractions and the salt-containing supernatant (SP)/eluent were collected for loading onto SDS-PAGE.

4.5.2 Single-step Purification Protein Activity Tests

Without the crRNA42-N_3 (PAM-dependent, positive control), less non-specific degradation was observed in the presence of the purified Cas12a (Lane 10, **Figure 4.6**), compared to the Cas12a lysate (Lane 7, **Figure 4.6**). When the crRNA42-N_3 was added, the amount of DNA cleavage products generated was higher for purified Cas12a (Lane 9, **Figure 4.6**) than Cas12a lysate (Lane 6, **Figure 4.6**), indicating the presence of the protein and the necessity of purification. However, digestion products contained a series of different DNA fragments, suggesting that the Cas12a-cleaved products were further digested possibly non-specifically by the remaining contaminants. Thus, Cas12a was extensively purified to obtain a very pure protein for the DNA cleavage assay.

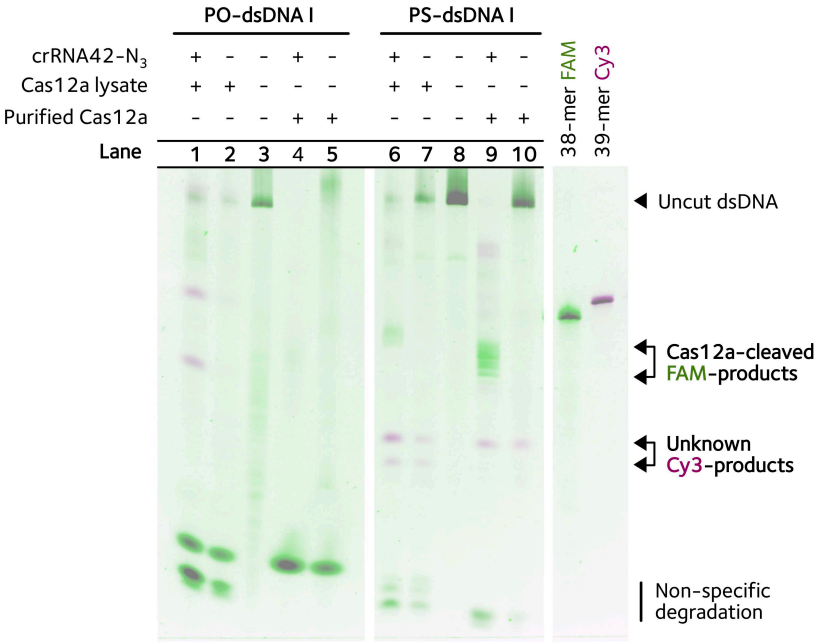


Figure 4.6 | *In vitro* DNA cleavage using one-step-purified AsCas12a. For Cas12a lysate, the reaction contained 5 µL Cas12a lysate (note that the amount of Cas12a could not be determined), 15 µM crRNA42-N_3, 0.75 µM PO-dsDNA I or PS-dsDNA I and a buffer containing 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/mL BSA and 5 mM DTT. For one-step-purified Cas12a, the reaction contained 15 µM Cas12a, 30 µM crRNA42-N_3, 1.5 µM PO-dsDNA I or PS-dsDNA I and a buffer containing 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, and 5 mM DTT. The reaction proceeded for 20 min at 37 °C and was loaded onto a 12.5 % denaturing PAGE gel. A 38-mer FAM-labelled ssDNA and the 39-mer Cy3-labelled ssDNA were also loaded as size markers. The oligonucleotide sequences can be found in **Appendix Table 18** and **20**.

4.5.3 Multi-step Purification

To improve His-tag purification, cobalt-charged resin was used, which has a lower binding capacity but greater specificity than nickel-charged resin for his-tagged proteins. It forms weaker non-specific interactions with contaminating proteins, therefore giving purer desired protein. Yet, almost all of protein was eluted during the first wash with high salt (1 M NaCl), suggesting interactions with cobalt were too weak (**Figure 4.7A**). Therefore, dialysis of the Cas12a-containing wash fraction was performed to reduce the salt. Since a significant amount of impurities were co-eluted with the Cas12a, the protein was then subject to further purification.

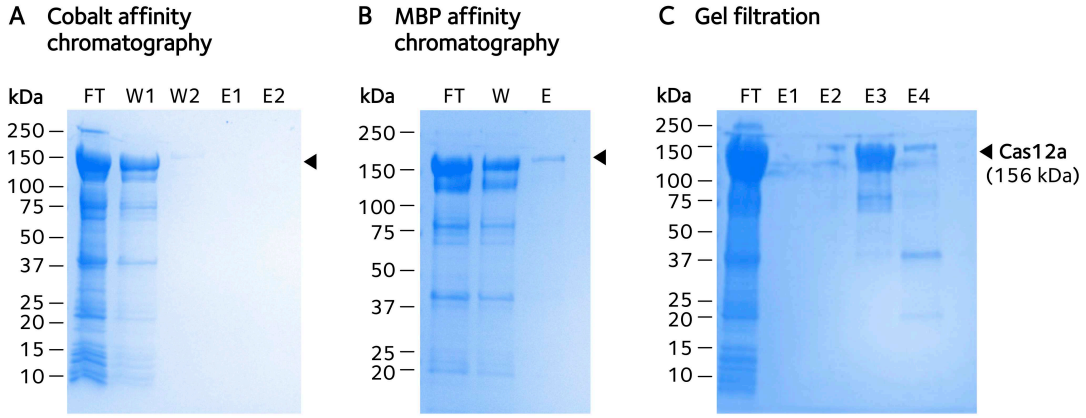


Figure 4.7 | Three-step purification of AsCas12a. (A) Cobalt affinity chromatography was used first to purify Cas12a. Coomassie Blue-stained 4–20% gradient SDS-PAGE gel showed different fractions obtained during the course of purification. Cas12a (156 kDa) appears as a thick band denoted by the black arrow at the top of the gel. FT = Flow-through fraction (the Cas12a and contaminants that did not bind to resin). W = Wash fractions (Cas12a and contaminants that weakly or non-specifically bound to resin). E = Elution fractions (Cas12a that specifically bound to the resin was eluted and collected in a series of fractions). Note that the number after W or E indicated the number of times this step was repeated. The wash fraction was subject to a further round of purification. (B) MBP affinity chromatography was the next purification approach. Coomassie Blue-stained 4–20% gradient SDS-PAGE gel indicated different fractions obtained during the course of purification. In this case, FT and W were combined and underwent further purification. (C) Gel filtration was the last method used for purification. Coomassie Blue-stained 4–20% gradient SDS-PAGE gel showed different fractions obtained during the course of purification. E3 was the purest fraction, thus it was concentrated and stored in 50% glycerol for *in vitro* DNA cleavage assays.

In addition to a His-tag, the recombinant Cas12a also contains a MBP fusion tag, between the six-histidine tag and the N-terminus of the protein, which was included to increase the solubility of Cas12a. Thus, the MBP could be exploited for an alternative purification approach. To this end, the Cas12a-containing wash fraction was passed through a column packed with Dextrin

Sepharose® resin that selectively binds to MBP, and pure Cas12a should be eluted using maltose. However, SDS-PAGE showed that the MBP-fused Cas12a did not bind efficiently to the resin as it was mainly present in the flow-through and wash fractions. This could be due to accessibility of the MPB-tag being hindered by its position between the His-tag and the Cas12a protein. Since Cas12a in both flow-through and wash fractions was reasonably pure (**Figure 4.7B**), they were combined and subject to further purification.

Cas12a was further purified by gel filtration in which the protein was separated from contaminants based on size. SDS-PAGE illustrated that the majority of contaminants as several bands below the main Cas12a band were eliminated (**Figure 4.7C**), giving a pure Cas12a which could be used for *in vitro* DNA cleavage assay.

4.5.4 Multi-step Purification Protein Activity Tests

The activity of Cas12a derived from multi-step purification was compared with that obtained previously from single-step purification. Surprisingly, when the positive control crRNA42-N_3 was present, only the single-step-purified Cas12a produced a significant amount of ~37-mer FAM-labelled DNA digested products (Lane 1, **Figure 4.8**) whereas the three-step-purified Cas12a mediated some DNA cleavage, but was barely detectable (Lane 3, **Figure 4.8**). This suggests that the protein had lost its activity during the extensive purification.

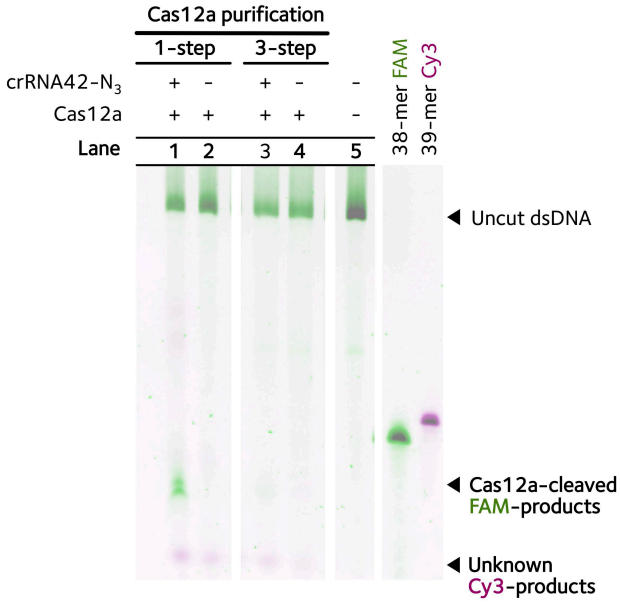


Figure 4.8 | Comparison of *in vitro* DNA cleavage by single- and multi-step-purified Cas12a. The reaction contained 3 µM Cas12a, 3 µM crRNA42-N_3, 1.5 µM PS-dsDNA I and a buffer containing 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mg/mL BSA and 5 mM DTT. The reaction proceeded for 1 h at 37 °C and was loaded onto a 10% denaturing PAGE gel. 38-mer FAM-labelled ssDNA and the 39-mer Cy3-labelled ssDNA were also loaded as size markers. The oligonucleotides used can be found in **Appendix Table 18** and **20**.

4.6 *In vitro* DNA Cleavage

4.6.1 Dual gRNA Activity

During this work, Cas12a became commercially available and was purchased from Integrated DNA Technologies Inc. and used for the remainder of the study. All dual gRNA constructs and controls were tested in the DNA cleavage assay with the expected outcomes listed in **Table 4.1**. The new dsDNA template was generated with ATTO 647N attached as the ATTO dye is more spectrally separated from FAM than Cy3 and brighter, enabling easier interpretation of the data.

gRNA	Substrate	Expected product
crRNA42-N_3	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	37-nt FAM/48-nt ATTO Full length 48-nt ATTO
crRNA42-alk	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	42-nt FAM/43-nt ATTO 42-nt FAM Full length
Dual gRNA 42-42	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	37-nt FAM/43-nt ATTO Full length 48-nt ATTO
crRNA44-N_3	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	37-nt FAM/48-nt ATTO Full length 48-nt ATTO
crRNA44-alk	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	46-nt FAM/39-nt ATTO 46-nt FAM Full length
Dual gRNA 44-44	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	37-nt FAM/39-nt ATTO Full length 48-nt ATTO
crRNA46-N_3	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	37-nt FAM/48-nt ATTO Full length 48-nt ATTO
crRNA46-alk	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	50-nt FAM/35-nt ATTO 50-nt FAM Full length
Dual gRNA 46-46	PO-dsDNA II PO-ssDNA-FAM PO-ssDNA-ATTO	37-nt FAM/35-nt ATTO Full length 48-nt ATTO
N/A	All substrates	Full length

Table 4.1 | List of expected cleavage products generated from commercial AsCas12a with different dual gRNAs, crRNA-N_3 and crRNA-alk. The template PO-dsDNA II was obtained from hybridisation of 5'-FAM-labelled ssDNA (PO-ssDNA-FAM) and 5'-ATTO 647N-labelled ssDNA (PO-ssDNA-ATTO). The oligonucleotide sequences can be found in **Appendix Table 18**, **19** and **20**.

With PO-dsDNA II as a substrate, there was only one minor non-specific DNA digestion product generated in the absence of the crRNA (Lane 10, **Figure 4.9A**). In the presence of the azido-crRNA (PAM-dependent, positive control) of various lengths (42 mer, 44 mer and 46 mer), the 48-mer ATTO-labelled DNA fragment was produced as anticipated (Lane 2, 5 and 8, ATTO channel, **Figure 4.9A**). However, there were three distinct FAM-labelled DNA fragments generated with lengths approximately between 35–38 mer (Lane 2, 5 and 8, FAM channel, **Figure 4.9A**). The 37-mer FAM-labelled DNA was the expected product whereas the other bands were not. These

unexpected DNA digestion products were consistent among different length of the azido-crRNA used, which implies the protein cleavage site-specificity is not perfect.

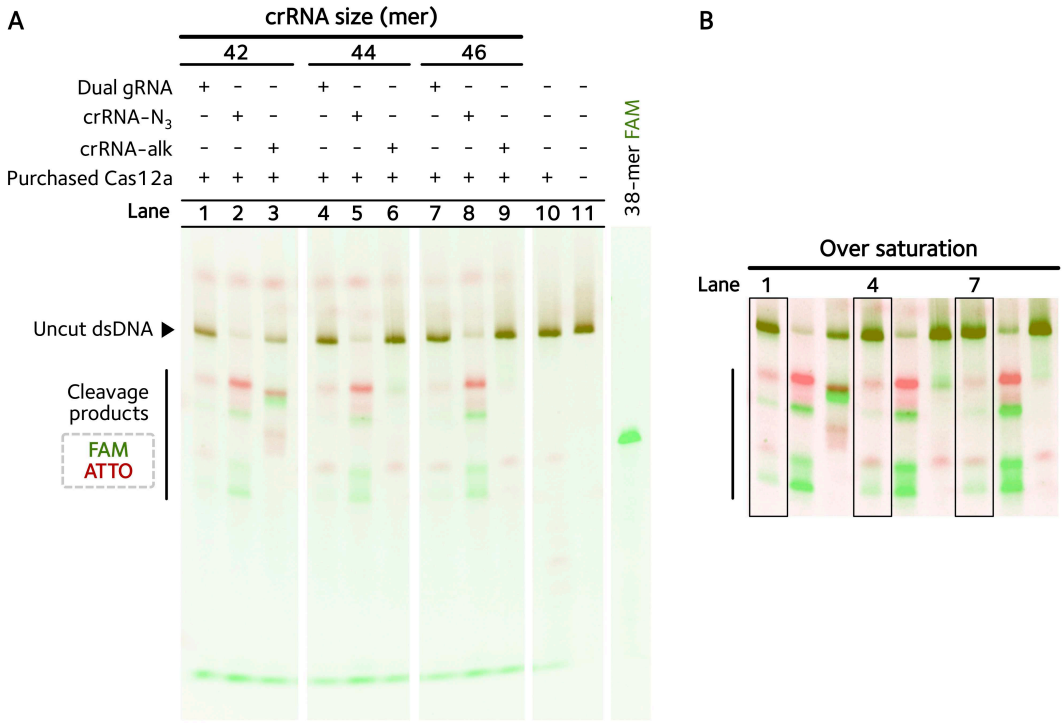


Figure 4.9 | *In vitro* cleavage of dsDNA using commercial AsCas12a. The reaction consisted of 1 µM commercial Cas12a, 2 µM dual gRNA, crRNA-N_3 or crRNA-alk (42, 44 or 46-mers as listed by the lane headings), 0.5 µM PO-dsDNA II and a buffer containing 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/mL BSA and 5 mM DTT. The reaction proceeded for 1 h at 37 °C and was loaded onto a 12.5 % denaturing PAGE gel. A 38-mer FAM-labelled ssDNA was also loaded as a size marker. The gel was imaged using normal saturation (A) and over saturation (B) – to observe the weakly fluorescent bands. The oligonucleotide sequences can be found in **Appendix Table 18**, **19** and **20** and the expected outcomes of the various cleavage reactions are listed in **Table 4.1**.

For the alkyne-crRNA (PAM-independent, negative control), DNA cleavage surprisingly occurred for crRNA42-alk, giving ~47-mer ATTO-labelled DNA fragment (Lane 3, ATTO channel, **Figure 4.9A**), with far weaker non-specific cleavage for crRNA44-alk and crRNA46-alk (Lane 6 and 9 cf. Cas12a-only Lane 10, ATTO channel, **Figure 4.9A**). Closer inspection of crRNA42-alk shows the sequence upstream is 5'-TGTG-3'. This is very similar to the AsCas12a PAM sequence 5'-TTTV-3' and hypothesised to be the reason for these off-target cleavage products.

When the dual gRNA was introduced, DNA cleavage products, which were identical to the use of the positive control crRNA, were observed but with far lower intensity (Lane 1, 4 and 7, FAM/ATTO channel, **Figure 4.9B**). Moreover, the hypothesised second PAM-independent cleavage event, which should give smaller ATTO fragments with dual gRNA 42 to 46 (**Table 4.1**), was not detected. This suggests that the presence of the second crRNA in the dual gRNA inhibited the first DNA cleavage by the PAM-dependent crRNA, possibly by sterically blocking Cas12a binding to the crRNA. This possibly due to: 1) the dual gRNA inhibiting Cas12a binding, 2) the dual gRNA altering the folding of the RNA hairpin recognised by Cas12a, 3) steric clashing of the Cas12a proteins that bind to the two gRNAs. More work is necessary to explore these possible reasons and their solution, with the simplest starting point being to drastically increase the spacing between the two crRNAs using multiple hexaglycol linkers during solid-phase oligonucleotide synthesis.

4.6.2 AsCas12a Non-specific DNase Activity

Another unexpected feature of **Figure 4.9** is that non-specific cleavage was still observed with purified protein. Therefore, the DNase activity of Cas12a on ssDNA was evaluated. Surprisingly, both ssDNAs were degraded non-specifically when mixed with Cas12a alone (Lane 4 and 9, **Figure 4.10**), demonstrating the ssDNase activity of the protein. In the presence of the PAM-dependent crRNA-N_3 (Lane 1-3, **Figure 4.10**) or PAM-independent crRNA-alk (Lane 6-8, **Figure 4.10**) and their respective complementary ssDNA, non-specific ssDNA degradation was observed. For PAM-independent crRNA-alk, the ssDNA is FAM labelled with the small fragments migrating at the bottom of the gel. For PAM-dependent crRNA-N_3, the ssDNA is ATTO-647N labelled. The fragments from cleavage release the dye that is positively charged attached to presumably mono- or di-nucleotides. This results in these small fragments migrating more slowly as a diffuse band above the expected position of the PO-ssDNA-ATTO strand.

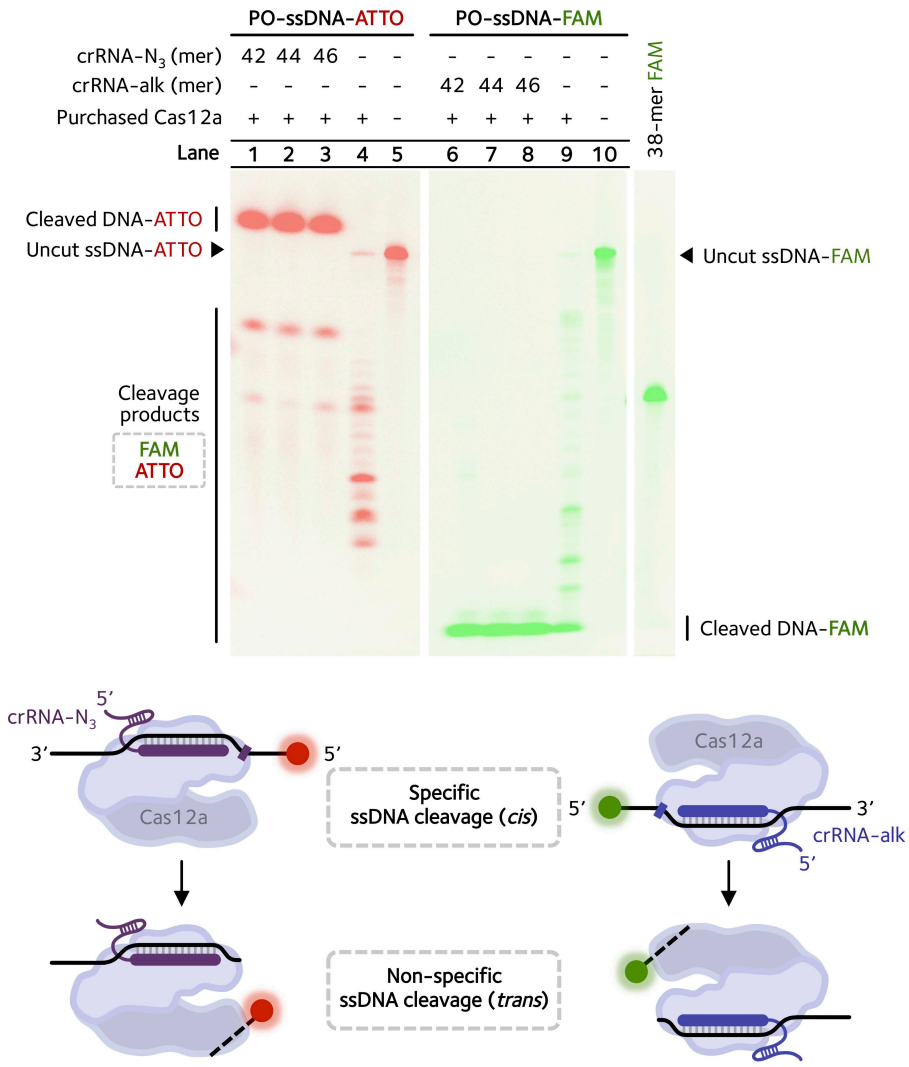


Figure 4.10 | *In vitro* cleavage of ssDNA using commercial AsCas12a. With PO-ssDNA-ATTO as a cleavage template, the reaction contained 1 µM purchased Cas12a, 2 µM crRNA (42, 44 or 46)-N_3, 0.5 µM PO-ssDNA-ATTO and a buffer containing 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 µg/mL BSA and 5 mM DTT. With PO-ssDNA-FAM as a cleavage template, the reaction contained 1 µM purchased Cas12a, 2 µM crRNA (42, 44 or 46)-alk, 0.5 µM PO-ssDNA-FAM and the same buffer composition as described above. The reaction proceeded for 1 h at 37 °C and was loaded onto a 12.5% denaturing PAGE gel. The 38-mer FAM-labelled ssDNA was also loaded as a size marker. The gel was imaged using normal saturation. The oligonucleotide sequences can be found in **Appendix Table 18** and **20**.

These were very surprising results and during the course of this work, the crRNA-dependent ssDNase activity was confirmed by literature published in Cell Research and Science[262,263]. In these reports "trans-cleavage" (i.e. cleavage of non-hybridised DNA) activity was observed in addition to usual "cis-cleavage" (i.e. specific on-target cleavage), when bound to short dsDNA or ssDNA.

The non-specific ssDNase activity in the absence of a crRNA, however, was not reported in these publications.

4.7 Conclusions and Future Work

The original aim of this work was to convert Cas12a from an enzyme that generates complementary staggered ends into one that creates non-complementary staggered ends in-order to alter/enhance genome editing. To this end, Cas12a was directed to cleave DNA using a dual gRNA generated by linking a PAM-dependent crRNA to a PAM-independent crRNA *via* CuAAC ligation. Once the PAM-dependent crRNA directs Cas12a to cleave target dsDNA, it was hypothesised that the other PAM-independent crRNA could recruit another Cas12a to cleave DNA at the adjacent site, destroying complementarity of the staggered ends. However, it was found that dual gRNA design inhibited DNA cleavage, possibly due to misfolding of the RNA secondary structure due to the close proximity of the second crRNA. This may prevent Cas12a from binding to the crRNA within the dual gRNA, inhibiting even on-target PAM-dependent cleavage.

In addition, ssDNA nuclease activity was discovered unexpectedly for AsCas12a, which was similar to the results from two recent publications[262,263] that showed the similar activity for other Cas12a variants (e.g. *Francisella novicida* and *Lachnospiraceae bacterium ND2006* derived Cas12a). The notable difference is that AsCas12a caused non-specific ssDNA digestion in the absence of a crRNA.

Nanopore STR Analysis
Towards
CRISPR Applications

5.1 Background

5.1.1 Short Tandem Repeats

STRs contain a repeating motif of two to seven nucleotides and the number of repeats is different among different individuals[264]. These repeats are globally distributed throughout the non-coding regions of the human genome. In forensic, genotyping the length and sequence variations is crucial. For example, the United States National DNA database known as Combined DNA Index System (CODIS) relies on 20 (previously 13, ref.[265]) core tetra-nucleotide repeat STR loci for human identification[266] (**Figure 5.1**).

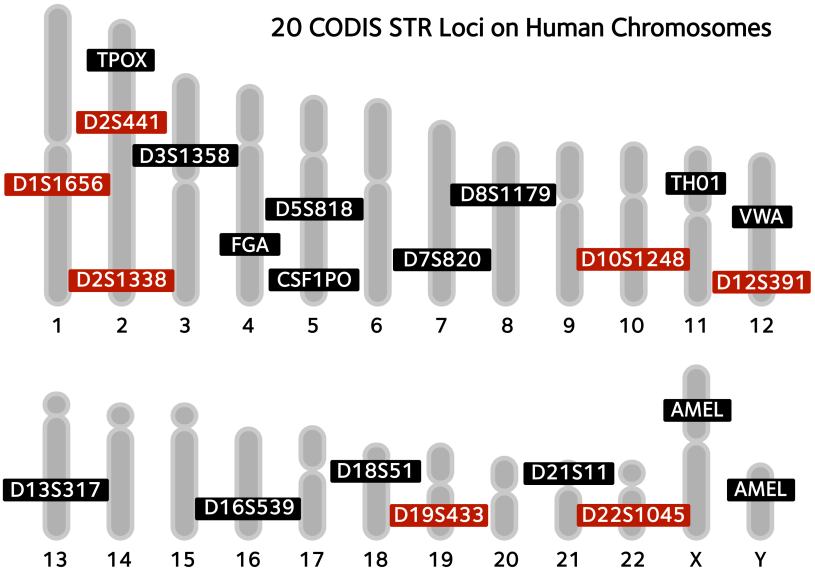


Figure 5.1 | The Combined DNA Index System (CODIS) core STR loci on human chromosomes. The original 13 CODIS STR core loci are labelled in black whereas the newly added loci (since 1st January 2017) are labelled in red[266]. Some STR loci are named by the gene name (e.g. TH01 within intron 1 of the Tyrosine Hydroxylase gene on chromosome 11, ref.[267], while others that lie outside a gene are assigned differently (e.g. D3S1358 where D = DNA, 3 = chromosome 3, S = single copy sequence and 1358 = 1358th locus on chromosome 3, ref.[267].

5.1.2 Current Analysis Methods

Due to the low sample quantity, STR analysis routinely relies on PCR amplification of the STR regions and there are commercially available primers for multiplexing this step. These amplicons can then be evaluated by either fragment length analysis using Capillary Electrophoresis (CE)[268] or Illumina sequencing[269,270]. CE allows the number of repeats within each STR to be determined. However, Illumina sequencing also provides information on the STR sequence enabling alleles with identical length but different sequences to be resolved, and Single Nucleotide Polymorphism (SNP) within or around STR to be identified.

PCR amplification can lead to artefacts called 'stutters' where the polymerase skips bases during replication. This results in the production of PCR amplicons with altered length and sequence composition leading to misinterpretation[271,272]. Moreover, Illumina sequencing requires an additional amplification step during the sequencing process ('bridge amplification') for signal

amplification, which makes this problem worse. To this end, Oxford Nanopore sequencing is an interesting amplification-free alternative.

5.1.3 Nanopore Sequencing and Ramified Rolling Circle Amplification

Oxford Nanopore sequencing is based on translocation of a ssDNA molecule through a nanopore. This generates nucleotide-dependent ionic current fluctuations that allow the nucleotides passing through the pore to be identified[273,274] (**Figure 5.2E**). Importantly, Oxford Nanopore sequencing offers a number of benefits over Illumina sequencing: 1) low cost, 2) portability of the MinION™ nanopore sequencer, 3) long-read sequencing (currently up to 882 kb[275] compared to 600 nt for Illumina sequencing), 4) fast library preparation, 5) real-time data acquisition and analysis. Collectively, this enables rapid sequencing to be achieved outside laboratory, ideal for forensics. However, a high error rate[276,277] caused by imperfect translocation of the nucleotide strand through the nanopore remains a big challenge for this technology.

To address this problem Ramified Rolling Circle Amplification (RAM, **Figure 5.2C**)[278], a technique that generates extremely long linear dsDNA products containing multiple copies of the template, could be used. The same template is copied in each round of RAM to form a concatemer. Splitting the nanopore signal for the concatemer into its individual parts and averaging the signal could potentially remove random polymerase-derived errors. In contrast, PCR does not use the original template for each subsequent amplification step, therefore inaccurately-amplified sequences will become templates in further cycles and the errors amplified throughout the process.

5.1.4 STR Extraction using CRISPR

Prior to circularisation, the original STR sequence must be excised from the genomic DNA. This could be achieved using CRISPR-Cas9 which has proved efficient for extracting STRs for Illumina Next-Generation sequencing[279]. Since 20 STR sequences need to be excised simultaneously, the simple split-and-click strategy developed in **Chapter 2** could be employed to generate a library of 20 click sgRNAs targeting STRs (**Figure 5.2**). This methodology potentially offers far greater flexibility in the scale and cost of sgRNA library.

5.2 Aims and Objectives

In this Chapter, the ultimate goal is to provide a simple workflow for STR analysis on the field (**Figure 5.2**). Initially, this requires validation of the hypothesis that Oxford Nanopore sequencing accuracy can be increased through the RAM amplification of STRs. It was hoped the Oxford Nanopore signal quality would be sufficiently enhanced to overcome its current barrier to use in STR analysis.

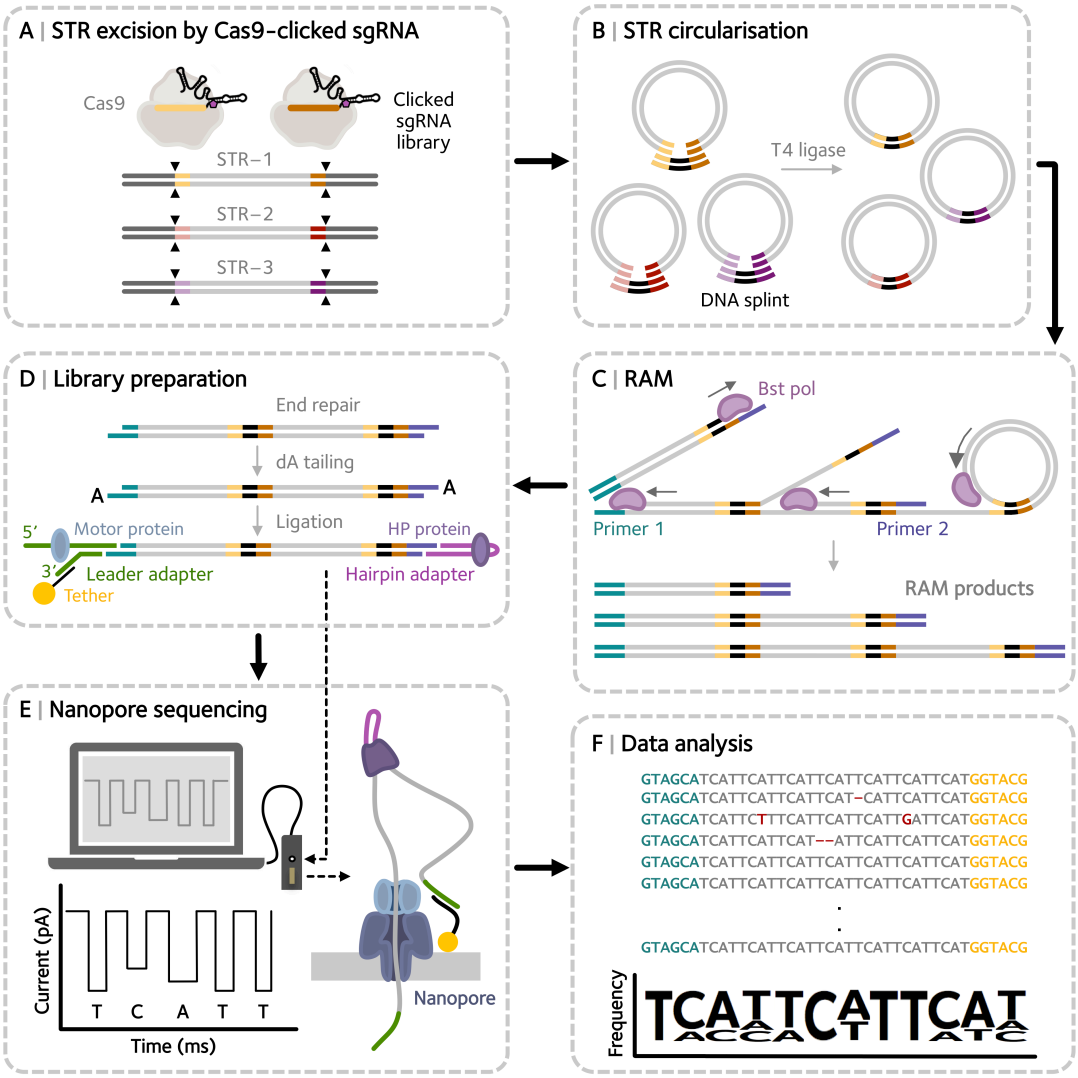


Figure 5.2 | Schematic representation of proposed workflow for STR analysis using CRISPR and Oxford Nanopore sequencing. (A) STR sequences are precisely excised from the human genome using Cas9 and a clicked sgRNA library. (B) STR sequences are circularised in the presence of a DNA splint yielding a circular dsDNA. (C) Upon denaturation, Primer 1 hybridises to the circular DNA template and allows Bst polymerase to perform the first round of amplification, giving a long ssDNA containing many copies of STRs. Following the hybridisation of Primer 2, the polymerase carries out DNA synthesis on the previously generated strand, yielding a laddered series of dsDNA. (D) The RAM products are ligated to the leader adapter composed of a motor protein/tether at one end and a hairpin adapter containing a 'hairpin protein' at the other end. (E) The DNA library then undergoes sequencing using a MinION™ nanopore sequencer. Fluctuations of ion current are caused by ssDNA translocation through the nanopore, which are subsequently transformed into four different nucleobase calls: A, C, G or T. (F) Splitting the signal of the concatemer into its individual parts and performing multiple sequence alignment allows polymerase and sequencer-derived errors to be averaged out, giving the final sequence with identified STR repeat and its number.

5.3 TH01 Ramified Rolling Circle Amplification

The work for this chapter began while designing the oligonucleotides for **Chapter 2**. Therefore, initially nanopore sequencing of RAM-amplified DNA was investigated (**Figure 5.2C–5.2F**).

5.3.1 Generation of Circular DNA Template

As a proof of principle, a 100-mer ssDNA containing seven TCAT repeats to mimic the TH01 STR locus in chromosome 11 was synthesised. This oligonucleotide contained two segments at the 5' and 3' termini that were complementary to a DNA splint. Upon hybridisation of the splint, the terminal 5'-phosphate and 3'-OH of the template were brought adjacent to each other and covalently joined using T4 DNA ligase in good yields (**Figure 5.3A**). Exonuclease I and III digestion, which removes single-stranded or double-stranded oligonucleotides respectively but not a circular DNA, confirmed the success of the cyclisation step (**Figure 5.3A**).

5.3.2 RAM

In order to be able to ligation to the available Oxford Nanopore adapters during the library preparation step for sequencing, the DNA sample must be double-stranded which could be generated by RAM. RAM required the addition of two primers (P1 and P2), one (P1) for the generation of a long ssDNA using the initial primer ('primary amplification') and another (P2) for the synthesis of the complementary strand to the ssDNA. Using thermostable, strand-displacing Bst DNA polymerase (large fragment) at 60 °C, a laddered series of dsDNA products, each ~100 bp (or one STR copy) larger in length than the previous, were obtained (**Figure 5.3B**). This confirmed cyclisation of the initial template was successful as laddered products are typical of RAM[278]. PCR of TH01 was also performed using commercial genomic DNA to investigate a real complex sample. The genomic DNA used has a define TH01 repeat size of 6 and 9.3 making it a good standard for evaluating nanopore sequencing accuracy (**Figure 5.3C**).

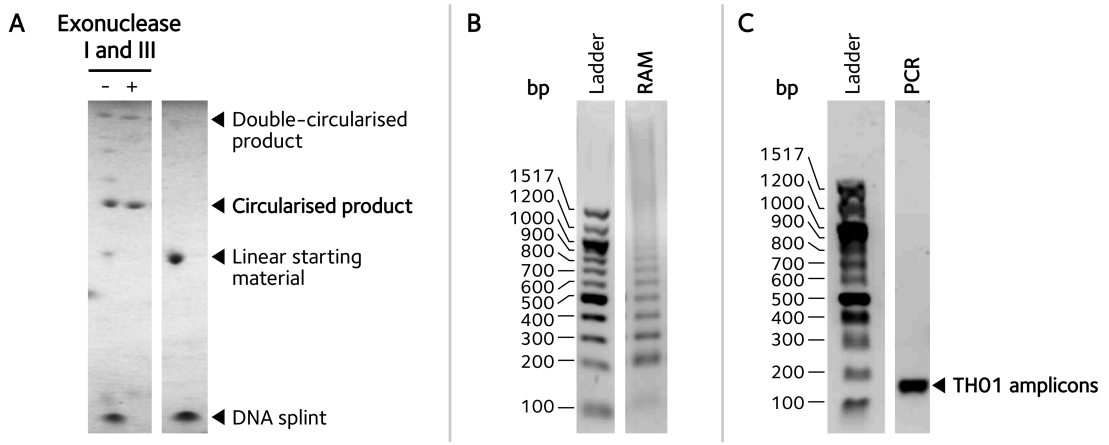


Figure 5.3 | Amplification of TH01 by RAM and PCR. (A) shows the circularisation of the synthetic TH01 template. The reaction consisted of 20 µM linear TH01-containing DNA, 20 µM DNA splint (Primer 1 for subsequent RAM), 0.1 U/µL T4 DNA ligase and a buffer containing 30 mM Tris-HCl, 10 mM $MgCl_2$, 10 mM DTT and 10 mM ATP. The ligation reaction proceeded for 1 h at 37 °C. Half of the reaction was treated with 0.4 U/µL exonuclease I and 0.4 U/µL exonuclease III for 1 h at 37 °C. Exonuclease digestion was inactivated for 5 min at 65 °C before loading onto an 8% denaturing PAGE gel. (B) illustrates the RAM reaction using the circularised DNA template from (A). The RAM reaction contained 10 µM circularised TH01-containing DNA, 0.75 µM primer P1, 0.1 µM primer P2, 0.13 U/µL Bst DNA polymerase (large fragment), 0.8 mM dNTP mix, 2% DMSO and a buffer containing 20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$ and 0.1% Triton® X-100. The reactions proceeded for 1 h at 60 °C and were loaded onto a 1.2% SYBR gold-stained agarose gel. (C) shows the PCR product of TH01-STR from the commercial genomic DNA with the expected size of 156-195 bp (Note that the TH01 contained 6 and 9.3 repeats according to the manufacturer's datasheet (Promega, 2800M Control DNA, cat. no. DD7101). The oligonucleotide sequences can be found in **Appendix Table 21** and **22**.

5.4 Nanopore Sequencing

5.4.1 Library Preparation

Prior to sequencing, the RAM or PCR product was ligated to sequencing adaptors according to the manufacturer's protocol (Oxford Nanopore Technologies, SQK-MAP004 kit, **Figure 5.2D**). This involved fragmentation of the extremely long RAM products, end-repair, dA-tailing at 3' ends, and adapter ligation at both double-stranded ends. The leader adapter at one end consists of a motor protein that unwinds the DNA and controls the rate of nucleotide translocating through the pore. The tether on the same adapter helps to present the DNA to the pore, increasing the yield of sequencing by ensuring that as many pores as possible are utilised. The hairpin at the other end connects the template strand to the complementary strand, enabling both strands to

be sequenced and thus leading to greater accuracy of sequencing. At the end of library preparation, ~20 ng/µL of DNA library was obtained, which is comparable to the suggested yields from the protocol. After sequencing overnight on a MinION™ flowcell, .fast5 files were generated with the data.

5.4.2 Data Analysis

The data consists of reads, where one read indicates one DNA translocation event per nanopore. All the reads were sent to the Metrichor basecaller, which transforms the electrical signals produced from five nucleotides occupying the pore[280,281] into individual base calls. Three 'types' of sequencing reads for each DNA molecule passing through a nanopore are produced: template, complement and two-direction (2D)[282]. The template read is generated when the first of the two strands is presented to the pore, with the aid of the leader adapter. If the hairpin is successfully ligated to the other end of the DNA molecule, the complementary strand will also be analysed. When both template and complement reads are available, a 2D read is obtained by the base-calling algorithm. Since the same piece of DNA is analysed twice, a 2D read will have a greater accuracy than both individual template and complement reads (which are considered to be 1D reads). In Metrichor basecalling high-quality 2D reads with a quality-score greater than the threshold value are determined[281]. Poretools[283] were used to extract the read statistics which are summarised in **Figure 5.4A** with only high-quality 2D reads chosen for further downstream analysis.

The base composition within the 2D reads was converted from a .fast5 format to a .fasta format using the poretools program. Next, the expected sequence (or 'reference') flanking the STR was compared to each read using the LALIGN software[284]. This programme looks for internal duplications of the 'reference' sequence within the target sequence in a pairwise manner. The regions that the 'reference' sequence encases correspond to the 'unknown' STR sequences.

A

Statistics	RAM	PCR
Median read length (bp)	379	220
Min read length (bp)	148	167
Max read length (bp)	3964	636
Total reads	306	252

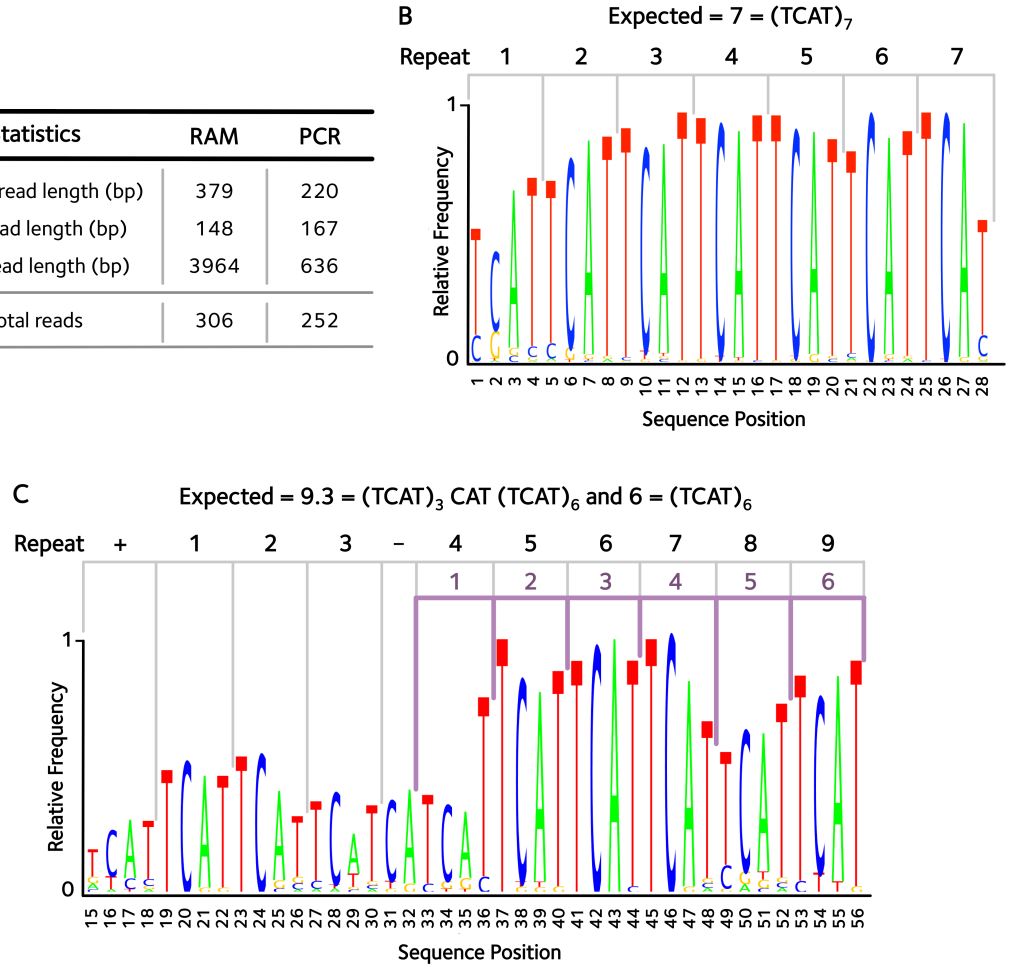


Figure 5.4 | Nanopore sequencing results of STR-RAM and STR-PCR products. (A) shows the read statistics of STR-RAM and STR-PCR sequenced products. (B) illustrates the alignment of the STR sequences from the RAM library preparation method. The Weblogo format indicates the relative frequency of each base found at a particular position (A = green, C = blue, G = yellow and T = red). The expected repeating unit is 5'-TCAT-3' and number of repeats is 7. (C) illustrates the alignment of the STR sequences from the PCR library preparation method and highlights the two different STR alleles found (labelled in black and purple). '+' indicates an inserted repeating unit while '–' indicates a truncated repeating unit. The expected repeating unit is 5'-TCAT-3' according to the HUMTH01 sequence from GenBank (accession number D00269) and numbers of repeats are 9.3 and 6 as described in the datasheet of the purchased human genomic DNA product (Promega, 2800M Control DNA, cat. no. DD7101). Note there are two expected alleles (6 and 9.3), one each from the chromosome 11 pair.

Using this information, these 'unknown' regions were isolated from the original read and reverse complemented where necessary using custom code written in MatLab by Dr. Arun Shivalingam in the Brown group. In the case of PCR, the 'reference' sequences are the two primer sequences. The 'unknown' regions were then subject to multiple alignment using the MUSCLE programme[285] to find a consensus motif for the STR region. Any position within the consensus

motif that has an observed base (A, C, G or T) accuracy below 20% was eliminated and the final sequence was displayed in the WebLogo format (**Figure 5.4B** and **5.4C**). This last step helps eliminate imperfect translation signal base calls and randomly generated polymerase errors.

Promisingly the 'unknown' region of TH01 from RAM shows seven TCAT repeats with high frequency of the expected bases making interpretation of the STR sequence and repeat length straightforward (**Figure 5.4B**). This is consistent with the 7xTCAT in the synthetic 100-mer oligonucleotide. Expected base frequency was lower towards the end of the sequences possibly due to the alignment programme. PCR of the human genomic DNA should generate two alleles as guaranteed by the manufacturer: 1) 6xTCAT repeats and 2) 9.3xTCAT repeats where the 0.3 is CAT. The first allele was observed (**Figure 5.4C**), but the first repeat was observed at a lower frequency than the following five. However, for the second allele, deletion of a few bases and/or insertion of an entire repeat were observed (10.2 repeats as against expected 9.3 repeats). The errors were possibly due to PCR or inherent to the sequencer. Therefore, these results suggest that RAM could help increase accuracy of the sequencing but a direct comparison of RAM from a genomic DNA sample is required to confirm this observation.

5.5 Conclusions and Future Work

Accuracy, speed and cost is vital for field-based forensic analysis. Compared to established sequencing technologies such as the Illumina method, Oxford Nanopore sequencing offers several benefits including portability and lower costs. However, it suffers from high error rates because signal detection and base-calling are imperfect, which is vital. To overcome this limitation, progress towards a workflow combining CRISPR and RAM before nanopore sequencing was made.

In this Chapter, isothermal RAM amplification was used to accurately sequence a synthetic TH01 STR by an Oxford MinION™ nanopore sequencer. Sequencing the dsDNA concatemer from RAM allows deconstruction of the concatemer signal into its individual parts and averaging of the signal to give a higher accuracy consensus motif. This reduces the random errors introduced by the sequencer, and polymerase errors from RAM that occur from replication of the original

circular template. Concurrently, STR analysis of a real genomic DNA sample was performed using PCR. Some errors relative to the expected manufacturer provided TH01 STR sequence was found, suggesting the RAM methodology could be useful.

This project began when Oxford Nanopore sequencing was first commercialised. Significant advances in the technology have since been made, however a RAM type approach could still be useful. For example, Rolling Circle Amplification (RCA) to generate a long ssDNA for sequencing has been reported to improve error rates[286]. The RAM protocol used here has the added benefit of producing dsDNA, which gives more accurate 2D reads. Future work will involve optimising the clicked sgRNA library methodology from **Chapter 2** for STR extraction from genomic DNA. When combined with the RAM-based nanopore sequencing developed here, the overall workflow for sequencing the CODIS 20 STR core loci from human genomic DNA should be evaluated.

Conclusions

6.1 Summary

In **Chapter 2**, a 'split-and-click' methodology was developed to offer a simple and cost-effective way of generating sgRNA. The sgRNA was split into two parts: a ~21-mer DNA-targeting crRNA and a 79-mer Cas9-binding tracrRNA, synthesised with 3'-propargyl and 5'-azide modifications respectively. Upon non-templated CuAAC ligation, a functional sgRNA bearing a biocompatible triazole linkage was generated with good activity *in vitro*. The construct was functional in cells and displayed no unexpected off-target activity, demonstrating for the first time that modified backbones can be tolerated in functionally critical parts of the sgRNA. Incorporating DNA nucleotides into tracrRNA did not impair DNA cleavage activity *in vitro* but unexpectedly the activity in cells was abolished. The feasibility of this approach for sgRNA library construction was also explored, with minimal bias in ligation of two different RNAs observed.

In **Chapter 3**, fluorescent sgRNAs (fsgRNAs) was chemically synthesised for live-cell imaging studies. The fsgRNA constructs were made by clicking three components: a telomere-targeting

crRNA, a fluorophore-containing DNA loop and a truncated tracrRNA (stem loop 2 deleted for improved RNA synthesis), together in a self-templated manner. However, the construct failed to label telomeres in cells as stem loop 2, which through further optimisation was found to be essential for dCas9 imaging, was absent, and the non-biocompatible artificial linkages used to form the construct may destabilise the binding further. An alternative fsgRNA construct based the split-and-click design of **Chapter 2** was also generated with a single fluorophore appended at the 5' end. Live-cell imaging was unsuccessful, again demonstrating the different requirements of gene editing and imaging. The stability of dCas9–fsgRNA or dCas9–fsgRNA–target DNA complex appears to be absolutely crucial for imaging. fsgRNA made from labelling IVT sgRNA with a fluorophore did allow telomere labelling, but strong cytoplasmic background fluorescence due to unbound/endosomally trapped fsgRNAs remained an issue. As a result, an RNase H-activated fgRNA with a cleavable quencher was developed. In this design the gRNA remains dim or invisible until the quencher is removed by nuclear RNase H, allowing the fgRNA to become fluorescent only within the nucleus. Removal of the quencher does not appear to be efficient in cells and further optimisation of the design is required.

In **Chapter 4**, attempts were made to manipulate Cas12a cleavage of dsDNA to generate non-complementary instead of complementary staggered ends, with the aim of altering/enhancing genome editing. To this end, two crRNAs: a PAM-dependent crRNA and a PAM-independent crRNA were linked *via* CuAAC ligation, yielding a dual gRNA. Hypothetically, the first crRNA in a dual gRNA should direct Cas12a to cleave target DNA, while the second crRNA should recruit another Cas12a to cleave the same target DNA at the adjacent site, causing the complementarity of the staggered ends to be destroyed. However, the dual gRNA design inhibited DNA cleavage, possibly due to the steric hindrance caused by the misfolding of the individual crRNAs that prevents Cas12a recognition. Moreover, unexpected ssDNA nuclease activity of AsCas12a was discovered which was consistent with recently published literature[262,263], with the notable exception that AsCas12a caused non-specific ssDNA cleavage in the absence of a crRNA.

In **Chapter 5**, a workflow combining CRISPR and RAM was proposed for Oxford Nanopore sequencing to provide rapid and highly accurate STR analysis in forensic applications. In brief, the

'split-and-click' approach from **Chapter 2** would be used to construct a sgRNA library to precisely excise 20 individual STR core loci, which will be subsequently circularised to serve as a template for RAM. Since RAM uses the same circular template for each round of amplification, the errors caused by polymerase should not be amplified (cf. PCR). Following RAM, a long dsDNA product containing multiple copies of STR sequence is sequenced multiple times by nanopore sequencing, with averaging of the individual repeats minimising the random errors potentially caused by the polymerase and the sequencer. As a proof-of-concept study, a ssDNA containing TH01 STR sequence was synthesised, circularised, RAM-amplified and subjected to nanopore sequencing. Nanopore sequencing of the PCR-amplified TH01 STR from human genomic DNA was also performed. It was found that RAM could improve the accuracy of STR sequencing.

6.2 Outlook

In 2012, the discovery that an RNA-guided CRISPR-Cas9 DNA nuclease could be used as a genome-editing tool caused massive interest in the CRISPR field. Subsequent research focussed on the biological modification of the protein and gRNA sequence to improve and expand the functionality of the system. Yet, by early 2015 when this project began, no publications explored the use of chemical modifications to achieve the same goals. Therefore, this became the focus of this thesis.

Solid-phase oligonucleotide synthesis is key to the generation of chemically modified gRNA. However, RNA monomers have bulky 2'-hydroxyl protecting groups. This reduces monomer coupling efficiency, which becomes progressively worse towards the 5' end as the length of the oligonucleotide increases. As a result, the direct chemical synthesis of 100-mer Cas9 sgRNAs, for example, is problematic. Yields are low and more importantly the DNA-targeting sequence at the 5' end is subject to more errors, potentially reducing target specificity. In this thesis, a common strategy is used to overcome this problem: the gRNA is split into smaller fragments, for purer synthesis, with functional groups attached that allow chemical ligation and reconstruction of the full-length gRNA.

In particular CuAAC and SPAAC 'click' ligation was explored. This requires the introduction of azide and alkyne functional groups into the RNA. Direct incorporation of the monomer during solid phase synthesis was the simplest approach (e.g. 3'-propargyl modifications on the resin), however azides are incompatible with phosphoramidite synthesis due to the Staudinger reaction[204]. Therefore, they were introduced either by post-synthetic solution-phase NHS ester labelling or post-synthetic solid phase conversion of the terminal 5'-hydroxyl group to a 5'-azide. Synthetically, 5' azide conversion is more attractive than azide-NHS ester labelling because: 1) the reaction occurs on 2'-protected RNA minimising degradation, 2) it requires fewer purification steps, and 3) it can give rise to a more biocompatible seven-bond triazole backbone linkage[182] (as will be discussed next). However, unlike NHS-ester coupling, 5' azide conversion specifically requires TBDMS-protected RNA and does not appear compatible with TC-RNA. Given that TC-RNA offers cleaner synthesis and deprotection than TBDMS-RNA, future work should aim to improve the synthesis of 5' azide-RNA using TC chemistry, for example, through the use of a 5' azide H-phosphonate monomer[205–207].

For the ligation step, extensive optimisation was required for non-templated CuAAC RNA-RNA ligation. Unlike DNA–DNA CuAAC ligation, the addition of more copper ions, a high percentage of DMSO (to denature structured RNA) and magnesium ions (to compete with and displace copper ions that bind to the RNA) were key to the success of the reaction. It was found that CuAAC was generally more efficient than SPAAC ligation, which suffers from the inactivation of the cyclic alkyne reactant during purification and annealing.

After construction of the clicked sgRNAs, the effect of artificial linkages on Cas protein activity was evaluated. All triazole linkages have negligible effect on sgRNA activity when placed far away from the protein and at the end of a hairpin (e.g. in 37–66 click sgRNAs for Cas9, **Chapter 2**). Larger triazole linkages become significantly unfavourable when introduced within the sgRNA duplex (e.g. fsgRNA II constructs for dCas9-3xGFP, **Chapter 3**). When placing the linkage within Cas9 interaction regions, a short artificial backbone mimetic design is necessary (e.g. Tz2 linkage in ~21–79 click sgRNAs for Cas9, **Chapter 2**). However, the Tz2 used linkage still causes a slight decrease in activity compared to the natural phosphodiester bond (~50% decrease in DNA

cleavage activity *in vitro* and indel formation in cells). This suggests there is still scope for investigation of other artificial backbones, such as an amide linkage[209] that is tolerated well by polymerases or a phosphoramidate linkage[208] that is a minimal modification of the phosphodiester backbone.

In general, the split-and-click approach to gRNA construction was successful synthetically, and functionally when the split point is carefully designed. The most promising example is the ~21–79 'clicked' sgRNAs, which allows synthesis of an invariable 79-mer tracrRNA (made once and in large scale) that can be coupled to a variable DNA-targeting 20-mer RNA (made when needed in higher purity and lower cost). Although *in vitro* transcription will likely be the method of choice when only a single or few sgRNAs are required, the split-and-click approach would benefit applications requiring rapid screening or sgRNA library construction (e.g. multiplex live-cell imaging, and forensic STR loci analysis).

Beyond the split-and-click strategy, direct chemical attachment of fluorophores to the sgRNA was also explored. Conceptually this approach is appealing as it allows the brightness (number of fluorophores) and colour (type of fluorophore) to be easily tailored for each DNA target. However, the key limitation was not sgRNA synthesis but its delivery into cells, a common problem to the field of nucleic acids. Endosomal/lysosomal trapping was observed, which generated large fluorescent cytoplasmic granules that could be confused with nuclear genomic foci in the absence of GFP labelled dCas9 as cross-validation. Attempts to overcome this issue were made by introducing a quencher into design, with nuclear RNase H-mediated release of the quencher generating fluorescence. Background signal quenching was efficient but activation of fluorescence requires more careful refinement of the gRNA design. Alternatively, intercalators can be appended to the sgRNA that become fluorescent only upon DNA binding. This strategy has been used in literature[287] for mRNA imaging and its use in CRISPR should also be explored.

Finally, other interesting biological observations were made. In **Chapter 4** intrinsic AsCas12a single-stranded DNase activity was discovered, while in **Chapter 3** it was found that the requirements of sgRNAs for live-cell imaging are different to those for gene editing. Stem loop 2

of the sgRNA was crucial for maintaining a stable dCas9–sgRNA–target DNA complex during live-cell imaging, but was not found to be necessary for gene editing[230].

At the end of this work, the number of publication that explore chemical modification of the sgRNA has increased to eleven[103,122,188,130–133,137–140]. However, these studies have mainly focused on improving the inherent properties of the sgRNA such as RNA stability and off-target activity. The work described here shows chemistry can also be used to facilitate CRISPR studies (e.g. library construction and access to chemically modified sgRNA) and could potentially be used to expand the applications of CRISPR (e.g. live-cell imaging and altering gene editing outcomes). It lays solid foundations for future studies.

Experimental
Methods and Materials

7.1 Oligonucleotide Synthesis

7.1.1 DNA Synthesis

DNA synthesis was performed on an Applied Biosystems 394 automated DNA/RNA synthesizer using a standard phosphoramidite cycle of detritylation, coupling, capping and oxidation using TCA (3% in dichloromethane), tetrazole (0.45 M in acetonitrile), Cap A (10% acetic anhydride in 10% lutidine and 80% tetrahydrofuran (THF))/Cap B (16% *N*-methylimidazole (NMI) in THF) and iodine (0.02 M in THF, pyridine and water) respectively on a 0.2 or 1.0 µmole scale. Appropriate solid supports listed in **Section 7.1.3** were packed into a twist column (Glen research) for synthesis. β-cyanoethyl protected phosphoramidites (dA-bz, dG-ib, dC-ac, dT, Sigma-Aldrich) were dissolved in anhydrous acetonitrile (0.1 M) immediately prior to use. The coupling time for dA, dC, dG and dT monomers was 1 min, and 10 min for modified monomers. Stepwise coupling efficiencies were determined by automated trityl cation conductivity monitoring and were >98% in all cases.

7.1.2 RNA Synthesis

RNA synthesis was performed on an Applied Biosystems 394 automated DNA/RNA synthesizer using a standard phosphoramidite cycle of detritylation, coupling, capping and oxidation on a 1.0 µmole scale. Appropriate solid supports listed in **Section 7.1.3** were packed into a twist column (Glen research) for synthesis.

For the TBDMS method, RNA phosphoramidites (A-tac, C-ac or C-tac, G-tac, U, Sigma-Aldrich) were 2'-*O*-TBDMS protected. Monomers were dissolved in anhydrous acetonitrile (0.1 M) immediately prior to use. Coupling, capping and oxidation reagents were benzylthiotetrazole (BTT) (0.3 M in acetonitrile, Link Technologies), fast deprotection Cap A (5% t-butylphenoxyacetyl acetic anhydride in THF)/Cap B (16% *N*-methylimidazole (NMI) in THF) and iodine (0.1 M in tetrahydrofuran (THF), pyridine and water) respectively. The coupling time for all monomers during RNA synthesis was 6 min. Stepwise coupling efficiencies were determined by automated trityl cation conductivity monitoring and in all cases were >97%.

For the TC method, RNA phosphoramidites (A-bz, C-ac, G-ib, U, Sigma-Aldrich) were 2'-*O*-TC protected. Monomers were dissolved in anhydrous toluene (0.1 M) immediately prior to use. Coupling, capping and oxidation reagents were identical to those used in DNA synthesis except a solution of ethylthiotetrazole (ETT) (0.25 M in acetonitrile, Link Technologies) was used as a coupling reagent. The coupling time for all monomers during RNA synthesis was 3 min. Stepwise coupling efficiencies were determined by automated trityl cation conductivity monitoring and in all cases were >97%.

7.1.3 Modified Monomers and Solid Supports

	Monomer	Supplier	Cat. no.
1	Amino-Modifier C6-dT-CE Phosphoramidite	Link Technologies	2135-F100
2	BHQ-2-dT-CE Phosphoramidite	Link Technologies	2157-F100
3	Chemical Phosphorylation Reagent (CPR) (5' Phosphate)	Link Technologies	2101-F100x10
4	Cyanine-3-CE Phosphoramidite (Cyanine 540)	Link Technologies	2520-F100X1
5	5'-Fluorescein-CE Phosphoramidite (6-FAM)	Link Technologies	2134-F050
6	5'-Hexynyl Phosphoramidite	Cambio	10-1908-90
7	5'-TFA-Amino-Modifier C6-CE Phosphoramidite	Link Technologies	2124-F100x10

7.1.4 Modified Solid Supports

	Solid support	Supplier	Cat. no.
1	3'-Alkyne-Modifier Serinol CPG	Cambio	20-2992-01
2	3'-Amino-Modifier C7 CPG 1000	Link Technologies	2350-P008
3	3'-*O*-propargyl Adenosine (n-bz) 2'-lcaa CPG 1000 Å	Chemgenes	N-9881-10
4	3'-*O*-propargyl Cytidine (n-bz) 2'-lcaa CPG 1000 Å	Chemgenes	N-9882-10
5	3'-*O*-propargyl Guanosine (n-ibu) 2'-lcaa CPG 1000 Å	Chemgenes	N-9883-10
6	2'-*O*-ethoxyazido-dT CPG	Prepared by Prof. Afaf El-Sagheer	

7.2 Solid-phase Oligonucleotide Modification

7.2.1 5'-Azide Conversion of RNA

After automated solid-phase synthesis, the fully protected resin-bound 5'-OH RNA was treated with methyltriphenoxyphosphonium iodide in anhydrous DMF (0.5 M, 2 mL) for 1 h at room temperature. The solid support was washed with dry DMF (3 x 1 mL) and dried with argon. A saturated sodium azide solution was then prepared by heating sodium azide (100 mg) resuspended in dry DMF (2 mL) for 10 min at 70 °C. After cooling to room temperature, the resin was treated with the solution for 5 h at 55 °C. The resin was then washed with DMF (3 x 1 mL),

acetonitrile (3 x 1 mL) and dried with argon. The 5'-azido RNA was cleaved from solid support and deprotected as described below.

7.2.2 Selective β-cyanoethyl Removal

Oligonucleotides bearing primary amines were treated with diethylamine (20% in anhydrous acetonitrile) for 20 min at room temperature. The resin was then washed with acetonitrile (3 x 1 mL) and dried with argon.

7.3 Oligonucleotide Deprotection

7.3.1 DNA Deprotection

DNA was cleaved from the solid support and deprotected by exposure to a concentrated solution of aqueous ammonia in a sealed vial for 5 h at 55 °C. After drying in vacuo, oligonucleotides were dissolved in water and subject to further purification.

7.3.2 2'-*O*-TBDMS RNA Deprotection

RNA was cleaved from the solid support and the nucleobases deprotected using concentrated aqueous ammonia:ethanol (3:1 v/v) in a sealed vial for 2 h at 55 °C unless stated otherwise. After cooling down to room temperature, the solution was filtered and then ammonia removed in vacuo. The ammonia-free solution was then freeze-dried, re-dissolved in a 1:1 mixture of dry DMSO (300 µL) and trimethylamine trihydrofluoride (300 µL) and heated for 2.5 h at 65 °C. After cooling down to room temperature, sodium acetate (3 M pH 5.2, 50 µL) and butanol (3 mL) were added and the RNA was stored for 30 min at -80 °C. The RNA was then pelleted by centrifugation (8,500 rpm, 30 min, 4 °C), the supernatant discarded and the pellet washed twice with 70% ethanol (750 µL). The pellet was then dried in vacuo, dissolved in water and desalted as described in the **Section 7.4.1** before further purification.

7.3.3 2'-*O*-TC RNA Deprotection

The solid support was exposed to dry ethylenediamine:toluene (1:1 v/v) for 6 h at room temperature, washed with toluene (3 x 1 mL) and dried using argon. The cleaved RNA was eluted with water and purified further.

7.4 Oligonucleotide Purification

7.4.1 Desalting

NAP™-10 columns (G.E. Healthcare, cat. no. GE17-0854-02) were used according to the manufacturer's instructions.

7.4.2 RP-HPLC

Oligonucleotides were purified using a Gilson HPLC system with ACE® C8 column (10 mm x 250 mm, pore size 100 Å, particle size 10 µm) with a gradient of buffer A (0.1 M TEAB, pH 7.5) to buffer B (0.1 M TEAB, pH 7.5 containing 50% v/v acetonitrile) and flow rate of 4 mL/min. For unmodified oligonucleotides (including 5'-azide converted oligonucleotides), the gradient was 20–30% buffer B over 21 min. For other oligonucleotides, the gradient was suitably adjusted.

7.4.3 Denaturing PAGE

Oligonucleotides was mixed with formamide (50% v/v) and loaded onto a denaturing 8–10% polyacrylamide gel (1 x TBE buffer containing 7 M urea, example recipes listed below, W x D x H = 18 x 0.2 x 24.4 cm) and separated at 20 W for 2–3 h. DNA bands were visualised under UV, excised, crushed, soaked in water (for DNA, ~15 mL) or buffer (for RNA, 50 mM Tris-HCl pH 7.5 containing 25 mM NaCl, ~15 mL) overnight at 37 °C with 900 rpm shaking. After gravity filtration to remove the gel, the oligonucleotide solutions were concentrated in vacuo and desalted using two NAP™-25 columns (G.E. Healthcare, cat. no. GE17-0852-01).

Reagent	Percentage of gel	
	8%	10%
Urea	29.4 g	29.4 g
Milli-Q H_2O	26 mL	22.5 mL
40% acrylamide:bis-acrylamide (29:1)	14 mL	17.5 mL
10x TBE (1 M Tris-base, 1 M boric acid, 20 mM EDTA, pH 8.3)	7 mL	7 mL
10% ammonium persulfate (APS)*	56 µL	56 µL
TEMED*	560 µL	560 µL

*Note that APS and TEMED were added to acrylamide mixture before casting.

7.5 Oligonucleotide Mass Spectrometry

All oligonucleotides were characterised by negative-mode electrospray using a UPLC-MS Waters XEVO G2-QTOF mass spectrometer and an Acquity UPLC system with a BEH C18 1.7 µm column (Waters). A gradient of methanol in TEA and hexafluoroisopropanol (HFIP) was used (buffer A, 8.6 mM TEA, 200 mM HFIP in 5% methanol/water (v/v); buffer B, 20% v/v buffer A in methanol). Buffer B was increased from 0–70% over 7.5 min or 15–30% over 12.5 min for normal oligonucleotides and 50–100% over 7.5 min for very hydrophobic oligonucleotides. The flow rate was set to 0.2 mL/min. Raw data were processed and deconvoluted using the deconvolution software MassLynx v4.1.

7.6 Post-synthetic Oligonucleotide Labelling

Freeze-dried oligonucleotide (20 nmol) was dissolved in $NaHCO_3$ buffer (0.5 M, pH 8.5, 30 µL) and mixed with the NHS ester (200 nmol, 20 µL) dissolved in DMSO. The reaction was then left for 4 h at 25 °C with 750 rpm shaking. After dilution with water, the samples were desalted using a NAP™-10 column (G.E. Healthcare, cat. no. GE17-0854-02) prior to RP-HPLC purification.

7.7 Click Oligonucleotide Ligation

7.7.1 Non-templated CuAAC

For final optimised conditions, 3'-alkyne RNA (750 pmol in 1 µL H_2O), 5'-azido RNA (500 pmol in 1 µL H_2O) were mixed with $MgCl_2$ (100 mM, 0.5 µL), triethylammonium acetate (TEAA) buffer (2 M, pH 7, 1 µL), DMSO (5 µL) and fresh ascorbic acid (125 mM, 1 µL). While degassing the oligonucleotide solution with argon in a 0.5 mL Eppendorf tube, a 55% DMSO solution of $CuSO_4$–THPTA complex (250 mM, 0.5 µL) was added and the reaction (final volume = 10 µL) was left for 1–2 h at room temperature. The sample was then diluted with water and desalted using a NAP™-5 column (G.E. Healthcare, cat. no. GE17-0853-01). Any variations in reaction conditions are described in the relevant figure caption.

7.7.2 Templated CuAAC

For final optimised conditions, 3'-alkyne RNA (250 pmol), 5'-azido RNA (250 pmol) and DNA splint (250 pmol) were dissolved in a NaCl solution (0.2 M, 8.13 µL) and annealed by heating for 5 min to 95 °C then cooling to 25 °C by 1 °C/min. Then TEAA buffer (2 M, pH 7, 1.25 µL), DMSO (1.25 µL) and fresh ascorbic acid (50 mM, 1.25 µL) were added to the reaction mixture. While degassing the oligonucleotide solution with argon in a 0.5 mL eppendorf, a 55% DMSO solution of $CuSO_4$–THPTA complex (100 mM, 0.63 µL) was added and the reaction (final volume = 12.5 µL) was left for 1–2 h at room temperature. The sample was then diluted with water and desalted using a NAP™-5 column (G.E. Healthcare, cat. no. GE17-0853-01). Any variations in reaction conditions are described in the relevant figure caption.

7.7.3 Non-templated SPAAC

For final optimised conditions, 3'-DBCO RNA (1250 pmol), 5'-azido RNA (500 pmol) were mixed in a NaCl solution (0.2 M, 7.9 µL) containing EDTA (0.5 M, 0.1 µL) and HEPES (0.1 M, 2 µL) and were either left for 1–2 h at room temperature, or heated for 5 min to 95 °C then cooled to 25 °C by 1 °C/min and left for 1–2 h at room temperature. Any variations in reaction conditions are described in the relevant figure caption.

7.7.4 Templated SPAAC

For final optimised conditions, 3'-DBCO RNA (250 pmol), 5'-azido RNA (250 pmol) and a DNA splint (250 pmol) were dissolved in a NaCl solution (0.2 M, 7.9 µL) containing EDTA (0.5 M, 0.1 µL) and HEPES (0.1 M, 2 µL) and annealed by heating for 5 min to 95 °C then cooling to 25 °C by 1 °C/min. Then, the reaction was left for 1–2 h at room temperature. Any variations in reaction conditions are described in the relevant figure caption.

After click ligation, samples were desalted using a NAP™-5 column (G.E. Healthcare, cat. no. GE17-0853-01) and purified using 10% denaturing PAGE unless stated otherwise.

7.8 *In vitro* Transcription

For transcription a dsDNA template was formed from mixing a single-stranded DNA template (50 µM, 25 µL), a single-stranded DNA complementary to the T7 promoter sequence (50 µM, 25 µL), water (40 µL) and hybridisation buffer (10 µL, 10x composed of 100 mM Tris-HCl pH 8.5, 500 mM NaCl and 10 mM EDTA). The MEGAScript™ T7 *in vitro* transcription kit (Thermo Fisher Scientific, AM1334) was used to generate sgRNA. The transcription reaction contained ATP (75 mM, 2 µL), CTP (75 mM, 2 µL), GTP (75 mM, 2 µL), UTP (75 mM, 2 µL), T7 enzyme mix (2 µL), reaction buffer (10x, 2 µL), dsDNA template (12.5 µM, 0.4 µL) and water (7.6 µL). The reaction was then left for 16 h at 37 °C before the addition of Turbo DNase (1 µL, 2 U/µL). After incubation for 1 h at 37 °C, reactions were then purified using the MEGAClear™ transcription Clean-up kit (Thermo Fisher Scientific, AM1908). The transcription reaction was mixed with Elution Solution (80 µL), Binding Solution (350 µL) and ethanol (100%, 250 µL) and then transferred to the filter cartridge. After centrifugation (13,300 rpm, 1 min), the flow through was discarded and Wash Solution (500 µL) was added to the filter cartridge, followed by centrifugation and removal of the flow through. This washing step was repeated once more. Next, Elution Solution (50 µL) was applied to the filter cartridge and heated for 10 min at 70 °C. After centrifugation (13,300 rpm, 1 min), the eluted *in-vitro* transcribed (IVT) RNA was collected.

7.9 IVT sgRNA Azido-dA Tailing

Reactions contained IVT sgRNA (2 nmol in 83 µL H_2O), 2'-azido-dATP (9.3 µM, 3.24 µL, TriLink Biotechnologies, cat. no. N-10451), Yeast Poly (A) polymerase (600 U/µL, 3 µL, Thermo Fisher Scientific, cat. no. 74225Z25KU), Poly (A) polymerase reaction buffer (5x, 25 µL) and water (10.76 µL). After incubation for 2 h at 37 °C, RNA was purified using the MEGAClear™ transcription Clean-up kit (Thermo Fisher Scientific, AM1908) as described in **Section 7.8**.

7.10 Azido-IVT sgRNA Labelling

2'-azido-IVT sgRNA (1.5 nmol in 200 µL H_2O) was mixed with Cy3-BCN (2 mM in DMSO, 75 µL) and the reaction left for 2 h at room temperature. The sample was then diluted with water and the excess dye was removed using a NAP™-10 column (G.E. Healthcare, cat. no. GE17-0854-02) prior to RP-HPLC purification.

7.11 Plasmid Construction

7.11.1 pSpCas9(BB)-2A-Puro-v2-Broccoli

The pSpCas9(BB)-2A-Puro (PX459) V2.0 plasmid (Addgene, plasmid number 62988, kindly provided by Dr. Feng Zhang[288]) was digested with BbsI and XbaI restriction enzymes to remove the gRNA insert. The digested plasmid was then ligated to a synthetic insert encoding the fluorescent F30-Broccoli aptamer (oligonucleotides listed in **Appendix Table 23**) to give the pSpCas9(BB)-2A-Puro-v2-Broccoli. The cloning of this plasmid was performed by Dr. Arun Shivalingam.

7.11.2 pDEST-his-dCas9-3xGFP

The vector was created from pDEST-hisMBP-AsCpf1-EC (Addgene, plasmid number 79007, kindly provided by Prof. Jin Soo Kim[245]) containing the tac promoter, MBP and a six-histidine tag and pHAGE-TO-dCas9-3xGFP (Addgene, plasmid number 64107, kindly provided by Prof. Thoru Pederson[222]) containing the dCas9-3xGFP gene.

The vector backbone of pDEST-hisMBP-AsCpf1-EC excluding AsCpf1 and the dCas9-3xGFP gene from pHAGE-TO-dCas9-3xGFP were amplified by PCR using the tailed primers listed in **Appendix Table 24**. The PCR reaction contained template plasmid DNA (1 ng/µL, 0.4 µL), forward primer (10 µM, 1 µL), reverse primer (10 µM, 1 µL), Q5® High-Fidelity DNA Polymerase (0.02 U/µL, 2 µL, NEB, cat. no. M0491S), dNTP (10 mM, 0.4 µL), Q5 reaction buffer (5x, 4 µL, NEB), Q5 high GC enhancer (5x, 2 µL, NEB) and water (9.2 µL). After heat activation (30 sec, 95 °C), 35 cycles of denaturation (10 sec, 98 °C), annealing (30 sec, 62 °C for pDEST-hisMBP-AsCpf1-EC vector and 66 °C for pHAGE-TO-dCas9-3xGFP vector) and extension (3.5 min, 72 °C) were performed before final extension (2 min, 72 °C). The samples were cooled for 30 min at 4 °C before 1 µL was taken for confirmation of PCR product formation using an Agilent Bioanalyzer DNA 12000 kit (Agilent Technologies, cat. no. 5067-1508 for DNA chips and cat. no. 6067-1509 for reagents), according to the manufacturer's protocol. The remaining mixture was mixed with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1 v/v) to remove the polymerase before the aqueous layer was taken for desalting using a NAP™-5 column (G.E. Healthcare, cat. no. GE17-0853-01) to remove dNTPs and salt. PCR amplicon concentration was determined using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific).

The two PCR amplicons were ligated using the Gibson Assembly® Cloning kit according to the manufacturer's protocol (NEB, cat. no. E5510S). The reaction mixture consisted of pDEST vector backbone (0.11 µM, 1.82 µL), dCas9-3xGFP insert (0.18 µM, 1.11 µL), water (0.07 µL) and Gibson Assembly Master Mix (2x, 3 µL). The reaction proceeded for 15 min at 50 °C and was then stored on ice prior to subsequent transformation.

7.12 Plasmid Production

7.12.1 Transformation

For Gibson assembly of pDEST-his-dCas9-3xGFP, 11 µL NEB 5-alpha Competent E. coli cells (supplied with the Gibson Assembly® Cloning kit, NEB, cat. no. E5510S) were thawed on ice before the addition of 0.44 µL of the assembly reaction to the cells. After mixing gently, the cells were placed on ice for 30 min prior to heat-shock transformation by incubation for 30 sec at 42

°C. SOC media (209 µL, provided with the kit) was added to the cells and the mixture was incubated for 1 h at 37 °C with 250 rpm shaking. The cells were then spread onto a Luria-Bertani (LB) agar selection plate containing 50 µg/mL ampicillin. Note that other plasmids from Addgene were provided in transformed E. coli cells. After incubation for 16 h at 37 °C, colonies were picked and grown in a small-scale culture (5 mL) containing LB media and 50 µg/mL ampicillin for 16 h at 37 °C with 250 rpm shaking.

7.12.2 Small-scale Plasmid DNA Preparation

The plasmid DNA was isolated from the bacterial cells using the QIAprep® Spin Miniprep Kit according to the manufacturer's protocol (QIAGEN, cat. no. 27104). The bacterial cells in the small-scale culture (5 mL) were harvested by centrifugation at 8,500 rpm for 3 min at room temperature. After discarding the supernatant, the cell pellet was resuspended in Buffer P1 (250 µL), Buffer P2 (250 µL) and Buffer N3 (350 µL) with gentle mixing between each addition of reagents. The mixture was centrifuged at 13,000 rpm for 10 min and the supernatant was transferred to a QIAprep 2.0 spin column (provided with the kit). After centrifugation at 13,000 rpm for 1 min, the flow-through was discarded. The QIAprep 2.0 spin column was washed by the addition of Buffer PE (750 µL) followed by centrifugation at 13,000 rpm for 1 min. This step was repeated a second time before the QIAprep 2.0 spin column was placed in a clean 1.5 mL eppendorf. After the addition of Buffer EB (50 µL, 10 mM Tris-HCl, pH 8.5) and incubation for 1 min at room temperature, the column was centrifuged at 13,000 rpm for 1 min to elute the plasmid. The plasmid DNA concentration was determined using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific).

7.12.3 Enzymatic Digestion of Plasmid DNA

To select for a plasmid with desired gene insertion, plasmids were digested using the desired enzyme(s) in the appropriate buffer (1 U total, total reaction volume = 10 µL, 30 min, 37 °C) and the cleavage products were visualised on a 0.8% SYBR gold-stained agarose gel.

7.12.4 Large-scale Plasmid DNA Preparation

Bacterial cells containing the plasmid of interest were grown in LB media (500 mL) containing 50 µg/mL ampicillin for 16 h at 37 °C with 250 rpm shaking prior to plasmid DNA extraction using the QIAGEN Plasmid Maxi Kit (QIAGEN, cat. no. 12162) according to manufacturer's protocol. Cells were harvested by centrifugation at 6,000 x g for 15 min at 4 °C. After the supernatant was removed, the cell pellet was resuspended in Buffer P1 (10 mL), before the addition of Buffer P2 (10 mL), thorough mixing and incubation for 5 min at room temperature. Next, pre-chilled Buffer P3 (10 mL) was added to the suspension, mixed and incubated on ice for 5 min before centrifugation at 25,000 x g for 30 min at 4 °C. In parallel, a QIAGEN-tip (a filtration column provided with the kit) was equilibrated using Buffer QBT (10 mL). The supernatant was applied to the empty QIAGEN-tip and the flow-through was discarded. The QIAGEN-tip was washed twice with Buffer QC before the plasmid DNA was eluted with Buffer QF (15 mL) into a clean 50 mL falcon tube. The plasmid DNA was precipitated by the addition of room-temperature absolute isopropanol (10.5 mL, 0.7 volume). Centrifugation at 9,000 x g for 30 min at 4 °C was performed and the supernatant was discarded. The DNA pellet was washed with 5 mL room-temperature 70% ethanol (5 mL), centrifuged at 9,000 x g for 10 min and air-dried for 10 min after removing the supernatant. The dry DNA pellet was re-dissolved in H_2O (0.5 mL) and quantified using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific) before diluting to 1 µg/µL.

7.13 Bacterial Protein Expression

pDEST-hisMBP-AsCpf1-EC (for AsCas12a expression) or pDEST-his-dCas9-3xGFP (for dCas9-3xGFP expression) was transformed into Rosetta™ competent cells (Novagen, cat. no. 70953) according to manufacturer's protocol. Competent cells (200 µL) were thawed on ice prior to the addition of plasmid DNA (10 ng/µL, 1 µL) to the cells. The mixture was incubated on ice for 5 min, heat-shocked for 30 sec at 42 °C and then incubated on ice for 2 min. Next, room-temperature SOC medium (80 µL, provided with the cells) was added and the mixture was incubated for 1 h at 37 °C with 250 rpm shaking prior to plating on an LB selection plate containing 50 µg/mL ampicillin. After incubation for 16 h at 37 °C, a single colony was picked and grown in a small-scale culture containing LB media (5 mL) and 50 µg/mL ampicillin for 16 h at 37 °C with 250 rpm shaking. The

small-scale culture was then transferred into a 2 L culture flask containing LB media (1 L) supplemented with 50 µg/mL ampicillin and incubated for 16 h at 37 °C with 250 rpm shaking until the optical density at 600 nm (OD_{600}) of the culture reached 0.6. Next, the culture was cooled down to 16 °C. IPTG with final concentration of 0.5 mM was used to induce protein expression for 16 h at 16 °C with 250 rpm shaking. Finally, the cells were harvested by centrifugation at 6,000 x g for 15 min at 4 °C. After removing the supernatant, the cell pellet was subject to protein purification immediately.

7.14 Protein Purification

7.14.1 dCas9-3xGFP

7.14.1.1 Nickel-Immobilised Metal Affinity Chromatography

The cell pellet was resuspended in 30 mL lysis buffer (20 mM HEPES pH 7.0, 500 mM NaCl, 1 mM $MgCl_2$, 1mM DTT and 10% glycerol) supplemented with 1 mg/mL lysozyme (Sigma) and one tablet of EDTA-free protease inhibitor cocktail (UltraCruz® Biotechnology cat. no. sc-2913). Cells were then lysed by sonication (5 sec on, 8 sec off and 14 cycles). After centrifugation at 25,000 rpm for 30 min at 4 °C, the supernatant (30 mL) was incubated with 0.5 mL nickel-charged resin (Ni-NTA agarose, QIAGEN, cat. no. 30210) for 30 min at 4 °C. The mixture was then applied to a gravity-flow chromatography column and washed with 10 mL wash buffer A (20 mM HEPES pH 7.0, 500 mM NaCl, 1 mM $MgCl_2$, 1mM DTT, 10% glycerol and 5 mM imidazole) followed by 2.5 mL wash buffer B (20 mM HEPES pH 7.0, 500 mM NaCl, 1 mM $MgCl_2$, 1mM DTT, 10% glycerol and 35 mM imidazole). The protein was then eluted with 14 mL elution buffer (20 mM HEPES pH 7.0, 500 mM NaCl, 1 mM $MgCl_2$, 1mM DTT, 10% glycerol and 300 mM imidazole) and collected in a series of 1.5 mL fractions. Protein concentration was estimated by measuring the absorbance at 280 nm using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific). Note that all flow-through samples (non-binding fractions, wash steps and eluted protein) were kept and analysed using 6% SDS-PAGE.

7.14.1.2 Size-exclusion Spin-column Buffer Exchange

Eluted fractions containing significant amounts of protein were combined and buffer-exchanged with buffer A for cation exchange chromatography (20 mM HEPES-KOH pH 7.5, 100 mM KCl and 10% glycerol) using an Amicon® Ultra-4 Centrifugal Filter unit with a molecular weight cut-off of 100 kDa (Merck, cat. no. UFC810008). According to manufacturer's protocol, 4 mL of protein solution was loaded onto the spin column and centrifuged at 7,500 x g for 15 min at 4 °C. The supernatant was removed and the desired buffer was then added to the protein. Centrifugation and re-addition of the buffer was repeated 5–6 times to give the protein is in the desired buffer and volume (~100 µL) for subsequent purification.

7.14.1.3 Cation Exchange Chromatography

A 1 mL HiTrap® SP HP column (G.E. Healthcare, cat. no. GE29-0513-24) was installed onto the ÄKTA protein purification system (G.E. Healthcare) and was equilibrated with 10 mL buffer A (20 mM HEPES-KOH, pH 7.5, 100 mM KCl and 10% glycerol) at a flow rate of 2 mL/min and a pressure limit of 0.3 MPa. Next, the protein was loaded onto the column, washed with 10 mL buffer A and eluted over a 25 min gradient of increasing (0–60%) buffer B (20 mM HEPES-KOH, pH 7.5, 1 M KCl and 10% glycerol). The eluted protein was collected in a series of 1 mL fraction. All peak fractions with a ratio of absorbances at 260 nm and 280 nm (A_{260}/A_{280}) < 0.68 were analysed on SDS-PAGE. The protein was concentrated and buffer-exchanged into storage buffer (20 mM HEPES-KOH pH 7.5, 150 mM KCl, 1 mM DTT and 10% glycerol) using an Amicon® Ultra-4 Centrifugal Filter unit with a molecular weight cut-off of 100 kDa (Merck, cat. no. UFC810008).

7.14.2 AsCas12a

7.14.2.1 Single-step Purification

The cell pellet was resuspended in 50 mL Lysis buffer (50 mM HEPES pH 7.0, 200 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT and 10 mM imidazole) supplemented with 1 mg/mL lysozyme (Sigma) and one tablet of EDTA-free protease inhibitor cocktail (UltraCruz® Biotechnology, cat. no. sc-2913). Cells were then lysed by sonication (5 sec on, 5 sec off, 6 cycles unless stated otherwise in Figure captions). After centrifugation at 16,000 x g for 30 min at 4 °C was performed and the cell lysate

(25 mL supernatant) was incubated with 5 mL nickel-charged resin (Ni-NTA agarose, QIAGEN, cat. no. 30210) for 1 h at 4 °C. The mixture was then applied to the gravity-flow chromatography column and washed with wash buffer (50 mM HEPES pH 7.0, 2 M NaCl, 5 mM $MgCl_2$, 1 mM DTT and 10 mM imidazole). The protein was then eluted with 12 mL elution buffer (50 mM HEPES pH 7.0, 200 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT and 250 mM imidazole) and collected in a series of 1.5 mL fractions. Protein concentration was estimated by measuring the absorbance at 280 nm using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific). Note that all flow-through samples (non-binding fractions, wash steps and eluted protein) were kept and analysed using 10% SDS-PAGE. One eluted fraction containing the greatest amount of protein was buffer-exchanged with storage buffer (50 mM HEPES pH 7.0, 200 mM NaCl, 1 mM DTT and 0.1 mM EDTA) using Amicon® Ultra-0.5 mL Centrifugal Filter unit with a molecular weight cut-off of 100 kDa (Merck, cat. no. UFC510008). According to manufacturer's protocol, 0.5 mL of protein solution was loaded onto the spin column and centrifuged at 14,000 x g for 15 min at 4 °C. The supernatant was removed and the remaining protein was added followed by centrifugation. Next, the storage buffer was then added to the protein. Centrifugation and re-addition of the buffer was repeated 5–6 times until the protein was in the desired buffer and volume (~50 µL) for use in subsequent *in vitro* DNA cleave assays. Protein concentration was estimated by measuring the absorbance at 280 nm using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific).

7.14.2.2 Three-step Purification

Cobalt-immobilised metal affinity chromatography. The cell pellet was resuspended in 30 mL lysis buffer (50 mM Tris-HCl pH 7.5, 200 mM NaCl, 5 mM $MgCl_2$, 1 mM β-mercaptoethanol, 10 mM imidazole and 10% glycerol) supplemented with 1 mg/mL lysozyme (Sigma) and one tablet of EDTA-free protease inhibitor cocktail (UltraCruz® Biotechnology, cat. no. sc-2913). Cells were then lysed by sonication (1 min on, 2 min off, 6 cycles). After centrifugation at 20,000 rpm for 45 min at 4 °C, the supernatant (25 mL) was applied to the gravity-flow chromatography column packed with 1.5 mL cobalt-charged resin (HisPur™ Cobalt Resin, Thermo Fisher Scientific, cat. no. 89964) and washed with 15 mL wash buffer (50 mM Tris-HCl pH 7.5, 1 M NaCl, 5 mM $MgCl_2$, 1 mM TCEP and 20 mM imidazole). The protein was then eluted with 12 mL elution buffer (50 mM Tris-HCl

pH 7.5, 1 M NaCl, 5 mM $MgCl_2$, 1 mM TCEP and 250 mM imidazole). Protein concentration was estimated by measuring the absorbance at 280 nm using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific).

Maltose binding protein affinity chromatography. The protein in the wash fraction from Cobalt-IMAC purification was buffer-exchanged to MBP-binding buffer (25 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM $MgCl_2$ and 1 mM TCEP using an Amicon® Ultra-4 Centrifugal Filter unit with a molecular weight cut-off of 100 kDa (Merck, cat. no. UFC810008). The 5 mL MBPTrap™ HP column (G.E. Healthcare, cat. no. GE28-9187-79) was connected to the Peristaltic Pump P-1 (G.E. Healthcare) and was equilibrated with 20 mL binding buffer. The sample was loaded onto the column with the flow rate of 2.5 mL/min, washed with 50 mL binding buffer and eluted with 15 mL elution buffer (20 mM Maltose, 25 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM $MgCl_2$ and 1 mM TCEP. The protein was concentrated/buffer-exchanged to a gel filtration buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM $MgCl_2$ and 1 mM TCEP).

Gel filtration. The protein (~500 µL) was loaded onto a Superose™ 6 10/300 GL column (G.E. Healthcare, cat. no. 29091596) that was installed onto the ÄKTA protein purification system (G.E. Healthcare). The protein was eluted with 120 mL gel filtration buffer at a flow rate of 1 mL/min and collected in a series of 2 mL fraction. The protein was concentrated using an Amicon® Ultra-4 Centrifugal Filter unit with a molecular weight cut-off of 100 kDa (Merck, cat. no. UFC810008). Protein concentration was estimated by measuring the absorbance at 280 nm using a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific) and mixed with an equal volume of storage buffer (20 mM Tris-HCl pH 7.5, 500 M NaCl, 1 mM DTT and 80% v/v glycerol).

Note that all samples (non-binding fractions, wash steps and eluted proteins from each of the above steps) were kept and analysed using a 4–20% Mini-PROTEAN® TGX™ precast SDS-PAGE gel (Bio-Rad, cat. no. 456-1093).

7.15 SDS-PAGE

Unless stated SDS-PAGE gels were prepared as described here. Stacking and resolving gel recipes are shown below –

Reagent	Stacking gel (5%)	Resolving gel (6%)
Milli-Q H_2O	1.98 mL	5.85 mL
40% acrylamide:bis-acrylamide (29:1)	0.32 mL	1.55 mL
1.5 M Tris-HCl, pH 8.8	–	2.5 mL
2 M Tris-HCl, pH 6.8	0.15 mL	–
10% SDS	25 µL	0.2 mL
10% ammonium persulfate (APS)*	12.5 µL	80 µL
TEMED*	2.5 µL	8 µL

*Note that APS and TEMED were added to acrylamide mixture before casting.

After applying the resolving gel solution into the gel apparatus, butanol was gently added to the surface of the polyacrylamide mixture. Once the resolving gel was polymerised, the butanol was removed and the stacking gel solution was loaded. Samples (3 µL) was mixed with loading buffer (2x, 3 µL, 0.08% Bromophenol blue, 0.64 M Tris-HCl pH 6.8, 20% glycerol, 4% SDS and 0.2 M DTT) and heated for 5 min at 95 °C before cooling on ice and loading. The SDS-PAGE gel was run at 200 V for 45–60 min in 1× Tris-glycine running buffer (25 mM Tris-base, 200 mM glycine and 0.1% SDS). Gel was washed briefly with water prior to staining with Bio-Safe™ Coomassie G-250 stain (Bio-Rad, cat. no. 161-0786) for 16 h at room temperature with gentle shaking. Gel was destained with water and then visualised using Gel-Doc XR+ system (Bio-Rad).

7.16 Native PAGE

Native PAGE gels were prepared as described below –

Reagent	Percentage of gel	
	5%	8%
Milli-Q H_2O	8.2 mL	7.45 mL
40% acrylamide:bis-acrylamide (29:1)	1.25 mL	2 mL
10x TBE (1 M Tris-base, 1 M boric acid, 20 mM EDTA, pH 8.3)	0.5 mL	0.5 mL
10% ammonium persulfate (APS)*	80 µL	80 µL
TEMED*	8 µL	8 µL

*Note that APS and TEMED were added to acrylamide mixture before casting.

Native gels (W x D x H = 80 mm × 1 mm × 75 mm) were run in 0.5× TBE buffer supplemented with $MgCl_2$ (5 mM) and samples were mixed with 6x loading buffer (0.3% Bromophenol blue, 0.3% Xylene and 30% glycerol).

7.17 Electrophoretic Mobility Shift Assay

7.17.1 Cas9–fsgRNA

An example binding reaction was composed of fsgRNA (1 µM, 1 µL), Cas9 (1 µM, 1 µL, NEB, cat. no. M0386S), Cas9 buffer (10x, 1 µL, NEB), BSA (10 mg/mL, 1 µL) and water (6 µL). After incubation for 15 min at 37 °C, samples were loaded onto a 5% native PAGE gel and visualised using Syngene G:Box imager. IVT sgRNA-Competitor (2, 20 and 200 equivalents relative to fsgRNA) was added and the concentration of Cas9 was altered as specified in the figure caption.

7.17.2 dCas9-3xGFP–IVT sgRNA

An example binding reaction was composed of dCas9-3xGFP (2.07 µM, 2.9 µL), IVT sgRNA (10 µM, 0.6 µL), Cy3-labelled dsDNA (10 µM, 0.6 µL) and buffer (15.7 µL of 20 mM HEPES-KOH pH 7.5, 150 mM KCl, 1 mM DTT, 10% glycerol) and BSA (10 mg/mL, 0.2 µL). After complex formation for 30 min at 37 °C, the reactions were loaded onto an 8% native PAGE gel and visualised using

Syngene G:Box imager. Competitor RNA was added to a final concentration of 30 μM as specified in the figure caption.

7.18 Agarose Gel Electrophoresis

Agarose gels were prepared by melting 1% w/v of agarose in 1× TBE buffer (100 mM Tris-base, 100 mM boric acid, 2 mM EDTA, pH 8.3). After the addition of 0.5x SYBR™ Gold Stain (supplied as 10,000x, Thermo Fisher Scientific, cat. no. S11494), the gel was allowed to solidify. Samples were then mixed with Blue gel loading buffer (6×, NEB, cat. no. B7021S) and loaded onto the gel before running at 126 V for 45–60 min in 1× TBE running buffer. Gels were visualised using a Syngene G:Box Imager.

7.19 *In vitro* Cleavage Assays

7.19.1 Cas9

pBR322 plasmid DNA (0.35 μM, 1.13 μL, NEB, cat. no. N3033S) was diluted with water (16.87 μL) and NEB buffer 3.1 (10x, 2 μL). The plasmid was then linearised by the adding PvuII (10 U/μL, 1 μL, NEB, R0151S) for 1 h at 37 °C. For the Cas9 cleavage assay, sgRNA (300 nM, 5 μL), Cas9 (1 μM, 0.3 μL, NEB, cat. no. M0386S), Cas9 buffer (10x, 1 μL, NEB), linearised pBR322 (20 nM, 1.5 μL) and H_2O (2.2 μL) were mixed together (final volume = 10 μL) and incubated for 1 or 16 h at 37 °C. The reaction was then terminated by the addition of proteinase K (20 mg/mL, 0.5 μL) for 1 h at 37 °C. The reaction (1 μL) was analysed using an Agilent Bioanalyzer DNA 12000 kit (Agilent Technologies, cat. no. 5067-1508 for DNA chips and cat. no. 6067-1509 for reagents), where appropriate, according to the manufacturer's protocol. The remaining reaction was mixed with Blue gel loading buffer (6x, 2 μL, NEB, B7703S) and loaded on a 1% agarose stained with 0.5x SYBR gold (1x TBE running buffer, 126 V).

7.19.2 AsCas12a Lysate

Reactions contained AsCas12a lysate (5 μL), crRNA (100 μM, 3 μL), DNA cleavage template (20 μM, 0.75 μL), CutSmart buffer (10x, 2 μL, NEB), DTT (100 mM, 1 μL) and water (8.25 μL). After incubation for 20 min at 37 °C, samples were mixed with an equal volume of formamide and

loaded onto a 12.5% denaturing PAGE gel. Any variations in reaction composition are described in the figure caption. Gels were imaged using a Syngene G:Box imager and fluorophore specific colouring was applied using ImageJ. Overlaid images were generated by applying 50% transparency to one of the two layers in ImageJ.

7.19.3 Purified AsCas12a

Reactions contained AsCas12a (10 µM, 3 µL), crRNA (100 µM, 0.3 µL), DNA cleavage template 20 µM, 0.75 µL), NEBuffer 3 (10x, 1 µL, NEB), DTT (100 mM, 1 µL) and water (3.95 µL). After incubation for 20 min at 37 °C, samples were mixed with an equal volume of formamide and loaded onto a 12.5% denaturing PAGE gel. Any variations in reaction composition are described in the figure caption. Gels were imaged using a Syngene G:Box imager and fluorophore specific colouring was applied using ImageJ. Overlaid images were generated by applying 50% transparency to one of the two layers in ImageJ.

7.19.4 Commercial AsCas12a

Reactions contained AsCas12a (63.3 µM, 0.32 µL, IDT, cat. no. 1081068), crRNA (3.33 µM, 12 µL), DNA cleavage template (5 µM, 2 µL), CutSmart buffer (10x, 2 µL, NEB), DTT (100 mM, 1 µL) and water (2.68 µL). After incubation for 60 min at 37 °C, samples were mixed with an equal volume of formamide and loaded onto a 12.5% denaturing PAGE gel. Gels were imaged using a Syngene G:Box imager and fluorophore specific colouring was applied using ImageJ. Overlaid images were generated by applying 50% transparency to one of the two layers in ImageJ.

7.20 Real-time Fluorescence Measurements

Diluted Cas9 (10 µM, 1 µL, NEB, cat. no. M0386T) was added to RNase H-activated fgRNAs (10 µM, 0.4 µL) in the RNase H buffer (10x, 0.8 µL, NEB) and water (6.8 µL). After incubation for 2 min, RNase H (5 U/µL, 1 µL, NEB, cat. no. M0297S) was added. After 2 min, the fluorescence resulting from RNase H cleavage of the RNA-DNA heteroduplex was monitored using a CFX96™ Real-Time PCR Detection System (Bio-Rad) every minute for 119 min.

7.21 Cell Culture

Human embryonic kidney (HEK) cell line HEK293T and human osteosarcoma cell line U2OS were cultured in DMEM (high glucose, HEPES-buffered, no phenol red, with glutamine, Thermo Fisher Scientific, cat. no. 21063029) supplemented with 10% (v/v) FBS (Thermo Fisher Scientific, cat. no. 10270098) in a humidified incubator at 37 °C with 5% CO_2.

7.22 Transfection

7.22.1 Click sgRNA Gene Editing

1.5 x 10^5 cells were seeded in 6-well plates. After 48 h, the media was replaced with fresh DMEM-FBS media and the cells were transfected using a mixture of Lipofectamine™ 3000 (3.75 µL, Thermo Fisher Scientific, cat. no. L3000001), P3000 (10 µL, provided with the Lipofectamine 3000 reagent) and pSpCas9(BB)-2A-Puro-v2-Broccoli (1 µg/µL, 5 µL) in colourless Opti-MEM™ reduced serum media (250 µL, Thermo Fisher Scientific, cat. no. 11058021) that was pre-incubated for 12 min at room temperature. After 12 h, the media was replaced with fresh DMEM-FBS media and the cells were transfected with a mixture of Lipofectamine™ RNAiMAX (7.5 µL, Thermo Fisher Scientific, cat. no. 13778030) and the appropriate sgRNA (10 µM, 2.5 µL) in colourless Opti-MEM™ media (250 µL) that was pre-incubated for 5 min at room temperature. After 12 h, the media was replaced with fresh DMEM-FBS media. Cells were then allowed to recover and grow for a further 72 h with media replaced when necessary.

7.22.2 Live-cell Imaging

U2OS cells were seeded on a Nunc™ Lab-tek™ II 8-well chambered slide (Thermo Fisher Scientific, cat. no. 155409PK) at a density of 8.9 x 10^4 cells/well. After 24 h, the growth media was replaced with colourless Opti-MEM™ reduced serum media (400 µL) and transfected with a mixture of Lipofectamine™ 2000 Transfection reagent (1 µL, Thermo Fisher Scientific, cat. no. 11668030) and pHAGE-TO-dCas9-3xGFP (1 µg/µL, 0.5 µL, Addgene, plasmid number 64107 kindly gifted by Prof. Thoru Pederson[222]) in colourless Opti-MEM™ media (50 µL, Thermo Fisher Scientific, cat. no. 11058021) that was pre-incubated for 5 min at room temperature. The

transfection mixture was then added to the cells. After 6 h at 37 °C, the media was replaced with fresh Opti-MEM™ for subsequent gRNA transfection.

For RNA transfection, sgRNA working solutions were prepared (10 µM in H_2O) while for crRNA-tracrRNA constructs, the crRNA was hybridised to the tracrRNA in a hybridisation buffer (100 mM Potassium Acetate and 30 mM HEPES pH 7.5) to a final concentration of 10 µM. Cells were transfected with a mixture of Lipofectamine™ RNAiMAX Transfection reagent (0.45 µL, Thermo Fisher Scientific, cat. no. 13778030) and gRNA (10 µM, 1.15 µL) in colourless Opti-MEM™ media (50 µL) that was pre-incubated for 5 min at room temperature. The transfection mixture was then added to the cells. After 16 h at 37 °C the media was replaced by fresh DMEM-FBS media prior to confocal microscopy.

7.22.3 AsCas12a Protein Expression

1.5 x 10^5 cells were seeded in a 6-well plate. After 24 h, the DMEM-FBS media was replaced with fresh Opti-MEM™ media and the cells were transfected using a mixture of Lipofectamine™ 2000 (5 µL) and pcDNA3.1-hAsCpf1 (1 µg/µL, 2.45 µL, Addgene, plasmid number 69982 kindly gifted by Prof. Jin Soo Kim[245]) in colourless Opti-MEM™ media (250 µL) that was pre-incubated for 5 min at room temperature. The media was replaced with fresh DMEM-FBS media after 8 h. Cells were then allowed to grow for a further 16 h before Cas12a protein extraction.

7.22.4 Lipid Transfection of Ribonucleoprotein Complex

U2OS cells were seeded on a Nunc™ Lab-tek™ II 8-well chambered slide at a density of 8.9 x 10^4 cells/well. After 24 h, the growth media was replaced with colourless Opti-MEM™ media (400 µL, Thermo Fisher Scientific, cat. no. 11058021) and transfected with a mixture of Lipofectamine™ RNAiMAX (2 µL in 25 µL colourless Opti-MEM™ media) and the ribonucleoprotein complex (0.49 µL of 22.63 µM dCas9-3xGFP and 1.1 µL of 10 µM IVT sgRNA in 25 µL colourless Opti-MEM™ media) that was pre-incubated for 5 min at room temperature. The transfection mixture was then added to the cells. After 6 h at 37 °C, the media was replaced with fresh DMEM-FBS media and the cells were allowed to grow for further 18 h prior to confocal microscopy.

7.22.5 Electroporation of Ribonucleoprotein Complex

Electroporation of dCas9-3xGFP–IVT sgRNA RNP complex was performed using an Amaxa® Cell Line Nucleofector® Kit V (Lonza, cat. no. VCA-1003) according to manufacturer's protocol. 1 x 10^6 cells were carefully resuspended in room-temperature Nucleofector® Solution (100 µL). The RNP complex (112 pmol, 5 µL dCas9-3xGFP and 135 pmol, 2.36 µL IVT sgRNA pre-mixed) was added to the 100 µL cell suspension and the mixture was transferred into the certified cuvette (provided with the kit). The cuvette was inserted into Amaxa™ Nucleofector™ II and electroporation was performed using the programme X-001. Pre-warmed DMEM media (500 µL) was immediately added to the cuvette and the sample (75 µL) was gently transferred to the DMEM-containing 8-well plate containing pre-warmed FBS-containing DMEM media (375 µL). Cells were then incubated for 16 h in a humidified 37 °C with 5% CO_2 incubator. After replacing the media with fresh DMEM, confocal microscopy was performed.

7.23 AsCas12a Cell Lysate Preparation

24h after transfection, HEK293T cells were washed with DPBS (Thermo Fisher Scientific, cat. no. 14190136) and detached from the plates using trypsin. After neutralising the trypsin with FBS-containing DMEM, the cell suspension was centrifuged at 1,000 rpm for 5 min at room temperature. The supernatant was then removed and the cell pellet was washed with DPBS (5 mL). Following centrifugation at 1,000 rpm for 5 min at room temperature and the removal of the supernatant, the cell pellet was resuspended in 300 µL lysis buffer (20 mM HEPES pH 7.5, 100 mM KCl, 1 mM DTT, 5% glycerol and 0.1% Triton X-100) supplemented with one tablet of EDTA-free protease inhibitor cocktail (UltraCruz® Biotechnology, cat. no. sc-2913). Cells were then lysed by sonication (30 sec on, 30 sec off, 2-10 cycles) using a Bioruptor® plus (Diagnode). Next, the cell lysate was centrifuged at 13,000 rpm for 15 min at 4 °C and the supernatant was kept for subsequent use for *in vitro* DNA cleavage assays.

7.24 Confocal Microscopy

Live-cell imaging was performed using Zeiss LSM 780 confocal microscope system with a 63x 1.4 Numerical Aperture (NA) oil immersion objective at 37 °C with 5% CO_2 (Zeiss Ltd). The samples

were excited using internal lasers and emission detected as listed below. Different channels were imaged using sequential excitation and detection of emission. Images were then processed using ImageJ, with overlaid images generated by applying 50% transparency on one layer.

For dCas9-3xGFP, GFP was excited at 488 nm and the emission recorded from 493–542 nm. For TAMRA and Cy3 containing oligonucleotides, the dyes were excited at 543 nm and the emission recorded from 548–604 nm. For ATTO 647N containing oligonucleotides, ATTO 647N was excited at 633 nm and the emission recorded from 639–742 nm. For GFP to Cy3 FRET, GFP was excited at 488 nm and the emission of Cy3 recorded from 590–650 nm.

7.25 Genomic DNA Extraction

Cells were washed with PBS, trypsinised and pelleted (200 rpm, 5 min). Genomic DNA was extracted from the cell pellet using the PureLink® genomic DNA mini kit (Invitrogen, cat. no. K1820-00). Briefly, the cells were resuspended in PBS (200 µL) and proteinase K (20 µL) added for 2 min. Next, RNase A (20 µL) was added followed by PureLink® genomic Lysis/Binding buffer (200 µL). The sample was then heated for 10 min at 55 °C before the addition of ethanol (200 µL). After applying the sample to the spin column and centrifuging (10,000 rpm, 1 min), the column was washed with Wash I (500 µL), centrifuged (10,000 rpm, 1 min), washed with Wash II (500 µL), and centrifuged (10,000 rpm, 1.5 min). Genomic DNA was then eluted using Elution buffer (30 µL) by centrifuging (13,300 rpm, 1min).

7.26 T7E1 Assay

PCR amplicons from genomic DNA were produced using primers listed in **Appendix Table 25**. Reactions contained genomic DNA (100 ng, 5 µL), forward primer (10 µM, 2.5 µL), reverse primer (10 µM, 2.5 µL), dNTPs (10 mM, 1 µL), Phusion Hot Start Flex polymerase (2 U/µL, 0.5 µL, NEB, cat. no. M0535S) and Phusion HF buffer (5x, 5 µL, NEB) and water (33.5 µL). The thermal cycling conditions consisted of denaturation (30 sec, 98 °C), amplification (34 cycles of 5 sec, 98 °C; 20 sec, 60 °C and 20 sec, 72 °C) and final extension (5 min, 72 °C). The amplicon was subsequently

purified using AMPure XP beads (90 µL, 1.8x beads to DNA ratio, Beckman Coulter, cat. no. A63880) according to the manufacturer's instructions.

The purified amplicons (200 ng, 17 µL) were mixed with NEBuffer 2 (10x, 2 µL, NEB) and annealed by heating for 5 min at 95 °C and cooling to 85 °C (ramp rate = –2 °C/sec) and then 25 °C (ramp rate = –0.1 °C/sec). Hybridised fragments were digested with T7 endonuclease I (10 U/µL, 1 µL, NEB, cat. no. M0302S) for 15 min at 37 °C before the addition of proteinase K (20 mg/mL, 1 µL) for 1 h at 37°C. The reaction (1 µL) was analysed using an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies, cat. no. 5067-1504) and the remaining reaction was mixed with Blue gel loading buffer (6x, 4 µL, NEB cat. no. B7703S) and loaded on a 2% agarose gel stained with 0.5x SYBR gold (1x TBE running buffer, 126 V).

7.27 Nanopore Sequencing

7.27.1 Circularisation of DNA Template

5'-phosphorylated DNA template (129 µM, 3.1 µL, **Appendix Table 21**), RAM DNA splint (95.24 µM, 2.1 µL, **Appendix Table 21**), T4 DNA ligase (3 U/µL, 0.66 µL, Promega, cat. no. M1801), T4 DNA ligase reaction buffer (10x, 2 µL, NEB with ATP) and water (12.14 µL) were mixed together and incubated for 1 h at 37 °C. Half of the reaction mixture was used for RAM amplification while the other half was treated with exonuclease I (20 U/µL, 0.4 µL, NEB, cat. no. M0293S) and exonuclease III (10 U/µL, 0.8 µL, NEB, cat. no. M0206S) for 1 h at 37 °C. The exonucleases were heat-inactivated for 5 min at 65 °C before loading onto an 8% denaturing PAGE gel to confirm cyclisation of the template.

7.27.2 Ramified Rolling Circle Amplification

The RAM reaction contained circularised DNA template (20 µM assuming full cyclisation, 10 µL), additional primer P1 (7.5 µM, 2 µL, **Appendix Table 21**), primer P2 (10 µM, 2 µL, **Appendix Table 21**), Bst DNA polymerase Large fragment (8 U/µL, 0.325 µL, NEB, cat. no. M0275S), dNTPs (10 mM, 1.6 µL), 2% DMSO, ThermoPol® Reaction Buffer (10x, 2 µL, NEB) and water (2.08 µL). The reaction proceeded for 1 h at 60 °C before analysis on a 1.2% agarose gel containing 0.5x SYBR gold.

7.27.3 Short Tandem Repeat PCR

Genomic DNA (Promega, cat. no. DD7101) was PCR amplified using forward and reverse primers listed in **Appendix Table 22**. The PCR reaction contained template DNA (0.25 ng/µL, 4 µL), forward primer (10 µM, 0.5 µL), reverse primer (10 µM, 0.5 µL), AmpliTaq Gold® Polymerase (0.05 U/µL, 1.25 µL, Thermo Fisher Scientific, cat. no. 4311806), dNTP (10 mM, 1 µL), $MgCl_2$ (25 mM, 1.5 µL), Gold buffer (10x, 2.5 µL, Thermo Fisher Scientific) and water (13.75 µL). After heat activation (11 min, 95 °C) and 10 cycles of denaturation (30 sec, 94 °C), annealing (30 sec, 60 °C) and extension (45 sec, 70 °C), a further 22 cycles of denaturation (30 sec, 90 °C), annealing (30 sec, 60 °C) and extension (45 sec, 70 °C) were performed prior to final extension (30 sec, 60 °C). The samples were cooled for 30 min at 4 °C before analysis on a 1.2% agarose gel containing 0.5x SYBR gold.

7.27.4 Library Preparation for Sequencing

The Oxford Nanopore MinION™ Genomic DNA Sequencing Kit, SQK-MAP004, was used to prepare the DNA library with slight modifications to the manufacturer's instructions. 10 µg of DNA quantified by a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific) was sheared in a g-Tube (Covaris, cat. no. 520079) in a total volume of 80 µL by centrifuging at 6,000 rpm for 1 min. This theoretically fragments DNA to ca. 8–10 kbp in length. The fragmented DNA was end-repaired in a reaction containing sheared DNA (80 µL), DNA CS (5 µL, provided with the SQK-MAP004 kit), NEBNext End Repair module (5 µL, NEB, cat. no. E6050S) and NEBNext End Repair buffer (10x, 10 µL, NEB) for 20 min at room temperature. Then the blunt-ended DNA was cleaned up using 1x volume Agencourt AMPure XP beads (100 µL, Beckman Coulter, cat. no. A63880), washed twice with 70% ethanol (200 µL) and eluted with water (25 µL). dA-tailing was performed in a reaction containing End-repaired DNA (25 µL), NEBNext dA-tailing module (2 µL, NEB, cat. no. E6053S) and NEBNext dA-tailing buffer (10x, 3 µL, NEB) for 30 min at 37 °C. Next, adapter ligation was performed in a reaction containing dA-tailed DNA (30 µL), Adapter mix (10 µL, provided with the SQK-MAP004 kit), HP adapter (2 µL, provided with the SQK-MAP004 kit), Blunt/TA Ligase Master Mix (2x, 50 µL, NEB, M0367S) and water (8 µL) for 30 min at room temperature. For clean up of adapter-ligated products, Dynabeads™ (10 µL, Thermo Fisher Scientific, cat. no. 10103D) that were washed with and resuspended in Wash buffer (2x, 100 µL, provided with the SQK-

MAP004 kit) were added to the adapter-ligation mixture (100 µL). After incubation for 5 min at room temperature, the beads were washed twice with diluted Wash buffer (1x, 250 µL, provided with the SQK-MAP004 kit) and the DNA was then eluted with Elution buffer (25 µL, provided with the SQK-MAP004 kit). Elution proceeded for 10 min at room temperature. The eluted DNA library was quantified by a NanoDrop™ 2000 Spectrophotometer (Thermo Scientific) before sequencing on a MinION™ nanopore sequencer (Oxford Nanopore Technologies).

7.27.5 Flowcell Preparation

The MinION™ device with the flowcell (FLO-MAP003) placed on top was connected to a laptop. The platform QC run was performed by selecting the MAP_Platform_QC.py in the MinION control software, MinKNOW version 0.49.2.9, to determine the number of active pores available for sequencing. Simultaneously, the mixture of Fuel mix (6.5 µL, provided with the SQK-MAP004 kit) and EP buffer (318.5 µL, provided with the SQK-MAP004 kit) was prepared. The mixture (150 µL) was loaded into the sample loading port using a P1000 pipette and left for 10 min for priming. This priming process was repeated once more.

7.27.6 Sample Loading

The DNA library (23.1 ng/µL, 6 µL) was mixed with EP buffer (141 µL, provided with the SQK-MAP004 kit) and Fuel mix (3 µL, provided with the SQK-MAP004 kit). The mixture was then loaded into the flowcell *via* the sample loading port. Sequencing was performed using a 6 h MAP_lambda_burn_in_Sequencing_Run.py protocol.

7.27.7 Data Analysis

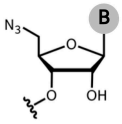
The MinKNOW software detected abrupt signal changes when five nucleotides passed through the nanopore and grouped them into reads. One read represented one strand of DNA translocating through the nanopore. Read event data were subsequently base-called by the Metrichor Agent software version 2.26 using 2D base calling workflow. Read data were generated as .fast5 files and classified into two folders: pass folder and fail folder. 2D reads in the pass folder were used in subsequent analyses with statistical information on the folder obtained using poretools[283]. .fast5 format files were converted into .fasta format files using poretools[283].

Sequences were then aligned to a 'reference' sequence using LALIGN[284] (version 36.3.7, parameters -r 5/-4 -g -12 -f -4 -E 2 -m 9i –n) which looks for internal duplications of the reference sequence. These reference sequences are primers for PCR and the known sequence around the STR for RAM products. For PCR, only reads with both forward and reverse primer regions were selected to extract the 'unknown' sequence between the primers. For RAM, only reads with multiple 'reference' regions were selected and the 'unknown' sequence between two 'reference' sequences was extracted. The 'unknown' sequences were then subject to multiple alignment using MUSCLE[285] (version 3.8.31) with default parameters. Frequency analysis was performed on the multiply-aligned sequences using Weblogo, resulting in a sequence logo that represents the most probable sequence of STR. All the steps from 2D read selection to frequency analysis were automated using the code written in MatLab 2017a by Dr. Arun Shivalingam.

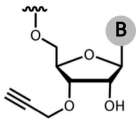
Appendix

Appendix Table 1 | Oligonucleotides used in optimisation of RNA CuAAC ligation. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
5407	Opt-azide (DNA)	5'-N_3-tagcacacaatctcacactctggaattcacactgacaatacttcagacaccagcacac	17645	17646
5471	Opt-alkyne (DNA)	ccgggagtccataagaagatcU-propargyl	6809	6808
6835	Opt-azide	5'-N_3-CAUCAGCACCAAGCGAUCGCCAUGGCUUCCAA	10189	10190
6833	Opt-alkyne	GGCAAUUUUAACUUUACUAAGGAGAAUUCACCAUGAAACAUCAA-propargyl	14087	14089
6837	DNA splint	gcgatcgcttggtgctgatgttgatgtttcatggtgaatt	12410	12411



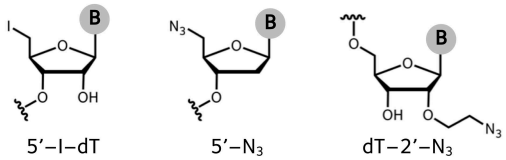
5'–N_3



propargyl

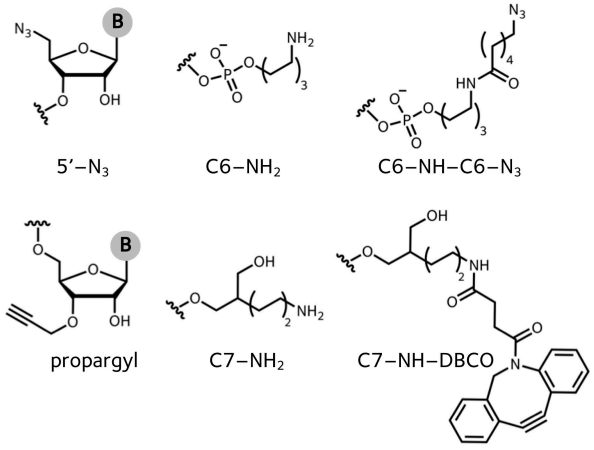
Appendix Table 2 | 2'-TC RNA and DNA oligonucleotides used for testing conversion of 5'-hydroxyl to 5'-azide. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
5742	TC-tracrRNA	CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	21200	21202
6205	AzideOpt DNA	5'-N_3-ggttttttttt	3055	3055
6313	AzideOpt TC-RNA	CAUCGAGUCUC-dT-2'-N_3	3816	3814
7550	IodoOpt TC-RNA	UUUUUUUUUUU	3304	3305
7267	5'-iodo-TR1-TC	5'-I-dT-UUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	25520	25517



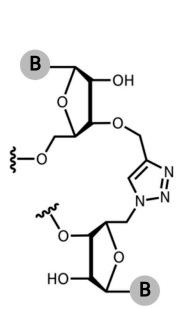
Appendix Table 3 | Azide and alkyne oligonucleotides used to generate click 37–66 sgRNA constructs. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
5938	crRNA-I	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAGAG CUAGA-propargyl	11999	11999
5743	crRNA-IIa	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAGAG CUAGA-C7-NH_2	12168	12170
5743-DBCO	crRNA-IIb	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAGAG CUAGA-C7-NH-DBCO	12456	12458
7121	tracrRNA-IIIa	NH_2-C6-CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC AACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	21378	21381
7121-LgN_3	tracrRNA-IIIb	N_3-C6-NH-C6-CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC AACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	21521	21523
6290	tracrRNA-IV	5'-N_3-CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC AACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	21225	21227
5941	DNA splint	tttaacttgctatttctagctctaaaac	8487	8487

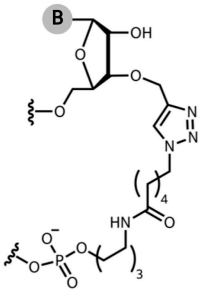


Appendix Table 4 | Click 37–66 sgRNA constructs and control IVT sgRNA. Uppercase bases are RNA nucleotides and the structures of modifications are listed at the bottom of the table. The precursor oligonucleotides for click ligation are listed in **Appendix Table 3** and the DNA templates for IVT sgRNA are in **Appendix Table 8**. ppp = triphosphate.

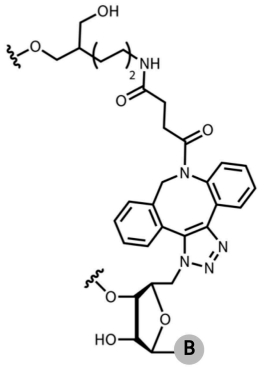
sgRNA Name	Oligo. Used	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
Construct 1	crRNA-I + tracrRNA-IV	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAG AGCUAGA-**Tz2**-CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUA UCAACUUGAAAAAGUGGCACCGAGUCGGUGC UUUU	33224	33225
Construct 2	crRNA-I + tracrRNA-IIIb	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAG AGCUAGA-**Linker 2**-UAGCAAGUUAAAAUAAGGCUAGUCCGUUAUC AACUUGAAAAAGUGGCACCGAGUCGGUGCUU UU	33518	33518
Construct 3	crRNA-IIb + tracrRNA-IV	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAG AGCUAGA-**Linker 3**-CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUA UCAACUUGAAAAAGUGGCACCGAGUCGGUGC UUUU	33681	33684
Construct 4	crRNA-IIb + tracrRNA-IIIb	GGGCGCUUGUUUCGGCGUGGGUAGUUUUAG AGCUAGA-**Linker 4**-CAUAGCAAGUUAAAAUAAGGCUAGUCCGUUA UCAACUUGAAAAAGUGGCACCGAGUCGGUGC UUUU	33977	33980
In vitro IVT sgRNA 1	*In vitro* sgRNA Template 1	pppGGGCGCUUGUUUCGGCGUGGGUAGUUU UAGAGCUAGACAUAGCAAGUUAAAAUAAGGC UAGUCCGUUAUCAACUUGAAAAAGUGGCACC GAGUCGGUGCUUUU	33464	33467



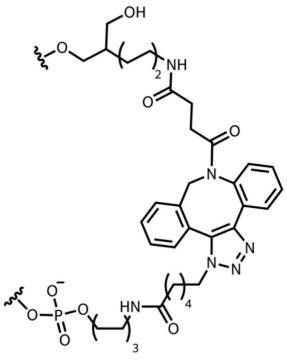
Tz2



Linker 2



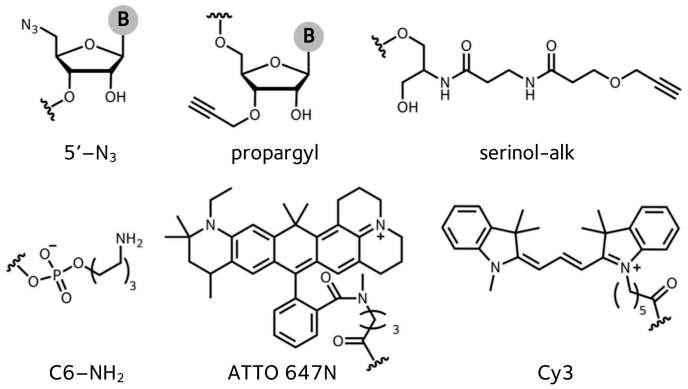
Linker 3



Linker 4

Appendix Table 5 | Azide and alkyne oligonucleotides used to generate click ~21–79 sgRNA constructs. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
7263-OH	CR1-OH	GGGCGCUUGUUUCGGCGUGGGUAG-propargyl	7811	7811
7317-NH_2	CR1-NH_2	NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-propargyl	7989	7991
7317-ATTO	CR1-ATTO	ATTO 647N-NH-C6-GGGCGCUUGUUUCGGCGUGGGUAG-propargyl	8618	8617
7317-Cy3	CR1-Cy3	Cy3-NH-C6-GGGCGCUUGUUUCGGCGUGGGUAG-propargyl	8431	8429
7760	CR2-LgAlk	NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-serinol-alk	8289	8287
7500	CR3- NH_2	NH_2-C6-GAGUCCGAGCAGAAGAAGAAG-propargyl	7100	7100
7499-NH_2	CR4- NH_2	NH_2-C6-UAGGGUUAGGGUUAGGGUUAG-propargyl	7066	7067 (crude)
7499-Cy3	CR4-Cy3	Cy3-NH-C6-UAGGGUUAGGGUUAGGGUUAG-propargyl	7506	7508
7499-ATTO	CR4-ATTO	ATTO 647N-NH-C6-UAGGGUUAGGGUUAGGGUUAG-propargyl	7695	7695
7385	TR1	5'-N_3-UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	25436	25437
7469	TR2	5'-N_3-UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAActtgaaaaaGtGgcaccGagtcggtgCtttt	25018	25021

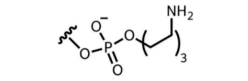


5′–N_3 propargyl serinol-alk

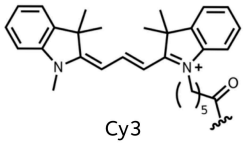
C6–NH_2 ATTO 647N Cy3

Appendix Table 6 | Clicked ~21–79 sgRNAs and IVT sgRNA for *in vitro* assays. Uppercase bases are RNA nucleotides lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. The precursor oligonucleotides for click ligation are listed in **Appendix Table 5** and the DNA templates for IVT sgRNA are in **Appendix Table 8**. ppp = triphosphate.

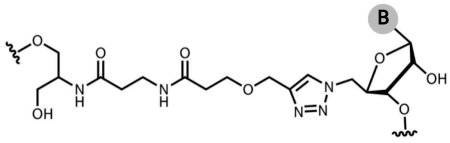
sgRNA Name	Oligo. Used	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
In vitro sgRNA-OH	CR1-OH + TR1	GGGCGCUUGUUUCGGCGUGGGUAG-**Tz2**-UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33247	33246
In vitro sgRNA-DNA-OH	CR1-OH + TR2	GGGCGCUUGUUUCGGCGUGGGUAG-**Tz2**-UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAActtgaaaaaGtGgcaccGagtcggtgCtttt	32829	32828
In vitro sgRNA-NH_2	CR1-NH_2 + TR1	NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-**Tz2**-UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33425	33426
In vitro sgRNA-DNA-NH_2	CR1-NH_2 + TR2	NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-**Tz2**-UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAACttgaaaaaGtGgcaccGagtcggtgCtttt	33007	33009
In vitro sgRNA-Cy3	CR1-Cy3 + TR1	Cy3-NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-**Tz2**-UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33867	33871
In vitro sgRNA-DNA-Cy3	CR1-Cy3 + TR2	Cy3-NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-**Tz2**-UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAACttgaaaaaGtGgcaccGagtcggtgCtttt	33446	33447
In vitro LongTz-sgRNA-NH_2	CR2-LgAlk + TR1	NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-**LongTz**-UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33725	33723
In vitro LongTz-sgRNA-DNA-NH_2	CR2-LgAlk + TR2	NH_2-C6-GGGCGCUUGUUUCGGCGUGGGUAG-**LongTz**-UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAACttgaaaaaGtGgcaccGagtcggtgCtttt	33307	33309
In vitro IVT sgRNA 2	*In vitro* sgRNA Template 2	pppGGGCGCUUGUUUCGGCGUGGGUAGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33540 (+2xMg)	33547



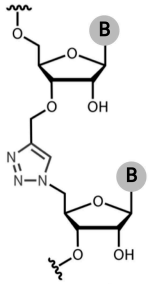
C6–NH_2



Cy3



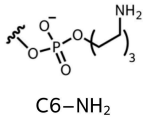
LongTz



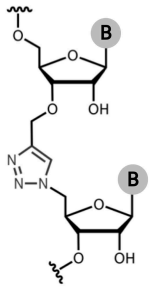
Tz2

Appendix Table 7 | Clicked ~21–79 sgRNAs and IVT sgRNA to target the *EMX1* gene in cell culture assays. Uppercase bases are RNA nucleotides lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. The precursor oligonucleotides for click ligation are listed in **Appendix Table 5** and the DNA templates for IVT sgRNA are in **Appendix Table 8**. ppp = triphosphate.

sgRNA Name	Oligo. Used	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
EMX1 sgRNA-NH_2	CR3-NH_2 + TR1	NH_2-C6-GAGUCCGAGCAGAAGAAGAAG-**Tz2**-UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	32536	32538
EMX1 sgRNA-DNA-NH_2	CR3-NH_2 + TR2	NH_2-C6-GAGUCCGAGCAGAAGAAGAAG-**Tz2**-UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAActtgaaaaaGtGgcaccGagtcggtgCtttt	32118	32119
EMX1 IVT sgRNA	*EMX1* sgRNA Template	pppGGAGUCCGAGCAGAAGAAGAAGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33020 (+3xMg)	33008



C6–NH_2



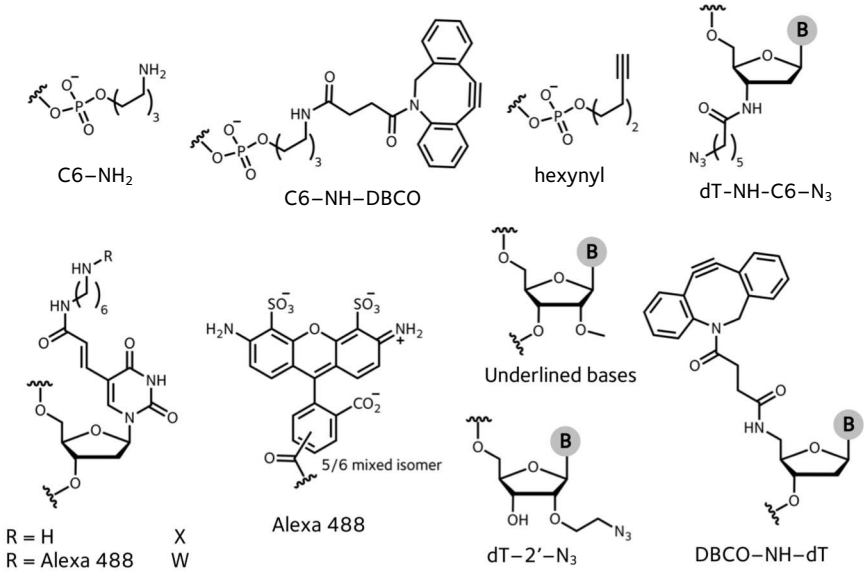
Tz2

Appendix Table 8 | DNA templates for IVT generation of sgRNAs for gene editing. Lower case bases are DNA nucleotides.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
5395	IVT Template Comple-ment-1	tctaatacgactcactatag	6060	6059
4752	IVT Template Comple-ment-2	gcggcctctaatacgactcactataggg	8574	8572
7233	*In vitro* sgRNA Template 1	aaaagcaccgactcggtgccacttttcaagttgataacggactag ccttattttaacttgctatgtctagctctaaaactacccacgccgaaa caagcgccctatagtgagtcgtattagaggccgc	39314	39316
7248	*In vitro* sgRNA Template 2	aaaagcaccgactcggtgccacttttcaagttgataacggactag ccttattttaacttgctatttctagctctaaaactacccacgccgaaac aagcgccctatagtgagtcgtattagaggccgc	39289	39289
7515	*EMX1* sgRNA Template	aaaagcaccgactcggtgccacttttcaagttgataacggactag ccttattttaacttgctatttctagctctaaaacttcttcttctgctcgga ctcctatagtgagtcgtattaga	36791	36790

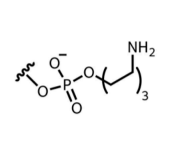
Appendix Table 9 | Oligonucleotides used to generate fsgRNA I and test DNA oligonucleotides to check ligation. Uppercase bases are RNA nucleotides, lowercase bases are DNA nucleotides and underlined bases are 2'-OMe modified nucleotides. The structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name. X = NH_2-C6-dT. W = Alexa 488-NH-C6-dT.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6014	f-crRNA Ia	UAG GGUUAGGGUUAGGGUUAGUUUUAGAGCUAAG C-dT-2'-N_3	11774	11775
6015	f-tracrRNA Ia	NH_2-C6-UAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCA ACUUGGCACCGAGUCGGUGCU	17228	17228
6015-DBCO	f-tracrRNA Ib	DBCO-NH-C6-UAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCA ACUUGGCACCGAGUCGGUGCU	17578 (+Cu)	17578
6013	f-DNA loop Ia	hexynyl-XacgcgtgtcacacgtattgXgcgtacgtgacgatatgcgXa gcatatcgXcacgtacgcacaatacgtgXgacacgcgtaagc -dT-2'-N_3	27235	27235
6013-Alexa 488	f-DNA loop Ib	hexynyl-WacgcgtgtcacacgtattgWgcgtacgtgacgatatgcgW agcatatcgWcacgtacgcacaatacgtgWgacacgcgtaa gc-dT-2'-N_3	29818	29818
5871-N_3	N_3-DNA	tagaggaagacagcctcgatc-dT-NH-C6-N_3	6907	6906
5873-DBCO	DBCO-DNA	DBCO-NH-dT-agcacacaatctcacactctggaattcacactgacaatactgcaa gacacaccactc	17907	17906
6044	DNA splint	gagattgtgtgctaagatcgaggctgtc	8700	8701

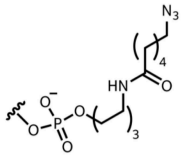

C6–NH2
C6–NH–DBCO
hexynyl
dT-NH-C6–N3
R = H
R = Alexa 488
X
W
Alexa 488
5/6 mixed isomer
Underlined bases
dT–2'–N3
DBCO–NH–dT

Appendix Table 10 | Oligonucleotides used to generate fsgRNA II-TAMRA and fsgRNA II-ATTO and the IVT sgRNA-Competitor for EMSA. Uppercase bases are RNA nucleotides, lowercase bases are DNA nucleotides and underlined bases are 2'-OMe modified nucleotides. The structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name. X = NH_2-C6-dT. Y = TAMRA-NH-C6-dT. Z = ATTO 647N-NH-C6-dT.

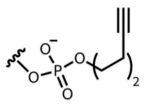
res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6310	f-crRNA IIa	UAGGGUUAGGGUUAGGGUUAGUUUUAGAGCU ACGCGCAUGCAACGCGUU-C7-NH_2	16080	16078
6310-N_3	f-crRNA IIb	UAGGGUUAGGGUUAGGGUUAGUUUUAGAGCU ACGCGCAUGCAACGCGUU-C7-NH-C6-N_3	16220	16217
6311	f-tracrRNA IIa	NH_2-C6-GCGCGUAGCAAGUUAAAAUAAGGCUAGUCCGU UAUCAACUUGGCACCGAGUCGGUGCU	18876	18874
6311-N_3	f-tracrRNA IIb	N_3-C6-NH_2-C6-GCGCGUAGCAAGUUAAAAUAAGGCUAGUCCGU UAUCAACUUGGCACCGAGUCGGUGCU	19016	19014
6263	f-DNA loop IIa	Hexynyl-tcXcacgtattgagcgtacgtgXcgatatgctXtgcatatcgac acgtacgcXcaatacgtgagaaacgcgtXgca-serinol-alk	24673	24674
6263-TAMRA	f-DNA loop IIb TAMRA	Hexynyl-tcYcacgtattgagcgtacgtgYcgatatgctYtgcatatcgac acgtacgcYcaatacgtgagaaacgcgtYgca-serinol-alk	26736	26736
6263-ATTO 647N	f-DNA loop IIb ATTO 647N	Hexynyl-tcZcacgtattgagcgtacgtgZcgatatgctZtgcatatcgac acgtacgcZcaatacgtgagaaacgcgtZgca-serinol-alk	27814	27813
5466	IVT template-Competitor	aaaagcaccgactcggtgccacttttttcaagttgataacggact agccttattttaacttgctatttctagctctaaaacgcgcttcgtta atacagatgccctatagtgagtcgtattagaggccgc	39356	39356
–	IVT sgRNA-Competitor	pppGGGCAUCUGUAUUAACGAAGCGCGUUUUA GAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAG UCCGUUAUCAACUUGAAAAAGUGGCACCGAGU CGGUGCUUUU	33522 (+3xMg)	33509



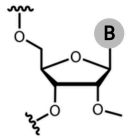
C6–NH_2



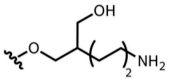
C6–NH–C6–N_3



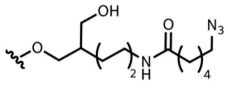
hexynyl



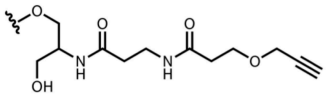
Underlined bases



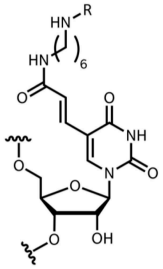
C7–NH_2

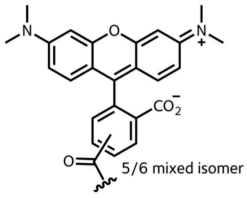


C7–NH–C6–N_3

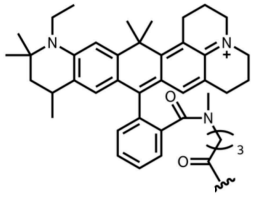


serinol-alk





TAMRA



ATTO 647N

R = H X
R = TAMRA Y
R = ATTO 647N Z

Appendix Table 11 | DNA templates for IVT generation of sgRNAs for live-cell imaging. Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name.

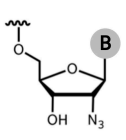
res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6519	IVT Template S1	agcaccgactcggtgccaagttgataacggactagccttattttaacttgctatttctagctctaaaactaaccctaaccctaaccctaccctatagtgagtcgtattaga	33967	33970
6491	IVT Template S2	aaaagcaccgactcggtgccacttttttcaagttgataacggactagccttattttaacttgctatttctagctctaaaactaaccctaaccctaacctaccctatagtgagtcgtattaga	37319	37324
6647	IVT Template S3	aaaagcaccgactcggtgccacttttttcaagttgataacggactagccttattttaacttgctatttctagctctaaaactaaccctaaccctaacctatagtgagtcgtattaga	35834	35836
6648	IVT Template S4	aaaagcaccgactcggtgccacttttttcaagttgataacggactagccttattttaacttgctatgctgtttccagcatagctctaaaactaaccctaaccctaaccctatagtgagtcgtattaga	38923	38926
6649	IVT Template S5	aaaagcaccgactcggtgccaagttgataacggactagccttatttttaacttgctatgctgtttccagcatagctctaaaactaaccctaaccctaaccctatagtgagtcgtattaga	36511	36513
6650	IVT Template S6	aaaagcaccgactcggtgccaagttgataacggactagccttatttttaacttgctacgcgctttcgcgcgtagctctaaaactaaccctaaccctaaccctatagtgagtcgtattaga	36513	36514
6809	IVT TemplateM 1	aaaagcaccgactcggtgccacttttttcaagttgataacggactagccttattttaacttgctatgctgatcgaacgcgatactgagtctttcgactcagaatcgcgttcgaacagcatagctctaaaactaaccctaaccctaaccctatagtgagtcgtattaga	51300	51304
6810	IVT TemplateM 2	aaaagcaccgactcggtgccacttttttcaagttgataacggactagccttattttaacttgctatgctgtttttaatcgaacgcgatactgagtctttcgactcagaatcgcgttcgaataaaaacagcatagctctaaaactaaccctaaccctaaccctatagtgagtcgtattaga	55005	55008
6811	IVT TemplateM 3	aaaagcaccgactcggtgccacttttttcaagttgataacggactagccttattttaacttgctatgctgtttttacgtgaatcgaacgcgatactgagtctttcgactcagaatcgcgttcgaatcacgtaaaaacagcatagctctaaaactaaccctaaccctaaccctatagtgagtcgtattaga	58093	58094

Appendix Table 12 | IVT sgRNAs used for live-cell imaging. Uppercase bases are RNA nucleotides and were generated from *in vitro* transcription of DNA templates in **Appendix Table 11**. ppp = triphosphate. * = untemplated addition of either C or U nucleotide (cannot be distinguished by mass spectrometry alone due to similar molecular weight. This is a known phenomenon in literature[147]).

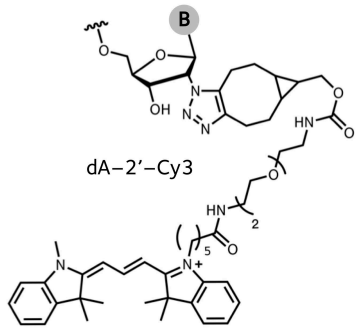
IVT Product Name	IVT Template Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
IVT sgRNA S1	IVT Template S1	pppGGGUAGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGGCACCGAGUCGGUGCU	30091 (+2xMg)	30100
IVT sgRNA S2	IVT Template S2	pppGGGUAGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33677 (+3xMg)	33663
IVT sgRNA S3	IVT Template S3	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	31981 (+2xMg)	31970
IVT sgRNA S4	IVT Template S4	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUGCUGGAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	35228 (+3xMg)	35213
IVT sgRNA S5	IVT Template S5	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUGCUGGAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGGCACCGAGUCGGUGCUUUU	32561 (+2xMg)	32549
IVT sgRNA S6	IVT Template S6	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUACGCGCGAAAGCGCGUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGGCACCGAGUCGGUGCUUUU	32591 (+2xMg)	32598
IVT sgRNA M1	IVT TemplateM 1	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUGCUGUUCGAACGCGAUUCUGAGUCGAAAGACUCAGUAUCGCGUUCGAUCAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	48367 (+C* +4xMg)	48368
IVT sgRNA M2	IVT TemplateM 2	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUGCUGUUUUUAUUCGAACGCGAUUCUGAGUCGAAAGACUCAGUAUCGCGUUCGAUUAAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	52180 (+C* +4xMg)	52176
IVT sgRNA M3	IVT TemplateM 3	pppGGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUGCUGUUUUUACGUGAUUCGAACGCGAUUCUGAGUCGAAAGACUCAGUAUCGCGUUCGAUUCACGUAAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	55377 (+C* +3xMg)	55367

Appendix Table 13 | Cy3-labelled IVT sgRNA used for live-cell imaging. Uppercase bases are RNA nucleotides and were generated from IVT sgRNA S2 from **Appendix Table 12**. The structures of the modifications are listed at the bottom of the table.

IVT Product Name	IVT sgRNA Precursor	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
N_3-IVT sgRNA	IVT sgRNA S2	pppGGGUAGGGUUAGGGUUAGGGUUAGUUUUA GAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGU CCGUUAUCAACUUGAAAAAGUGGCACCGAGUCG GUGCUUUU-dA-2'-N_3	34013 (+3xMg)	34019
Cy3-IVT sgRNA	N_3-IVT sgRNA	pppGGGUAGGGUUAGGGUUAGGGUUAGUUUUA GAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGU CCGUUAUCAACUUGAAAAAGUGGCACCGAGUCG GUGCUUUU-dA-2'-Cy3	34839 (+3xMg)	34845



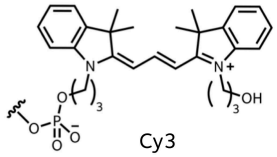
dA–2′–N_3



dA–2′–Cy3

Appendix Table 14 | DNA templates for EMSA using dCas9-3xGFP. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of the modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6482	EMSA Non-target DNA	Cy3-acctcgctgatgagcatagggttagggttagggttagggccgcagctgcctcgcgc	17875	17878
6483	EMSA Target DNA	gcgcgaggcagctgcggccctaaccctaaccctaaccctatgctcatcagcgaggt	17116	17117



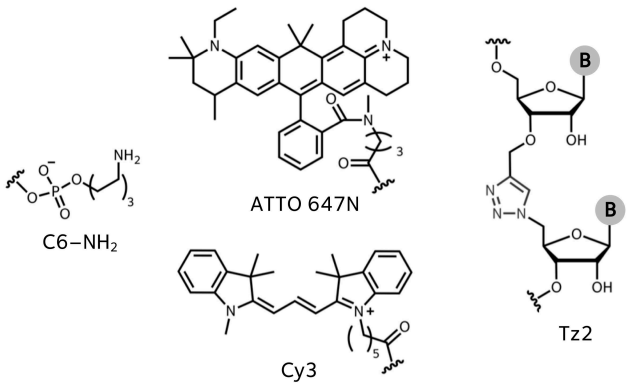
Cy3

Appendix Table 15 | Primers for sequencing of pDEST-his-dCas9-3xGFP. Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6742	Primer 1	gatgaagccctgaaagacgcgcag	7421	7420
6743	Primer 2	ataagccgcgaggcaggtacc	6465	6465
6744	Primer 3	ccggatgcggcaaaaatgataccc	7356	7355
6745	Primer 4	gcaccggtaagatcctggaacagc	7372	7371
6746	Primer 5	ggtccgtgatctgaacggcgtttg	7416	7415

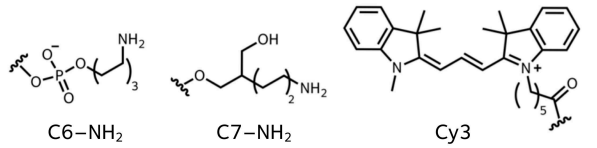
Appendix Table 16 | Clicked ~21–79 sgRNAs for live-cell imaging. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. The precursor oligonucleotides for click ligation are listed in **Appendix Table 5**.

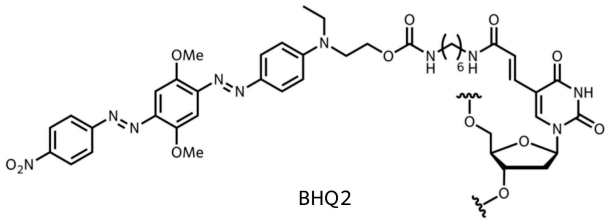
sgRNA Name	Oligo. Used	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
Telomeric click sgRNA-NH_2	CR4-NH_2 + TR1	NH_2-C6-UAGGGUUAGGGUUAGGGUUAG -**Tz2**- UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	32502	32505
Telomeric click sgRNA-DNA-NH_2	CR4-NH_2 + TR2	NH_2-C6-UAGGGUUAGGGUUAGGGUUAG -**Tz2**- UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAActtgaaaaaGtGgcaccGagtcggtgCtttt	32084	32085
Telomeric click sgRNA-Cy3	CR4-Cy3 + TR1	Cy3-NH-C6-UAGGGUUAGGGUUAGGGUUAG -**Tz2**- UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	32942	32944
Telomeric sgRNA-DNA-Cy3	CR4-Cy3 + TR2	Cy3-NH-C6-UAGGGUUAGGGUUAGGGUUAG -**Tz2**- UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAActtgaaaaaGtGgcaccGagtcggtgCtttt	32524	32523
Telomeric click sgRNA-ATTO	CR4-ATTO + TR1	ATTO 647N-NH-C6-UAGGGUUAGGGUUAGGGUUAG-**Tz2**- UUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUU	33131	33134
Telomeric click sgRNA-DNA-ATTO	CR4-ATTO + TR2	ATTO 647N-NH-C6-UAGGGUUAGGGUUAGGGUUAG-**Tz2**- UUUUAGAGCUAgaaatagcAAGUUAAAAUAAGGCUAGUCCGUUAUCAActtgaaaaaGtGgcaccGagtcggtgCtttt	32713	32714



Appendix Table 17 | crRNA and tracrRNA oligonucleotides used for generation of RNase H-activated fgRNAs. Uppercase bases are RNA nucleotides and lowercase bases are DNA nucleotides. The structures of modifications are listed at the bottom of the table. *Accurate mass could not be determined for this oligonucleotide.

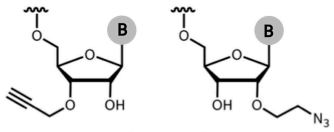
res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6997	Telomeric crRNA-NH_2	GGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUG CUGUUUUUAatcgactggca-C7-NH_2	16810	16810
6997	Telomeric crRNA-Cy3	GGGUUAGGGUUAGGGUUAGUUUUAGAGCUAUG CUGUUUUUAatcgactggca-C7-NH_2-Cy3	17248	17249
7023	tracrRNA-0Q	NH_2-C6-UGCCAGUCGAUUAAAAACAGCAUAGCAAGUUAA AAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGU GGCACCGAGUCGGUGCUUUU	27842	27843
7018	tracrRNA-1Q	BHQ2-UGCCAGUCGAUUAAAAACAGCAUAGCAAGUUAA AAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGU GGCACCGAGUCGGUGCUUUU	28626	28658*
7024	tracrRNA-2Q	2xBHQ2-UGCCAGUCGAUUAAAAACAGCAUAGCAAGUUAA AAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGU GGCACCGAGUCGGUGCUUUU	29589	29591





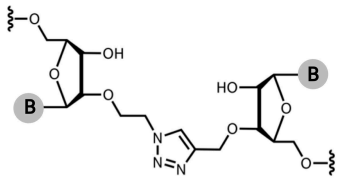
Appendix Table 18 | AsCas12a azide- and alkyne-crRNAs used to generate 3'–3' dual gRNAs. Uppercase bases are RNA nucleotides and the structures of modifications are listed at the bottom of the table. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
5912	crRNA42-N_3	AAUUUCUACUCUUGUAGAUCCGCAGCUGCCUCG CGCUGUCU-dT-2'-N_3	13318	13317
5835	crRNA42-alk	AAUUUCUACUCUUGUAGAUUCAGAGUGUAUCAC CGUCAUCAC-propargyl	13313	13312
5913	crRNA44-N_3	AAUUUCUACUCUUGUAGAUCCGCAGCUGCCUCG CGCUGUCUGG-dT-2'-N_3	14009	14006
5834	crRNA44-alk	AAUUUCUACUCUUGUAGAUUGUGUCAGAGUGUA UCACCGUCAUC-propargyl	13982	13978
5914	crRNA46-N_3	AAUUUCUACUCUUGUAGAUCCGCAGCUGCCUCG CGCUGUCUGGUG-dT-2'-N_3	14660	14659
5833	crRNA46-alk	AAUUUCUACUCUUGUAGAUUGCAUGUGUCAGAG UGUAUCACCGUCA-propargyl	14656	14654



Appendix Table 19 | AsCas12a clicked 3'–3' dual gRNAs. Uppercase bases are RNA nucleotides and the structures of the modifications are listed at the bottom of the table. Note that the 5' to 3' sequence direction is inverted for part of the construct. The precursor oligonucleotides for click ligation are listed in Appendix Table 18.

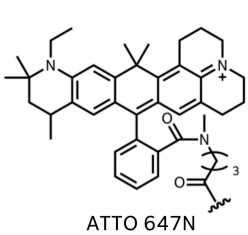
Code	Oligo. Used	Sequence	Mol. Weight (g/mol)	
			Expected	Found
Dual gRNA 42-42	crRNA42-N_3 + crRNA42-alk	5'-AAUUUCUACUCUUGUAGAUCCGCAGCU GCCUCGCGCUGUCUT-3' -InvLink- 3'-CACUACUGCCACUAUGUGAGACU UAGAUGUUCUCAUCUUUAA-5'	26631	26632
Dual gRNA 44-44	crRNA44-N_3 + crRNA44-alk	5'- AAUUUCUACUCUUGUAGAUCCGCAGCU GCCUCGCGCUGUCUGGT-3' -InvLink- 3'-CUACUGCCACUAUGUGAGACU GUGUUAGAUGUUCUCAUCUUUAA-5'	27991	27992
Dual gRNA 46-46	crRNA46-N_3 + crRNA46-alk	5'- AAUUUCUACUCUUGUAGAUCCGCAGCU GCCUCGCGCUGUCUGGUGT-3' -InvLink- 3'-ACUGCCACUAUGUGAGACU GUGUACGUUAGAUGUUCUCAUCUUUAA-5'	29316	29311

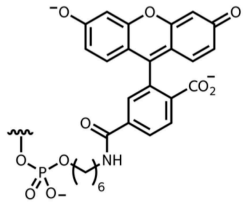


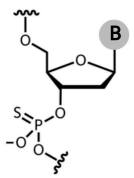
InvLink

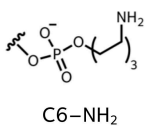
Appendix Table 20 | DNA templates for AsCas12a DNA cleavage. Lowercase bases are DNA nucleotides and bold bases have 3'-phosphorothioate modifications. The structures of the modifications are listed at the bottom of the table.

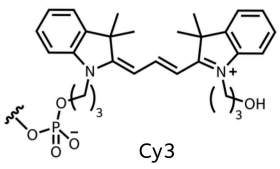
dsDNA	Code (res)	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
PO-dsDNA I	PO-ssDNA-FAM (res 6040)	FAM-acctcgctgatgagctttaccgcagctgcc tcgcgctgtctggtgatgacggtgatacactctgacacatgcagc tcccggagacgatca	28194	28195
	PO-ssDNA-Cy3 (res 6041)	Cy3-tgatcgtctccgggagctgcatgtgtcaga gtgtatcaccgtcatcaccagacagcgcgaggcagctgcggtaa agctcatcagcgaggt	28342	28343
PS-dsDNA I	PS-ssDNA-FAM (res 6047)	FAM-**acctc**gctgatgagctttaccgcagctgcc tcgcgctgtctggtgatgacggtgatacactctgacacatgcagc tcccggagac**gatca**	28338	28338
	PS-ssDNA-Cy3 (res 6048)	Cy3-**tgatc**gtctccgggagctgcatgtgtcaga gtgtatcaccgtcatcaccagacagcgcgaggcagctgcggtaa agctcatcagc**gaggt**	28486	28487
PO-dsDNA II	PO-ssDNA-FAM (res 6040)	FAM-acctcgctgatgagctttaccgcagctgcc tcgcgctgtctggtgatgacggtgatacactctgacacatgcagc tcccggagacgatca	28194	28195
	PO-ssDNA ATTO (res 7363)	ATTO 647N-NH-C6-tgatcgtctccgggagctgcatgt gtcagagtgtatcaccgtcatcaccagacagcgcgaggcagctg cggtaaagctcatcagcgaggt	28643	28640
–	38-mer FAM (res 6042)	FAM-acctcgctgatgagctttaccgcagctgcctcgcgctg	12119	12118
–	39-mer Cy3 (res 6043)	Cy3-gttcgtctccgggagctgcatgtgtcagagtgtatcacc	12512	12513


ATTO 647N


FAM


Bold bases


$C6–NH_2$


Cy3

Appendix Table 21 | Synthetic TH01 STR-containing DNA template, splint and primer for Ramified rolling circle amplification (RAM). Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name. 'p' = a phosphate group.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
4419	TH01 template	p-ttggcctgttcctcccgctttgcactttagg cgagttcgattattatttccctcattcattcattcattcattcattc accatggagtctgtgttc	30549	30552
4148	RAM Primer P1/Splint	gggaggaacaggccaagaacacagactccatg	9910	9910
4395	RAM Primer P2	gctttgcactttaggcgagttcgatta	8296	8296

Appendix Table 22 | PCR primers for genomic DNA amplification of TH01 STR sequence. Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name. 'p' = a phosphate group.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
4648	TH01 forward primer	gtgattcccattggcctgttc	6379	6378
4649	TH01 reverse primer	p-attcctgtgggctgaaaagctc	6828	6829

Appendix Table 23 | F30-broccoli oligonucleotides used for construction of pSpCas9(BB)-2A-Puro-v2-Broccoli vector. Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
7113	F30-Broccoli-Template	caccttgccatgtgtatgtgggagacggtcgggtccagatattcgtatctgtcgagtagagtgtgggctcccacatactctgatgatccttcgggatcattcatggcaacttttttgtttttcgt	38580	38581
7114	F30-Broccoli-Compl.	ctagacgaaaaacaaaaaagttgccatgaatgatcccgaaggatcatcagagtatgtgggagcccacactctactcgacagatacgaatatctggacccgaccgtctcccacatacacatggcaa	38473	38475

Appendix Table 24 | Tailed primers used for Gibson assembly of pDEST-his-dCas9-3xGFP from pDEST-hisMBP-AsCpf1-EC and pHAGE-TO-dCas9-3xGFP vectors. Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
6689	pDEST-hisMBP-AsCpf1-EC Forward	gcatcacgcatggtatggatgaactgtacaaactcgagtagactggccgtcgttttacaac	18812	18813
6690	pDEST-hisMBP-AsCpf1-EC Reverse	cgcttccgcttccgaattcgaccttgcgtttcttttttaggggcaccacttccgtgatggtgatggtgatgg	21847	21845
6691	pHAGE-TO-dCas9-3xGFP Forward	gattatgaaaatccatcaccatcaccatcacggaagtggtgcccctaaaaagaaacgcaag	18745	18745
6692	pHAGE-TO-dCas9-3xGFP Reverse	gggttttcccagtcacgacgttgtaaaacgacggccagtctactcgagtttgtacagttcatc	19363	19364

Appendix Table 25 | PCR primers for amplification of *EMX1* gene fragment from genomic DNA for the T7E1 assay. Lowercase bases are DNA nucleotides. Oligonucleotides have both a unique identifier ('res') and a code name.

res	Code	Sequence (5'–3')	Mol. Weight (g/mol)	
			Expected	Found
7513	*EMX1* PCR Forward	ggagcagctggtcagagggg	6288	6289
7514	*EMX1* PCR Reverse	ccatagggaaggggacactgg	6874	6876

References

1. Rothstein, R. J. One-Step Gene Disruption in Yeast. *Methods Enzymol.* **101**, 202–211 (1983).

2. Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A. & Kucherlapati, R. S. Insertion of DNA sequences into the human chromosomal β-globin locus by homologous recombination. *Nature* **317**, 230–234 (1985).

3. Thomas, K. R., Folger, K. R. & Capecchi, M. R. High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419–428 (1986).

4. Capecchi, M. R. Altering the genome by homologous recombination. *Science* **244**, 1288–1292 (1989).

5. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* **7**, 2902–2906 (2008).

6. Lin, F.-L., Sperle, K. & Sternberg, A. N. Recombination in mouse L cells between DNA introduced into cells and homologous chromosomal sequences (homologous recombination/thymidine kinase gene/gene transfer/DNA rearrangement). *Biochemistry* **82**, 1391–1395 (1985).

7. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* **14**, 8096–8106 (1994).

8. Bassing, C. H. & Alt, F. W. The cellular response to general and programmed DNA double strand breaks. *DNA Repair* **3**, 781–796 (2004).

9. Seligman, L. M. *et al.* Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic Acids Res.* **30**, 3870–3879 (2002).

10. Sussman, D. *et al.* Isolation and characterization of new homing endonuclease specificities at individual target site positions. *J. Mol. Biol.* **342**, 31–41 (2004).

11. Rosen, L. E. *et al.* Homing endonuclease I-CreI derivatives with novel DNA target specificities. *Nucleic Acids Res.* **34**, 4791–4800 (2006).

12. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci.* **93**, 1156–1160 (1996).

13. Klug, A. & Rhodes, D. Zinc fingers: A novel protein fold for nucleic acid recognition. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 473–482 (1987).

14. Liu, Q., Segal, D. J., Ghiara, J. B. & Barbas, C. F. Design of polydactyl zinc-finger proteins for

unique addressing within complex genomes. *Proc. Natl. Acad. Sci.* **94**, 5525–5530 (1997).

15. Bibikova, M. *et al.* Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases. *Mol. Cell. Biol.* **21**, 289–297 (2001).

16. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (2003).

17. Chandrasegaran, S. & Carroll, D. Origins of Programmable Nucleases for Genome Engineering. *J. Mol. Biol.* **428**, 963–989 (2016).

18. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512 (2009).

19. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (2009).

20. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757–761 (2010).

21. Zhang, F. *et al.* Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.* **29**, 149–153 (2011).

22. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148 (2011).

23. Li, T. *et al.* Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Res.* **39**, 6315–6325 (2011).

24. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39**, e82 (2011).

25. Holkers, M. *et al.* Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.* **41**, e63 (2013).

26. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

27. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).

28. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).

29. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433 (1987).

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

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

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

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

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

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

36. Makarova, K. K. S. *et al.* A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7 (2006).

37. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

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

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

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

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

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

43. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60 (2005).

44. Makarova, K. S., Brouns, S. J. J., Horvath, P., Sas, D. F. & Wolf, Y. I. Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477 (2012).

45. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39**, 9275–9282 (2011).

46. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2579--E2586 (2012).

47. Cong, L., Ran, F., Cox, D., Lin, S. & Barretto, R. Multiplex Genome Engineering Using CRISPR / Cas Systems. *Science* **339**, 819–823 (2013).

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

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

50. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).

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

52. Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867–871 (2016).

53. Hirano, S., Nishimasu, H., Ishitani, R. & Nureki, O. Structural Basis for the Altered PAM Specificities of Engineered CRISPR-Cas9. *Mol. Cell* **61**, 886–894 (2016).

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

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

56. Abudayyeh, O. O. *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).

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

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

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

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

61. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: Gene regulation or autoimmunity? *Trends Genet.* **26**, 335–340 (2010).

62. Westra, E. R. *et al.* Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. *PLoS Genet.* **9**, e1003742 (2013).

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

64. Abudayyeh, O. O. *et al.* RNA targeting with CRISPR-Cas13. *Nature* **550**, 280–284 (2017).

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

66. Jiang, F. & Doudna, J. A. CRISPR–Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46**, 505–529 (2017).

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

68. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **322**, 1843–1845 (2010).

69. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

70. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci.* **111**, 9798–9803 (2014).

71. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).

72. Josephs, E. A. *et al.* Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res.* **43**, 8924–8941 (2015).

73. Dahlman, J. E. *et al.* Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. *Nat. Biotechnol.* **33**, 1159–1161 (2015).

74. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).

75. Chen, J. S. & Doudna, J. A. The chemistry of Cas9 and its CRISPR colleagues. *Nat. Rev. Chem.* **1**, 0078 (2017).

76. Wright, A. V. *et al.* Rational design of a split-Cas9 enzyme complex. *Proc. Natl. Acad. Sci.* **112**, 2984–2989 (2015).

77. Mekler, V., Minakhin, L., Semenova, E., Kuznedelov, K. & Severinov, K. Kinetics of the CRISPR-Cas9 effector complex assembly and the role of 3'-terminal segment of guide RNA. *Nucleic Acids Res.* **44**, 2837–2845 (2016).

78. Shechner, D. M., Hacisüleyman, E., Younger, S. T. & Rinn, J. L. CRISPR Display: A modular method for locus-specific targeting of long noncoding RNAs and synthetic RNA devices in vivo. *Nat. Methods* **12**, 664–670 (2015).

79. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

80. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).

81. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

82. O'Connell, M. R. *et al.* Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* **516**, 263–266 (2014).

83. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).

84. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proc. Natl. Acad. Sci. USA* **110**, 15644–15649 (2013).

85. Kim, E. *et al.* In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nat. Commun.* **8**, 14500 (2017).

86. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).

87. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

88. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

89. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

90. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).

91. Vakulskas, C. A. *et al.* A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).

92. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).

93. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).

94. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

95. Gaudelli, N. M. *et al.* Programmable base editing of A • T to G • C in genomic DNA without DNA cleavage. *Nat. Publ. Gr.* **551**, 464–471 (2017).

96. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).

97. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

98. Gilbert, L. a *et al.* CRISPR-Mediated Modular RNA-Guided Regualtion of Transcription in Eukaryotes. *Cell* **154**, 442–451 (2013).

99. Maeder, M. L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10**, 977–979 (2013).

100. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* **10**, 973–976 (2013).

101. Kearns, N. A. *et al.* Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).

102. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).

103. Hendel, A. *et al.* Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33**, 985–989 (2015).

104. Cameron, P. *et al.* Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).

105. Liang, X. *et al.* Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J. Biotechnol.* **208**, 44–53 (2015).

106. Yin, H. *et al.* Therapeutic genome editing by combined viral and non-viral delivery of CRISPR system components in vivo. *Nat. Biotechnol.* **34**, 328–333 (2016).

107. Ramakrishna, S. *et al.* Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res.* **24**, 1020–1027 (2014).

108. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).

109. Fonfara, I., Richter, H., BratoviÄ, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).

110. Zetsche, B. *et al.* Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**, 31–34 (2017).

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

112. Dong, D. *et al.* The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* **532**, 522–526 (2016).

113. Gao, P., Yang, H., Rajashankar, K. R., Huang, Z. & Patel, D. J. Type v CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell Res.* **26**, 901–913 (2016).

114. Kim, H. K. *et al.* In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods* **14**, 153–159 (2017).

115. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

116. Berg, J. M., Tymoczko, J. L. & Stryer, L. Biochemistry. in *Biochemistry* Chapter 27.1 (2002).

117. Chapter 2 DNA and RNA Structure. in *Nucleic Acids in Chemistry and Biology* 13–76 (The Royal Society of Chemistry, 2006).

118. Drew, H. R. *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci.* **78**, 2179–2183 (1981).

119. Luo, Z., Dauter, M. & Dauter, Z. Phosphates in the Z-DNA dodecamer are flexible, but their P-SAD signal is sufficient for structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 1790–1800 (2014).

120. Shivalingam, A. Small Molecule Probes for the Elucidation of G-quadruplex Function.

(Imperial College London, 2015).

121. Chandrasekaran, R. & Arnott, S. 2.4.5 References for 2.4. in *Crystallographic and Structural Data II* (ed. Saenger, W.) 169–170 (Springer Berlin Heidelberg, 1989).

122. Rueda, F. O. *et al.* Mapping the sugar dependency for rational generation of a DNA-RNA hybrid-guided Cas9 endonuclease. *Nat. Commun.* **8**, 1610 (2017).

123. Rich, A. & Zhang, S. Z-DNA: The long road to biological function. *Nat. Rev. Genet.* **4**, 566–572 (2003).

124. Wyatt, J. R., Puglisi, J. D. & Tinoco, I. RNA folding: Pseudoknots, loops and bulges. *BioEssays* **11**, 100–106 (1989).

125. Crick, F. H. C. Codon—anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).

126. Hoogsteen, K. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.* **16**, 907 (1963).

127. Kim, S. H. *et al.* Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185**, 435–440 (1974).

128. Robertus, J. D. *et al.* Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **250**, 546–551 (1974).

129. Batey, R. T., Rambo, R. P. & Doudna, J. A. Tertiary motifs in RNA structure and folding. *Angew. Chemie Int. Ed.* **38**, 2326–2343 (1999).

130. Yin, H. *et al.* Structure-guided chemical modification of guide RNA enables potent non-viral in vivo genome editing. *Nat. Biotechnol.* **35**, 1179–1187 (2017).

131. Rahdar, M. *et al.* Synthetic CRISPR RNA-Cas9–guided genome editing in human cells. *Proc. Natl. Acad. Sci.* **112**, E7110–E7117 (2015).

132. Mir, A. *et al.* Heavily and Fully Modified RNAs Guide Efficient SpyCas9-Mediated Genome Editing. *bioRxiv Biochem.* **9**, 2641 (2018).

133. Basila, M., Kelley, M. L. & Smith, A. V. B. Minimal 2'-O-methyl phosphorothioate linkage modification pattern of synthetic guide RNAs for increased stability and efficient CRISPR-Cas9 gene editing avoiding cellular toxicity. *PLoS One* **12**, e0188593 (2017).

134. Kawasaki, A. M. *et al.* Uniformly Modified 2'-Deoxy-2'-fluoro Phosphorothioate Oligonucleotides as Nuclease-Resistant Antisense Compounds with High Affinity and Specificity for RNA Targets. *J. Med. Chem.* **36**, 831–841 (1993).

135. Lesnik, E. A. *et al.* Oligodeoxynucleotides Containing 2'-O-Modified Adenosine: Synthesis and Effects on Stability of DNA:RNA Duplexes. *Biochemistry* **32**, 7832–7838 (1993).

136. Ikeda, H. *et al.* The effect of two antipodal fluorine-induced sugar puckers on the conformation and stability of the Dickerson-Drew dodecamer duplex [d(CGCGAATTCGCG)]2. *Nucleic Acids Res.* **26**, 2237–2244 (1998).

137. Yin, H. *et al.* Partial DNA-guided Cas9 enables genome editing with reduced off-target activity. *Nat. Chem. Biol.* **14**, 311–316 (2018).

138. Cromwell, C. R. *et al.* Incorporation of bridged nucleic acids into CRISPR RNAs improves Cas9 endonuclease specificity. *Nat. Commun.* **9**, 1448 (2018).

139. Ryan, D. E. *et al.* Improving CRISPR-Cas specificity with chemical modifications in single-guide RNAs. *Nucleic Acids Res.* **46**, 792–803 (2018).

140. Lee, K. *et al.* Synthetically modified guide RNA and donor DNA are a versatile platform for CRISPR-Cas9 engineering. *Elife* **6**, e25312 (2017).

141. Vannini, A. & Cramer, P. Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries. *Mol. Cell* **45**, 439–446 (2012).

142. Shan, Q. *et al.* Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 686–688 (2013).

143. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471 (2013).

144. Ranganathan, V., Wahlin, K., Maruotti, J. & Zack, D. J. Expansion of the CRISPR-Cas9 genome targeting space through the use of H1 promoter-expressed guide RNAs. *Nat. Commun.* **5**, 4516 (2014).

145. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/cas-mediated genome engineering. *Cell* **154**, 1370–1379 (2013).

146. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).

147. Milligan, J. F., Groebe, D. R., Witherell, G. W. & Uhlenbeck, O. C. Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. *Nucleic Acids Res.* **15**, 8783–8798 (1987).

148. Pleiss, J. A., Derrick, M. L. & Uhlenbeck, O. C. T7 RNA polymerase produces 5' end heterogeneity during in vitro transcription from certain templates. *RNA* **4**, 1313–1317 (1998).

149. Melton, D. A. *et al.* Efficient in vitro synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.* **12**, 7035–7056 (1984).

150. Adhya, S., Basu, S., Sarkar, P. & Maitra, U. Location, function, and nucleotide sequence of a promoter for bacteriophage T3 RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 147–151 (1981).

151. Hwang, W. Y. *et al.* Efficient In Vivo Genome Editing Using RNA-Guided Nucleases. *Nat. Biotechnol.* **31**, 227–229 (2013).

152. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).

153. Kim, D. *et al.* Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).

154. Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P. & Lu, T. K. Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human

Cells. *Mol. Cell* **54**, 698–710 (2014).

155. Gao, Y. & Zhao, Y. Self-processing of ribozyme-flanked RNAs into guide RNAs in vitro and in vivo for CRISPR-mediated genome editing. *J. Integr. Plant Biol.* **56**, 343–349 (2014).

156. Xie, K., Minkenberg, B. & Yang, Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci.* **112**, 3570–3575 (2015).

157. Helm, M., Brulé, H., Giegé, R. & Florentz, C. More mistakes by T7 RNA polymerase at the 5′ ends of in vitro-transcribed RNAs. *RNA* **5**, 618–621 (1999).

158. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites-A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22**, 1859–1862 (1981).

159. Usman, N., Pon, R. T. & Ogilvie, K. K. Preparation of ribonucleoside 3'-O-phosphoramidites and their application to the automated solid phase synthesis of oligonucleotides. *Tetrahedron Lett.* **26**, 4567–4570 (1985).

160. Wu, X. & Pitsch, S. Synthesis and pairing properties of oligoribonucleotide analogues containing a metal-binding site attached to β-D-allofuranosyl cytosine. *Nucleic Acids Res.* **26**, 4315–4323 (1998).

161. Pitsch, S., Weiss, P. & Jenny, L. Reliable Chemical Synthesis of Oligoribonucleotides (RNA) with 2'-O-methyl (2'-O-tom)-Protected Phosphoramidites. *Helv. Chim.* **84**, 3773–3795 (2001).

162. Scaringe, S. A., Wincott, F. E. & Caruthers, M. H. Novel RNA synthesis method using 5'-O-silyl-2'-O-orthoester protecting groups. *J. Am. Chem. Soc.* **120**, 11820–11821 (1998).

163. Scaringe, S. A., Kitchen, D., Kaiser, R. J. & Marshall, W. S. Preparation of 5'-silyl-2'-orthoester ribonucleosides for use in oligoribonucleotide synthesis. *Curr. Protoc. Nucleic Acid Chem.* **16**, 2.10.1–2.10.16 (2004).

164. Dellinger, D. J. *et al.* Streamlined process for the chemical synthesis of RNA using 2'-O-thionocarbamate-protected nucleoside phosphoramidites in the solid phase. *J. Am. Chem. Soc.* **133**, 11540–11556 (2011).

165. Gasparutto, D. *et al.* Chemical synthesis of A biologically active natural tRNA with its minor bases. *Nucleic Acids Res.* **20**, 5159–5166 (1992).

166. Westmanu, E. & Stromberg, R. Removal of f-butyldimethylsilyl protection in RNA-synthesis. Triethylamine trihydrofluoride (TEA, 3HF) is a more reliable alternative to tetrabutylammonium fluoride (TBAF). *Nucleic Acids Res.* **22**, 2430–2431 (1994).

167. Ogilvie, K. K., Usman, N., Nicoghosian, K. & Cedergren, R. J. Total chemical synthesis of a 77-nucleotide-long RNA sequence having methionine-acceptance activity. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 5764–5768 (1988).

168. Sproat, B. *et al.* An Efficient Method for the Isolation and Purification of Oligoribonucleotides. *Nucleosides and Nucleotides* **14**, 255–273 (1995).

169. Welz, R. & Müller, S. 5-(Benzylmercapto)-1H-tetrazole as activator for 2'-O-TBDMS phosphoramidite building blocks in RNA synthesis. *Tetrahedron Lett.* **43**, 795–797 (2002).

170. Shivalingam, A. & Brown, T. Synthesis of chemically modified DNA. *Biochem. Soc. Trans.* **44**,

709–715 (2016).

171. Efcavitch, J. W. & Heiner, C. Depurination as a yield decreasing mechanism in oligodeoxynucleotide synthesis. *Nucleosides and Nucleotides* **4**, 267 (1985).

172. Septak, M. Kinetic studies on depurination and detritylation of CPG-bound intermediates during oligonucleotide synthesis. *Nucleic Acids Res.* **24**, 3053–3058 (1996).

173. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).

174. Huisgen, R. 1,3-Dipolar Cycloadditions. Past and Future. *Angew. Chem. Int. Ed.* **2**, 565 (1963).

175. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chemie Int. Ed.* **40**, 2004–2021 (2001).

176. Rostovtsev, V. V., Green, L. G., Fokin, V. V. & Sharpless, K. B. A stepwise huisgen cycloaddition process: Copper(I)-catalyzed regioselective 'ligation' of azides and terminal alkynes. *Angew. Chemie Int. Ed.* **41**, 2596–2599 (2002).

177. Chan, T. R., Hilgraf, R., Sharpless, K. B. & Fokin, V. V. Polytriazoles as copper(I)-stabilizing ligands in catalysis. *Org. Lett.* **6**, 2853–2855 (2004).

178. Tornøe, C. W., Christensen, C. & Meldal, M. Peptidotriazoles on solid phase: [1,2,3]-Triazoles by regiospecific copper(I)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *J. Org. Chem.* **67**, 3057–3064 (2002).

179. Worrell, B. T., Malik, J. A. & Fokin, V. V. Direct evidence of a dinuclear copper intermediate in Cu(I)-catalyzed azide-alkyne cycloadditions. *Science* **340**, 457–460 (2013).

180. El-Sagheer, A. H. & Brown, T. Synthesis and polymerase chain reaction amplification of DNA strands containing an unnatural triazole linkage. *J. Am. Chem. Soc.* **131**, 3958–3964 (2009).

181. Dallmann, A. *et al.* Structure and dynamics of triazole-linked DNA: Biocompatibility explained. *Chem. - A Eur. J.* **17**, 14714–14717 (2011).

182. El-Sagheer, A. H., Sanzone, A. P., Gao, R., Tavassoli, A. & Brown, T. Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in Escherichia coli. *Proc. Natl. Acad. Sci.* **108**, 11338–11343 (2011).

183. El-Sagheer, A. H. & Brown, T. Efficient RNA synthesis by in vitro transcription of a triazole-modified DNA template. *Chem. Commun. (Camb).* **47**, 12057–12058 (2011).

184. Birts, C. N. *et al.* Transcription of click-linked DNA in human cells. *Angew. Chem. Int. Ed.* **53**, 2362–2365 (2014).

185. El-Sagheer, A. H. & Brown, T. New strategy for the synthesis of chemically modified RNA constructs exemplified by hairpin and hammerhead ribozymes. *Proc. Natl. Acad. Sci.* **107**, 15329–15334 (2010).

186. Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J. Am. Chem. Soc.* **126**, 15046–15047 (2004).

187. Chen, X., El-Sagheer, A. H. & Brown, T. Reverse transcription through a bulky triazole linkage in RNA: Implications for RNA sequencing. *Chem. Commun.* **50**, 7597–7600 (2014).

188. He, K., Chou, E. T., Begay, S., Anderson, E. M. & van Brabant Smith, A. Conjugation and Evaluation of Triazole-Linked Single Guide RNA for CRISPR-Cas9 Gene Editing. *ChemBioChem* **17**, 1809–1812 (2016).

189. Shechner, D. M., Hacisuleyman, E., Younger, S. T. & Rinn, J. L. Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. *Nat. Methods* **12**, 664–670 (2015).

190. Usman, N., Ogilvie, K. K., Jiang, M. Y. & Cedergren, R. J. Automated Chemical Synthesis of Long Oligoribonucleotides Using 2'-O-silylated Ribonucleoside 5'-O-phosphoramidites on a Controlled-pore Glass Support: Synthesis of a 43-nucleotide Sequence Similar to the 3'-half Molecule of an Escherichia Coli Formylmeth. *J. Am. Chem. Soc.* **109**, 7845–7854 (1987).

191. Scaringe, S. A., Francklyn, C. & Usman, N. Chemical synthesis of biologically active oligoribonucleotides using beta-cyanoethyl protected ribonucleoside phosphoramidites. *Nucleic Acids Res* **18**, 5433–5441 (1990).

192. Wu, T., Ogilvie, K. K. & Pon, R. T. Prevention of chain cleavage in the chemical synthesis of 2'-silylated oligoribonucleotides. *Nucleic Acids Res.* **17**, 3501–3517 (1989).

193. Stawinski, J., Strömberg, R., Thelin, M. & Westman, E. Studies on the t-butyldimethylsilyl group as 2'-O-protection in oligoribonucleotide synthesis via the H-phosphonate approach. *Nucleic Acids Res.* **16**, 9285–9298 (1988).

194. Reddy, M. P., Hanna, N. B. & Farooqui, F. Fast cleavage and deprotection of oligonucleotides. *Tetrahedron Lett.* **35**, 4311–4314 (1994).

195. Wincott, F. *et al.* Synthesis, deprotection, analysis and purification of RNA and ribosomes. *Nucleic Acids Res.* **23**, 2677–2684 (1995).

196. Bellon, L. Oligoribonucleotides with 2'-O-(tert-Butyldimethylsilyl) Groups. *Curr. Protoc. Nucleic Acid Chem.* **1**, 3.6.1–3.6.13

197. Sigma-Aldrich. Date accessed: 01-09-2018. *https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/SAFC/Posters/1/tc-rna-phosphoramidites.pdf*

198. Lumiprobe. Date accessed: 01-09-2018. *https://www.lumiprobe.com/protocols/click-chemistry-dna-labeling*

199. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).

200. Miller, G. P. & Kool, E. T. Versatile 5'-Functionalization of Oligonucleotides on Solid Support: Amines, Azides, Thiols, and Thioethers via Phosphorus Chemistry. *J. Org. Chem.* **69**, 2404–2410 (2004).

201. Zanetti-Domingues, L. C., Tynan, C. J., Rolfe, D. J., Clarke, D. T. & Martin-Fernandez, M. Hydrophobic Fluorescent Probes Introduce Artifacts into Single Molecule Tracking Experiments Due to Non-Specific Binding. *PLoS One* **8**, e74200 (2013).

202. Cradick, T. J., Fine, E. J., Antico, C. J. & Bao, G. CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41**, 9584–9592 (2013).

203. Tsai, S. Q. *et al.* CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).

204. Jawalekar, A. M. *et al.* Conjugation of nucleosides and oligonucleotides by [3+2] cycloaddition. *J. Org. Chem.* **73**, 287 (2008).

205. Wada, T. *et al.* Synthesis and properties of 2-azidodeoxyadenosine and its incorporation into oligodeoxynucleotides. *Tetrahedron Lett.* **42**, 9215–9219 (2001).

206. Ötvös, L., Bajor, Z., Kraicsovits, F., Sági, G. & Tegyey, Z. Synthesis and enzymatic characterization of P1-thio-P2-oxo trideoxynucleoside diphosphates having AZT, FdU, or dT at the 3'-position. *Nucleosides, Nucleotides and Nucleic Acids* **21**, 79–92 (2002).

207. Kiviniemi, A., Virta, P. & Lönnberg, H. Utilization of intrachain 4'-C-azidomethylthymidine for preparation of oligodeoxyribonucleotide conjugates by click chemistry in solution and on a solid support. *Bioconjug. Chem.* **19**, 1726–1743 (2008).

208. El-Sagheer, A. H. & Brown, T. Single tube gene synthesis by phosphoramidate chemical ligation. *Chem. Commun.* **53**, 10700–10702 (2017).

209. Shivalingam, A., Tyburn, A. E. S., El-Sagheer, A. H. & Brown, T. Molecular requirements of high-fidelity replication-competent DNA backbones for orthogonal chemical ligation. *J. Am. Chem. Soc.* **139**, 1575–1583 (2017).

210. Zink, D., Sadoni, N. & Stelzer, E. Visualizing chromatin and chromosomes in living cells. *Methods* **29**, 42–50 (2003).

211. Martin, R. M., Leonhardt, H. & Cardoso, M. C. DNA labeling in living cells. *Cytom. Part A* **67**, 45–52 (2005).

212. Bauman, J. G. J., Wiegant, J., Borst, P. & van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp. Cell Res.* **128**, 485–490 (1980).

213. Singer, R. H. & Ward, D. C. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 7331–7335 (1982).

214. Amano, T. *et al.* Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* **16**, 47–57 (2009).

215. Speicher, M. R., Ballard, S. G. & Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat. Genet.* **12**, (1996).

216. Schrock, E. *et al.* Multicolor spectral karyotyping of human chromosomes. *Science* **273**, 494–497 (1996).

217. Schnell, U., Dijk, F., Sjollema, K. A. & Giepmans, B. N. G. Immunolabeling artifacts and the need for live-cell imaging. *Nature Methods* **9**, 152–158 (2012).

218. Lindhout, B. I. *et al.* Live cell imaging of repetitive DNA sequences via GFP-tagged polydactyl zinc finger proteins. *Nucleic Acids Res.* **35**, e107 (2007).

219. Miyanari, Y., Ziegler-Birling, C. & Torres-Padilla, M. E. Live visualization of chromatin dynamics with fluorescent TALEs. *Nat. Struct. Mol. Biol.* **20**, 1321–1324 (2013).

220. Ma, H. *et al.* CRISPR-Cas9 nuclear dynamics and target recognition in living cells. *J. Cell Biol.* **214**, 529–537 (2016).

221. Anton, T., Bultmann, S., Leonhardt, H. & Markaki, Y. Visualization of specific DNA sequences in living mouse embryonic stem cells with a programmable fluorescent CRISPR/Cas system. *Nucleus* **5**, 163–172 (2014).

222. Ma, H. *et al.* Multicolor CRISPR labeling of chromosomal loci in human cells. *Proc. Natl. Acad. Sci.* **112**, 3002–3007 (2015).

223. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159**, 635–646 (2014).

224. Qin, P. *et al.* Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9. *Nat. Commun.* **8**, 14725 (2017).

225. Ma, H. *et al.* Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nat. Biotechnol.* **34**, 528–530 (2016).

226. Nielsen, S., Yuzenkova, Y. & Zenkin, N. Mechanism of eukaryotic RNA polymerase III transcription termination. *Science* **340**, 1577–1580 (2013).

227. Ma, H. *et al.* Pol III promoters to express small RNAs: Delineation of transcription initiation. *Mol. Ther. - Nucleic Acids* **3**, e161 (2014).

228. Deng, W., Shi, X., Tjian, R., Lionnet, T. & Singer, R. H. CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells. *Proc. Natl. Acad. Sci.* **112**, 11870–11875 (2015).

229. Fu, Y. *et al.* CRISPR-dCas9 and sgRNA scaffolds enable dual-colour live imaging of satellite sequences and repeat-enriched individual loci. *Nat. Commun.* **7**, 11707 (2016).

230. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **55**, 333–339 (2014).

231. Janssen, A. & Medema, R. H. Genetic instability: Tipping the balance. *Oncogene* **32**, 4459–4470 (2013).

232. Hughes, L. D., Rawle, R. J. & Boxer, S. G. Choose your label wisely: Water-soluble fluorophores often interact with lipid bilayers. *PLoS One* **9**, e87649 (2014).

233. Chesnoy, S. & Huang, L. Structure and Function of Lipid-DNA Complexes for Gene Delivery. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 27–47 (2000).

234. Hirko, A., Tang, F. & Hughes, J. A. Cationic lipid vectors for plasmid DNA delivery. *Curr. Med. Chem.* **10**, 1185–1193 (2003).

235. Liu, D., Ren, T. & Gao, X. Cationic transfection lipids. *Curr. Med. Chem.* **10**, 1307–1315 (2003).

236. Yuhong Xu and Francis C. Szoka, J., Transfection, C., Xu, Y. & Szoka, F. C. J. Mechanism of DNA Release from Cationic Liposome / DNA Complexes Used in. *Biochemistry* **35**, 5616–5623 (1996).

237. Zelphati, O. & Szoka, F. C. Mechanism of oligonucleotide release from cationic liposomes. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 11493–11498 (1996).

238. Hafez, I. M., Maurer, N. & Cullis, P. R. On the mechanism whereby cationic lipids promote intracellular delivery of polynucleic acids. *Gene Ther.* **8**, 1188–1196 (2001).

239. Hoekstra, D., Rejman, J., Wasungu, L., Shi, F. & Zuhorn, I. Gene delivery by cationic lipids: in and out of an endosome. *Biochem. Soc. Trans.* **35**, 68–71 (2007).

240. Ma, H. *et al.* CRISPR-Cas9 nuclear dynamics and target recognition in living cells. *J. Cell Biol.* **214**, 529–537 (2016).

241. Jorjani, H. *et al.* An updated human snoRNAome. *Nucleic Acids Res.* **44**, 5068–82 (2016).

242. Cardarelli, F. *et al.* The intracellular trafficking mechanism of Lipofectamine-based transfection reagents and its implication for gene delivery. *Sci. Rep.* **6**, 25879 (2016).

243. Kirschman, J. L. *et al.* Characterizing exogenous mRNA delivery, trafficking, cytoplasmic release and RNA-protein correlations at the level of single cells. *Nucleic Acids Res.* **45**, e113 (2017).

244. Kierzek, R., Burkard, M. E. & Turner, D. H. Thermodynamics of single mismatches in RNA duplexes. *Biochemistry* **38**, 14214–14223 (1999).

245. Hur, J. K. *et al.* Targeted mutagenesis in mice by electroporation of Cpf1 ribonucleoproteins. *Nat. Biotechnol.* **34**, 807–808 (2016).

246. Anders, C. & Jinek, M. In vitro enzymology of cas9. *Methods Enzymol.* **546**, 1–20 (2014).

247. Zuris, J. A. *et al.* Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat. Biotechnol.* **33**, 73–80 (2015).

248. McNaughton, B. R., Cronican, J. J., Thompson, D. B. & Liu, D. R. Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins. *Proc. Natl. Acad. Sci.* **106**, 6111–6116 (2009).

249. Ferry, Q. R. V., Lyutova, R. & Fulga, T. A. Rational design of inducible CRISPR guide RNAs for de novo assembly of transcriptional programs. *Nat. Commun.* **8**, 14633 (2017).

250. Shen, W. *et al.* Dynamic nucleoplasmic and nucleolar localization of mammalian RNase H1 in response to RNAP I transcriptional R-loops. *Nucleic Acids Res.* **45**, 10672–10692 (2017).

251. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair (Amst).* **7**, 1765–1771 (2008).

252. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3**, e04766 (2014).

253. Song, J. *et al.* RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. *Nat. Commun.* **7**, 10548 (2016).

254. Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat. Biotechnol.* **33**, 538–542 (2015).

255. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

256. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339–344 (2016).

257. Savic, N. *et al.* Covalent linkage of the DNA repair template to the CRISPR-Cas9 nuclease enhances homology-directed repair. *Elife* **7**, e33761 (2018).

258. Aird, E. J., Lovendahl, K. N., Martin, A. St., Harris, R. S. & Gordon, W. R. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Commun. Biol.* **1**, 54 (2017).

259. Carlson-Stevermer, J. *et al.* Assembly of CRISPR ribonucleoproteins with biotinylated oligonucleotides via an RNA aptamer for precise gene editing. *Nat. Commun.* **8**, 1711 (2017).

260. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).

261. Bothmer, A. *et al.* Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nat. Commun.* **8**, 13905 (2017).

262. Chen, J. S. *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439 (2018).

263. Li, S. Y. *et al.* CRISPR-Cas12a has both cis- and trans-cleavage activities on single-stranded DNA. *Cell Res.* **28**, 491–493 (2018).

264. Butler, J. M. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* **51**, 253–265 (2006).

265. Budowle, B., Moretti, T. R., Baumstark, a L., Defenbaugh, D. a & Keys, K. M. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *J. Forensic Sci.* **6**, 1277–1286 (1999).

266. Hares, D. R. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* **17**, 33–34 (2015).

267. Butler, J. M. J. M. Forensic DNA typing: biology, technology, and genetics of STR markers. in *Chemistry & biodiversity* 17–31 (2005).

268. Butler, J. M. Short tandem repeat typing technologies used in human identity testing. *Biotechniques* **43**, Sii–Sv (2007).

269. Bornman, D. *et al.* Short-read, high-throughput sequencing technology for STR genotyping. *Biotechniques* 1–6 (2012).

270. Van Neste, C., Vandewoestyne, M., Van Criekinge, W., Deforce, D. & Van Nieuwerburgh, F. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci. Int. Genet.* **9**, 1–8 (2014).

271. Walsh, P. S., Fildes, N. J. & Reynolds, R. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Res.* **24**, 2807–2812 (1996).

272. Brookes, C., Bright, J. A., Harbison, S. & Buckleton, J. Characterising stutter in forensic STR

multiplexes. *Forensic Sci. Int. Genet.* **6**, 58–63 (2012).

273. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).

274. Ip, C. L. C. *et al.* MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* **4**, 1075 (2015).

275. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

276. Kilianski, A. *et al.* Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. *Gigascience* **4**, 12 (2015).

277. Ashton, P. M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.* **33**, 296–300 (2015).

278. Beals, T. P., Smith, J. H., Nietupski, R. M. & Lane, D. J. A mechanism for ramified rolling circle amplification. *BMC Mol. Biol.* **11**, 94 (2010).

279. Shin, G. *et al.* CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat. Commun.* **8**, 14291 (2017).

280. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12**, 351–356 (2015).

281. Urban, J. M., Bliss, J., Lawrence, C. E. & Gerbi, S. A. Sequencing ultra-long DNA molecules with the Oxford Nanopore MinION. *bioRxiv* (2015).

282. Quick, J., Quinlan, A. R. & Loman, N. J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).

283. Loman, N. J. & Quinlan, A. R. Poretools: A toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399–3401 (2014).

284. Huang, X. & Miller, W. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* **12**, 337–357 (1991).

285. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1–19 (2004).

286. Li, C. *et al.* INC-Seq: Accurate single molecule reads using nanopore sequencing. *Gigascience* **5**, 34 (2016).

287. Okamoto, A. ECHO probes: A concept of fluorescence control for practical nucleic acid sensing. *Chem. Soc. Rev.* **40**, 5815–5828 (2011).

288. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281–2308 (2013).