Chemoselective Conjugation Strategies for the Programmable Detection of Epigenetic Cytosine 5-Modifications with Transcription Activator-Like Effectors

Dissertation

Submitted for the degree of Doctor of Natural Sciences

(Dr. rer. nat.)

Presented by

Mario David Gieß

at the

fakultät für chemie
und chemische biologie

of the

technische universität
dortmund

Dortmund, MMXIX

“To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science.” – Albert Einstein

This work was prepared from April 2016 to July 2019 in the group of Prof. Dr. Daniel Summerer at the TU Dortmund University. Funding was provided by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the Priority Program entitled “Chemical Biology of native Nucleic Acid Modifications” (SPP 1784) and TU Dortmund University.

Acknowledgements

I want to thank first and foremost Prof. Dr. Daniel Summerer for the invitation and opportunity to work as a research scientist in his group and to be able to pursue my doctoral degree. Furthermore, I want to thank him for his open-minded attitude, the trust and creative freedom given to me as a graduate student and for the constructive discussion during my thesis.

I would also like to thank Prof. Dr. Daniel Rauh for kindly agreeing to be the second evaluator of this work.

I want to also thank the collaborators Dr. Oliver Koch and Julia Jasper whom I want to give credit to for providing computational TALE structure models for my first publication. Next, I would like to thank all the current members of the Dehmelt and Summerer lab Álvaro Muñoz-López, Anna Witte, Anne Jung, Benjamin Buchmuller, Brinja Kosel, Christoph Hoppe, Dominic Kamps, Jan Wolffgramm, Nadine Schmidt, Dr. Shubhendu Palei, Suchet Nanda, Tzu-Chen Lin and of the former members of the Summerer group especially Dr. Grzegorz Kubik and Katharina Kuhr for their contributions to my work and/or great time spent in- and outside of the lab. My thanks also include the student assistant Leonie Fleige for helping with everyday laboratory tasks and my bachelor students Patrick Günther and Kim Fischer, who both assisted my work with their theses. I want to extend my thanks to the staff of the department for chemical biology, namely Maria Sergani, Martina Reibner, Petra Alhorn and Ulrich Schoppe for their work and help regarding bureaucratic affairs and all things involving general lab management. Furthermore, I want to thank the Rauh lab for letting me use some of their equipment and Andreas Arndt for his help. I want to acknowledge the Max-Planck-Institute Dortmund for providing me with a guest status to allow me to use some of their equipment during my thesis. I want to thank the NMR facility of the TU Dortmund for their fast and excellent service and Christiane Heitbrink and Jens Warmers for their help with mass spectrometry.

I want to acknowledge funding from the TU Dortmund university and the priority program SPP1784 of the DFG.

To my friends and family, I can only express nothing but my deepest gratitude for their constant support, encouragement and interest in my work. Mi aprecio especial va a Marisabel González Ocanto por su amor, apoyo y motivación.

Table of Contents

	Acknowledgements	III
	List of Figures	VII
	List of Tables	XI
	List of Publications	XII
	Abbreviations	XIII
	Abstract	1
	Zusammenfassung	2
1.	Introduction	3
1.1.	The Chemical Biology of Epigenetic 5-Modified Cytosines	3
1.1.1.	Structure and Function of DNA	3
1.1.2.	Epigenetic DNA Cytosine 5-Modifications	9
1.2.	Detection and Analysis of Epigenetic Cytosine 5-Modifications	14
1.2.1.	Overview	14
1.2.2.	Analytical Platforms and Direct Detection Methods	15
1.2.3.	Chemical-based Detection Methods	16
1.2.3.1.	Bisulfite-based Methods	16
1.2.3.2.	Bisulfite-free Chemical-based Detection Methods	20
1.2.3.2.1.	5-Methylctosine	21
1.2.3.2.2.	5-Hydroxymethylcytosine	22
1.2.3.2.3.	5-Formylcytosine	24
1.2.3.2.4.	5-Carboxylcytosine	26
1.2.4.	Protein-based Approaches	27
1.3.	Transcription Activator-Like Effectors	29
1.3.1.	Structure and binding mode of TALE proteins	29
1.3.2.	The DNA recognition code of TALE repeats	33
1.3.3.	Design of TALE repeats and their application	35
1.4.	Protein engineering with non-canonical amino acids	37
1.4.1.	Strategies for site-selective non-canonical amino acid incorporation into proteins	37
1.4.2.	Applications of non-canonical amino acids	38
2.	Aim	39
3.	Results and Discussion	41

3.1. Complete, Programmable Decoding of Oxidized 5-Methylcytosine Nucleobases in DNA by Chemoselective Blockage of Universal TALE Repeats 41
3.1.1. Design of Universal TALE Repeats 41
3.1.2. Cytosine 5-modification-selective Blockage of TALE Binding 50
3.1.3. Complete decoding of oxidized 5mC nucleobases 57
3.2. Programmable Protein-DNA Crosslinking for the Direct Capture and Quantification of 5-Formylcytosine 60
3.2.1. Design of TALE Repeats bearing para-Acetylphenylalanine 60
3.2.2. Programmable crosslinking of TALE repeats with pAcF and 5fC 64
3.2.3. Detection and quantification of single 5fC-sites in genomic DNA 77
4. Summary and Outlook 80
5. Materials and Methods 82
5.1. Materials 82
5.1.1. Table 1: List of Services 82
5.1.2. Table 2: List of Software and Tools 83
5.1.3. Table 3: List of Lab Equipment 85
5.1.4. Table 4: List of Disposables and Glass Ware 88
5.1.5. Table 5 List of Consumables 91
5.1.6. Table 6: List of Commercial Kits and Master Mixes 93
5.1.7. Table 7: List of Enzymes 94
5.1.8. Table 8: List of Chemicals 95
5.1.9. Table 9: List of Buffers and Stock Solutions 99
5.1.10. Table 10: List of Oligonucleotides 102
5.1.11. Table 11: List of Plasmids 110
5.1.11. Table 12: List of TALE Proteins 112
5.1.12. Table 13: List of Strains/Cell Lines 114
5.2. Methods 115
5.2.1. Biological/Biochemical Methods 115
5.2.1.1. Plate Culture 115
5.2.1.2. Liquid Overnight Culture 115
5.2.1.3. Preparation of Chemical-competent GH317 E.coli bacteria 115
5.2.1.4. Preparation of Chemical-competent BL21 E.coli bacteria 115
5.2.1.5. Preparation of Electro-competent GH371 or BL21 E.coli 116

5.2.1.6. Transformation by Heat Shock 116
5.2.1.7. Transformation by Electroporation 117
5.2.1.8. Cassette Mutagenesis 117
5.2.1.9. Site-directed Mutagenesis (SDM) by PCR 118
5.2.1.10. Agarose Gel Electrophoresis 118
5.2.1.11. Colony PCR 119
5.2.1.12. TALE Assembly via Golden Gate Reaction 119
5.2.1.13. SDS-PAGE 122
5.2.1.14. TALE Expression 122
5.2.1.15. Circular Dichroism Spectroscopy 123
5.2.1.16. Electromobility Shift Assay 123
5.2.1.17. Structure Modelling 124
5.2.1.18. Luciferase Assay 124
5.2.1.19. Phenol/Chloroform Extraction and Ethanol Precipitation of DNA 125
5.2.1.20. Enzymatic Glucosylation 125
5.2.1.21. Preparation of Oligonucleotides for MALDI-TOF MS 126
5.2.1.22. Crosslinking Efficiency Analysis 126
5.2.1.23. Preparation of Spike-In DNA 127
5.2.1.24. Purification of DNA via Agarose Gel Electrophoresis 128
5.2.1.25. Preparation of Genomic DNA (for chapter 3.1.3) 128
5.2.1.26. Quantitative PCR (qPCR) 129
5.2.1.27. Enrichment (for chapter 3.1.3) 129
5.2.1.28. Preparation of Genomic DNA (for chapter 3.2.3) 130
5.2.1.29. Enrichment (for chapter 3.2.3) 131
5.2.2. Chemical Methods 131
5.2.2.1. Oxime Formation 131
5.2.2.2. Amide Formation 132
5.2.2.3. Linker Synthesis (modified from Holder, Francis 2007386) 133
5.2.2.4. Crosslinking Reaction 139
6. Supplementary Data 137
7. Eidesstattliche Versicherung (Affidavit) 170
8. References 171

List of Figures

Figure 1	Deoxyribonucleotides and structure of the DNA double helix.	4
Figure 2	The organization of DNA into three levels of chromatin structure	5
Figure 3	The flow of the genetic information.	6
Figure 4	Structures, names and three letter and single letter code for the 20 proteinogenic amino acids grouped according to chemical properties.	7
Figure 5	DNMT-mediated methylation of Cytosine with SAM.	9
Figure 6	Active demethylation of 5-methylcytosine involves the oxidation and repair of 5mC.	11
Figure 7	Methods for the analysis of cytosine 5-modifications.	14
Figure 8	Bisulfite conversion of C, 5mC and ox5mCs and transition of converted nucleobases during PCR.	16
Figure 9	Analysis of 5hmC, 5fC and 5caC with bisulfite-sequencing (BS-Seq) strategies.	18
Figure 10	Exemplary Bisulfite-free strategies for the selective modification of cytosine 5-modifications.	20
Figure 11	Protein-based strategies for the analysis of cytosine modifications.	27
Figure 12	Structure and modular organization of natural TALE proteins.	30
Figure 13	Structure of the cryptic repeats -1 and 0 of the PthXo1 TALE NTR and interaction of W232 with the initial T of the target DNA.	31
Figure 14	Repeat structures in complex with DNA nucleobases.	33
Figure 15	Concept of chemoselective blockage of TALE binding for complete decoding of oxidized 5mC nucleobases in DNA.	39
Figure 16	Concept of 5fC-specific crosslinking with TALE proteins bearing a size-reduced repeat containing *p*AcF through oxime condensation.	40
Figure 17	The modular organization of TALEs and design of size-reduced TALE repeats.	42

Figure 18	TALE expression and circular dichroism spectroscopy.	43
Figure 19	EMSA with TALEs and target DNA with C opposite the indicated repeat.	44
Figure 20	Binding profiles of CDKN2A TALEs with target DNA bearing C, 5mC, 5hmC, 5fC or 5caC opposite the variable repeat.	46
Figure 21	Models of TALEs bound to targets with C or 5caC opposite of the variable repeat position.	47
Figure 22	Binding curves of TALE_SG*GG with CDKN2A target DNA.	48
Figure 23	Interaction of CDKN2A TALEs with target DNA bearing C, A, G or T opposite the variable repeat.	49
Figure 24	5hmC-selective blockage of TALE binding.	51
Figure 25	*p*PDA-catalzyed oxime formation of 5fC-DNA with hydroxylamines 3a-d.	52
Figure 26	5fC-selective blockage of universal TALE binders.	53
Figure 27	PyAOP-mediated amide formation with 5caC-DNA with amines 5a-d.	55
Figure 28	5caC-selective blockage of universal TALE binders.	56
Figure 29	Complete decoding of oxidized 5mC nucleobases in a genomic DNA background with a TALE-based enrichment assay.	57
Figure 30	Engineering size-reduced TALE repeats bearing the non-canonical amino acid (ncAA) *para*-acetylphenylalanine (*p*AcF, **1**).	61
Figure 31	Interaction of TALEs with BRCA1a target DNA bearing one of the five cytosines opposite of the variable repeat as indicated.	62
Figure 32	Crosslinking of 5fC with 1 via linker L1-L5 under mild acidic conditions.	64
Figure 33	5fC-specific DNA-TALE-crosslinking with TALE proteins bearing the ncAA *p*AcF.	66

Figure 34	Test for protein-protein crosslinking of BRCA1b_TALEs and off-target crosslinking to CDKN2A-DNA with 5fC as indicated.	67
Figure 35	Quantification of crosslinking reactions with indicated TALE-linker combinations by SDS-PAGE analysis using ^{32}P-labeled DNA.	68
Figure 36	TALEs for positional resolution of 5fC-crosslinking.	70
Figure 37	SDS-PAGE images of positional resolution of TALE-5fC-DNA crosslinking for BRCA1b.	71
Figure 38	Quantification of positional resolution of TALE-5fC-DNA crosslinking for BRCA1b using ^{32}P-labeled DNA.	72
Figure 39	Programmable TALE-DNA crosslinking with engineered last repeats.	74
Figure 40	Test for protein-protein crosslinking of BRCA1c_TALEs and off-target crosslinking to CDKN2A-DNA with 5fC as indicated.	75
Figure 41	Quantification of crosslinking reaction yields with indicated TALE-linker combinations by SDS-PAGE analysis using ^{32}P-labeled DNA.	76
Figure 42	Covalent enrichment of 5fC-containing DNA sequences from mammalian genomic DNA backgrounds.	78
Figure SI1a	Module, array and backbone vectors for TALE assembly.	137
Figure SI1b	TALE assembly via two-step Golden Gate reaction.	138
Figure SI1c	Exemplary pET vector for TALE expression.	139
Figure SI2a	Protein sequence and molecular mass of CDKN2A_TALEs.	140
Figure SI2b	Protein sequence and molecular mass of BRCA1_TALEs.	141
Figure SI3	ESI-MS spectra of dsDNA before and after glycosylation.	142
Figure SI4	Representative EMSA data for differential TALE interaction with their cognate targets before and after treatment as shown in Fig. 24d, Fig. 26a, and Fig. 28a.	143
Figure SI5	MALDI-TOF MS spectra of dsDNAs used for 5fC conversion with hydroxylamines and 5caC conversion with amines.	144

Figure SI6	Oxime formation of 5fC-DNA with Alexa Fluor 488 hydroxylamine.	146
Figure SI7	Superimposed models of DNA containing benzyloxime (4c) or tert-butyloxime (4d) bound by TALE_SG*GG.	147
Figure SI8	Model of DNA containing benzylamide (6c) (grey sticks) clashing with non-minimized structure of repeat_SG*GG.	147
Figure SI9	Spike-ins and genomic DNA for enrichment.	148
Figure SI10	Plasmid maps for vectors containing genes for a) *p*AcF tRNA/aaRS, for b) BRCA1a and BRCA1b or c) BRCA1c genes.	149
Figure SI11a	Protein sequence and molecular mass of BRCA1a_TALEs.	150
Figure SI11b	Protein sequence and molecular mass of BRCA1b_TALEs.	151
Figure SI11c	Protein sequence and molecular mass of BRCA1c_TALEs.	152
Figure SI12	SDS-PAGE images of TALE proteins.	153
Figure SI13	^{1}H and ^{13}C NMR Spectra of L1-5	154
Figure SI14	High Resolution Mass Spectra of L1-5	159
Figure SI15	5 % agarose gels of 5'-FAM labeled RefSeq_5fC12 oligonucleotide after incubation under crosslinking conditions do not show crosslinking between 5fC-bearing oligonucleotides.	164
Figure SI16	MALDI-TOF mass spectrometry analysis of reference oligonucleotides with 5fC after incubation under crosslinking conditions without or with linkers L1-5.	165
Figure SI17	EMSA assays of BRCA1a TALEs with pAcF and its target DNA under crosslinking reaction conditions.	166
Figure SI18	SDS-PAGE gel image for TALE-DNA crosslinking as indicated from Fig. 35.	167
Figure SI19	SDS-PAGE gel images for positional resolutions of TALE-DNA crosslinking as indicated from Fig. 38.	168
Figure SI20	Spike-ins and genomic DNA for enrichment.	169

List of Tables

Table 1	List of Services	82
Table 2	List of Software and Tools	83
Table 3	List of Lab Equipment	85
Table 4	List of Disposables and Glass Ware	88
Table 5	List of Consumables	91
Table 6	List of Commercial Kits	93
Table 7	List of Enzymes	94
Table 8	List of Chemicals	95
Table 9	List of Buffers and Stock Solutions	99
Table 10	List of Oligonucleotides	102
Table 11	List of Plasmids	110
Table 12	List of TALE Proteins	112
Table 13	List of Strains/Cell Lines	114
Table 14	SDM-PCR reagent concentrations in 25 µl	118
Table 15	SDM-PCR reaction times and temperatures	118
Table 16	Colony PCR reagent concentrations in 25 µl	119
Table 17	Colony PCR reaction times and temperatures	119
Table 18	Golden Gate 1 reagent concentrations in 25 µl	120
Table 19	Golden Gate 1 reaction times and temperatures	120
Table 20	Golden Gate 2 reagent concentrations in 25 µl	121
Table 21	Golden Gate 2 reaction times and temperatures	121
Table 22	PCR reagent concentrations for the preparation of spike-in DNA in 50 µl	127
Table 23	PCR settings for the preparation of spike-in DNA	128
Table 24	qPCR reaction times and temperatures	129
Table SI1	Theoretical and observed peak list of BRCA1_5hmC before and after glycosylation by ESI-TOF mass spectrometry.	142
Table SI2	Observed and calculated masses for 5fC- and 5caC-ODN from data of Fig. SI5.	14

List of Publications

(1) M. Gieß, Á. Muñoz-López, B. Buchmuller, G. Kubik, D. Summerer, *JACS* **2019**, 141, 9453

(2) M. Gieß, A. Witte, J. Jasper, O. Koch, D. Summerer, *JACS* **2018**, *140*, *5904*.

(3) S. Maurer, M. Gieß, O. Koch, D. Summerer, *ACS Chem. Biol.* **2016**, *11*, 3294.

(4) S. Flade, J. Jasper, M. Gieß, M. Juhasz, A. Dankers, G. Kubik, O. Koch, E. Weinhold, D. Summerer, *ACS Chem. Biol.* **2017**, *12*, 1719.

(5) D. Summerer, M. Gieß, G. Kubik, S. Maurer, EP3214183 (B1), filed 2016 by the University of Konstanz, published **2018**.

Abbreviations

*	Deletion
1	*para*-Acetylphenylalanine
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
A	Adenine, Adenosine
aa	Amino acid
aaRS	aminoacyl-RNA Synthetase
ACE-seq	APOBEC-coupled epigenetic sequencing
AD	Activation domain
ADD	ATRX–DNMT3–DNMT3L
AI	1,3-indandione
Ala, A	Alanine
AP	Alkaline phosphatase
Arg, R	Arginine
ARP	Aldehyde reactive probe
Asn, N	Asparagine
Asp, D	Aspartate, Aspartic acid
azi-BP	3-azido-N-(2-(cyanomethyl)benzo[d]thiazol-6-yl)propanamide
BER	Base excision repair
BGT	β-glucosyltransferase
bp	Base pair
bg	Background
BS	Bisulfite
C	Cytosine, Cytidine
CAB	Chemical modification-assisted bisulfite
CAPS	Chemical-assisted pyridine borane sequencing
CBAN	2-(5-chlorobenzo[d]thiazol-2-yl)acetonitrile
CD	Circular dichroism
CET	C-to-T transition
cf.	confer (*lat.)* – compare to/with

COBRA	combined bisulfite restriction analysis
CRD	Central repeat domain
CTCF	CCCTC-binding factor
CTR	C-terminal domain
Cys, C	Cysteine
DBCO	Dibenzocycloocytne
DBD	DNA-binding domain
DHU	Dihydroxyuracil
DIP	DNA immunoprecipitation
DNA	Deoxyribonucleic acid
DNMT	DNA methyl transferases
DNMT3L	DNMT3-like
dNTP	Deoxyribonucleoside triphosphates
DPC	DNA-protein crosslink
ds	Double-stranded
e.g.	exempli gratia (*lat.*) – for example
ECL	Electrogenerated chemiluminescence
EDC	1-ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride
ELISA	Enzyme-linked immunosorbent assay
EMSA	Electro mobility shift assay
EPR	Electron paramagnetic resonance
ESI	Electrospray ionization
Fig.	Figure
FRET	Förster resonance energy transfer
fCAB	5fC-selective chemical modification-assisted bisulfite
G	Guanine, Guanosine
GC	Gas chromatography
gDNA	Genomic DNA
GE	Gel electrophoresis
GFP	Green fluorescent protein
Gln, Q	Glutamine
Glu, E	Glutamate, Glutamic acid
Gly, G	Glycine
H	Histone

His, H	Histidine
HPLC	High-performance liquid chromatography
Ile, I	Isoleucine
IR	Infrared
J	β-d-glucosyl-hydroxymethyluracil
JBP1	J-binding protein 1
K_D	Dissociation constant
L	Linker
LCR	Ligation chain reaction
Leu, L	Leucine
LR	Last repeat
Lys, K	Lysine
MAB	M.SssI methylase-assisted BS
*m*AcF	*meta*-Acetylphenylalanine
MALDI	Matrix-assisted laser desorption/ionization
MAX	Myc-associated factor X
MBD	Methyl-binding domain
mESC	Mouse embryonic stem cell
min	Minute(s)
mRNA	Messenger RNA
MS	Mass spectrometry
MSP	Methylation-specific PCR
MV	Module vector
MW	Molecular weight
ncAA	Non-canonical amino acid
NLS	Nuclear localization sequence
NMR	Nuclear magnetic resonance
nt	Nucleotide
NTA	Nitrilotriacetic acid
NTR	N-terminal domain
ORF	Open reading frame
ox5mC	oxidized 5-methylcytosine
oxBS	Oxidative BS
p	Indicates 3’-5’-phosphodiester bond

*p*AcF	*para*-Acetylphenylalanine
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
pdb	Protein database
Phe, F	Phenylalanine
Pro, P	Proline
PTM	Post-translational modification
PyAOP	(7-azabenzotriazol-1-yloxy)tripyrrolidino-phosphonium hexafluorophosphate
qPCR	Quantitative PCR
R	Residues
redBS	Reduced BS
ref	Reference
RLA	Relative luciferase activity
RNA	Ribonucleic acid
RP	Reverse phase
rpm	Rounds per minute
RRBS	Reduced representation BS
rRNA	ribosomal RNA
RVD	Repeat divariable residue
SAM	S-adenosylmethionine
scRRBS	single cell RRBS
SDS	Sodium dodecyl sulfate
Seq	Sequencing
Ser, S	Serine
SERS	Surface-enhanced Raman spectroscopy
SMRT	Single molecule real-time
SNP	Single nucleotide polymorphism
ss	Single-stranded
STR	Short tandem repeat
T	Thymine, Thymidine
TAB	Tet-assisted BS
TALE	Transcription activator-like effector
TAPS	TET-assisted pyridine borane sequencing
TDG	Thymine DNA glycosylase

TE	Transposable element
TEMPO	2,2,6,6-tetramethylpiperidine-1-oxyl
TET	Ten-eleven translocation
TEV	TALE expression vector
TF	Transcription factor
Thr, T	Threonine
TLC	Thin layer chromatography
Tm	Melting temperature
TMI	2,3,3-trimethylindole
TOF	Time of flight
tRNA	Transfer RNA
Trp, W	Tryptophan
TSS	Transcription start site
Tyr, Y	Tyrosine
U	Uracil, Uridine
UAA	Unnatural amino acid
UDP	Uridine diphosphate
UHRF1	ubiquitin-like plant homeodomain and RING finger domain 1
Un	Unit
Val, V	Valine
vdW	van-der-Waals
WGA	Whole genome amplification
XL	Crosslink formation
ZF	Zinc finger

Abstract

Methylation at the cytosine (C) carbon-5 position in DNA is a reversible regulatory element of transcription in mammalian cells involved in development and disease. An active demethylation pathway through iterative oxidation of 5-methylcytosine (5mC) has been identified that leads to abasic sites at which unmodified C are restored via the base excision repair (BER) pathway. This process yields 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) as partially stable intermediates. An increasing amount of data indicate that these 5-modified cytosines can also act as epigenetic regulatory elements and are involved in developmental and pathological processes. However, the function of epigenetic 5-modified cytosines are not fully understood and require sensitive typing and profiling approaches with high resolution.

Engineered transcription activator-like effector (TALE) proteins have been established as programmable DNA-binders for the detection of epigenetic 5-modified cytosines. However, previous efforts to engineer new TALE selectivities did either not provide full selectivity for epigenetic 5-modified cytosines or had low affinity to target DNA. To overcome these hurdles, two novel TALE-based strategies based on chemical conjugation reactions were developed.

In the first approach, a size-reduced TALE repeat with universal binding to the four canonical nucleobases and four epigenetic 5-modified cytosines was engineered. Chemoselective conjugation of 5hmC, 5fC or 5caC to dedicated blocking groups was found to abolish TALE binding. This enabled complete decoding of the three oxidized 5mC derivatives (ox5mCs) at single positions within target DNA.

In the second approach, the non-canonical amino acid (ncAA) *para*-acetylphenylalanine (*p*AcF) was incorporated into dedicated positions of the universal TALE repeat. TALEs bearing *p*AcF maintained universal binding to cytosine 5-modified nucleobases. Furthermore, *p*AcF-bearing TALE repeats and 5fC-bearing target DNA could be crosslinked through oxime condensation via bifunctional hydroxylamine linkers. This enabled robust and selective enrichment of 5fC-bearing target DNA from genomic DNA (gDNA) backgrounds.

Taken together, the addition of selective conjugation chemistries to TALE-based detection methods expand the toolbox for the selective, programmable analysis of epigenetic 5-modified cytosines.

Zusammenfassung

DNA-Methylierung der Carbon-5-Position von Cytosin (C) ist ein reversibles regulatorisches Element der Transkription in Säugetierzellen, dass in Entwicklungsprozessen und Krankheitsbildern involviert ist. Die aktive Demethylierung durch iterative Oxidation von 5-Methylcytosin (5mC) ist eine Möglichkeit C über Basen-Exzisionsreparatur (BER) von dabei entstehenden abasischen Positionen wiederherzustellen. Dabei werden 5-Hydroxymethylcytosin (5hmC), 5-Formylcytosin (5fC) und 5-Carboxylcytosin (5caC) gebildet. Daten deuten vermehrt darauf hin, dass diese 5-modifizierten Cytosine epigenetische regulatorische Elemente sein können und in Entwicklungs- und pathologischen Prozessen involviert sind. Die Funktion von 5-modifizierten Cytosinen ist jedoch noch nicht vollständig geklärt und erfordert neue sensible Detektionsmethoden.

Transkriptionsaktivator-ähnliche Effektor-Proteine (TALE) wurden als programmierbare DNA-Bindeproteine für den Nachweis von epigenetischen 5-modifizierten Cytosinen etabliert. Diese TALEs boten jedoch entweder nicht die vollständige Selektivität für einzelne epigenetische 5-modifizierte Cytosine oder hatten eine niedrige Gesamtaffinität. Um diese Hürden zu überwinden, wurden zwei neue TALE-basierte Strategien entwickelt, die auf chemischen Konjugationsreaktionen beruhen.

Im ersten Projekt wurden universell bindende TALE-Repeats konstruiert. Daraufhin konnte über chemo-selektive Konjugation von 5hmC, 5fC oder 5caC mit dedizierten Blockierungsgruppen die TALE-Bindung aufgehoben werden. Dies ermöglichte die vollständige Dekodierung der drei oxidierten 5mC-Derivate an einzelnen Positionen innerhalb definierter DNA-Sequenzen.

Im zweiten Projekt wurde die nicht-kanonische Aminosäure (ncAA) *para*-Acetylphenylalanin (*p*AcF) an dedizierten Positionen des universellen TALE-Repeat eingebaut. Diese TALE-Repeats mit *p*AcF banden weiterhin universell an epigenetische 5-modifizierte Cytosine. Des Weiteren konnten sie mit DNA-Sequenzen mit 5fC durch Oximkondensation über bifunktionelle Hydroxylamin-Linker vernetzt werden. Dies ermöglichte eine robuste und selektive Anreicherung von DNA-Sequenzen mit 5fC aus genomischer DNA (gDNA).

Diese Strategien erweitern die Toolbox von bisher existierenden TALE-basierten Nachweismethoden für die selektive, programmierbare Analyse epigenetischer 5-modifizierter Cytosine.

1. Introduction

1.1. The Chemical Biology of Epigenetic 5-Modified Cytosines

1.1.1. Structure and Function of DNA

Deoxyribonucleic acid (DNA) is known as the blueprint of life.[1] DNA encodes the essential instructions for development and function of an organism.[2] The entirety of DNA sequence information of a cell is referred to as the genome.

DNA is a linear polymer consisting of nucleotide monomers linked via a phosphodiester bond (**Figure** (**Fig.) 1a, b**).[3] Nucleotides in DNA consist of a 2-dexoyribose, a 5'-phosphate group and one of the four nucleobases cytosine (C), thymine (T), guanine (G) and adenine (A) (**Fig. 1c**).[4]

In the DNA, C and G as well as A and T interact via hydrogen bonds and form specific base pairs (**Fig. 1c**).[5] This complementary base-pairing is a major stabilizing factor for the formation of the antiparallel orientation of double-stranded (ds) DNA (**Fig. 1d**).

The DNA sugar-phosphate backbone adopts a helical structure of the dsDNA which is mainly stabilized via π-π stacking interactions of the aromatic nucleobases.[6] A characteristic structural feature of the DNA double helix is the formation of minor and major grooves (**Fig. 1e**).[3] Depending on the sequence and environmental conditions the conformation of the DNA double helix can vary (**Fig. 1f**).[7]

In Eukaryotes, the dsDNA folds into structures called chromosomes. Here, the DNA is wrapped around a DNA-binding protein complex consisting of two copies each of the histone proteins H2A, H2B, H3, and H4 to form nucleosomes (**Fig. 2**).[8]

The resulting chromatin 10 nm fiber can be compacted further through interaction with DNA-binding proteins into a fiber of approximately 30 nm and further into a 60–130 nm chromonema fiber (**Fig. 2**).[3,9]

The structure of DNA molecules is essential for its functionality. They provide the basis for a hereditary mechanism as both strands can act as a template for the synthesis of new DNA molecules. The synthesis of DNA from another DNA template in a cellular environment is called replication. This process is spatially and temporally tightly regulated and carried out by a complex machinery.[10]

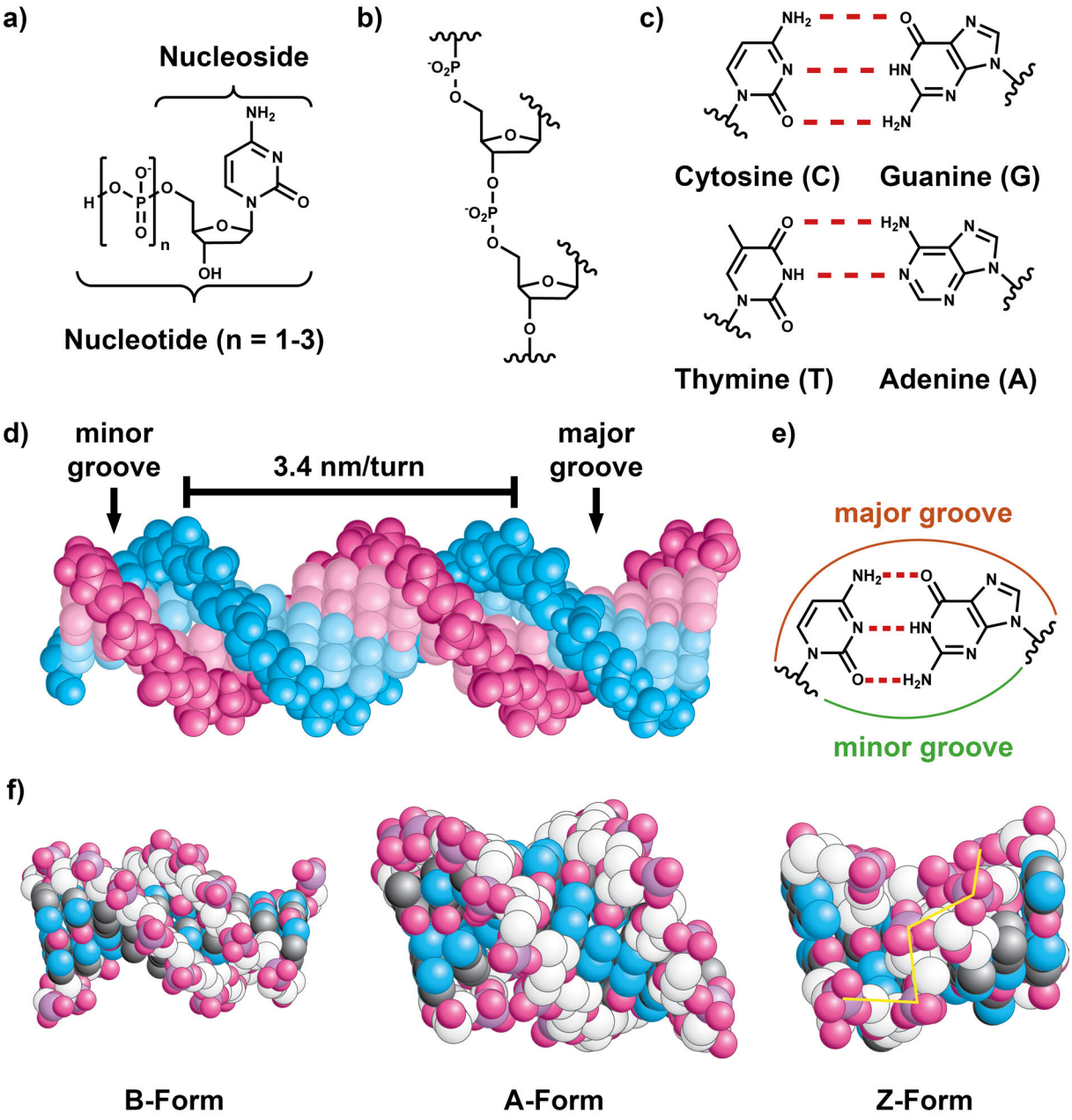


Figure 1: Deoxyribonucleotides and structure of the DNA double helix.[a]

a) Structure of 2'-deoxyribonucleoside n-phosphates (n = mono, di, tri). **b)** The DNA backbone is linked via a 3'-5' phosphodiester bond. c) Watson-Crick base pairs between pyrimidines (C/T) and purines (G/A). **d)** Watson-Crick model of dsDNA with the two antiparallel forward and reverse strands with the sugar-phosphate-backbone on the outside (pink and blue) and the base pairs facing (light blue and pink) towards the inside. **e)** Major and minor groove site in the Watson-Crick DNA double helix. **f)** DNA helix depending on sequence and environmental conditions.

[a] Figures in d) and f) reprinted (adapted) with permission from reference (ref.) [3]. Copyright 2019 Springer Nature.

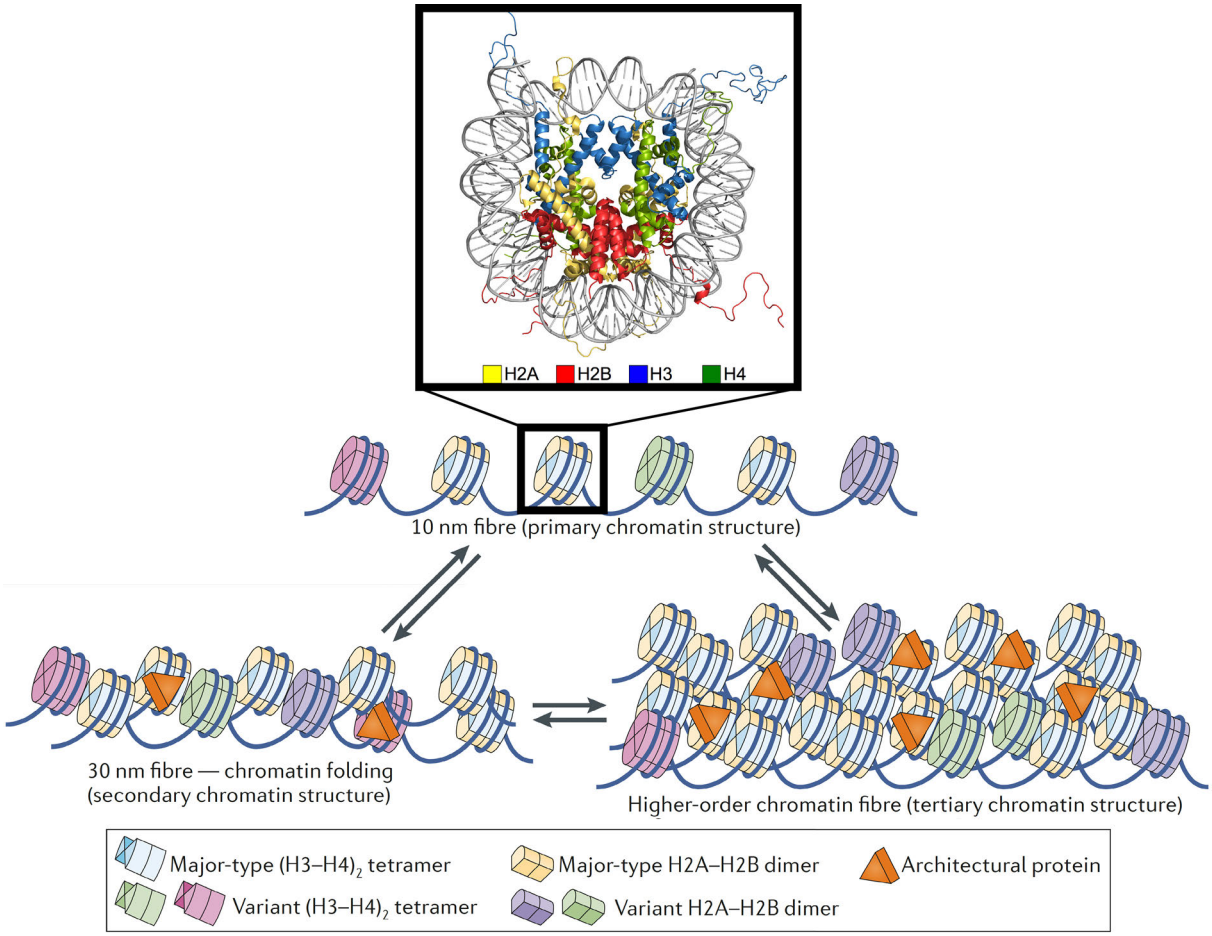


Figure 2: The organization of DNA into three levels of chromatin structure.[b]

The flow of genetic information (**Fig. 3**) is almost the same in all forms of life.[3] It is based on the expression (transcription) of specific sequence fractions termed genes. During gene expression, the DNA is first transcribed to ss ribonucleic acid (RNA), which is a polynucleotide with the nucleobases A, G, C and uracil (U) instead of T.[3,13] Usually, the transcription machinery consists of a RNA polymerase for the synthesis of the RNA, and additional transcription factors (TFs) for regulation of initiation, elongation and termination.[14]

Essentially, the transcription machinery assembles near the transcription start sites (TSS) of genes at promotors and synthesizes RNA from the DNA template.[15]

After transcription, the messenger RNA (mRNA) can be used as template for the translation of the mRNA into proteins.

[b] reprinted (adapted) with permission from ref.[11] and ref.[12]. Copyright 2019 Okinawa Institute of Science and Technology Graduate University and copyright 2019 Springer Nature.

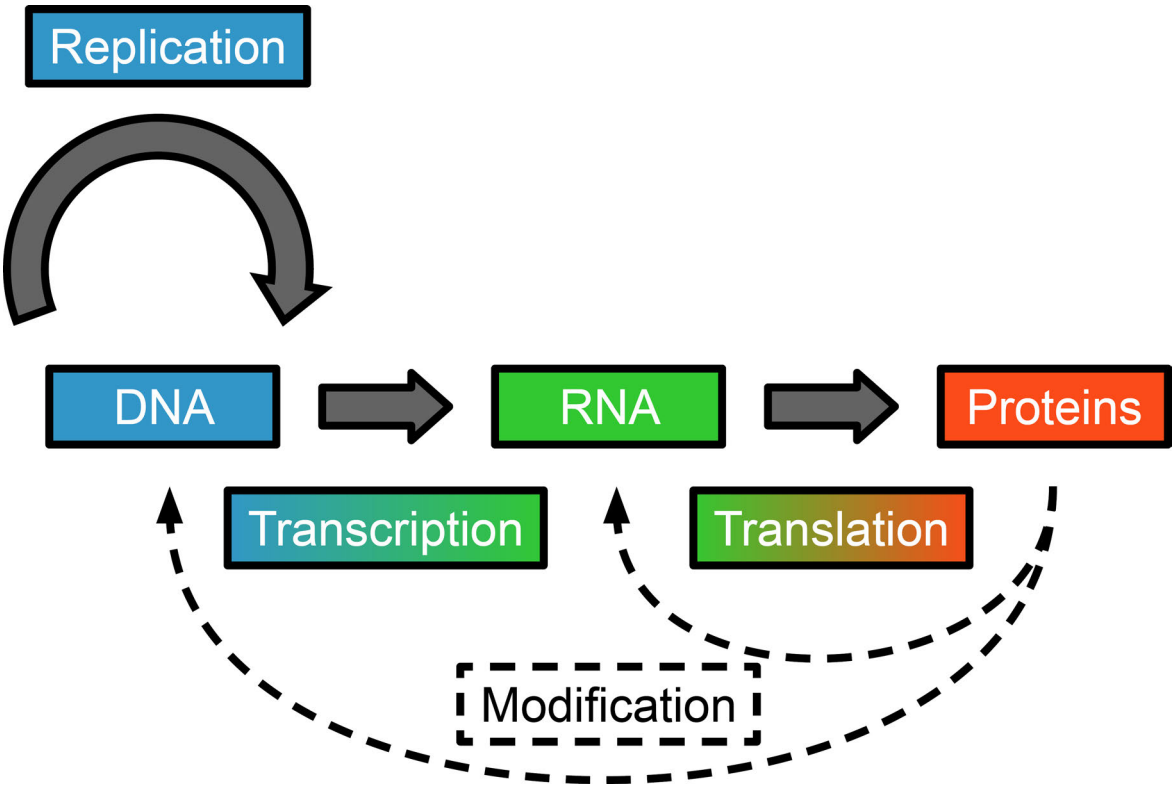


Figure 3: The flow of the genetic information.

Proteins are biopolymers consisting of amino acids which are linked via amide bonds to a polypeptide chain. Proteins catalyze most biochemical reactions in biological processes.[9] There are 20 standard proteinogenic amino acids with different α-substituents and distinct chemical properties (**Fig. 4**).[3] The translation of mRNA into proteins follows a non-overlapping, degenerate code where one amino acid is encoded by one or multiple sequences of nucleotide triplets (codon) (**Fig. 4**).[3]

Codons are recognized through interaction of a complementary nucleotide triplet (anticodon) of the transfer RNAs (tRNA).[3] tRNAs are aminoacylated with a specific amino acid through ATP activation by high-fidelity aminoacyl-RNA synthetases (aaRS).[16] The aminoacylation of the tRNA depends on interaction of a cognate tRNA/aaRS pair.[16]

Translation is catalyzed by ribosomes. These complexes consist of ribosomal RNA (rRNA) and specific proteins.[17]

The codon AUG serves as the start codon for translation and encodes methionine (Met, M). At the beginning of protein synthesis, the synthesis is initiated through the binding of the aminoacylated tRNA-Met in the ribosome at the start codon of an open reading frame (ORF).

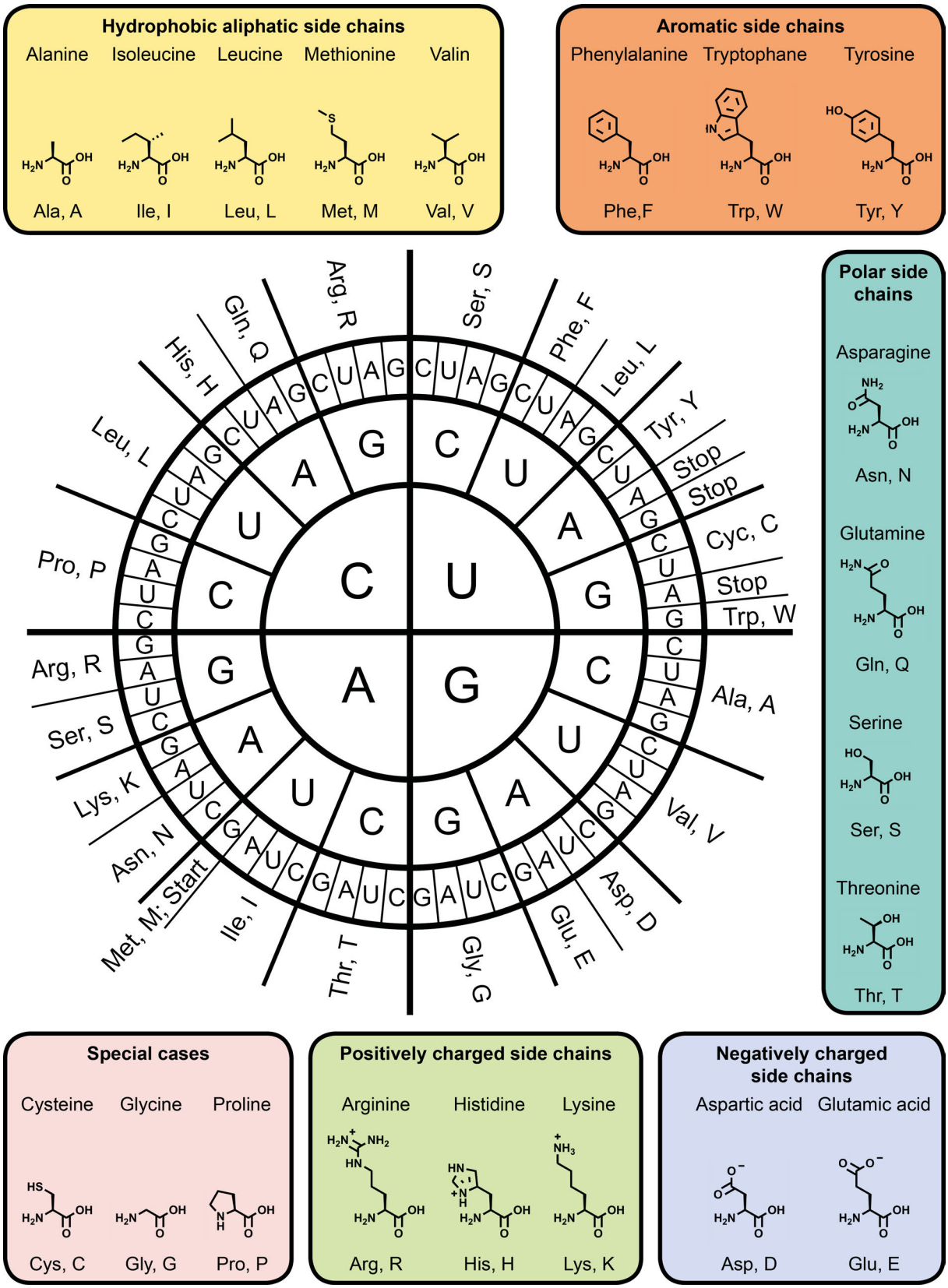


Figure 4: Structures, names and three letter and single letter code for the 20 proteinogenic amino acids grouped according to chemical properties. The codon diagram displays the nucleotide triplet sequences encoding the corresponding amino acids.

Then, the next cognate elongating aminoacyl-tRNA can be delivered to the ribosome.[18] There, elongation factors promote the transfer of the amino acid to the growing peptide chain and deacylated tRNAs dissociate from the mRNA.[19]

The three codons UAA, UAG, and UGA are designated as termination codons, which are also referred to as ochre (UAA), amber (UAG) and opal (UGA) stop codons.[20]

Throughout the life cycle of organisms, not all genetic information is constitutively expressed. Instead, distinct set of genes are required. Spatial and temporal control of gene expression can be influenced through epigenetic mechanisms. That is through structural modification but without alternating the genetic sequence of chromosomal DNA.[21] For example, the modification of histone molecules can lead to the formation of decondensed euchromatin accessible for the transcription machinery or condensed heterochromatin that “silences” gene expression.[22]

Another mechanism is the modification of the DNA. The major groove contains a nucleobase-specific layer of chemical information. This enables interaction of regulatory proteins without dissociating overall Watson-Crick base pairing. Therefore, modification of nucleobases in the major groove can affect binding of regulatory DNA-binding proteins and shape the structure and function of the genome.[23]

1.1.2. Epigenetic DNA Cytosine 5-Modifications

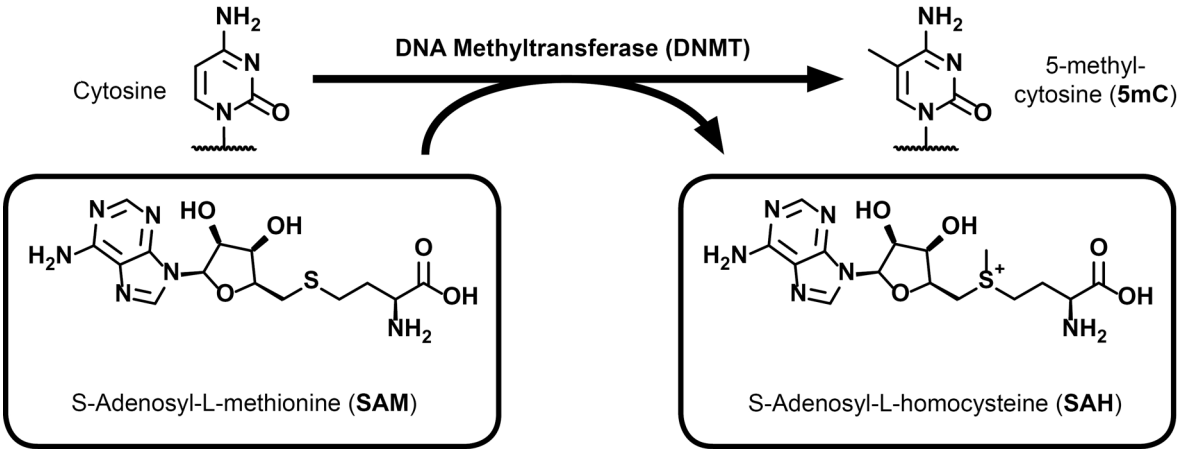


Figure 5: DNMT-mediated methylation of Cytosine with SAM.

Methylation at the carbon-5 position of cytosines is the most abundant epigenetic mark in the human genome (**Fig. 5**). Cytosine 5-methylation is catalyzed by DNA methyl transferases (DNMT) through the formation of a covalent DNA-protein intermediate with *S*-adenosylmethionine (SAM) (**Fig. 5**).[24,25]

Of all cytosines in the human genome, approximately 1-4 % are methylated.[26,27] 5-Methylcytosine (5mC) is associated with transcriptional regulation, recruitment of transcription factors, and modulation of chromatin structure.[28–30] Aberrant methylation can also be a hallmark of diseases like cancer or neurological disorders.[31,32]

5mC could be an important part of a complex genome-wide crosstalk system.[33,34] For example, 5mC can recruit histone-modifying proteins through interaction with methyl-CpG-binding domain (MBD) .[35] Vice versa, transcriptional activity can modify chromatin structure or -associated proteins to recruit writers of cytosine 5-methylation.[36]

Between the dinucleotide combinations CpG, CpA, CpT and CpC more than 99 % of cytosine 5-methylation can be found in CpG sequences.[32,37] Despite its rarity, non-CpG-methylation can also be a prevalent epigenetic mark of embryonic stem cells and aging brain cells.[37,38]

Approximately 70 to 80 % of genomic CpGs are methylated.[36] Most unmethylated CpG sequences are in CpG-rich regions (CpG-islands) of regulatory regions such as promotors or first exons.[29,39] These regions are involved in transcription and often occupied by CXXC domain containing proteins and TFs, which protect the DNA from methylation.[34] Cytosine 5-methylation in promotor sequences (~10% of CpGs) is

usually associated with long-term, stable gene repression, cell differentiation, X-chromosome inactivation and genomic imprinting.[40]

In contrast, 5mCpGs are predominantly found in repetitive sequences, gene bodies, and intergenic regions.[39] An example for extensively methylated repetitive regions in the human genome are transposons. Transposon-derived sequences can affect genome integrity and need to be tightly regulated.[29] Thereby, cytosine 5-methylation is likely responsible for stable heritable transcriptional repression and silencing in these regions.[23,29]

Cytosine 5-methylation levels and patterns in other genome sequences are often highly variable and can depend on tissue type or developmental stages.[29,34] For example, 5mC levels in mammalian cells are depleted after fertilization and during germ cell development.[23] Additionally, transcriptional regulation by 5mC need to be dynamic. Therefore, parts of the epigenome are continuously “re-written”. Demethylation can occur passively over multiple cycles of DNA replication.[41] Alternatively, 5mCpG-sites can be removed through active demethylation in an iterative enzymatic process that yields oxidized 5mC derivatives (ox5mC) (**Fig. 6**).[42] The oxidation of 5mC is catalyzed by dioxygenases of the Ten-eleven translocation (TET) protein family, in both processive and non-processive manner.[43,44] In mammals, there are three homologs (TET1, TET2 and TET3) with cell/organ-specific expression.[45] These enzymes oxidize the 5-group of cytosine with α-ketoglutarate (α-KG) as cofactor at an Fe(II)-center with activated dioxygen (O_2) via an Fe(IV)-oxo intermediary state.[46,47]

During this process 5mC is first oxidized to 5-hydroxymethylcytosine (5hmC), then consecutively oxidized to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by TET proteins.[48–51]

5fC and 5caC, but not 5hmC can be actively removed with thymine DNA glycosylase (TDG), resulting in abasic sites.[50,52] Unmodified cytosines are restored at these sites via the base excision repair (BER) pathway.[53] Additionally, in mouse embryonic stem cells (mESCs) 5fC and possibly 5caC can undergo C–C bond cleavage for the direct removal of the 5-substituent.[54,55] It could be possible that iterative oxidation is an evolutionary solution to the challenging direct removal of the methyl group which is a poor leaving group.

Additionally, ox5mCs have distinct chemical properties which could add another layer of accessible chemical information to the DNA.

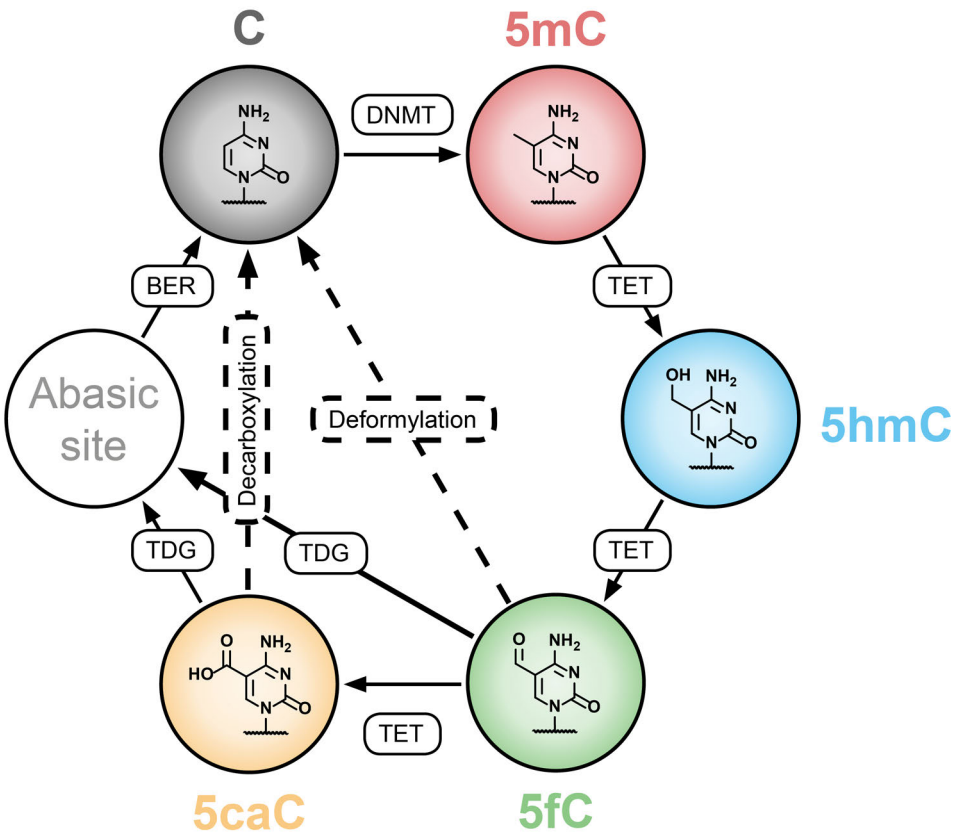


Figure 6: Active demethylation of 5-methylcytosine involves the oxidation and repair of 5mC. 5hmC: 5-Hydroxylmethylcytosine; 5fC: 5-Formylcytosine; 5caC: 5-Carboxylcytosine; TET: Ten-eleven translocation enzyme; TDG: Thymine DNA glycosylase; BER: Base excision repair.

Crystal structure show a rotational dimorphism in the major groove of the 5-hydroxymethyl group of 5hmC and exhibits with two distinct conformations.[56,57] In contrast, the formyl group of 5fC is coplanar with the cytosine ring and forming an intramolecular hydrogen bond with the exocyclic N4-amine.[58,59] Similarly, the carboxylate group of 5caC also can form such an intramolecular hydrogen bond.[59,60] Cytosine 5-modifications do not seem to affect overall DNA conformation significantly, but effects on local conformation, flexibility and base pair stability have been reported.[57,59,61–63] This could facilitate local DNA unwinding and the base-flipping of 5hmC, 5fC and 5caC by TET and TDG enzymes for example.[40]

The contents of ox5mCs in genomic DNA (gDNA) are highly dynamic and cell/tissue type dependent but in general levels of 5hmC are 2-3 and 3-4 orders of magnitude higher than levels of 5fC and 5caC, respectively.[49,64] A reason can be the affinity and

activity of TET and TDG. After 5hmC is established in gDNA, it is not readily oxidized to 5fC and 5caC.[58]

Because of this stability, 5hmC is considered the sixth base in the genome. Elevated levels of 5hmC can be found especially in neurons and stem cells, whereas levels are depleted in cancer cells.[65–68] 5hmC is often accumulated at TSS and poised or active regulatory regions in the genome marked by histone modifications associated with transcriptional repression, activation or enhancement.[33,45] Moreover, factors like chromatin configuration or protein binding can affect the distribution of 5hmC in the genome during cell differentiation and specification.[45]

At 5hmC-sites, the polar hydroxyl group can influence interactions with TFs or MBD proteins, which indicates a role of 5hmC as *cis*-element in the regulation of gene expression or facilitator of alternative splicing.[35,43,45,54] Furthermore, the presence of 5hmC has been associated with neurodevelopment and oxidative stress response.[69–71]

In contrast, 5fC and 5caC can be rapidly removed due to highly efficient excision by TDG.[52,72] To a certain extent however, 5fC can be a stable epigenetic mark in mammalian genomes.[73–75] 5fC is enriched in important genetic regions like poised enhancers.[76–80] Moreover, 5fC has been shown to form DNA-protein crosslinks (DPCs) *in vitro* and *in vivo* via Schiff-base formation under physiological conditions with histones.[81–84] This could be a mechanism for nucleosome orientation, increased chromatin structure stability and regulation of transcription.[85]

Owing to the rarity of 5caC relatively little is understood about its distribution and function. Absence or depletion of TDG leads to an accumulation of 5fC and 5caC which points towards 5caC as site of active demethylation.[72,77,86] 5caC accumulates at promoters that are demethylated during differentiation of neuronal stem cells and can also be detected in specific tissue and cell types including cancer.[87–89] Furthermore, specific interaction of 5caC with the RNA polymerase II elongation complex, TFs like myc-associated factor X (MAX) or CCCTC-binding factor (CTCF) indicate a role in transcriptional regulation.[90–92]

Taken together, the 5-substituents of 5mC, 5hmC, 5fC and 5caC add an additional layer of information to genomic functionality and provide chemical handles with unique reactivity and specific hydrogen-bond acceptors and donors. Amongst specific readers of 5hmC, 5fC and 5caC are repair proteins, transcriptional and chromatin regulators.[93,94] Furthermore, the occurrence of 5mC and ox5mCs can provide

mechanistic information about progression or treatment of diseases and could be used for clinical disease diagnosis and therapy.[95]

However, even though many advancements for an improved understanding of 5mC and ox5mCs as epigenetic marks have been made, many questions remain. Therefore, sensitive detection methods for the analysis of 5mC and ox5mCs are required.

1.2. Detection and Analysis of Epigenetic Cytosine 5-Modifications

1.2.1. Overview

Because cytosine 5-modifications reside in the major groove of the dsDNA and do not alter the Watson-Crick base pairing information in a useful manner, conventional sequencing techniques cannot be used for the detection of epigenetic cytosine 5-modifications. Most methods aim at exploiting the characteristic chemical properties of the 5-modification and can be grouped into one of three categories (**Fig. 7**). There are analytical platforms, chemical-based and protein-based approaches. To improve sensitivity and facilitate analysis of epigenetic cytosine 5-modifications combinational strategies can be employed.

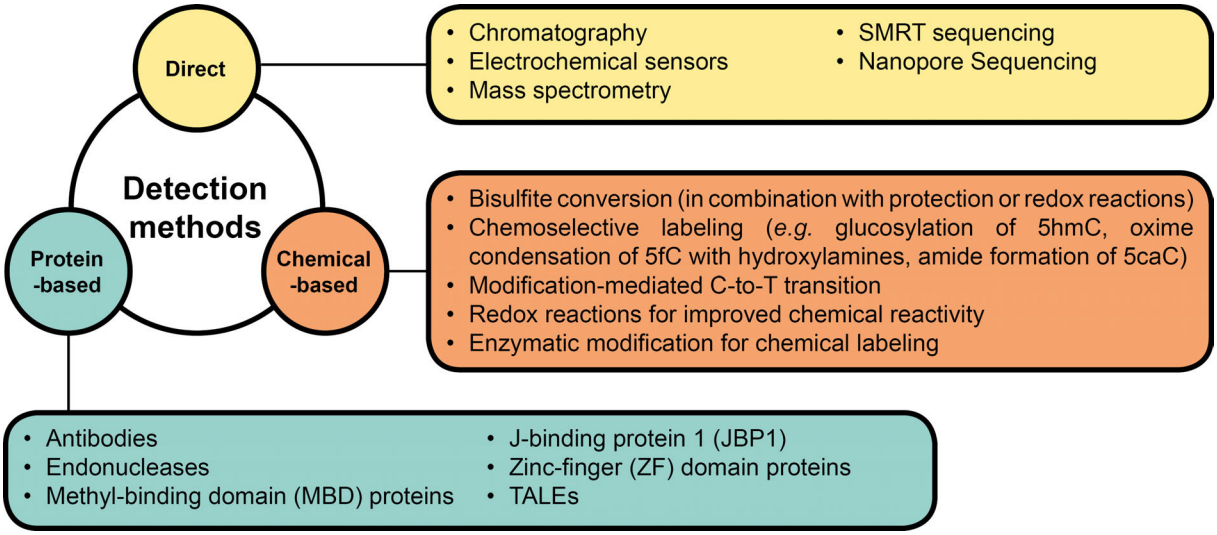


Figure 7: Methods for the analysis of cytosine 5-modifications.

1.2.2. Analytical Platforms and Direct Detection Methods

A facile method for the overall detection and quantification of methylated cytosines and oxidized derivatives is the analysis of single nucleotides with two-dimensional (2D) thin layer chromatography (TLC)[48–50,65,96,97] and capillary electrophoresis.[98–100] However, the accuracy of both methods can be limited due to variabilities in matrix or buffer compositions.[98]

Highly sensitive detection methods of C and 5mC based on electrochemical oxidation peak potentials or of C, 5mC and 5hmC with surface-enhanced Raman spectroscopy (SERS) have been developed, however the application scope is still quite limited.[101–105]

Standard highly accurate methods include gas chromatography (GC)[106] and reverse-phase (RP) high-performance liquid chromatography (HPLC),[107,108] which have been combined with mass spectrometry (MS).[98,109] Several LC-MS strategies use chemical derivatization to increase the information yield of the analyte with polar,[110] bulky[76] or easily chargeable moieties.[111–115]

MS-based strategies are considered the “gold standard” for quantitative analysis of 5mC but usually require highly trained personnel or expensive equipment.[98,109,116] Moreover, MS-based strategies usually requires the degradation of DNA which can lead to sample loss.

Emerging third-generation sequencing technologies for the detection of epigenetic 5-modified cytosines, like nanopore sequencing[117–121] or single molecule real-time sequencing (SMRT-Seq)[122,123] enable analysis with long reads and base resolution in single molecules without any DNA amplification or treatment.[124–126]

A drawback of the third-generation sequencing technologies is the relatively high error rate.[127,128] To improve the detection of epigenetic 5-modified cytosines, chemical derivatization strategies have been implemented.[129–133] However, accurate typing and profiling of epigenetic 5-modified cytosines is still challenging especially in repetitive or CG-rich regions.[125,128]

1.2.3. Chemical-based Detection Methods

1.2.3.1. Bisulfite-based Methods

Bisulfite treatment of dsDNA at stable temperature and pH leads to deamination of cytosine to uracil (**Fig. 8**).[134–136]

However, under standard reaction conditions up to 96 % of the input DNA are subject to degradation due to decomposition of pyrimidines and formation of abasic sites.[137]

Methylation of the cytosine 5-position prevents bisulfite conversion (**Fig. 8**).[138] This enables a binary readout for the detection of cytosine 5-DNA methylation with analytical platforms such as HPLC,[139] MS[140] or electrochemical arrays[141,142].

Bisulfite conversion can influence enzyme performance or PCR product formation. This enables 5mC analysis in assays such as the combined bisulfite restriction analysis (COBRA)[143] or methylation specific PCR (MSP).[144–146]

Additionally, analysis of bisulfite conversion-induced single nucleotide mismatches have been used in highly sensitive methylation detection assays with Förster resonance energy transfer (FRET)[147] and SERS.[148,149]

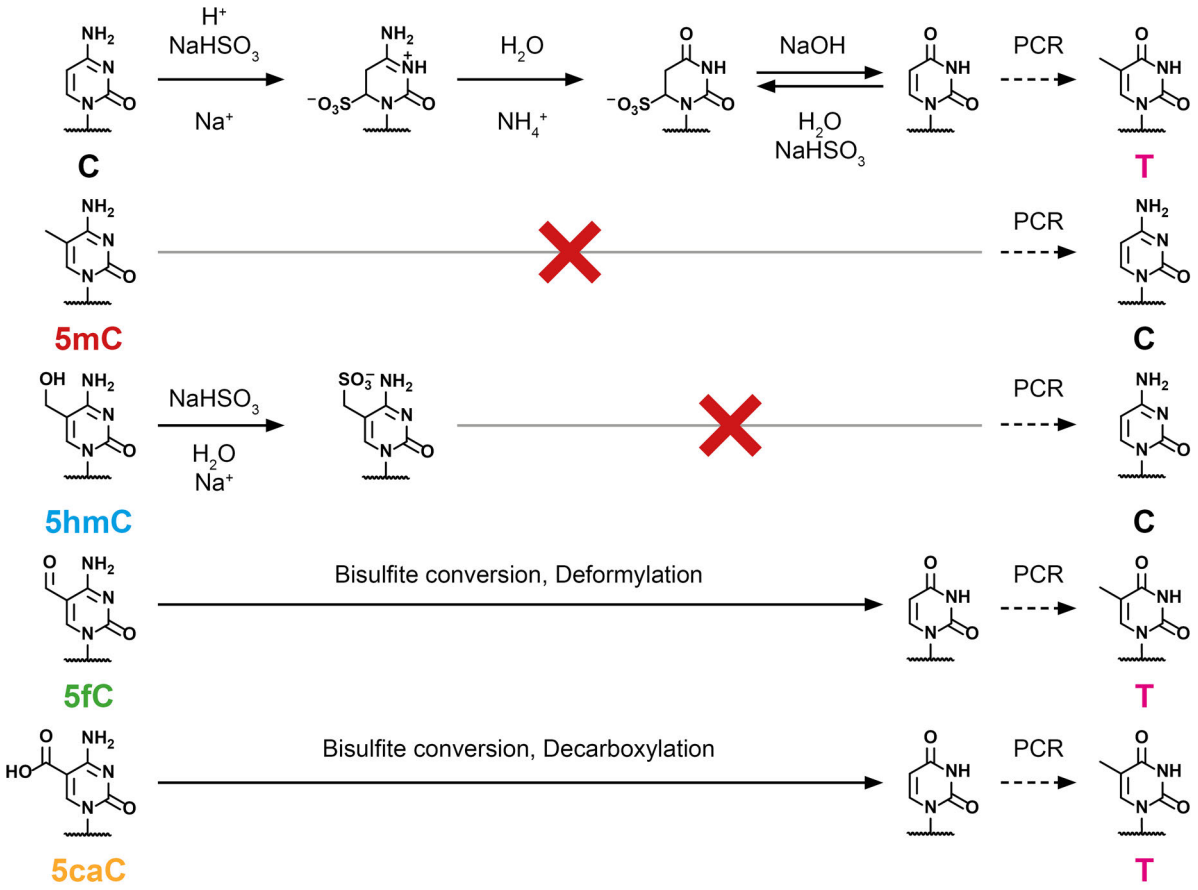


Figure 8: Bisulfite conversion of C, 5mC and ox5mCs and transition of converted nucleobases during PCR.

The gold standard for assaying cytosine 5-methylation is bisulfite sequencing (BS-Seq).[150,151] BS-Seq offers analysis with sequence coverage at single base-resolution. However, 5hmC reacts with bisulfite to cytosine-5-methylsulfonate (CMS) and cannot undergo deamination (**Fig. 8**).[152] In contrast, 5fC and 5caC are deformylated and decarboxylated, respectively, and then deaminated.[50,153]

Therefore, BS-Seq cannot discriminate between 5mC and 5hmC as well as C, 5fC and 5caC (**Fig. 8; Fig. 9**).[154]

In oxidative BS-Seq (oxBS-Seq), reaction with potassium perruthenate ($KRuO_4$) can be used for oxidation of 5hmC to 5fC.[153] This enables discrimination between 5mC and 5hmC (**Fig. 9**).

in TET-assisted-BS-Seq (TAB-Seq) the enzymatic glucosylation[155,156] protects 5hmC from TET-oxidation while 5mC and 5fC are oxidized to 5caC (**Fig. 9**).[157] This enables discrimination between 5mC and 5hmC as well(**Fig. 9**).

Reduced bisulfite sequencing (redBS-Seq) enables discrimination of 5fC in the context of the other cytosine nucleobases (**Fig. 9**).[78] In this strategy, 5fC is reduced to 5hmC with sodium borohydride ($NaBH_4$) and detected in the same manner as 5mC while the other bases are not affected.

Alternatively, a 5fC-selective tagging strategy termed chemical modification-assisted bisulfite sequencing (fCAB-Seq) protects the base from the bisulfite-mediated deformylation and deamination with aminooxy-functionalized reactants (**Fig. 9**).[77] Analog, chemical modification-assisted bisulfite sequencing (CAB-Seq) protects 5caC decarboxylation and deamination with a primary amine in 1-ethyl-3-[3-dimethylaminopropyl]carbodiimide (EDC)-catalyzed amide bond formation (**Fig. 9**).[158]

In M.SssI methylase-assisted bisulfite sequencing (MAB-Seq)[159] unmodified cytosine in CpGs are enzymatically methylated and consequently detected in the same manner as 5mC (**Fig. 9**). 5fC can be discriminated from 5caC with $NaBH_4$ reduction prior to M.SssI treatment (caMAB-Seq).[159] However, MAB-Seq and caMAB-Seq cannot be applied to cytosines outside of a CpG context as they are not methylated by M.SssI (**Fig. 9**).[159]

Bisulfite-based approaches allow detection and quantification of epigenetic 5-modified cytosine nucleobases at single nucleobase resolution in (single cell) reduced-representation bisulfite sequencing (RRBS[160], scRRBS[161]) or across whole genomes.[150]

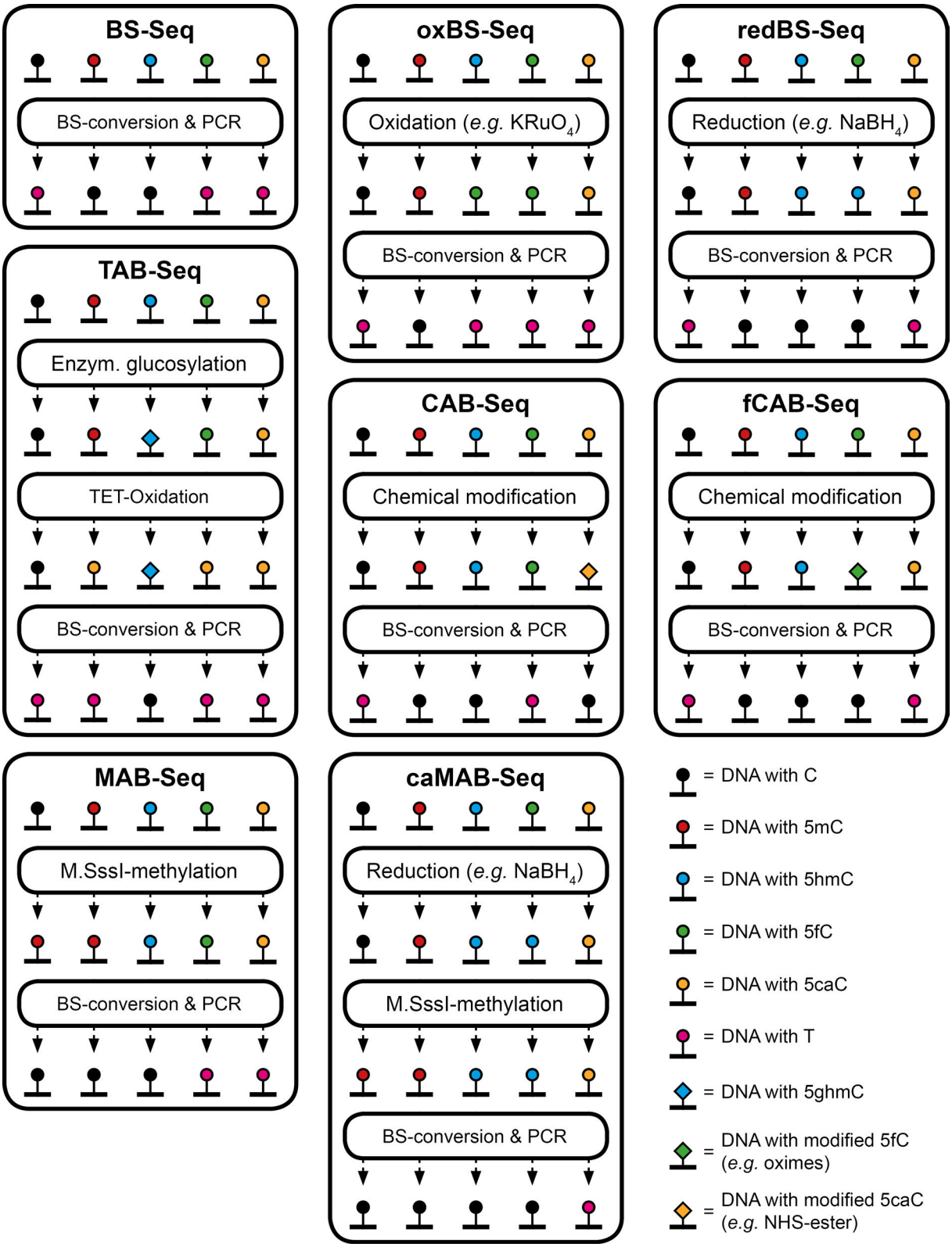


Figure 9: Analysis of 5hmC, 5fC and 5caC with bisulfite-sequencing (BS-Seq) strategies. ox: oxidative; red: reductive; TAB: TET-assisted BS; Enzym.: enzymatic; CAB: chemical-modification-assisted BS; fCAB: 5fC-specific-CAB; MAB: M.SssI-assisted BS; caMAB: 5caC-specific-MAB.

However, BS-Seq also requires a relatively high amount of template.[136] Additionally, error-proneness of the reaction due to incomplete conversions or bias in conversion steps and PCRs mean that outcomes can be unreliable.[162]

1.2.3.2. Bisulfite-free Chemical Modification-based Detection Methods

Other methods aim to target the unique chemical handle of the epigenetic cytosine 5-modifications, preferentially under non-denaturing reaction conditions (**Fig. 10**). An advantage over bisulfite sequencing is that only target nucleobases are derivatized.

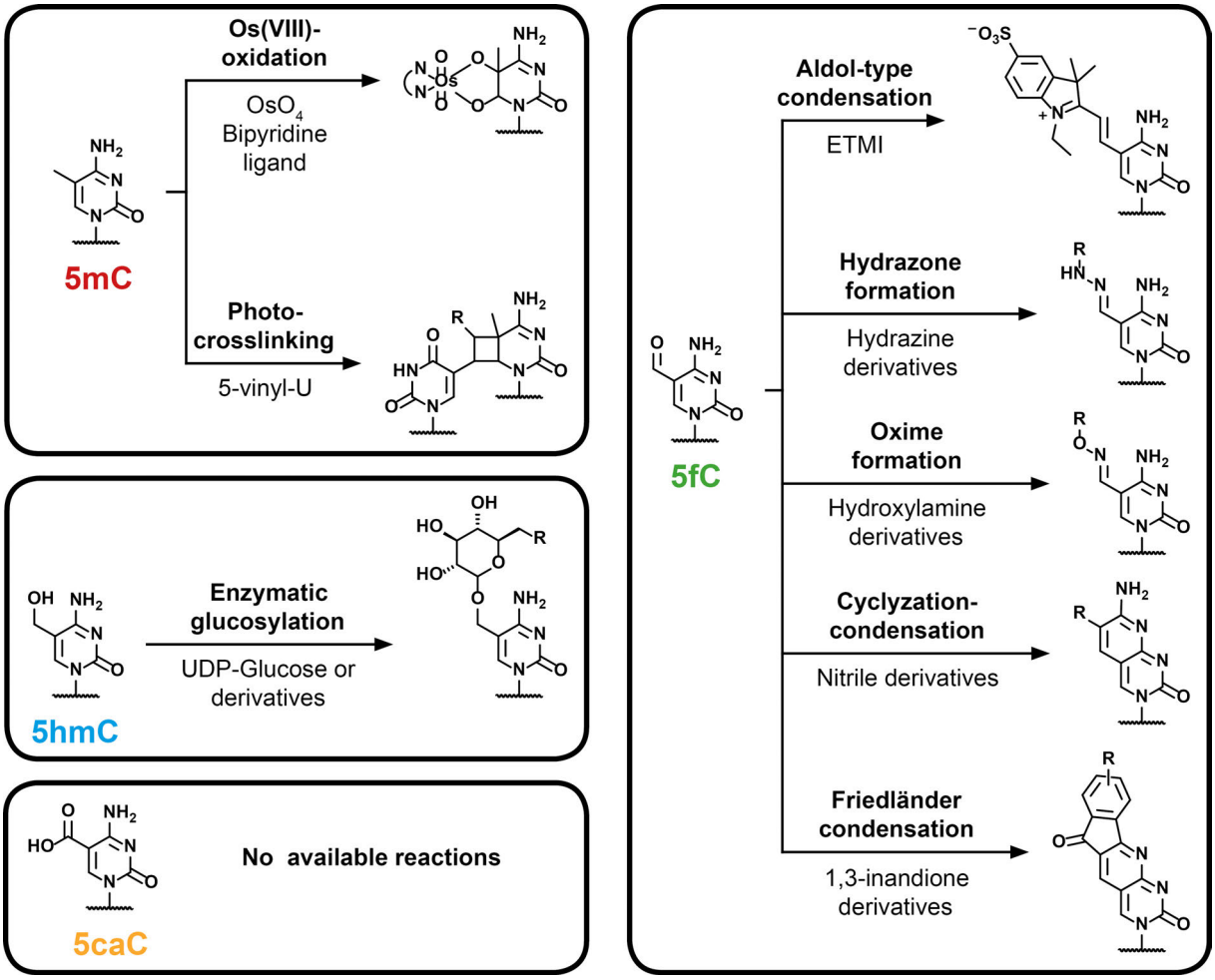


Figure 10: Exemplary Bisulfite-free strategies for the selective modification of cytosine 5-modifications. The derivatization of the nucleobases can lead to distinct chemical properties (*e.g.* susceptibility to chemical cleavage, fluorescence, C-to-T transition in PCR) that enable identification of the converted nucleobase. Alternatively, chemical handles (*e.g.* azides) for further tagging reactions can be provided with the reagents through functional residues (R) for fluorescence or pull-down probes, for example. UDP: Uridine diphosphate; ETMI: 1-ethyl-2,3,3-trimethyl-3H-indole-5-sulfonate.

1.2.3.2.1. 5-Methylctosine

C and 5mC can be distinguished via derivatization of cytosine nucleobases with O-allylhydroxylamine as the adduct configurations can affect base pairing.[163]

5mC can be selectively oxidized with osmium tetroxide (OsO_4) which enables interstrand crosslinking or complex formation with bipyridine ligands.[164,165] The adduct formation can be susceptible to hot piperidine cleavage or can be used to introduce reporter probes for the detection of 5mC-sites. Alternatives to OsO_4 are vanadium pentoxide (V_2O_5) or sodium periodate ($NaIO_4$) in combination with lithium bromide (LiBr).[166,167]

Alternatively, oligonucleotide-based strategies for the detection of 5mC including interstrand photocrosslinking have been reported.[151,168,169] [151,170–175]

A common issue of these assays is that they cannot directly distinguish between 5mC and the ox5mCs, or that it has not been conclusively tested.

Information about the symmetry of cytosine 5-methylation at CpGs can be provided by methyltransferase M.SssI-mediated labeling of unmodified cytosines with S-adenosylmethionine analogues bearing an amine or azide groups.[176,177] A chemical handle for Huisgen cycloaddition with alkyne-bearing probes allows detection of non-methylated CpGs in gDNA at single base resolution.[178]

A novel TET-assisted pyridine borane sequencing (TAPS) can be used for base-level resolution detection of 5mC and other cytosine derivatives.[179] TAPS induces C-to-T transition in PCR through TET-mediated oxidation of 5mC and 5hmC to 5caC and subsequent pyridine borane reduction of 5fC and 5caC to dihydroxyuracil (DHU).[179] Enzymatic glucosylation of 5hmC prevents TET-oxidation and reduction with pyridine borane. TAPSβ enables differentiation between 5mC and 5hmC.[179]

1.2.3.2.2. 5-Hydroxymethylcytosine

5hmC can be oxidized to 5fC with reagents such as $KRuO_4$ or Cu(II)/2,2,6,6-Tetramethylpiperidine-1-oxyl (TEMPO). 5fC can then be labeled with amine, hydroxylamine or hydrazine derivatives containing biotin,[180] fluorescence,[181,182] ruthenium(II) complex-tagged electrogenerated chemiluminescence (ECL)[183] and Förster resonance energy transfer (FRET) probes.[184] Alternatively, oxidation of 5hmC to 5fC with $KRuO_4$ or potassium ruthenate (K_2RuO_4) and labeling of 5fC with an azido derivative of 1,3-indandione or benzothiazole enable single-base resolution analysis. Both reagents can participate in a Friedländer intramolecular bicyclic addition reaction, which leads to C-to-T transition during PCR amplification due to the loss of the exocyclic 4-amino group after adduct formation.[185,186]

Selective oxidation of 5hmC to 5fC with $KRuO_4$ has also been used in chemical-assisted pyridine borane sequencing (CAPS).[179] In this method, the oxidation of 5hmC to 5fC makes it susceptible for reduction with pyridine borane to DHU which enables base-resolution analysis of 5hmC.

The enzymatic glucosylation with T4 β-glucosyltransferase (T4 BGT) is a direct, 5hmC-selective tagging strategy,[95] which has served as an early method to quantify genomic 5hmC via scintillation measurements of radiolabeled glucose.[187]

Enzymatic glucosylation of 5hmC can protect the base from reduction with pyridine borane (*cf*. chapter 1.4.2.2.1). This enables detection and quantification of 5hmC with single base resolution.[179] Similarly, glucosylation of 5hmC prevents enzymatic deamination of cytosine nucleobases which induces C-to-T transition for sites with C or 5mC in APOBEC-coupled epigenetic sequencing (ACE-seq).[188]

Boronic acids can make pH-sensitive reversible esters in alkaline solutions with 1,2- or 1,3-diol groups of carbohydrates.[189] Several conjugation strategies of β-glucosyl-5hmC (5ghmC) with boric acid-functionalized probes or signal generation via boronic acid chemistries have been used to detect 5hmC-sites.[190–195]

Oxidation of vicinal hydroxyl groups of the glucosyl moiety to aldehydes after glucosylation of 5hmC with $NaIO_4$ provide chemical handles for oxime formation with aldehyde reactive probes (ARPs).[196] This enables tagging with fluorescent molecules to detect 5hmC in the DNA.[197] Alternatively biotin-labeled 5ghmC can be enriched for downstream analysis.[196,198] However, the reaction cannot differentiate between 5hmC- and endogenous 5fC-sites.

Other multi-step labeling approaches use selective bio-orthogonal labeling of 5hmC with an 6-azido-glucose derivative for Huisgen cycloaddition ("click chemistry").[199] This enables conjugation with pull-down probes and selective enrichment for downstream applications such as genome-wide 5hmC-profiling.[130,199–203] Moreover, click chemistry enables introduction of reporter molecules for global coverage of 5hmC-sites, or to detect complementary 5mCpG/5hmCpG-sites with FRET signals.[204–206]

The bacterial DNA methyltransferase M.HhaI can catalyze nucleophilic condensation at 5hmC with thiol compounds such as cysteamine.[207] The amino-derivatized DNA can be subsequently labeled with amine-reactive biotin for example for the recruitment of reporter enzymes.[208,209]

Additionally, 5hmC can be labeled in a sulfinate based reaction that is similar to the formation of CMS during BS-Seq but with a cleavable, biotinylated probe for pull-down and downstream analysis of enriched 5hmC-DNA.[210,211]

5hmC can be oxidized with peroxotungstate and deaminated to yield a trihydroxylated thymine derivative in an one-pot reaction.[212] This causes C-to-T transition that enable detection of 5hmC with single-base resolution.[213]

1.2.3.2.3. 5-Formylcytosine

The reduction of 5fC to 5hmC with $NaBH_4$ enables enzymatic glucosylation with an azido-glucose moiety and click chemistry analog to 5hmC (*c.f.* 1.2.3.2.2.).[77]

The 5-formyl group of 5fC can be readily functionalized with active nucleophiles like hydroxylamines and hydrazines to give rise to oxime or hydrazone conjugates.[95] Based on these condensation chemistries,[74,75] a multitude of direct labelling, enrichment and crosslinking approaches have been reported. Strategies for the detection of 5fC-sites employ specific pull-down,[76,214] fluorogenic,[215–217] or oligonucleotide probes.[75] Additionally, labels for supramolecular reversible guest-host complexes[218,219] or so-called "roadblocks"[217] for the inhibition of protein activity have been reported.

Using hydroxylamines or hydrazines has the advantage of convenient, well-established and comparatively mild acidic reaction conditions for quantitative, rapid labeling.[215,220] A drawback of ARP with hydroxylamine and hydrazine groups can be their cross-reactivity towards abasic sites or 5-formyluridine (5fU).[214,220,221] Depending on the reaction conditions, labeling of 5fC but not abasic sites have been shown, however the reactivity towards 5fU often remains elusive.[215,218]

5fU-sites often react more readily with reagents than 5fC.[222] To rule out false-positive 5fC-sites, a possible strategy is selectively reacting 5fU-sites first with dedicated moieties that will not interfere with the probe for analysis or enrichment. The probe can then be reacted selectively with 5fC-sites.[214]

An aldol-type condensation of 2,3,3-trimethylindole (TMI) derivatives can be used for differentiation between 5fC and 5fU, as the extended conjugated π-systems of the adducts have distinct photophysical characteristics.[223] Moreover, labeling of 5fC-sites with TMI was found to inhibit or delay DNA polymerase elongation *in vitro*.[223]

Schiff base formation of the carbonyl of 5fC with aliphatic amine is another 5fC-labeling reaction.[224] Reaction of 5fC and partially of abasic sites can occur under non-denaturing conditions without the need for a catalyst.[215,224] However, the reaction is reversible due to hydrolysis.[225] In contrast, oxime and hydrazone formation can be also reversible but products are generally much more stable than imines due to the negative inductive effect of the additional heteroatom.[226]

The Friedländer condensation with an azido derivative of 1,3-indandione (AI) is a highly selective 5fC-selective labeling reaction.[227] The reaction yields a bicyclic adduct that can be used to detect and quantify 5fC in genomic DNA with single base resolution.[227] The azide functional group can also be used for pull-down and enrichment with azido-

reactive disulfide biotin probes.[227] A similar strategy with malonitrile (CLEVER-Seq) has been used for 5fC profiling at single base resolution in single cells.[80]

Other cyclization reactions of 5fC with 2-(5-chlorobenzo[*d*]thiazol-2-yl)acetonitrile (CBAN) and 3-azido-N-(2-(cyanomethyl)benzo[*d*]thiazol-6-yl)propanamide (azi-BP) also enable base resolution analysis.[228,229] Additionally, the cyclization reaction of CBAN with 5fC also leads to an extensive fluorescence enhancement of the adduct which can be used to quantify global 5fC-levels.[228] Azi-BP can participate in copper-free click-chemistry with dibenzocycloocytne (DBCO) reagents which act as a "roadblock" for DNA polymerases.[229]

1.2.3.2.4. 5-Carboxylcytosine

To the best knowledge, no bisulfite-free chemical detection method has been reported yet for 5caC.

Liu et al. suggested that EDC-mediated coupling could be used to protect 5caC from borane reduction.[179] This would enable differentiation between 5fC and 5caC similar to 5mC and 5hmC in CAPS.[179]

However, conjugation of carboxylic acids to reactive probes can be difficult in aqueous solutions and under non-denaturing conditions.[230,231] Moreover, chemical-based detection strategies for 5caC require very high sensitivity given the low 5caC-levels in DNA.[49,232]

1.2.4. Protein-based Approaches

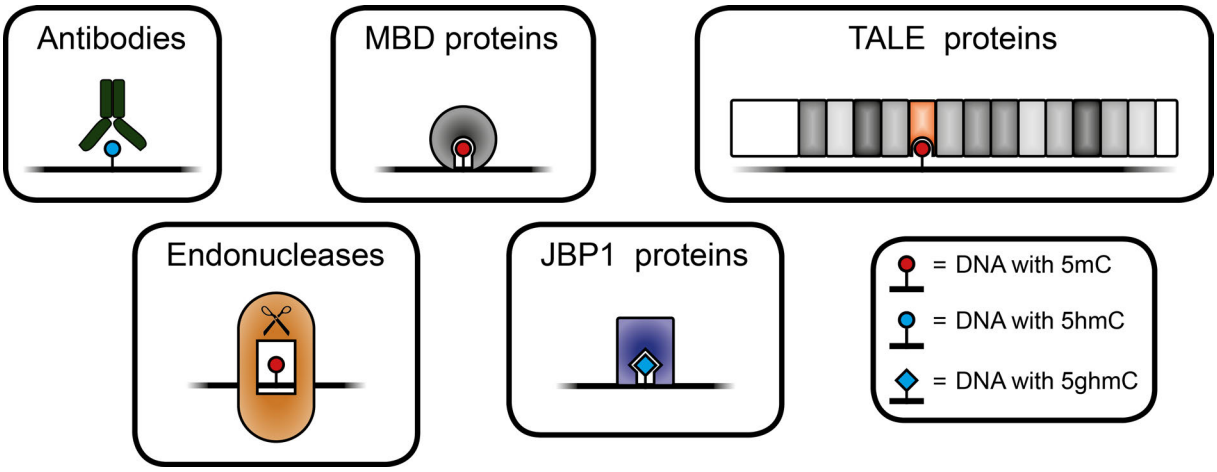


Figure 11: Protein-based strategies for the analysis of cytosine modifications.

Detection of cytosine modifications through direct interaction of proteins with DNA (**Fig. 11**) is the main alternative to the chemical derivatization strategy.[151] A main advantage of this approach is the simple, direct analysis of samples both *in vitro* and *in vivo* assays.

A commonly used protein-based approaches for the genome-wide distribution of 5mC and ox5mCs in DNA are immunoassays. Immunostaining,[48,66,233,234] immunosensors,[102,151] enzyme-linked immunosorbent assay (ELISA)[116] or DNA immunoprecipitation (DIP)[72,235,236] are simple and cost-effective. Antibodies have been raised for the highly selective detection of specific epigenetic 5-modified cytosines or CMS.[72,196,237] However, sequence or modification bias of antibodies as well as cross-reactivity of IgG antibodies for short tandem repeat (STR) sequences has been reported.[238–240] Moreover, antibodies lack sequence specificity and detection of cytosine 5-modifications can be limited by the density of CpGs.[154,241]

Selective cytosine 5-modification-specific enzymatic digestion of DNA enables genome-wide mapping of cytosine 5-modifications. Sensitive restriction enzymes for 5mC or 5hmC have been identified and used for a multitude of high throughput assays and electrochemical DNA methylation detection platforms.[102,124,151,242] Similarly, 5ghmC can also regulate enzymatic digestion.[243–245] Sensitive endonucleases for 5fC or 5caC have not yet been reported to the best knowledge. Instead, 5fC can be reduced to 5hmC with $NaBH_4$.[246]

A major drawback of restriction endonucleases is the non-programmable specificity of recognition sites. Also resolution is limited to restriction sites.[150,154]

MBD proteins have been used for isolation, live-cell imaging and single molecule analysis of cytosine 5-methylated DNA.[247–252] Similarly, β-d-glucosyl-hydroxymethyluracil (J)-binding protein 1 (JBP1) can be used to isolate 5hmC-DNA after enzymatic glucosylation.[253] Both proteins enable quantification and genome wide profiling in combination with high throughput sequencing of 5mC or 5hmC.[254–258]

No approaches for the profiling of 5fC or 5caC with non-sequence-specific proteins are available to the best knowledge. A potential candidate for sequence-non-specific 5caC-profiling could be a N157D (N = asparagine; D = aspartic acid) mutated variant of thymine DNA glycosylase, which has been shown to selectively processes 5caC in gDNA.[259]

In general, a drawback of these protein-based approaches is no or limited target DNA selectivity.

Zinc finger (ZF) proteins are sequence-specific, modular DNA-binding proteins consisting of concatenated units with highly conserved ββα configurations.[260] Several amino acids localized within the α-helix motif of each unit interact with sequences of typically three base pairs (bps) in the major groove of DNA.[260] Assembly of ZF domains into modular arrays enable “programming” of custom DNA-binding scaffolds with user-defined sequence-specificity.[261]

Additionally, several ZF domains have been shown to recognize epigenetic cytosine 5-modifications via R (arginine) and E (glutamate) residues.[262] In theory, this would enable programmable design of DNA-binding domains with sensitivity towards 5mC or ox5mCs, but so far no recombinant ZF proteins for the direct detection of 5mC or ox5mCs have been reported to the best knowledge.

Instead, a combination of MBDs and ZF domains has been used.[263,264] In this method, a functional fluorescent domain can only assemble if MBD binds to 5mC, and the ZFs binds in vicinity to the 5mC at a user-defined target DNA sequence.[263,264] Alternatively, synthetic peptides resembling three ZF domains bearing a phosphotyrosine have been used as a proof-of-concept for direct sequence-specific detection of 5mC in the context of C and 5hmC in oligonucleotide sequences.[265,266]

However, ZF-based methods are limited in sequence-selectivity due to the confined recognition mode of ZFs which reportedly suffer from significant off-target effects.[267,268] Alternatively, programmable DNA-binding scaffolds with a single-nucleobase recognition mode in the form of transcription activator-like effectors (TALEs) have recently emerged as sequence-specific sensors for 5mC and ox5mCs.[151,267]

1.3. Transcription Activator-Like Effectors

1.3.1. Structure and binding mode of TALE proteins

TALEs were first discovered in Gram-negative *Xanthomonas* bacteria and essential factors to bacterial virulence, proliferation, and dissemination.[269] Natural TALE proteins often consist of three parts (**Fig. 12a**).[269]

The N-terminal region (NTR) contains the type III secretion signal.[270] The C-terminal region (CTR) contains a nuclear localization sequence (NLS) and an acidic activation domain (AD).[271–273] The central part constitutes the main DNA-binding domain (DBD).[274] The DBD is a repetitive sequence and referred to as central repeat domain (CRD). Repeats of the CRD predominantly have 34 aa, except for the last repeat (LR) which is truncated at 20 amino acids.[269] TALEs with 2 to 34 repeats exists but the most frequent number of repeats in naturally occurring TALEs is 18.[269] While at least 7 repeats are necessary to induce target gene expression,[275] the maximum specificity of the CRD for a DNA sequence ranges from 15 to 19 repeats.[276]

One repeat of the CRD interacts with one nucleobase of the target DNA strand.[277] The specificity of each repeat is determined by a repeat variable di-residue (RVD) at aa position 12 and 13 of each repeat (**Fig. 12b, c**).[275,278] The remaining amino acids in a repeat sequence are typically highly conserved (**Fig. 12b**). The RVD is located in a loop region which faces the target nucleobase via the major groove of the DNA duplex (**Fig. 12c, d**).[279,280] The loop is flanked by two α-helices. Two repeats are linked via a loop region that is located on the outside of the TALE repeat structure. The core of single repeats consists of small aliphatic residues (**Fig. 12c**).[280] The small aliphatic residues make extensive van der Waals (vdW) contacts and enable close stacking of the two helices within each repeat against each other.[281] In contrast, the interface between repeats contains distinct polar residues that interact with each other.[280]

The sequential packing and association of tandem repeats with the DNA double helix appears to be facilitated by P27 (P = proline), which interrupts the second helix and is followed by a left-handed sequence leading to the right handed α-helix of the next repeat.[280] Together, the concatenated repeats make a right-handed superhelical structure that wraps around the DNA double helix (**Fig. 12d**).[279,280]

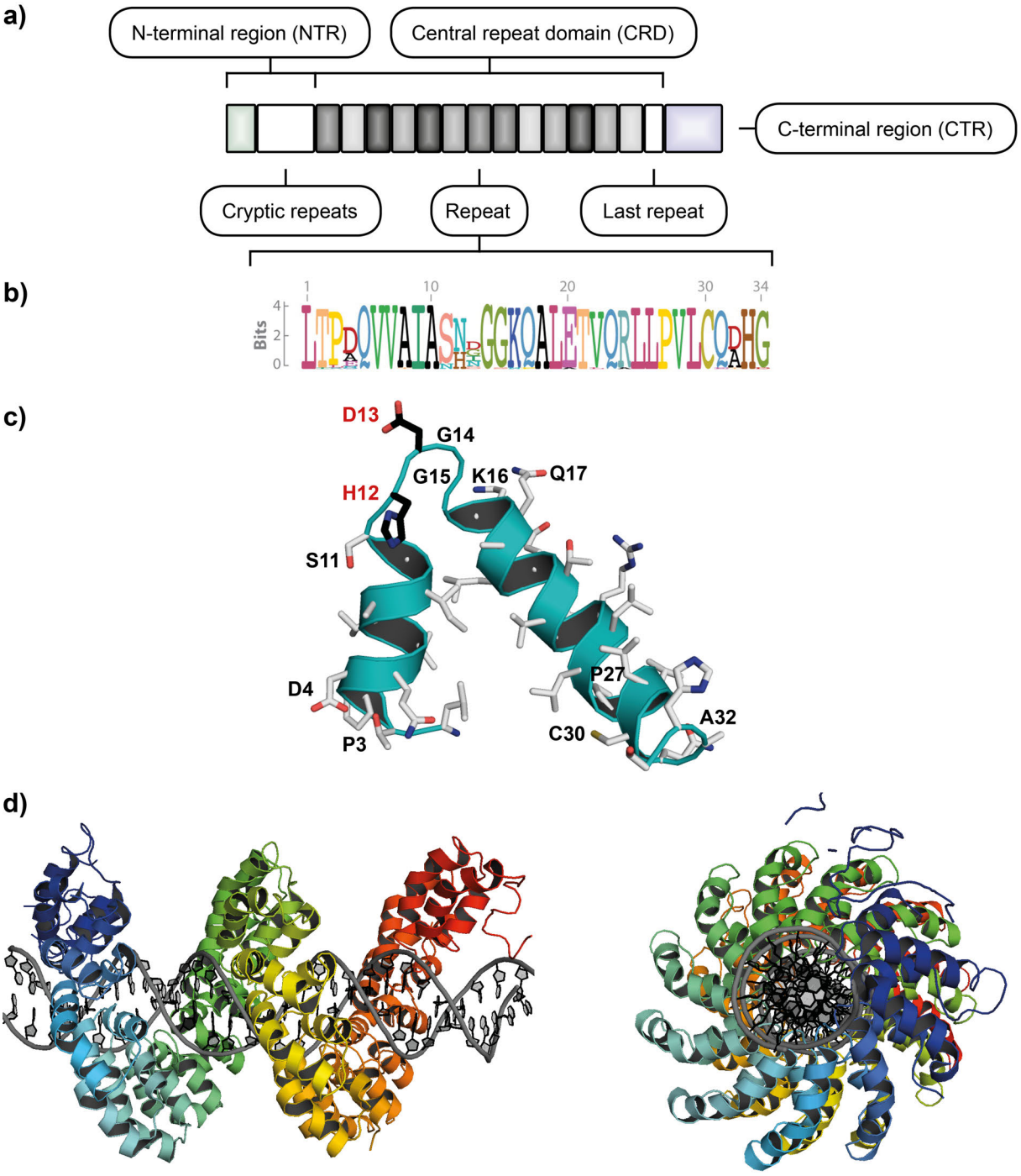


Figure 12: Structure and modular organization of natural TALE proteins.[c]

a) Schematic representation of the modular organization of TALE proteins with main regions and domains. **b)** Comparison of natural TALE repeat sequences show the variability at aa position 12 to 13 but an overall conserved sequence. **c)** A TALE repeat structure from PthXo1 (protein database (pdb): 3UGM) with the RVD highlighted. Important amino acids involved in structure formation and DNA binding have been labeled. **d)** The PthXo1 TALE bound to the target DNA sequence (pdb: 3UGM).

[c] Figure in b) reprinted (adapted) with permission from ref.[269]. Copyright 2019 Annual Reviews.

The process behind the formation of the TALE-DNA complex is still not completely understood. DNA-binding proteins can identify target sites through one-dimensional (1D) diffusion along the DNA helix via sliding and hopping motions.[282] A similar motion has been observed for TALEs, although with a rotationally decoupled two-state mode of target search and binding events in which the TALE follows a direct trajectory along the template.[283,284]

During target search, TALEs are loosely wrapped around the DNA duplex. This conformation minimizes electrostatic interactions between the positively charged patches along the inner surface of the CRD and negatively charged DNA backbone, and enables the TALE to rapidly hop and slide along the DNA.[279,285] In between, binding events enable the CRD to test local DNA sequences for compatibility with its array. Upon encountering target DNA sequences, TALE proteins undergo a conformational compaction.[281,286] This aligns RVDs and other residues including aa position 4, 30 and 32 of the repeat in the CRD array for energetically favorable TALE-DNA interactions.[283,287,288]

Essential for the transition from the search to the binding state is the recognition of an initial nucleobase by a cryptic repeat of the NTR. In most TALEs this occurs through a interaction between the indole ring of W232 (W = tryptophan) and the thymine 5-methyl group (**Fig. 13**).[280]

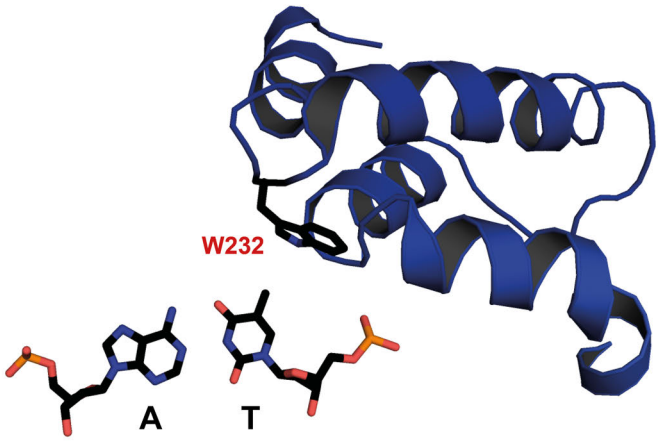


Figure 13: Structure of the cryptic repeats -1 and 0 of the PthXo1 TALE NTR and interaction of W232 with the initial T of the target DNA. (pdb: 3UGM)

Usually, the single T is preceding naturally occurring target DNA and is required for TALE activity.[275,278] Moreover, repeat number, order, and RVD at position 1 can also influence the sensitivity for the initial T.[289] Furthermore, the substitution of W232R/Q231S (Q = glutamine; S = serine) in the NTR enabled accommodation of an initial G.[290]

In total, the NTR contains four cryptic repeats (-3,-2,-1,0) which directly precede the CRD with repeat at position 0 packing tightly against the first canonical repeat of the CRD.[285] Similar to the array of the CRD, the repeats of the NTR consists of a loop flanked by two α-helices and are arranged in a right-handed superhelical structure.[285] However, crystal structure show that the intervening loops of the cryptic repeats are degenerated except for the repeat at position -1 which makes the contact with the initial T.[280,285]

All four cryptic repeats contain additional positively charged amino acids that contribute significantly to the overall basic charge of the protein and enable electrostatic interactions to DNA backbone phosphates.[285,291] Therefore, the NTR likely acts as nucleation site for TALE-DNA binding.[285,291] However, the first 152 aa of the natural NTR are not required for full activity of TALEs.[288,292]

The role of the CTR is uncertain. It has been reported that the removal of the CTR and even the removal of the truncated last repeat might affect structural compactness but does not affect TALE activity significantly.[288,293,294] Accordingly, it has been reported that the NTR and N-terminal repeats of the CRD contribute more to TALE–DNA recognition than the CTR and C-terminal repeat.[288,295–297]

1.3.2. The DNA recognition code of TALE repeats

Several amino acids in a repeat contribute to either stabilizing the repeat structure or to DNA binding. Amino acid residues K16 (K = lysine) and Q17 contribute to the electrostatic potential and coordinate the backbone phosphate of the forward strand DNA through direct or water-mediated hydrogen-bonds.[281] These two amino acids account for the majority of the total binding energy of TAL proteins to DNA.[298] Additionally, G14 (glycine) and G15 bind to DNA phosphates through water-mediated H-bonds.[280,281,299]

The four canonical RVDs NI (I = isoleucine), HD (H = histidine), NN and NG have different affinity to the four canonical nucleobases (**Fig. 14**).[300] The four canonical RVD-nucleobase pairs are HD and C, NG and T, NI and A, NN and G or A.[275,278]

In canonical TALE repeat, the 12th position is usually either H or N (**Fig. 12b**).[269] Crystal structure show, that the residue faces away from the DNA towards the α-helix of the TALE repeat preceding the loop and acts as H-bond donor toward backbone carbonyl oxygen of aa position 8 in each repeat.[280] This results in a helix break allowing the formation and stabilization of the RVD-containing loop from aa position 12 to 15.[280,281,298]

Base discrimination and molecular recognition of the nucleobase is solely provided by aa position 13 which can also interacts with the preceding DNA base.[280,281]

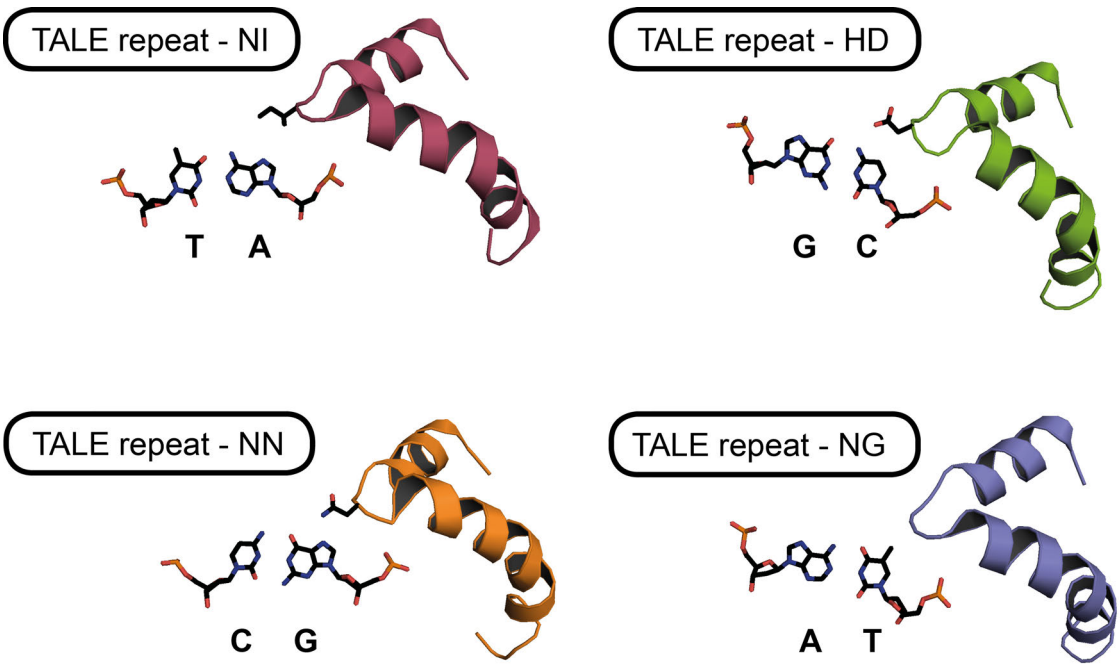


Figure 14: Repeat structures in complex with DNA nucleobases. (pdb: 3UGM)

Within RVD HD, the side chain of D13 makes vdW interaction with Cytosine whereas the carboxylate moiety on D13 forms a stabilizing interaction with the amino group on C4 (β) of C.[280,281] Models predicting a steric clash of D13 with the methyl group on carbon-5 (α) of T and interactions with N7 (γ) of G or A that destabilize the complex.[298] RVD NG makes a vdW interaction between the Cσ of the 5-methylgroup of thymine and G13.[280,281] G13 does not provide stabilizing interaction with another nucleobase and other residues at that position would likely introduce steric clash with the 5-methyl group of thymine.[281] The RVD NI is the second most common RVD in naturally occurring TALEs but is comparatively weak.[280,301] In a crystal structure the aliphatic side chain of Ile in the RVD NI faces nucleobases without favorable contact.[299] I13 likely clashes sterically with the C4 carbonyl oxygen and the 5-methyl group of T.[298] A model showed that the aliphatic side chain of I13 can interact with C8 (and N7) of the adenine purine ring or carbon-5 of the cytosine pyrimidine ring.[298] Between the two bases RVD NI seems to prefer adenine on the basis of charge distribution, dipole moment and steric discrimination due to the C4 (β) amino group of C.[280,298] The RVD NN interacts with both purine nucleobases. The side chain of N13 can form a H-bond with the pyridine-like nitrogen N7 (γ).[280] Interactions with pyrimidine nucleobases are not favorable due to steric clashes with groups on C4/C5.[298] There are natural occurring RVDs that are more selective for G such as RVD NH or NK, however both RVDs decrease TALE binding.[292,301,302]

In repeats with N* (* = deletion, del) the RVD is replaced by a single asparagine. These repeats exhibit a size-reduced loop in which the subsequent amino acid G14 does not take the place of the deleted 13th position.[269,280] As the loop extends less deeply into the major groove, the steric demand is reduced and both pyrimidine nucleobases can be accomodated.[278]

1.3.3. Design of TALE repeats and their application

The arrangement of repeats into a CRD array technically allows to target virtually any desired DNA sequence. There are different assembly methods such as Golden Gate cloning,[303] fast ligation-based automatable solid-phase high-throughput assembly,[304] ligation-independent cloning[305] or isothermal assembly[306] for "programming" TALEs. TALE fusion proteins can be used for genome engineering,[307] transcriptional regulation,[300,308–310] epigenetic writing[311] and erasing[312] as well as fluorescent probes.[264,313–315] These applications require high activity and specificity of TALE proteins which can be influenced by sequence context, neighboring effects or RVD selection.[289,295,297,316] For example, RVD NN binds both purine nucleobases but contributes to strong TALE-DNA interaction.[301] RVD NI is selective for A but displays comparatively weak binding.[301] Hence, using only the four "canonical" repeats HD, NG, NN and NI can limit TALE function. Three studies addressed this issue by testing all possible amino acid combinations of the RVD and revealed new selectivity profiles of novel RVDs such as RVD RH for G.[302,316,317]

5mC and ox5mCs can be found in regulatory regions that are often chosen as targets in TALE-based technologies.[318] The presence of cytosine 5-modifications can obstruct TALE-DNA binding, similar to steric clash between D13 and the 5-methyl group of T.[298,318–321] In contrast, repeats NG or N* can accommodate 5mC.[280,321,322] Furthermore, in the context of C, 5mC and 5hmC, N* is able to discriminate between 5mC and 5hmC, while RVD NG shows positive selectivity for 5mC.[323,324] However, RVD N* shows similar high affinity towards C, 5mC and 5fC and weak affinity to 5hmC and 5caC.[324,325]

Designing novel size-reduced repeats with a deletion at aa position 13 and variations at aa position 12 revealed diverse selectivity profiles for C, 5mC and ox5mCs, which are often similar to RVD N*.[324,325] Interestingly, repeats with G12 or aa residues bearing hydroxyl groups (G*, S*, T* (T = threonine)) exhibited strong binding to all five cytosine nucleobases.[324,325] A repeat with proline at aa position 12 (P*) was negatively selective for 5caC.[325]

Notably, the same study showed that double deletions of aa residue 13 and 14 led to significantly reduced binding in all cases.[325] Variations of aa residue 12 did not restore affinity.[325] Deletions at aa position 12 and 15 reduced binding even further.[326] However, a repeat lacking aa position 12 to 15 and with R at aa position 11 showed weak but selective recognition of 5caC.[326]

Another study provided a complete assessment of the efficiencies and specificities of amino acid combinations in addition to combinations of a deletion at either aa position 12 or 13 with all amino acids for C, 5mC and 5hmC.[327] RVD HA (A = alanine) or NA were identified as positive selective for 5mC in the context of C, 5mC and 5hmC.[327] However, no information regarding the other oxidized cytosine derivatives were provided, but previous studies had already shown a weak binding and selectivity of these RVDs towards other canonical nucleobases.[302,316,317,327]

Another strategy focused on mutation of aa position 11 to 14 of the TALE repeat aiming at identifying structures with improved methylation-discrimination ability in living cells.[328] According to this report, RVD NG is insufficient to completely regulate TALE binding in a methylation-dependent manner. Instead, the study reported that TALE repeat with amino acids ASAA at aa position 11 to 14 (instead of SNGG) shows improved methylation-dependent base recognition but weaker affinity.[328]

Enhanced 5mC-selectivity of TALEs due to deletion of selected TALE-DNA-phosphate interactions has been observed for substitution of aa K16 and Q17 to A.[329] This effect could be further enhanced through substituting key basic residues of the NTR likely responsible for backbone interactions in the DNA-bound state with A.[329] However, both studies showed that incorporation of two engineered repeats decreased DNA binding drastically.[328,329]

Despite extensive efforts, no cytosine 5-modification selective TALE repeat with high affinity has been found as of writing this thesis.

It has been shown, that reduction of 5fC to 5hmC enables discrimination of 5fC from 5mC with the partly selective TALE repeat N*.[324] This indicates that chemoselective modification of carbon-5 groups can offer a strategy for complete, programmable decoding of oxidized 5-methylcytosine nucleobases with non-selective TALE repeats.

Additionally, engineering strategies for TALE repeats had been limited to the canonical amino acids. With the advances regarding genetic code expansion and incorporation of artificial amino acid at user-defined positions, utilization of unnatural amino acids presents a promising strategy for selective, sensitive detection of ox5mCs as well.

1.4. Protein engineering with non-canonical amino acids

1.4.1. Site-selective non-canonical amino acid incorporation into proteins

The use of non-canonical amino acids (ncAA) can provide precise chemical properties and enable site-selective bio-orthogonal modification of proteins. The bottleneck for the applicability can be the incorporation at desired sites. Today, the most commonly used method site-selective non-canonical amino acid incorporation into proteins is genetically encoding ncAA via amber codon suppression due to its fidelity, efficiency, minimal influence on structure and natural environment during translation.[330]

Site-specific incorporation of ncAAs can be achieved by exploiting the functionality of nonsense suppressor tRNAs in prokaryotic translation systems.[331–334] Using an unique nonsense codon in the mRNA, the ncAA will be directed to the specific position encoding for the targeted amino acid. The amber stop codon UAG is rarely used as termination signal and has an overall low frequency in model organisms such as *Escherichia coli (E. coli*) or *Saccharomyces cerevisiae* (*S. cerevisiae*).[333,335] Moreover, growth rates of certain *E. coli* that are naturally able to suppress amber codons were not significantly affected during incorporation of canonical amino acids.[336,337]

To ensure high specificity of the incorporation, the aminoacylation of the nonsense suppressor tRNA needs to be specific for the ncAA but do not interfere or compete with any of the endogenous aaRS/tRNA pairs in the host organism. The orthogonality in amber codon suppression can be achieved by natural amber suppression aaRS/tRNA pairs from a domain of life different to that of the extract's origin as the recognition of an tRNA by its cognate aaRS is often species specific.[338–340] However, fully specific and orthogonal aaRS/tRNA pairs often require evolution through structure-based mutagenesis and multiple rounds of selection.[341–347]

To date, highly efficient amber suppressions with more than 100 structurally diverse non-canonical amino acids *in vivo* have been reported.[348]

1.4.2. Applications of non-canonical amino acids

Site-selective non-canonical amino acid incorporation into proteins can be used to quantitatively introduce post-translational modifications (PTM) without sequence restrictions to study structural and biochemical implications.[330,348,349]

ncAA also enable engineering of proteins with new or enhanced catalytic function or as highly sensitive probes.[330,349,350] These probes can be used to study structural properties, conformational changes, interactions or cellular processes of proteins with X-ray crystallography,[351,352] infrared (IR),[353] nuclear magnetic resonance (NMR),[354] electron paramagnetic resonance (EPR),[355] and fluorescence spectroscopy or microscopy *in vitro* or prokaryotic or eukaryotic cells.[356–358]

Other applications include the spatial and temporal control of protein function and reversible activation *in vitro* or *in vivo* with photoswitchable ncAA,[359,360] or photo-crosslinking between proteins or between proteins and nucleic acids.[361,362]

The use of ncAA in the generation of therapeutic proteins is also an emerging field.[363]

But the arguably largest area of applications for genetically encoded chemically reactive ncAAs is the selective conjugation of proteins to a variety of molecules, such as oligonucleotides, synthetic polymers, and other proteins.[364]

A commonly used reactions to modify chemically reactive ncAAs with non-proteinogenic moieties is oxime formation.[365] One advantage is the stable product formation under relatively mild acidic reaction conditions.[364] A disadvantage is the incompatibility with pH-sensitive proteins, live cells and potential cross-reactivity.[364]

Keto groups have been introduced into proteins by genetically encoding the amino acids *para*-acetylphenylalanine (*p*AcF) or *meta*-acetylphenylalanine (*m*AcF).[365–367] These amino acids and others have been used as a chemical handle to attach small molecules for a variety of applications such as protein conjugation.[333,364,368]

Taken together, adding new chemistries via genetically encoding ncAA offers new possibilities to engineer protein structures and define protein function beyond the canonical amino acids.

2. Aim

The aim of this study is to expand the toolbox for the direct detection of specific epigenetic nucleobases in user-defined DNA sequences with two approaches utilizing cytosine 5-modification-selective chemical conjugation reactions.

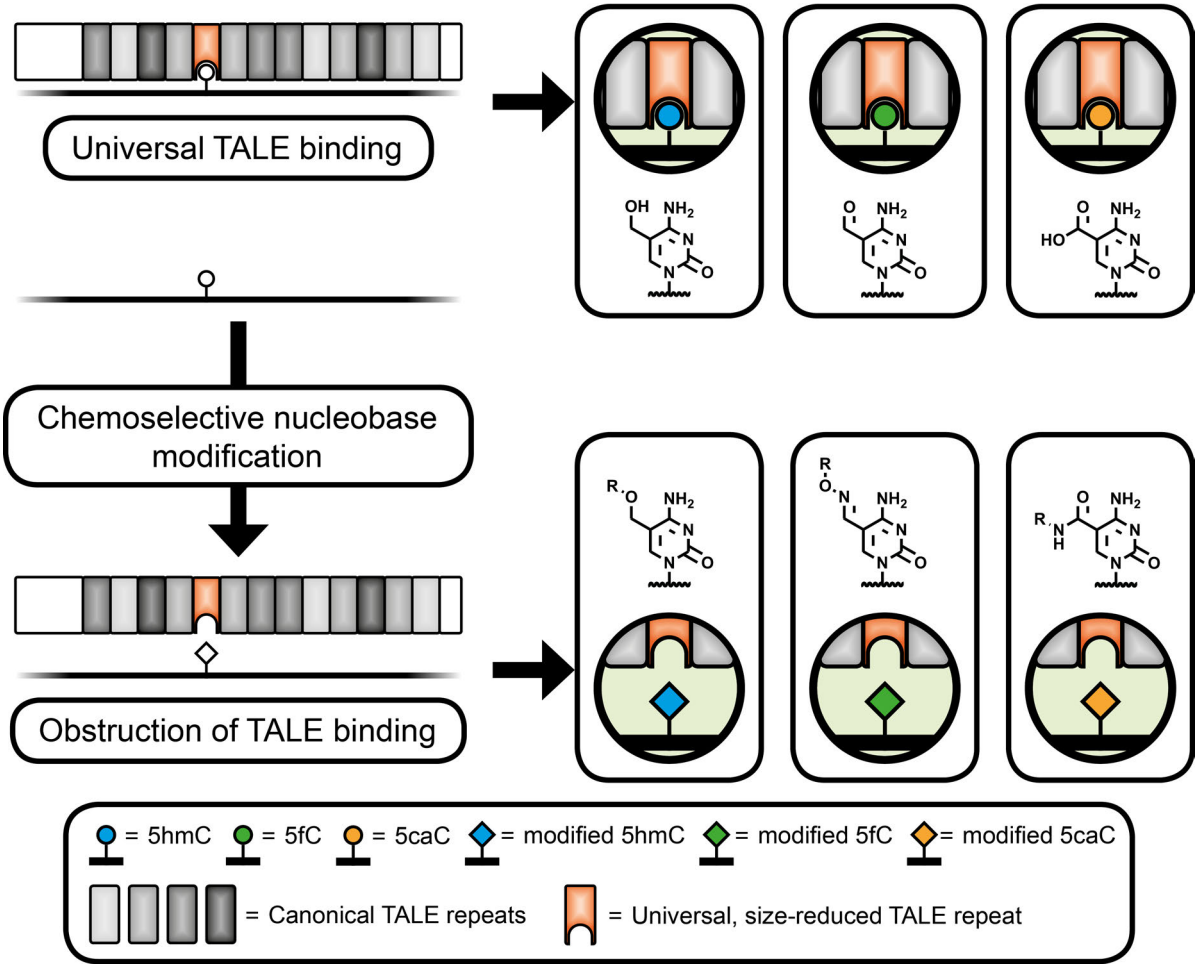


Figure 15: Concept of chemoselective blockage of TALE binding for complete decoding of oxidized 5mC nucleobases in DNA.

First, a universal TALE repeat is sought as the basis for the accommodation of the four canonical and the four epigenetic 5-modified cytosine nucleobases. Next, modification-specific chemical derivatization strategies for 5hmC, 5fC or 5caC will be tested to reintroduce nucleobase-selectivity to the universal TALE repeat. The selective exclusion of a modified nucleobase will provide the basis for the sequence-specific detection of 5hmC, 5fC and 5caC in DNA (**Fig. 15**). This approach will be extended to the detection of 5hmC, 5fC and 5caC in user-defined target DNA within a complex human genomic DNA background.

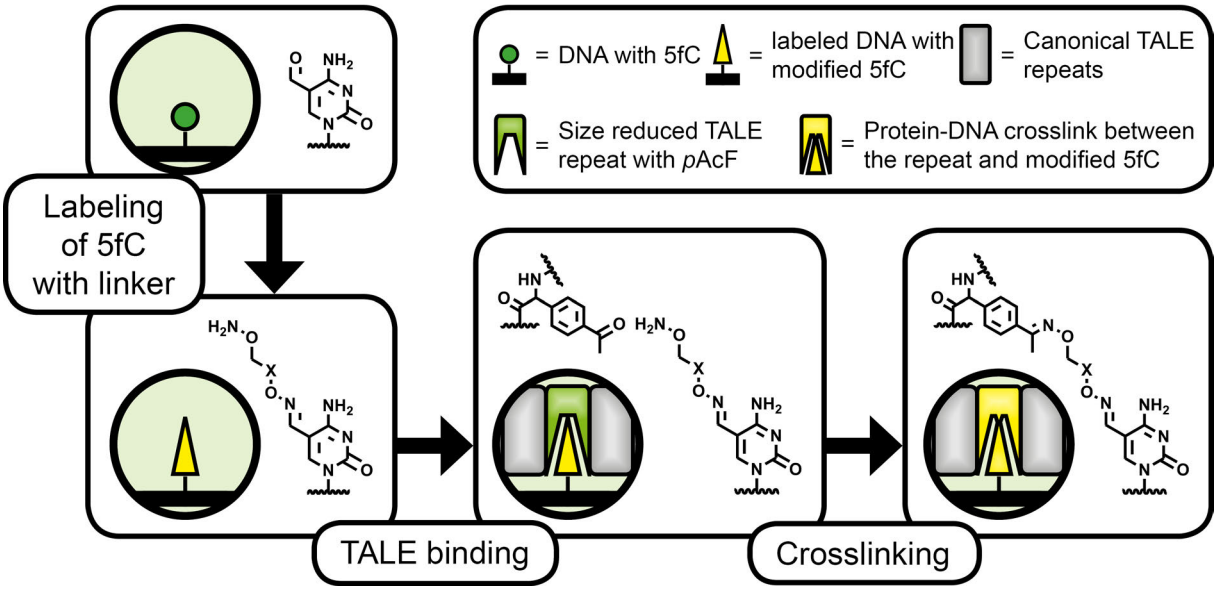


Figure 16: Concept of 5fC-specific crosslinking with TALE proteins bearing a size-reduced repeat bearing *p*AcF through oxime condensation.

Second, 5fC-selective crosslinking with TALE repeats bearing the ncAA *p*AcF will be established to provide a robust strategy to enrich user-defined target DNA with 5fC (**Fig. 16**). First, different amino acid positions in the size-reduced universal repeat will be screened for accommodation of *p*AcF without interfering with TALE-DNA binding. Next, a set of bifunctionalized linkers with hydroxylamine moieties will be tested for non-denaturating sequence-specific crosslinking. The crosslinking yield will be quantified. Selected TALE-linker combinations will then be used for covalent enrichment of target DNA with 5fC from complex genomic DNA backgrounds. Finally, this advanced TALE-based enrichment strategy will be to quantify 5fC levels of target DNA in genomic DNA backgrounds.

3. Results and Discussion[d]

3.1. Complete, Programmable Decoding of Oxidized 5-Methylcytosine Nucleobases in DNA by Chemoselective Blockage of Universal TALE Repeats

3.1.1. Design of Universal TALE Repeats

TALE proteins are programmable DNA-binders than can be used as sensors of epigenetic cytosine 5-modifications. However, positive recognition of single cytosine 5-modifications with full selectivity or high affinity has not been achieved yet with TALE proteins. The rational of this work was to engineer universal TALE repeats with equally high affinity towards the eight human nucleobases. Then, selectivity towards single 5-modifcations could be reintroduced by selectively blocking TALE binding via chemoselective labeling of cytosine 5-modifications to achieve complete, programmable decoding of epigenetic 5-modified cytosine nucleobases.

The deletion of aa position 13 in the TALE repeat N* reduces steric demand and enables accommodation of epigenetic 5-modified nucleobases (*c.f.* 1.3.3.).[320,324] This inspired the design of size-reduced loop structures for unhindered accommodation of nucleobases in the TALE-DNA complex, including cytosine 5-modifications. Therefore, TALE repeats with S or N at aa position 11 were designed by minimalizing the loop of the TALE repeat through the deletion (*) of amino acids at positions 12-15 (**Fig. 17a, b**). During the course of this work, two studies from our lab reported uniform accommodation of cytosine 5-modifications with size-reduced TALE repeats in which the RVD was replaced with single, small amino acids (G, S or T).[324,325]

The original and mutated module vectors provided by A. J. Bogdanove and D. F. Voytas via Addgene were included in Golden Gate assembly of TALE proteins (**Fig. SI1a-c**). These proteins had size-reduced repeats positioned opposite the C of the single CpG dinucleotide (underlined) of either a 26 nucleotide (nt) sequence in the human tumor suppressor gene CDKN2A (TCAGCCGAAGGCTCCATGCTGCTCCC, human chr9: 21974786-21974811) or a 26 nt sequence of a promotor sequence associated with the DNA repair protein BRCA1

[d] The contents of this section have been (partly) published in Gieß et al. (2018) and Gieß et al. (2019).[369,370] Reprinted (adapted) with permission from ref.[369] and ref.[370] Copyright 2019 American Chemical Society.

(TCTTTCCTTTTACGTCATCCGGGGGC, human chr17: 43125551-43125576, BRCA1a) (**Fig. 17c**).

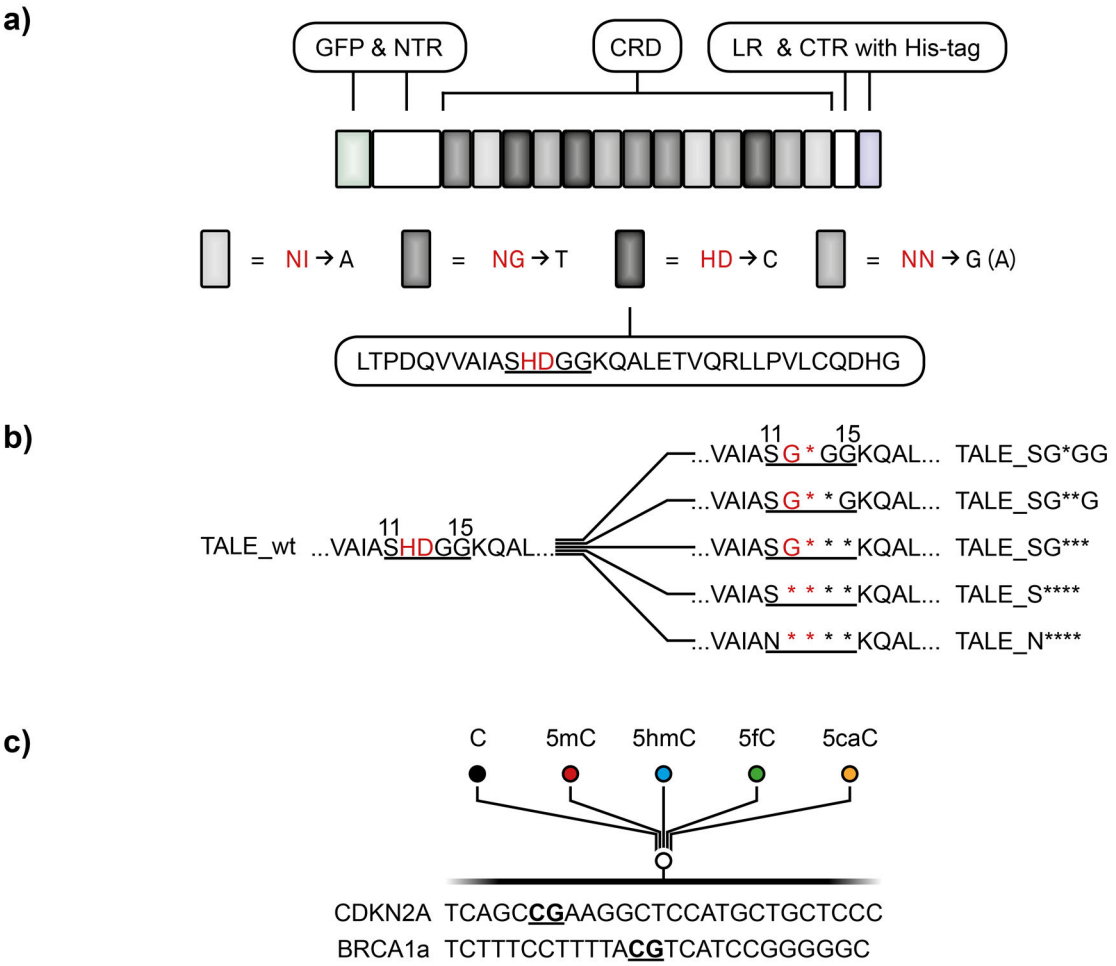


Figure 17: The modular organization of TALEs and design of size-reduced TALE repeats. a) Schematic illustration of a modular recombinant TALE protein with major domains. The RVD of repeats targeting one of the four canonical nucleobases are highlighted in red. GFP: Green fluorescent protein. **b)** Mutagenesis of the loop region (underlined) to generate new size-reduced TALE repeats. *: Deletion. **c)** Target DNA of two subsets of TALEs containing a variable repeat of **b)** opposite of the bold, underlined C, which contains one of the five indicated cytosine nucleobases.

The recombinant TALEs with N-terminal GFP domain, shortened AvrBs3-type NTR and a His6-tag after the CTR were expressed in *E.coli* from a pET-vector, isolated and purified via Ni-Nitrilotriacetic acid (NTA) affinity chromatography (**Fig. SI2a, b**).[303,324] High yields of TALE proteins with molecular weights (MW) at approximately ~130 kDa were isolated with good solubility (**Fig. 18a, b**).

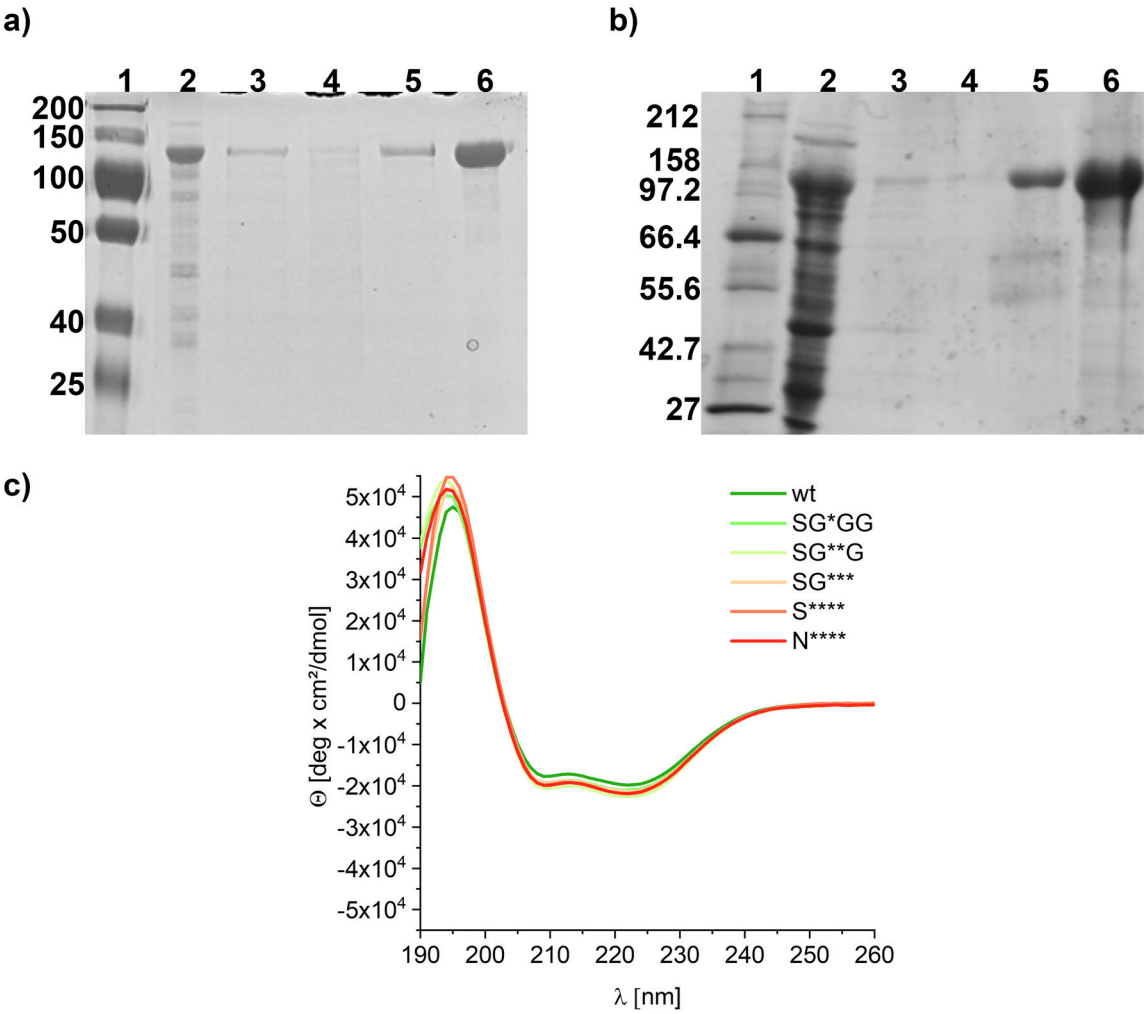


Figure 18: TALE expression and circular dichroism spectroscopy. a) Exemplary sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) image of TALE_wt CDKN2A. Lane 1: Protein Ladder. Lane 2: Flow-through after incubation of lysate with Ni-NTA beads. Lane 3: Washing with 4x PBS. Lane 4: Washing with buffer supplemented with 20 mM Imidazole. Lane 5: Washing with buffer containing 50 mM Imidazole. Lane 6: Purified first elution of TALE_wt CDKN2A. **b)** Exemplary SDS-PAGE analysis of TALE_wt BRCA1. Lane 1: Protein Ladder. Lane 2: flow-through after incubation of lysate with Ni-NTA beads. Lane 3: Washing with 4x PBS. Lane 4: Washing with buffer supplemented with 20 mM Imidazole. Lane 5: Washing with buffer containing 50 mM Imidazole. Lane 6: Purified first elution of TALE_wt BRCA1. **c)** Circular Dichroism Spectra of CDKN2A TALEs.

Circular dichroism (CD) spectra of expressed CDKN2A-TALE proteins displayed a defined maximum at ~195 nm and two minima at ~210 nm and ~225 nm (**Fig. 18c**). These elements are typically observed for α-helix secondary structures.[371] Other features that would indicate β-sheet or disordered structures cannot be observed. The

comparable CD spectra of the tested TALEs indicate that the folding of TALE proteins is not (significantly) affected by the incorporation of size-reduced TALE repeats. Therefore, it can be assumed that the overall structure is comparable between all TALE proteins.

Next, an electromobility shift assays (EMSA) with oligonucleotide duplexes containing the TALE target sequence was used to analyze TALE-DNA binding.

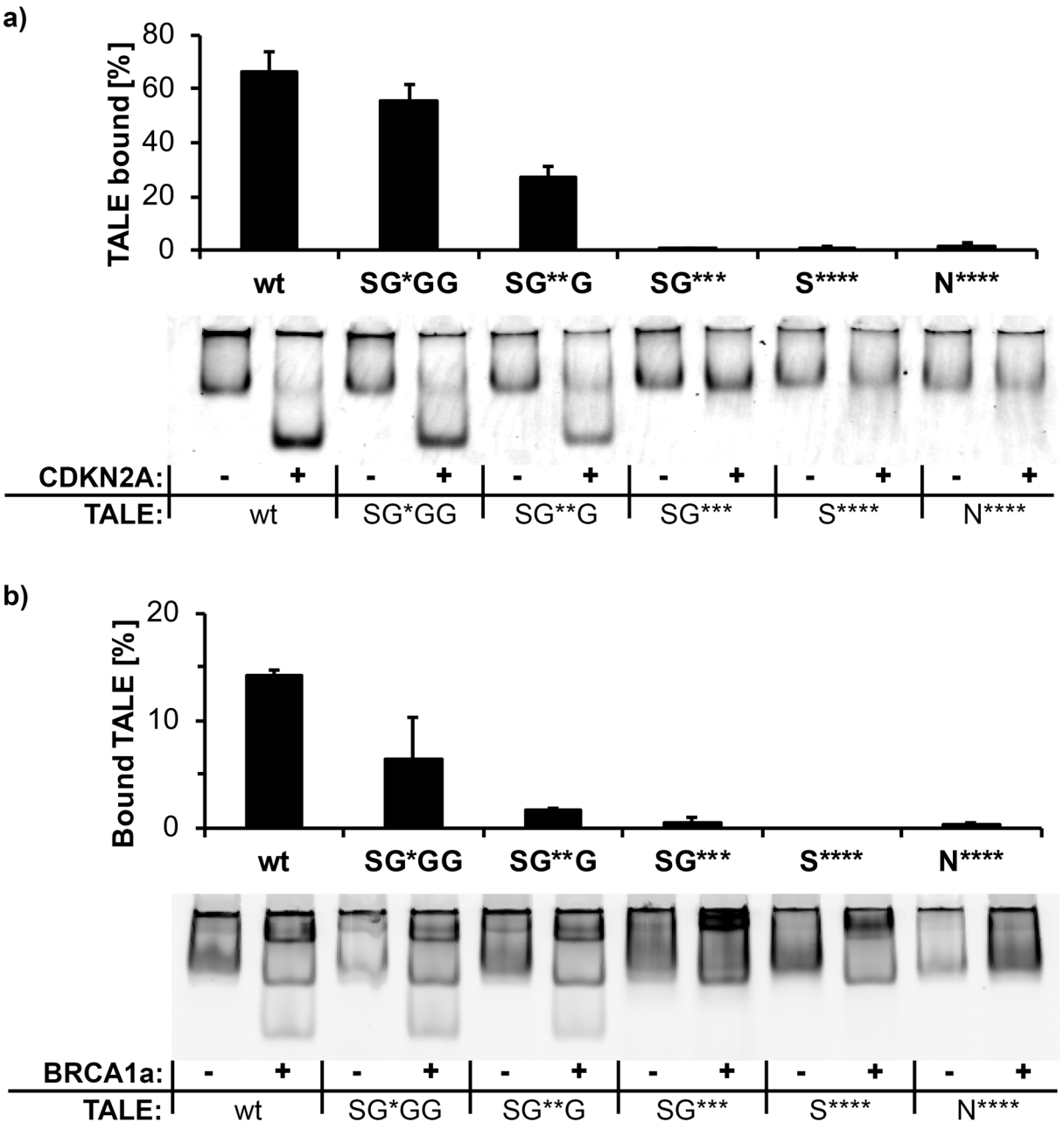


Figure 19: EMSA with TALEs and target DNA with C opposite the indicated repeat. a) Binding profiles of 150 nM CDKN2A TALEs and **b)** 150 nM BRCA1a TALEs to their target DNA.

Herein, the N-terminal GFP-domain is used for detection of protein bands after native polyacrylamide gel electrophoresis (PAGE). A shift down in the gel signals binding to DNA. In comparison, the shift of CDKN2A TALE_wt, TALE_SG*GG and SG**G was significantly higher to the BRCA1a TALEs indicating a stronger binding of CDKN2A TALEs to target DNA with unmodified cytosines at this concentration in EMSA (**Fig. 19a**, **b**). Furthermore, a similar high fraction of TALE_wt and TALE_SG*GG shifted in EMSA. This indicate comparable binding of TALE_wt and TALE_SG*GG with unmodified cytosine.

The strong binding of repeat_SG*GG indicates that this interaction is not affected by the single deletion of aa position 13 similar to repeat N*. It is possible that interactions could be facilitated due to the absence of the steric demand imposed through residue 13 and the overall more flexible loop structure due to the N/H12G substitution.

Amino acid position 14-17 interact with the backbone phosphate of the target DNA. For both sets of TALE proteins, the deletion of aa position 14 reduced the binding for TALE_SG**G to target DNA (**Fig. 19a**, **b**). Strikingly, more than 2 deletions of repeat residue 12-15 abolished binding completely (**Fig. 19a**, **b**). This shows that repeats with residue 11 to 17 consisting of SG***KQ, S****KQ or N****KQ instead of SHDGGKQ are insufficient for TALE-DNA interactions, even though the two amino acids which account for the majority of the total binding energy of TAL proteins to DNA (K16 and Q17)[298] are still present in these repeats.

Next, further binding studies with CDKN2A TALE_wt, TALE SG*GG and TALE SG*GG against 5-modified cytosines were conducted (**Fig. 20**). EMSA data showed the expected selectivity profile for TALE-wt. Strong binding was only observable for C but not for 5mC or ox5mCs, even at a five-fold excess of DNA to TALE (**Fig. 20a, b**).

In contrast, TALE_SG*GG displayed uniform and strong binding to the five 5-modified cytosines (**Fig. 20a, b**). Binding of TALE_SG**G to target DNA was somewhat selective for 5mC and 5fC opposite of the variable repeat, but overall lower than for TALE_SG*GG (**Fig. 20a, b**). Both repeat_SG*GG and repeat_SG**G lack loop stabilizing aa at position 12 which could cause a more flexible orientation of the loop in the major groove of the DNA. The CD spectra indicate that deletions in the loop structure of a single repeat would not compromise the overall structure of the TALE, but the different binding profiles suggest that deletions influence local flexibility of the repeat. Moreover, multiple deletions could tension the structure, restrict conformations and also influence the structure of neighboring repeats which could affect overall TALE

binding Additionally, the deletion of aa position 13 removes possible steric hindrance from a nucleobase-contacting aa side chain for universal accommodation of cytosine 5-modifications. Presence of G residues could enable vdW interaction between the Cα carbons of G residues and 5mC or 5fC. The rotational dimorphism observed for 5hmC[56,57] could be a reason for decreased TALE binding. This could explain why 5mC and 5fC are generally bound with better affinity in size reduced repeats.[324,325]

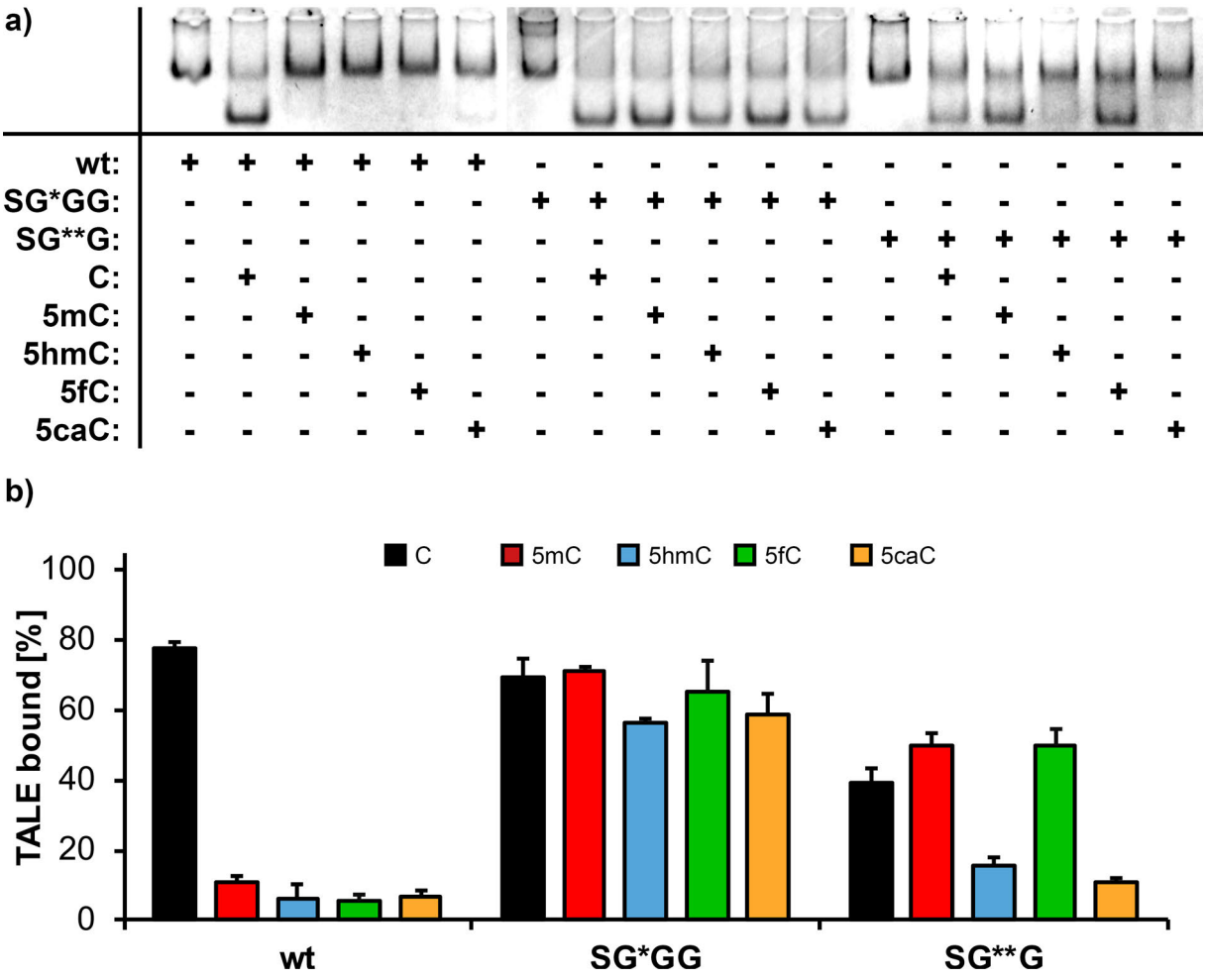


Figure 20: Interaction of CDKN2A TALEs with target DNA bearing C, 5mC, 5hmC, 5fC or 5caC opposite the variable repeat. a) Native PAGE images from EMSA for 150 nM TALEs and DNA as indicated. **b)** Quantified TALE binding from EMSA with TALE_wt, TALE_SG*GG and TALE_SG**G.

A computational modeling approach (credit to J. Jasper and O. Koch) of TALE structures with repeat_wt, repeat_SG*GG and repeat_SG**G projected a similar TALE-DNA complex. Furthermore, the repeat_SG*GG and repeat_SG**G form cavities with increasing size, which can accommodate even the arguably sterically

most demanding 5caC. (**Fig. 21**, employed modeling approach was not applicable to repeats with >2 deletions). This is in good accordance with the experimental data.

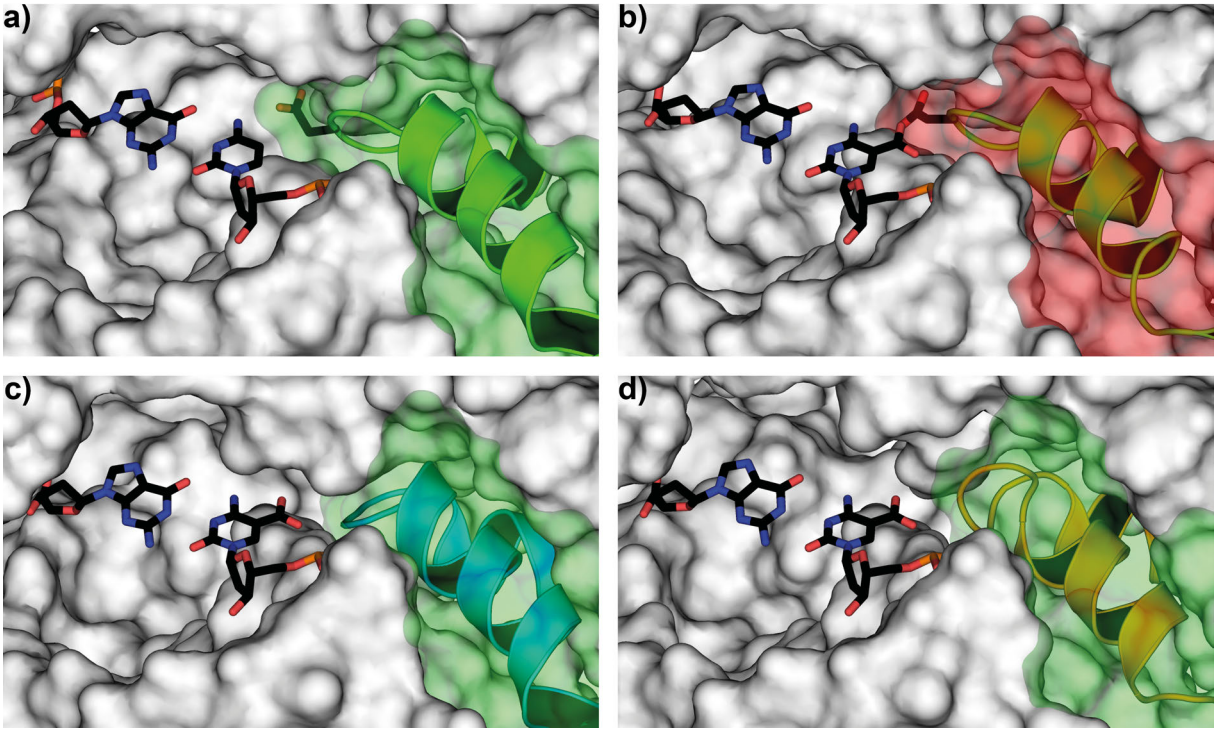
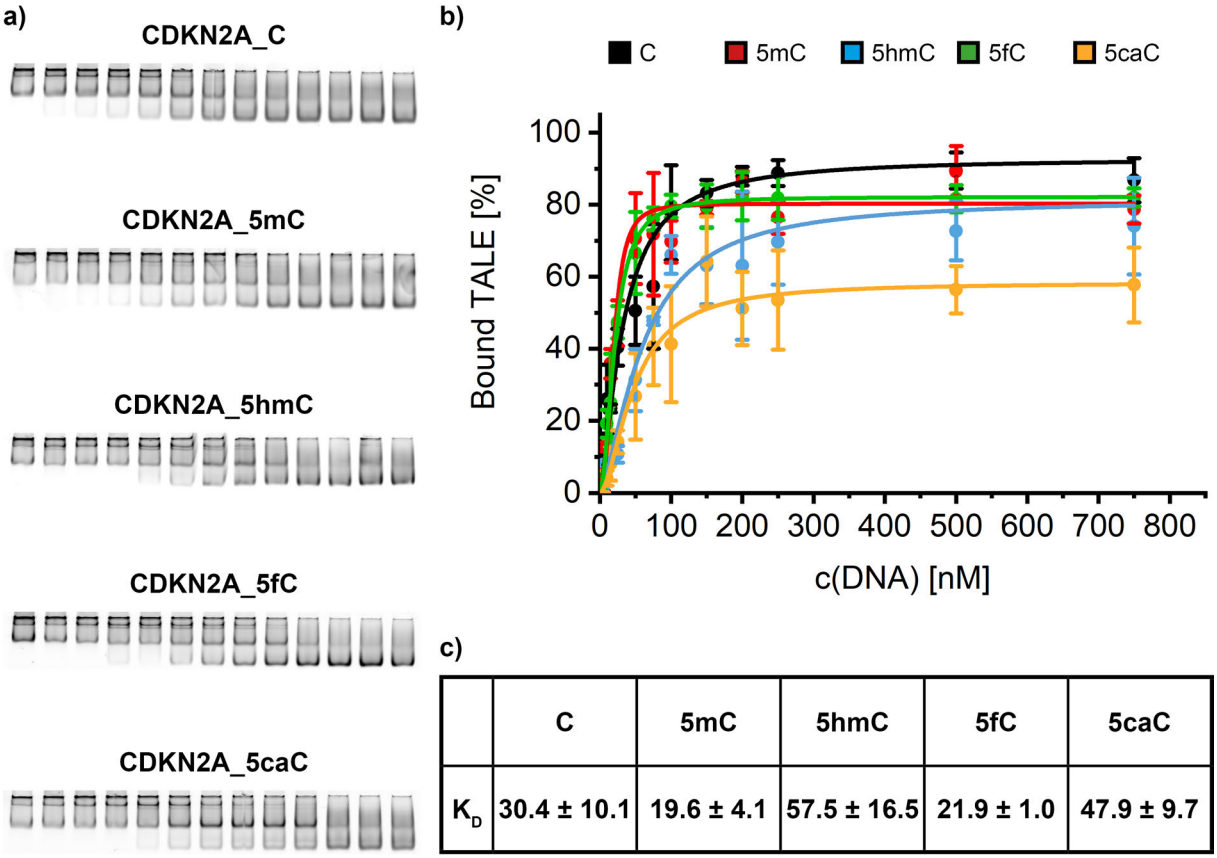


Figure 21: Models of TALEs bound to targets with C or 5caC opposite of the variable repeat position. a) Modelled repeat_SHDGG with C. **b)** Modelled repeat_SHDGG with 5caC with the steric clash of D13 and the 5-carboxyl. **c)** Modelled repeat_SG*GG with a cavity for accommodation of 5caC. **d)** Modelled repeat_SG**G has a similar cavity for accommodation of 5caC but has a somewhat twisted conformation.

TALE-DNA binding at single concentrations cannot accurately define the binding ability of the TALE proteins. Therefore, TALE binding curves were determined for TALE_SG*GG as the strongest binder and the five cytosine nucleobases (**Fig. 22a**, **b**). TALE binding was approximately saturated at 80 % for target DNA with C, 5mC or ox5mCS, except for TALE binding to target DNA with 5caC which was approximately saturated at 50 %. Dissociation constants (K_D) calculated from the fitted binding curves confirmed a strong, universal binding of TALE_SG*GG to the five cytosine nucleobases within one order of magnitude (**Fig. 22c**).



	C	5mC	5hmC	5fC	5caC
K_D	30.4 ± 10.1	19.6 ± 4.1	57.5 ± 16.5	21.9 ± 1.0	47.9 ± 9.7

Figure 22: Binding curves of TALE_SG*GG with CDKN2A target DNA. a) EMSA for titration of 150 nm TALE_SG*GG with target DNA bearing one of the five cytosine nucleobases opposite of repeat_SG*GG. **b)** Data points from **a)** with corresponding binding curves obtained from non-linear dose response fit. **c)** Calculated K_D-values in nM for TALE binding.

It would be of interest if the universal binding of TALE_SG*GG also applies to the other three canonical nucleobases A, G and T. This could be of use for targeting or bypassing variations in the consensus target DNA in the form of single nucleotide polymorphisms (SNPs) during epigenetic nucleobase analysis or in vivo (epi-)genome engineering and transcription control. EMSA data with TALE_SG*GG and target DNA bearing A, G or T opposite of repeat_SG*GG as well as a transcription activation assay with a luciferase reporter and TALE_SG*GG fused to the transcriptional activator VP64 in HEK293T cells (credit to A. Witte) revealed similar binding to all four canonical nucleobases (**Fig. 23a**, **b**). However, the relative luciferase activity (RLA) triggered through expression of TALE_SG*GG-VP64 was only half as strong as for TALE_wt-VP64 and the canonical target DNA. This suggests a difference in binding between the two proteins in live cell experiments. To analyze potential differences and affinity in

binding of TALEs in cellular systems, further *in vivo* studies with multiple cell lines are currently underway in the Summerer lab as of writing this thesis.

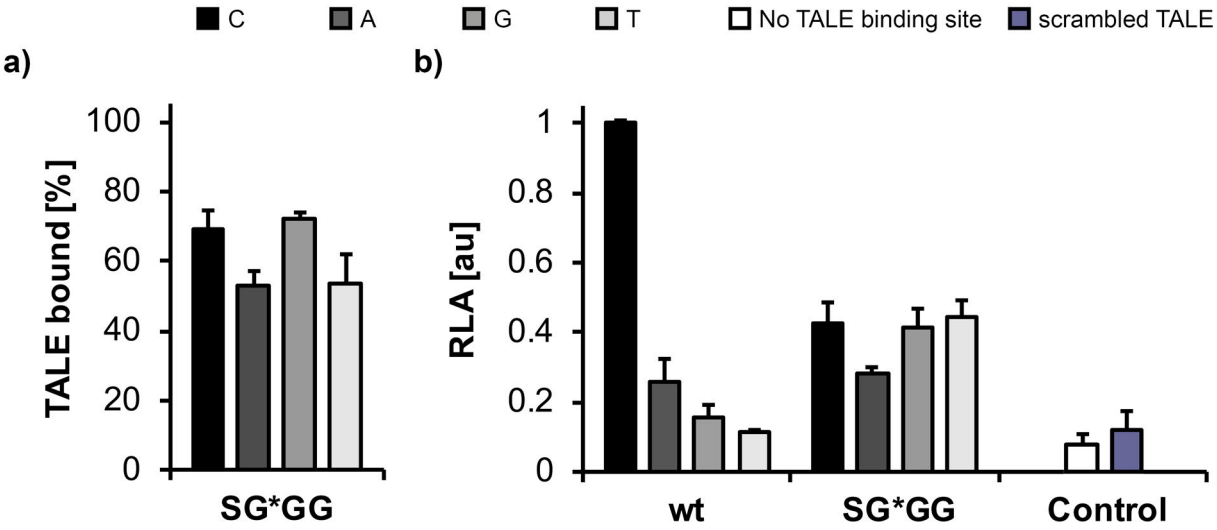


Figure 23: Interaction of CDKN2A TALEs with target DNA bearing C, A, G or T opposite the variable repeat. a) EMSA of 1.5 pmol TALE_SG*GG and DNA. **b)** Luciferase transcription activation using TALE-VP64 constructs with indicated repeat opposite single C, A, G, or T in the target. Relative luciferase activity (RLA) normalized to TALE_wt binding to target site with C. Scrambled TALE control represents a binding of a TALE with a shuffled TALE_wt CRD sequence.

Taken together, these results are in accordance with previous binding studies that reported similar binding profiles for TALEs with repeat_SG*GG for different target DNA sequences.[324,325] This indicates that these selectivity profiles of TALE-DNA binding are not significantly influenced by the DNA context and/or protein sequence.

Most importantly, these results revealed the first repeat design that enables universally binding to all eight human DNA nucleobases.

It is likely, that negative discrimination of repeat_SG*GG applies to at least certain nucleobases. Unlike HD-repeats for C, the molecular structure of the loop does not provide chemical handles for positive recognition of nucleobases like C, 5caC, G or A. For the other bases however, repeat-SG*GG could provide the molecular scaffold for vdW interaction with 5-modification of 5mC, 5fC and T. This could explain the higher affinity of TALEs with engineered repeats with a deletion at aa position 13 for 5mC and 5fC compared to 5hmC and 5caC.

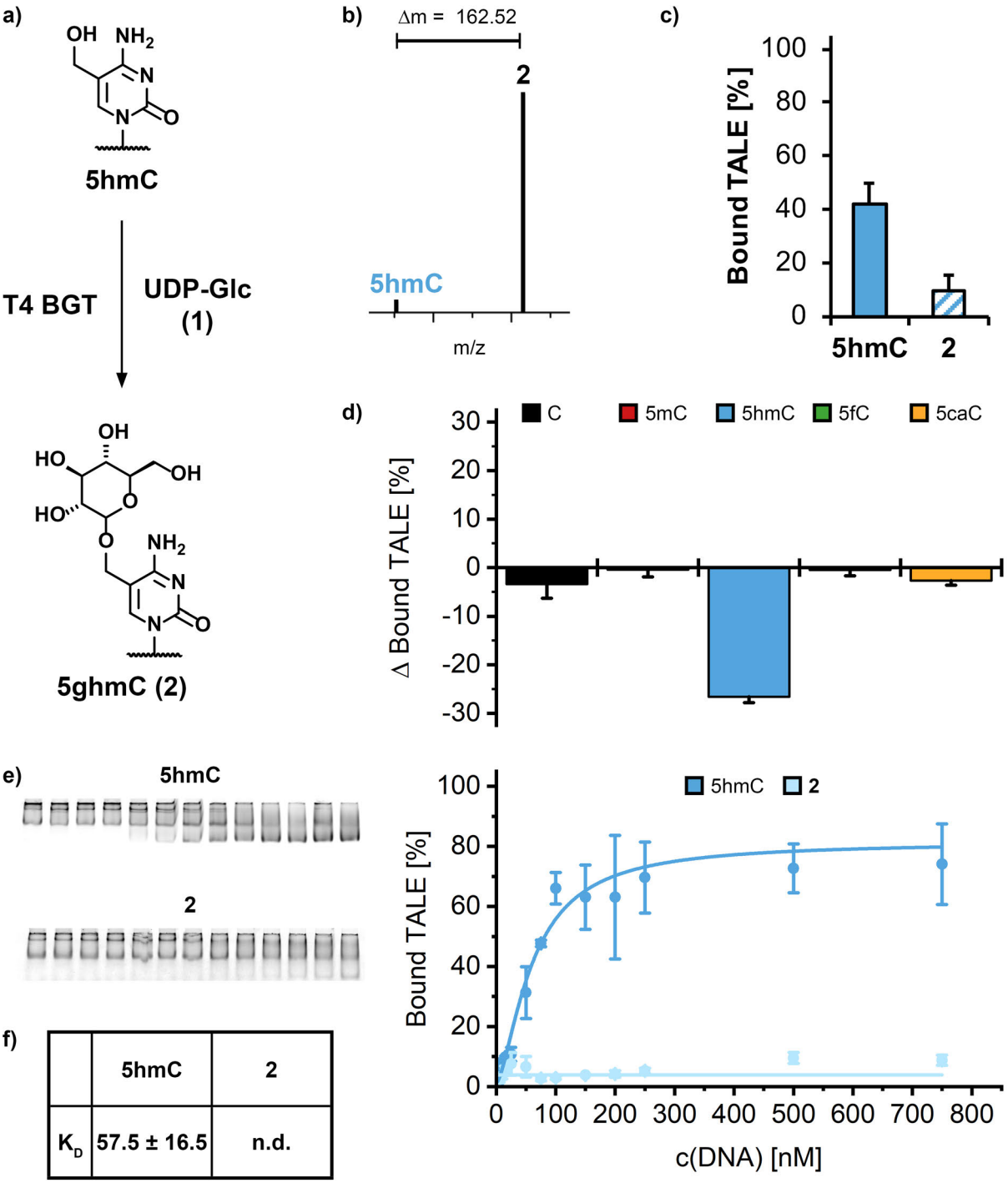
3.1.2. Cytosine 5-modification-selective Blockage of TALE Binding

Chemoselective conjugation of cytosine 5-modifications to block protein functions has been used to detect epigenetic 5-modified cytosines (*cf*. chapter 1.2.3.2.). To investigate the possibility of TALE binding reintroducing selectivity of for ox5mCs through this strategy, uridine diphosphate (UDP)-glucose (**1**)-dependent enzymatic glucosylation with T4 BGT was chosen (**Fig. 24a**). It is a well-established and facile 5hmC-selective modification reaction (*cf*. chapter 1.2) which had been shown to obstruct protein function such as endonuclease activity of MspI for example.[243] Multiple reaction times were tested (data not shown) and product formation was analyzed through electrospray ionization (ESI)-MS. The overnight reaction indicated efficient glucosylation with a high yield of 5ghmC (**2**) formation with a rest of 5hmC remaining detectable (**Fig. 24b; Fig. SI3; Table SI1**).

TALE_SG*GG binding to oligonucleotides bearing 5ghmC opposite of repeat_SG*GG was evaluated in EMSA studies. In contrast to unmodified 5hmC, binding was strongly reduced for 5hmC-DNA subjected to enzymatic glucosylation. No blocking effects were observed for the other four cytosine nucleobases subjected to enzymatic glucosylation (**Fig. 24c, d; Fig. SI4**). Additionally, TALE binding in EMSA to different concentrations of **2**-DNA did not exceed 10 %. No K_D could be determined from TALE_SG*GG binding curves for 5ghmC-DNA, confirming the 5hmC-selective strong blocking effect (**Fig. 24e, f**). This indicates that the indicated conformational flexibility of repeat_SG*GG is limited and cannot evade steric clash with the large hydrophilic glucose moiety in **1**. This suggests that labeling of 5fC and 5caC with similarly sterically challenging moieties can also obstruct TALE binding.

Therefore, oxime formation with hydroxylamines varying in size, flexibility, branching degree and polarity was tested for 5fC-selective blockage. Oxime formation is a well-established mild reaction. Moreover, oxime formation with supramolecular moieties have been used for 5fC-selective inhibition of protein function (*cf*. chapter 1.2.3.2.).[218,219]

Different reaction conditions and catalysts (data not shown) for DNA modification were tested with hydroxylamines including O-ethylhydroxylamine (**3a** to **4a**), O-(tetrahydro-2H-pyran-2-yl)hydroxylamine (**3b** to **4b**), O-*ortho*-benzylhydroxyl-amine (**3c** to **4c**) and O-tert-butylhydroxylamine (**3d** to **4d**) (**Fig. 25a**).



	5hmC	2
K_D	57.5 ± 16.5	n.d.

Figure 24: 5hmC-selective blockage of TALE binding. a) Glucosylation of 5hmC with UDP-Glc (1) to 5ghmC (2) by T4-β-glucosyltransferase (BGT). **b)** Extracted mass peaks from Fig. SI3 and Table SI1 indicating highly efficient glucosylation of 5hmC. **c)** Effect of 2 on TALE_SG*GG binding. **d)** Interaction of TALE_SG*GG with DNA containing nucleobase as indicated. EMSA data using each 150 nM TALE and reacted DNAs were subtracted from data of unreacted DNA and plotted as % difference (Δ). **e)** Binding curves of TALE_SG*GG with unreacted 5hmC-containing DNA and reacted 5hmC (**2**)-containing DNA. **f)** KD values from binding curves in **e)**.

Eventually, matrix-assisted laser desorption/ionization (MALDI) time of flight (TOF) MS indicated high oxime yields under mild acidic conditions in a phosphate buffer with *para*-phenylenediamine (*p*PDA) as catalyst (**Fig. 25a, b; Fig. SI5; Table SI2**).

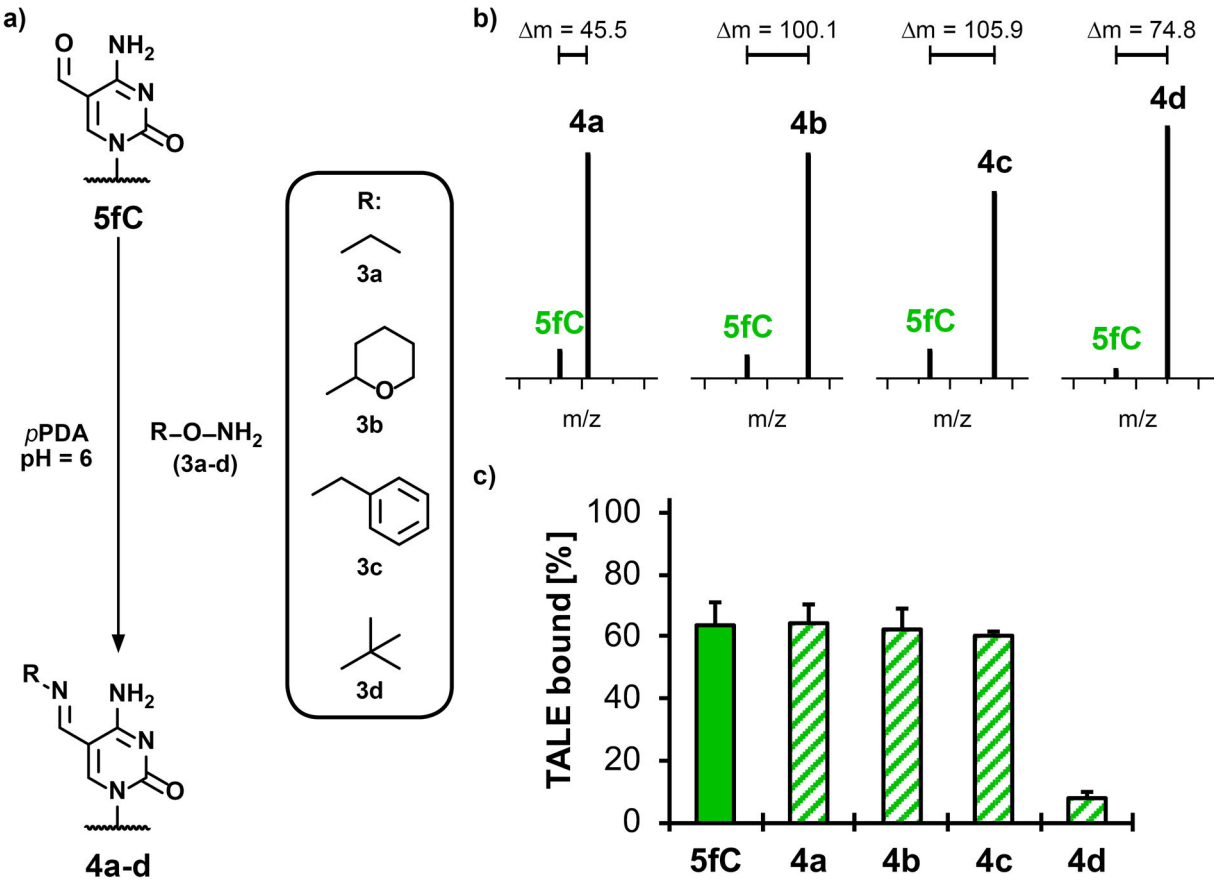
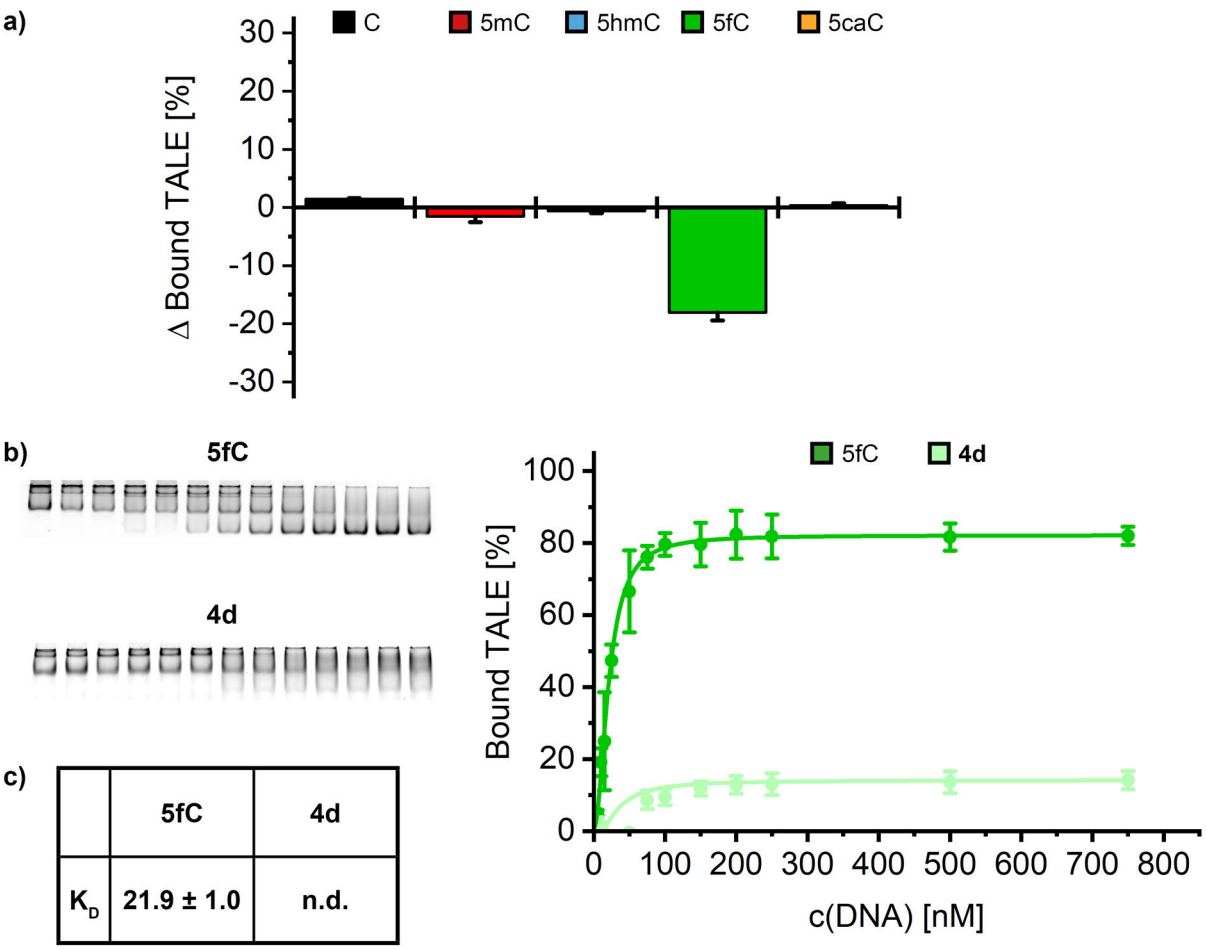


Figure 25: pPDA-catalzyed oxime formation of 5fC-DNA with hydroxylamines 3a-d. a) Oxime reaction scheme of 5fC to **4a-d** with *para*-phenylenediamine (*p*PDA). **b)** Extracted mass peaks from Fig. SI5 and Table SI2 indicating highly efficient oxime formation. **c)** Effect of **3a-d** opposite of repeat_SG*GG on TALE binding.

EMSA for the evaluation of TALE_SG*GG binding to 5fC-DNA after oxime formation revealed that binding was strongly reduced to DNA with **4d** but was still able to bind to DNA with **4a**-**c** (**Fig. 25c**). A possible explanation of this binding profile could be a fixed orientation of oxime substituents due to the hydrogen bond of the oxime nitrogen with the 4-amino group similar to the orientation of the formyl group of 5fC in DNA.[59] Crystal structures of TALE-DNA complexes indicate that TALE proteins do not completely occupy the major groove of the DNA.[280,281] This might create a “blind spot” that can accommodate bulky substituents without interfering with TALE binding.[372]



c)

	5fC	4d
K_D	21.9 ± 1.0	n.d.

Figure 26: 5fC-selective blockage of universal TALE binders. a) Interaction of TALE_SG*GG with DNA containing nucleobase as indicated. EMSA data using each 150 nM TALE and reacted DNAs with **3d** were subtracted from data of unreacted DNA and plotted as % difference (Δ). **b)** Binding curves of TALE_SG*GG with unreacted 5hmC-containing DNA and reacted 5fC-containing DNA (**4d**). **c)** KD values from binding curves in **b)**.

This leaves the possibility that structurally-related glucose groups like the tetrahydropyranyl group of **4b** or aromatic residues like the rigid benzyl group of **4c** and even bulkier substituents like an Alexa Fluor 488 probe (**Fig. SI6**) can be accommodated without interference. In contrast, the strong effect of **4d** on TALE binding indicates, that the quaternary sp^3-center of the tert-butyl residue was critical. Therefore, the orientation and conformational flexibility of the residue from the 5-substituent after conjugation and not mere size of the conjugated residue appears to be an essential factor for effective blocking. Indeed, models (credit to J. Jasper and O. Koch) suggest that DNA with **4c** for example can be favorably oriented to evade steric clash with the TALE_SG*GG repeat, whereas DNA with **4d** cannot (**Fig. SI7**).

Similar to 5hmC-specific blocking, reaction of DNA with the five cytosine nucleobases and **3d** only led to an effect for TALE binding in EMSA for 5fC-DNA (**Fig. 26a, Fig. SI4**). TALE binding to **4d**-DNA was saturated <15 % in titration experiments with EMSA (**Fig. 26b, c**). Again, no K_D-value could be determined for TALE_SG*GG with **4d**-DNA (**Fig. 26c, d**).

Carboxyl groups can form hydroxamates with hydroxylamines, but generally need to be activated to obtain good yields.[373] Under the employed conditions, no blocking effect cannot be observed for 5caC. This indicates that either no hydroxamate between 5caC and **3d** is formed or that the formation does not significantly interfere with TALE binding. Also, no to very little differences in binding of TALE_SG*GG to unreacted and reacted DNA indicates that the DNA is not significantly denatured or otherwise affected in a way that could prevent TALE binding to the other nucleobases. Together, these results demonstrate a strong 5fC-selective blocking effect of TALE binding.

Reactions with 5caC have only been reported in one instance for an EDC-catalyzed NHS ester formation and subsequent coupling with primary amines.[158] Testing different reaction conditions and catalysts (data not shown), efficient amide bond formation of 5caC with 7-azabenzotriazol-1-yloxy)tripyrrolidino-phosphonium hexafluorophosphate (PyAOP) under mild basic conditions in MES buffer was identified (**Fig. 27a**). To the best knowledge, amide formation with PyAOP has not been previously reported for 5caC. This can be an alternative to amide formation with EDC for other chemical modification-based approaches.

MALDI-TOF MS of coupling reaction with different amines indicated good yields for ethylamine (**5a** to **6a**), THP-methylamine (**5b** to **6b**) and benzylamine (**5c to 6c**) (**Fig. 27b; Fig. SI5; Table SI2**). In contrast, tert-butylamine (**5d** to **6d**) reacted poorly (**Fig. 27b; Fig. SI5; Table SI2**).

Evaluation of TALE binding to target DNA bearing 5caC after amide coupling with **5a-d** in EMSA revealed that only formation of **6c** led to effective blocking (**Fig. 27c**). Due to the orientation of the carboxyl group of 5caC with the carbonyl oxygen of 5caC making a hydrogen bond to the exocyclic C4-amino group,[59] the substituents of **6a**-**d** are likely differently positioned than substituents of **4a**-**d**. This could explain the different structure-activity relationships compared to 5fC. Herein, the similar sized, but more flexible sp^3-linked THP group of **6b** but not the benzyl group of **6c** could evade clash with the repeat backbone (**Fig. SI8**). Additionally, likely due to poor reactivity of **5d** no significant effect on TALE binding can be observed for **6d**.

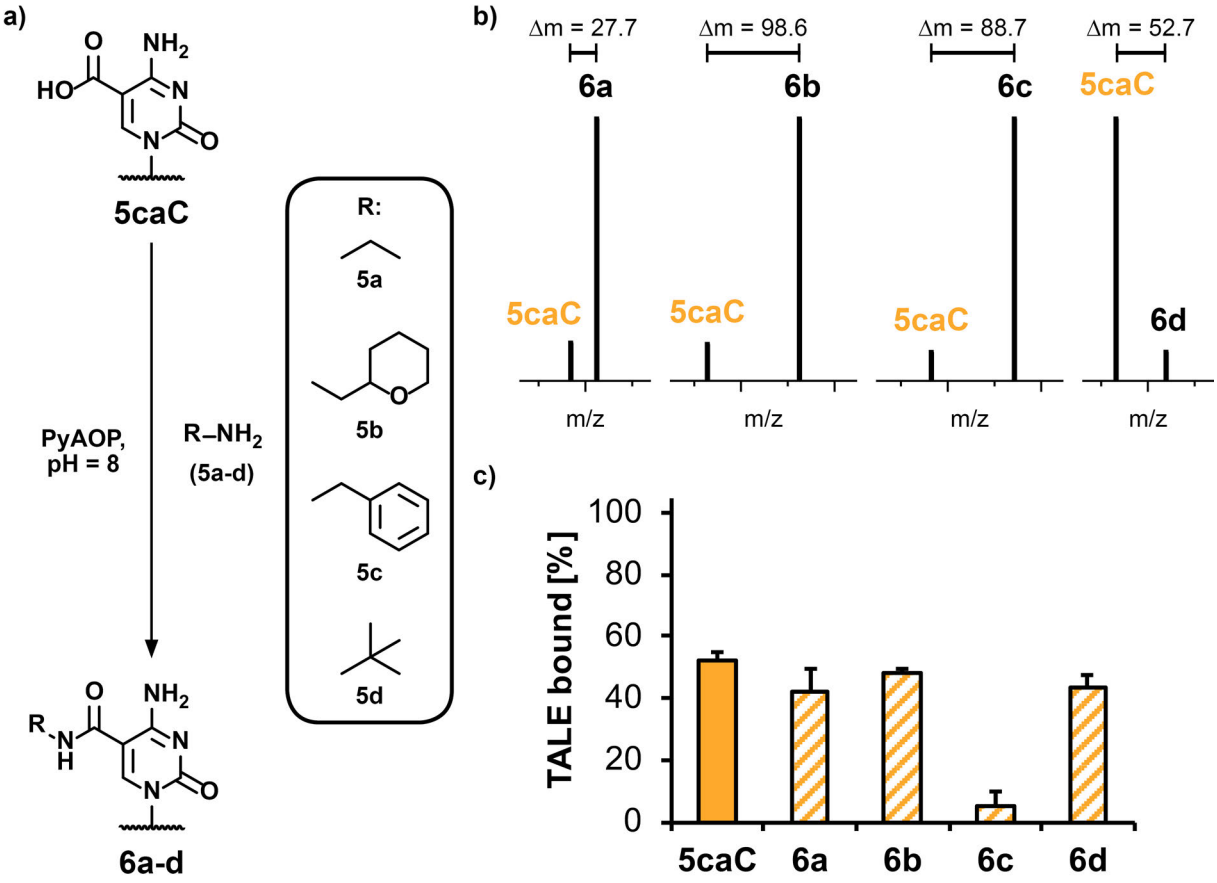
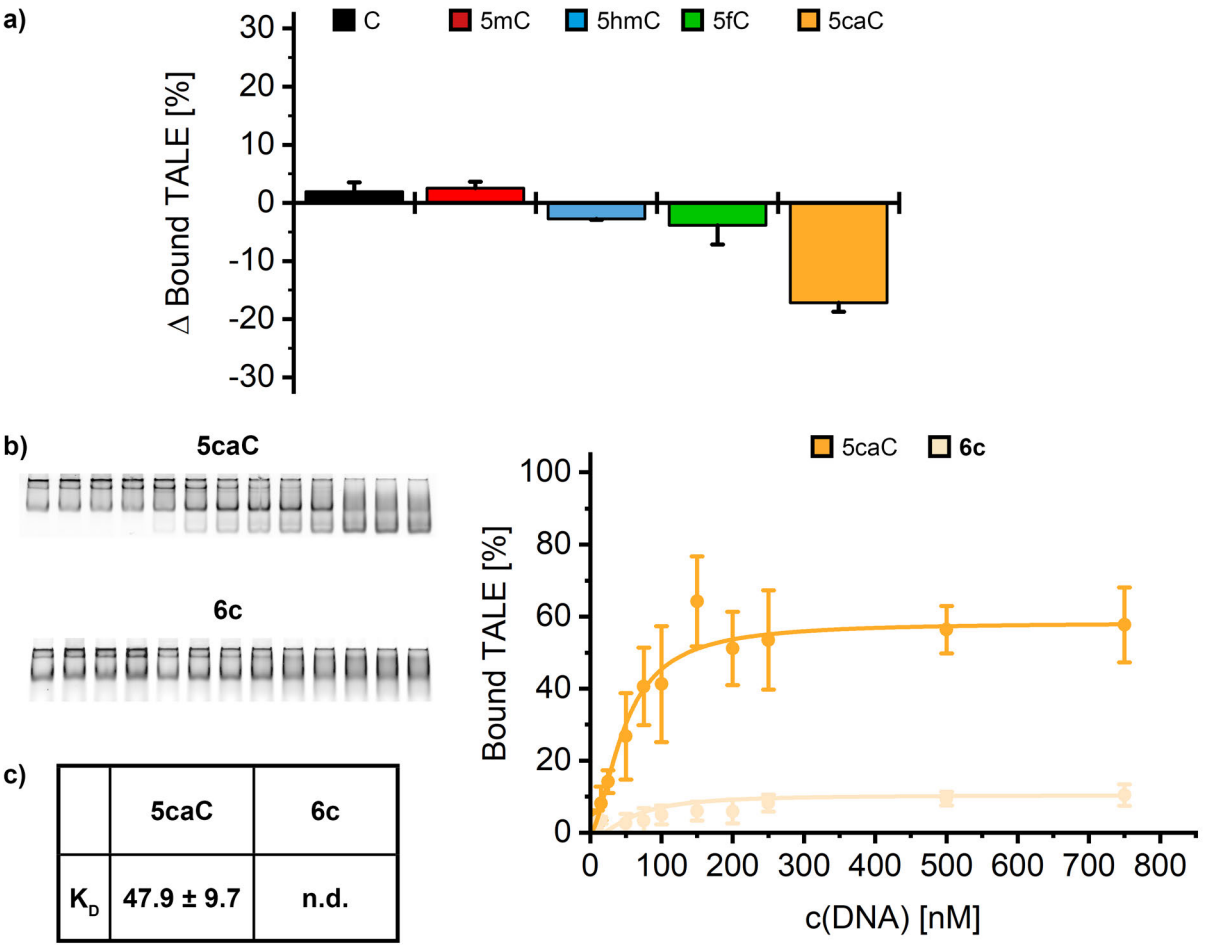


Figure 27: PyAOP-mediated amide formation with 5caC-DNA with amines 5a-d. a) Amide reaction scheme of 5caC to **6a-d** with (7-azabenzotriazol-1-yloxy)tripyrrolidino-phosphonium hexafluorophosphate (PyAOP). **b)** Extracted mass peaks from Fig. SI5 and Table SI2 indicating highly efficient amide formation for **5a-c** but not for **5d**. **c)** Effect of **5a-d** opposite of repeat_SG*GG on TALE binding.

Similar to glucosylation for 5hmC and oxime formation for 5fC, the effect of amide coupling conditions with **5c** was chemoselective for 5caC (**Fig. 28a; Fig. SI4**).

Again, as DNA with other cytosines did not display strong differences in binding after the reaction, this indicates that the DNA is not significantly denatured or otherwise affected in a way that could prevent TALE binding to the other nucleobases. Under basic conditions, 5fC can participate in Schiff base formation.[295] However, potentially due to the different position of the benzyl residue similar to **4c** or low yield, no pronounced effect can be observed.

TALE binding for DNA with **6c** saturated <10 % in titration experiments with EMSA (**Fig. 28b, c**). As for DNA with **2** and **4d**, no K_D could be determined from fitted binding curves. (**Fig. 28c, d**). This demonstrates strong 5caC-selective blocking of TALE binding.



c)

	5caC	6c
K_D	47.9 ± 9.7	n.d.

Figure 28: 5caC-selective blockage of universal TALE binders. a) Interaction of TALE_SG*GG with DNA containing nucleobase as indicated. EMSA data using each 150 nM TALE and reacted DNAs with **5c** were subtracted from data of unreacted DNA and plotted as % difference (Δ). **b)** Binding curves of TALE_SG*GG with unreacted 5hmC-containing DNA and reacted 5caC-containing DNA (**6c**). **c)** KD values from binding curves in **b)**.

Taken together, three facile non-denaturing chemical modification approaches were identified that obstruct binding of TALEs with an engineered universal repeat. This enable cytosine 5-modification-selective blockage of TALE binding for 5hmC, 5fC and 5caC for the complete, programmable decoding of epigenetic 5-modified cytosines.

3.1.3. Complete decoding of oxidized 5mC nucleobases

Finally, the approach was tested for detection of epigenetic nucleobases at single, user-defined positions of 5hmC, 5fC and 5caC in a TALE-based genomic enrichment assay[324] with human genomic DNA background (**Fig. 29a**).

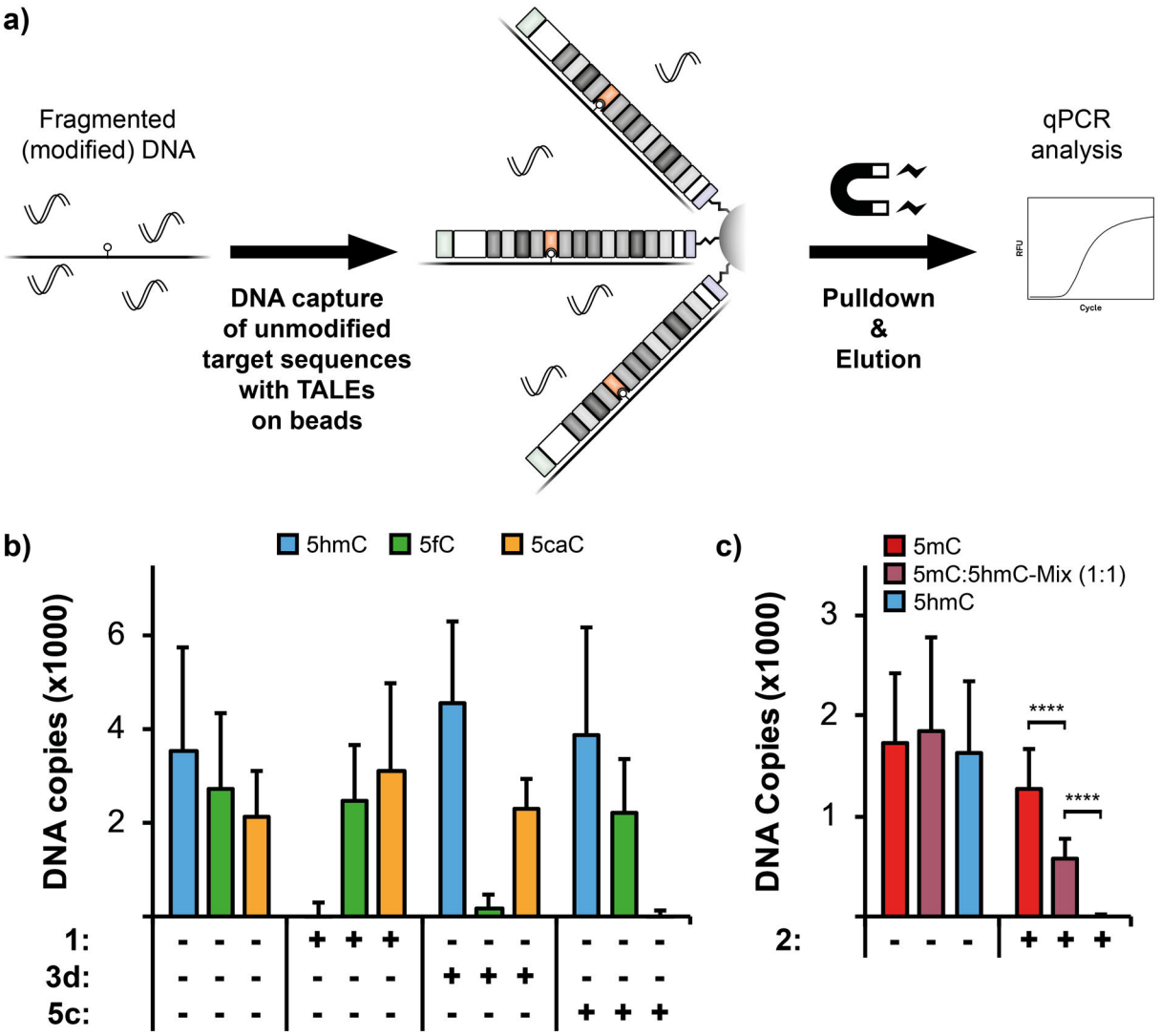


Figure 29: Complete decoding of oxidized 5mC nucleobases in a genomic DNA background with a TALE-based enrichment assay. a) Affinity enrichment workflow. **b)** qPCR quantification of affinity enrichments with TALE_SG*GG and DNAs bearing single oxidized 5mC reacted as indicated. Copies of control enrichments without TALE were subtracted from enrichments with TALE. **c)** Enrichment of gDNA with 5mC, 5hmC or 1:1 mix as in b). ****: $p \leq 0.00001$; $n \geq 5$.

To benchmark the chemoselective blocking strategy in the genomic enrichment assay, a defined modification status of target DNA is required. Therefore, 421 bp long sequences with the CDKN2A TALE target DNA and a single 5hmC, 5fC or 5caC

opposite repeat_SG*GG were generated (**Fig. SI9a**). Interestingly, efficient synthesis of spike-ins with 5caC in the target DNA was considerably more challenging with standard PCR protocols compared to spike-ins with 5hmC of 5fC in the target DNA. A plausible reason could be that G-5caC pairs have been reported to stimulate DNA polymerase exonuclease activity.[374] This could apply to the 3′→5′ exonuclease activity of the utilized Q5 DNA polymerase. Therefore, the G-5caC pair could be detected as DNA lesion and affect polymerase activity. Accordingly, good processivity of 5caC in PCR with a polymerase without 3′→5′ exonuclease activity had been reported.[375] Additionally, Q5 polymerase is fused to the non-specific DNA-major groove binding protein Sso7d. It is used for better processivity, but can be sensitive to 5-modifications of a pyrimidine.[376] 5caC with the arguably sterically most challenging 5-modification could therefore have an impact on binding and hence Q5 polymerase processivity. Eventually, running a gradient PCR with low annealing temperatures resolved this issue. Still, understanding underlying mechanism could be an interesting topic for further studies.

Synthesized amplicons were spiked into a fragmented (~300-700 bp) non-modified genome background (bg) DNA at a concentration of a single genome copy (**Fig. SI9b**). The bgDNA was generated through whole genome amplification (WGA) from gDNA of a human individual (ENCODE sample NA18507). To remove false-positive enrichment, bgDNA was digested within the corresponding spike-in sequence including the TALE target DNA (**Fig. SI9c**).

Next, bgDNA with spike-ins bearing 5hmC, 5fC or 5caC were subjected to glucosylation, oxime formation or amide formation. Then, purified reacted and non-reacted DNA samples were employed in TALE-based affinity enrichments[324] with TALE_SG*GG.

Non-reacted DNA and reacted DNA bearing off-target nucleobases were enriched at comparable levels from the genomic DNA background (**Fig. 29b**). In contrast, DNA with the target nucleobases were only enriched at background levels, confirming the previous results for cytosine 5-modification-selective blockage of TALE binding and advancing the approach from short oligonucleotides to a complex genomic DNA background (**Fig. 29b**).

Next, the sensitivity of the approach was analyzed. Therefore, target sites with a heterogeneous modification status at a single CpG-site were spiked into bgDNA.

Heterogeneity of naturally occurring 5-modifications at single CpGs in gDNA is an important aspect due to dynamic process of the iterative oxidation of 5mC.

Enrichment of TALE targets with pure or equimolar mixed 5mC/5hmC-sites showed an inverse proportional TALE binding to 5hmC levels in reacted samples (**Fig. 29c**). This demonstrates differentiation between two epigenetic cytosine 5-modification based on chemoselective blockage in a complex genomic DNA background.

Notably, the loss of DNA in both enrichment procedures were surprisingly high at ~95 % and comparable to the denaturing conditions of bisulfite sequencing.[137] It is possible that the genomic DNA behaves differently than short oligonucleotides under the different conjugation reaction conditions. Another option is the relatively short binding time of TALE to target DNA in the complex gDNA environment meaning that a high percentage of TALEs could be non-specifically bound to bgDNA searching for target DNA. During washing, this fraction of sample would then be lost. Additionally, the dense immobilization of TALEs on agarose beads could lead to restriction of their binding to DNA.

In conclusion, these results demonstrate fully selective modification and sensitive detection of single 5hmC, 5fC and 5caC with TALEs in a genomic assay with high confidence which expands the toolbox of detection strategies for ox5mCs.

3.2. Programmable Protein-DNA Crosslinking for the Direct Capture and Quantification of 5-Formylcytosine

3.2.1. Design of TALE Repeats bearing *para*-Acetylphenylalanine

5fC is a highly dynamic, cell-type specific cytosine 5-modification. 5fC can be a stable modification but relatively little is known about the function of 5fC in gDNA.

Several probes have been developed that enable enrichment and downstream analysis of genome wide 5fC-sites after chemical derivatization (cf. chapter 1.2.3.2). In one instance an oligonucleotide probe with a chemical handle for crosslinking of 5fC within the complementary sequence had been developed during the work of this thesis.[75] A particular challenge for detection methods are the low genomic levels of 5fC.

No TALE repeats for positive selective recognition of 5fC are available. Oxime formation of 5fC with O-tert-butylhydroxylamine enables detection of target DNA bearing 5fC in gDNA background via chemoselective blockage of TALE repeat_SG*GG (**Fig. 29b**). However, this approach does not allow enrichment of 5fC.

Based on the universal TALE repeat_SG*GG structure, novel TALE repeats were designed that bear the non-canonical amino acid *p*AcF (**1**) at aa position 11, 12 or 14 (**Fig. 30a**-**c**). The ketone functionality enables crosslinking of TALE proteins to user-defined 5fC-DNA sequences via oxime condensation with diaminooxy-linkers.

First, the original and mutated module vectors were assembled via Golden Gate reaction into a set of TALE proteins (**Fig. SI1a-c**) targeting a 26 nt sequence of a promotor sequence associated with the DNA repair protein BRCA1 (TCTTTCCTTTTACGTCATCCGGGGGC, human chr17: 43125551-43125576, BRCA1a) (**Fig. 30d**).

Size-reduced repeats bearing **1** were positioned opposite the target C of the single CpG dinucleotide (underlined) (**Fig. 30d**). TALE proteins (TALEa_**1**G*GG, TALEa_S**1***GG and TALEa_SG***1**G) were expressed from a pET-vector. *p*AcF was incorporated into these TALEs via an orthogonal *Methanocaldococcus jannaschii* (*M. jannaschii*) tyrosyl-tRNA synthetase/tRNA pair[365] in response to the amber codon in *E. coli* (**Fig. 30e**; **Fig. SI10; Fig. SI11; Fig. SI12a, b**).

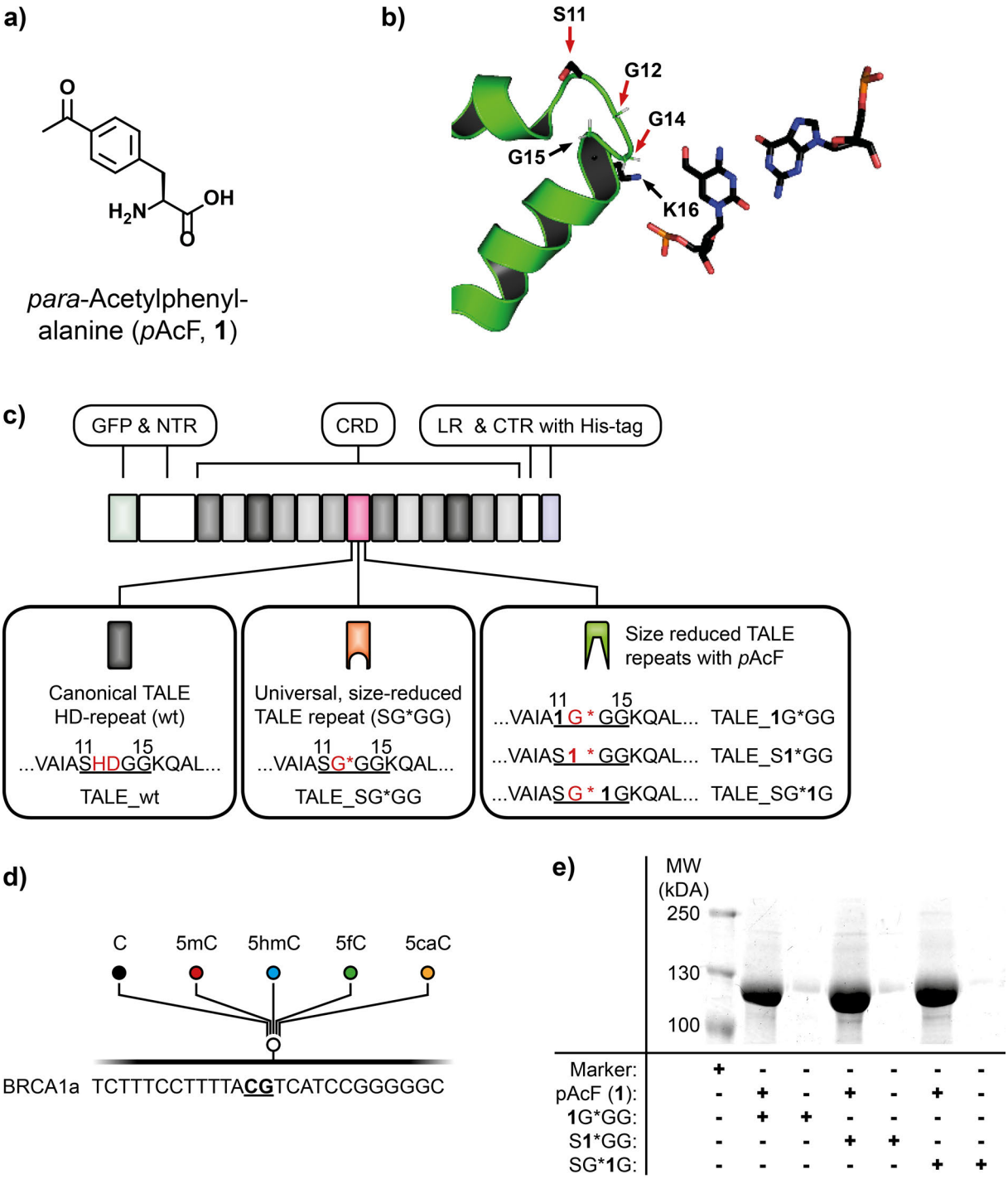


Figure 30: Engineering size-reduced TALE repeats bearing the non-canonical amino acid (ncAA) para-acetylphenylalanine (*p*AcF, 1). a) Chemical structure of **1**. **b)** Modelled TALE repeat_SG*GG with 5fC (design based on **Fig. 21c**). Positions for incorporation of **1** marked with red arrows. **c)** Design of TALEs for the **d)** BRCA1a target DNA with variable repeats opposite of the underlined CpG. The CpG contains one of the five cytosines as indicated. **e)** SDS-PAGE image of purified TALEa_**1**G*GG, TALEa_S**1***GG, TALEa_SG***1**G after expression either with or without pAcF (see **Fig. SI12a** for full gel).

Next, binding of the three TALE proteins to oligonucleotide duplexes containing the BRCA1a sequence with a single C, 5mC, 5hmC, 5fC or 5caC opposite the engineered repeat was evaluated in EMSA and compared to binding of TALEa_wt and TALEa_SG*GG (**Fig. 31**).

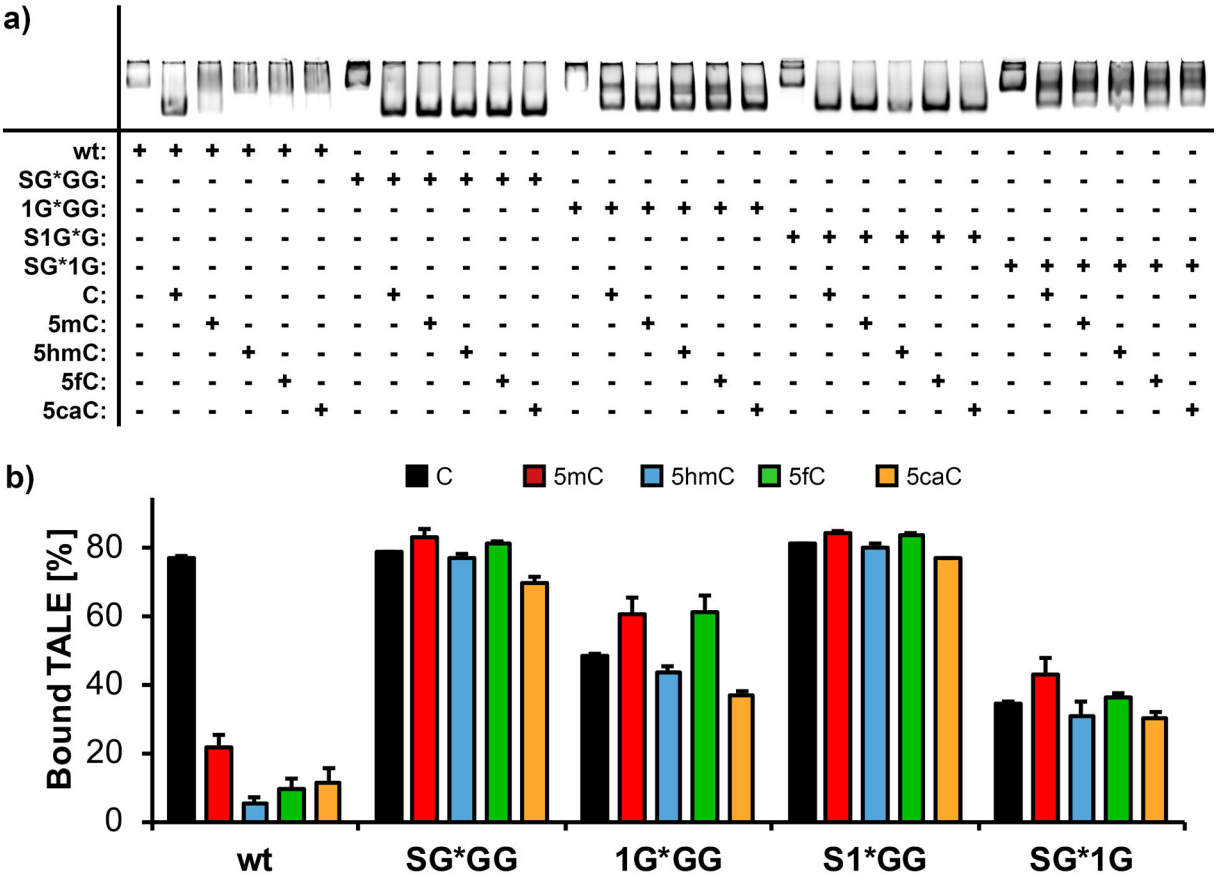


Figure 31: Interaction of TALEs with BRCA1a target DNA bearing one of the five cytosines opposite of the variable repeat as indicated. a) Native PAGE gel images after EMSA of the five TALEs with the different target DNA as indicated. **b)** Binding profiles of the five TALEs with the different target DNA as indicated (0.75 pmol TALE to 7.5 pmol DNA).

Expectedly, high binding of TALEa_wt to DNA was observed only to sequences with unmodified C, whereas TALE_SG*GG bound strongly to DNA with all five cytosine bases opposite the repeat_SG*GG.

For TALEs with **1**, binding of the three TALEs were observed to all five cytosines but compared to TALEa_SG*GG, the incorporation of the ncAA slightly reduced binding of TALEa_S**1***GG and TALEa_**1**G*GG, whereas binding of TALEa_SG***1**G was reduced by ~50 % compared to TALEa_SG*GG (**Fig. 31**). Reduced binding of TALEs with

repeat_**1**G*GG indicate a hindrance of TALE binding. This might be possible due to an unfavorable position of **1** towards the sterically challenging 5hmC and 5caC, as binding to 5mC and 5fC is comparable to TALEa_S**1***GG. In contrast, the overall reduced binding of TALEa_SG***1**G indicate a local impact on the repeat structure, rather than overall TALE folding or complex formation. Interestingly, the binding pattern of the TALE with repeat_S**1***GG is significantly higher than of TALEs with the chemically similar aa F or Y in repeat_SF*GG or repeat_SY*GG,[325] which suggests that **1** can either stabilize a more favorable loop structure for biding of 5-modified cytosines or the keto-group participates in TALE-DNA binding.

3.2.2. Programmable crosslinking of TALE repeats with *p*AcF and 5fC

Next, the three TALEs were reacted with a set of diaminooxy-linkers with one to six ethylene glycol moieties that had been labeled to 5fC-DNA (**Fig. 32a**, **L1**-**L5**). The linkers had been synthesized (credit to G. Kubik) according to a previously published protocol via a Mitsunobu reaction with N-hydroxyphthalimide and subsequent hydrazinolysis (**Fig. 32a**).[377] Both ESI-MS and NMR (1H and 13C) confirmed product formation (**Fig. SI13; Fig. SI14**).

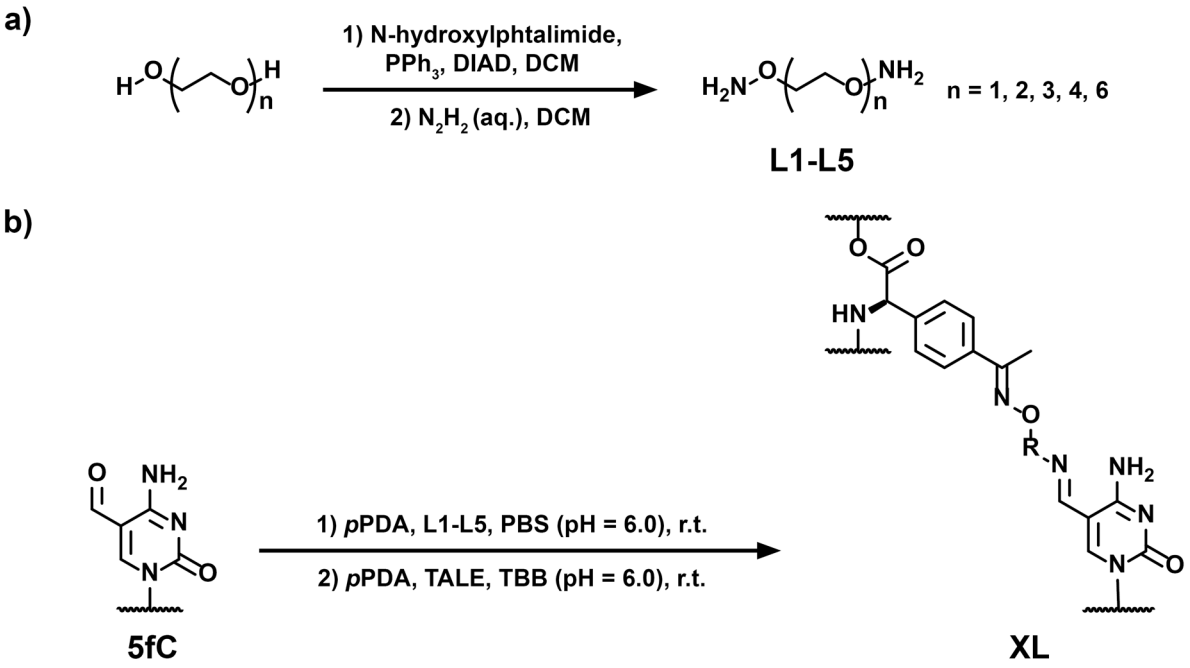


Figure 32: Crosslinking of 5fC with 1 via linker L1-L5 under mild acidic conditions. a) Synthesis scheme of linker **L1-L5** with different number (n) of ethylene glycol moieties (see methods for full protocol). **b)** Crosslinking of 5fC in a *p*PDA-catalyzed two-step reaction to TALE proteins bearing **1**. XL: Crosslink formation yields a 5fC-linker-**1** adduct.

For the crosslinking reaction, conditions for oxime formation compatible with DNA hybridization from chapter 3.1.2. were utilized. Interestingly, previous experiments under similar conditions were not able to provide TALE-DNA-crosslinks after TALE had been incubated with DNA (data not shown). In that instance, access of the reagents to 5fC-sites and **1** could have been restricted after formation of the TALE-DNA complex. Therefore, a different approach of first reacting the DNA with the linker, then reacting TALE with DNA was employed. Briefly, 100 pmol DNA in 20 mM phosphate buffer at

pH = 6.0 was reacted with 2 µmol *p*PDA and 10 µmol linker **L1**-**L5** (**Fig. 32b**). First, *p*PDA-catalyzed crosslink formation was analyzed by MS and gel electrophoresis (GE). MS data indicated high yields of 5fC-labeling with **L1-L5**, whereas DNA-DNA crosslinks of 5fC-containing 5'-FAM-labeled oligonucleotides could neither be detected with agarose GE nor with MALDI-TOF MS indicating that the linker excess under the reaction conditions led to quasi quantitative labeling of 5fC (**Fig. SI15; Fig. SI16**).

In the next step, 100 pmol BRCA1a TALEs with **1** at aa position 11, 12 or 14 in repeat 12 in 20 mM Tris-HCl buffer at pH = 6.0 was added to the DNA. Crosslink analysis via denaturing sodium dodecyl sulfate (SDS) PAGE revealed a series of new bands corresponding to increased weight compared to unreacted TALE proteins only for reactions with linker present (**Fig. 33a**). Interestingly, the position of **1** in the three TALEs apparently also influenced band formation and intensity for different linker.

Especially strong bands were visible for the longest linker **L4**, **L5** with 4 and 6 ethylene glycol moieties for all three TALEs (**Fig. 33a**). This suggests that during or after binding of TALE proteins the linker can be guided best to **1**. However, these molecules need to be also accommodated within the TALE-DNA complex. In this regard, the chain could be positioned in a "blind spot". This could be similar to linker of the Alexa Fluor 488, which position at a 5fC-site in target DNA did not affect TALE binding (**Fig. SI6**).

Notably, crosslinking of TALEa_SG*1G to 5fC-DNA via **L2**-**L5** resulted in comparable strong bands. Following the orientation of aa side chains, **1** would face towards a similar direction as the second position of the RVD due to deletion of aa position 13. This would position the keto-group of **1** towards the modified 5fC and could explain why the crosslinking appears to be the most efficient for TALEa_SG*__1__G. Accordingly, from the orientation in the repeat_SHDGG (**Fig. 12c**) and modelled repeat_SG*GG (**Fig. 30b**) the position of **1** at aa position 11 or 12 would likely move the keto-group away from the labeled nucleobase and could explain why longer linker are required for crosslinking of TALEa_**1**G*GG and TALEa_S**1***GG than for TALEa_SG*__1__G. Interestingly, no crosslinking could be observed for TALEs and **L1** (**Fig. 33a**). It is unlikely that this is a result of poor reactivity in the labeling reaction of 5fC with **L1**. It could be a result of **1** not being proximal enough to the labeled 5fC-site, or in case of TALEa_SG*__1__G the modified 5fC cannot be efficiently and stably accommodated at the repeat for oxime formation between the hydroxylamine and **1**. The consequence would be a dissociation of the TALE-DNA complex similar to the chemoselective blockage of **4d** or due to denaturation prior to the SDS-PAGE.

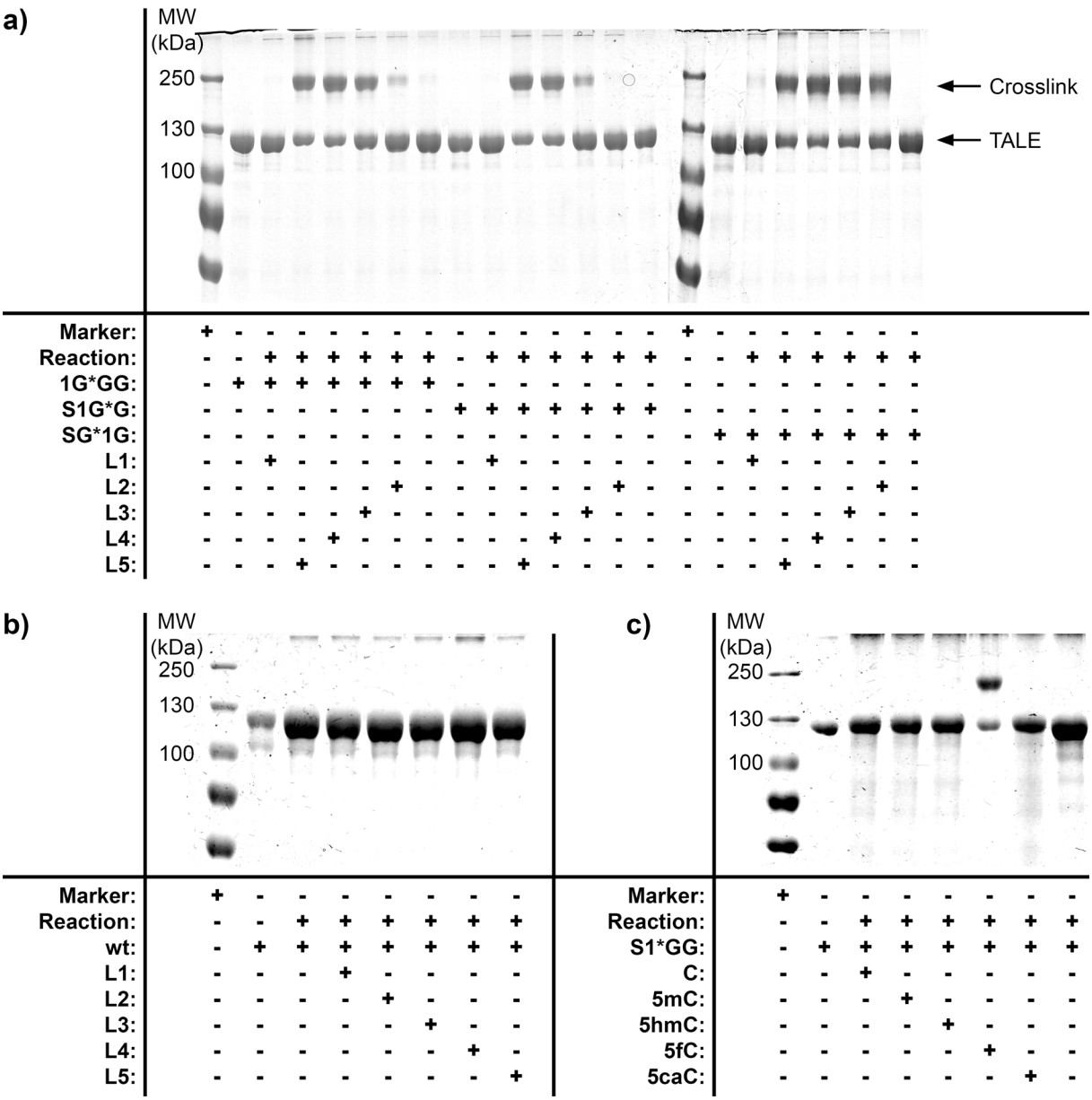


Figure 33: 5fC-specific DNA-TALE-crosslinking with TALE proteins bearing the ncAA *p*AcF. a) Denaturing SDS-PAGE image after crosslinking reaction with L1-L5, 5fC-DNA and TALEs as indicated. **b)** Denaturing SDS-PAGE image after crosslinking reaction with **L1-L5**, 5fC-DNA and TALEa_wt. **c)** Denaturing SDS-PAGE image after crosslinking reaction with **L4**, TALEa_S**1***GG and DNA bearing either C, 5mC, 5hmC, 5fC or 5caC opposite the repeat with *p*AcF.

As no or only background level bands could be observed for samples without linker or without DNA and linker, the band can be assigned to the formation of DNA-TALE-crosslinks. Additionally, no crosslinking with TALEa_**1**G*GG and linker **L4** could be observed for DNA containing C, 5mC, 5hmC or 5caC opposite the reactive repeat or of TALEa_wt with **L1**-**L5** and 5fC-DNA (**Fig. 33b**, **c**). This confirms that the TALE-DNA crosslink is dependent on the presence of 5fC and **1**. Furthermore, TALE-TALE-

crosslinking (expected bands at ~260 kDa) could not be confidently detected for the three TALEs when DNA was present in the sample and might only be observed at background levels for reactions without DNA where free linker was still present in the sample (**Fig. 34**). Remarkably, no detectable crosslinking occurred between TALEs and an oligonucleotide duplex bearing a 5fC in the CDKN2A off-target sequence from chapter 3.1. (**Fig. 34**).

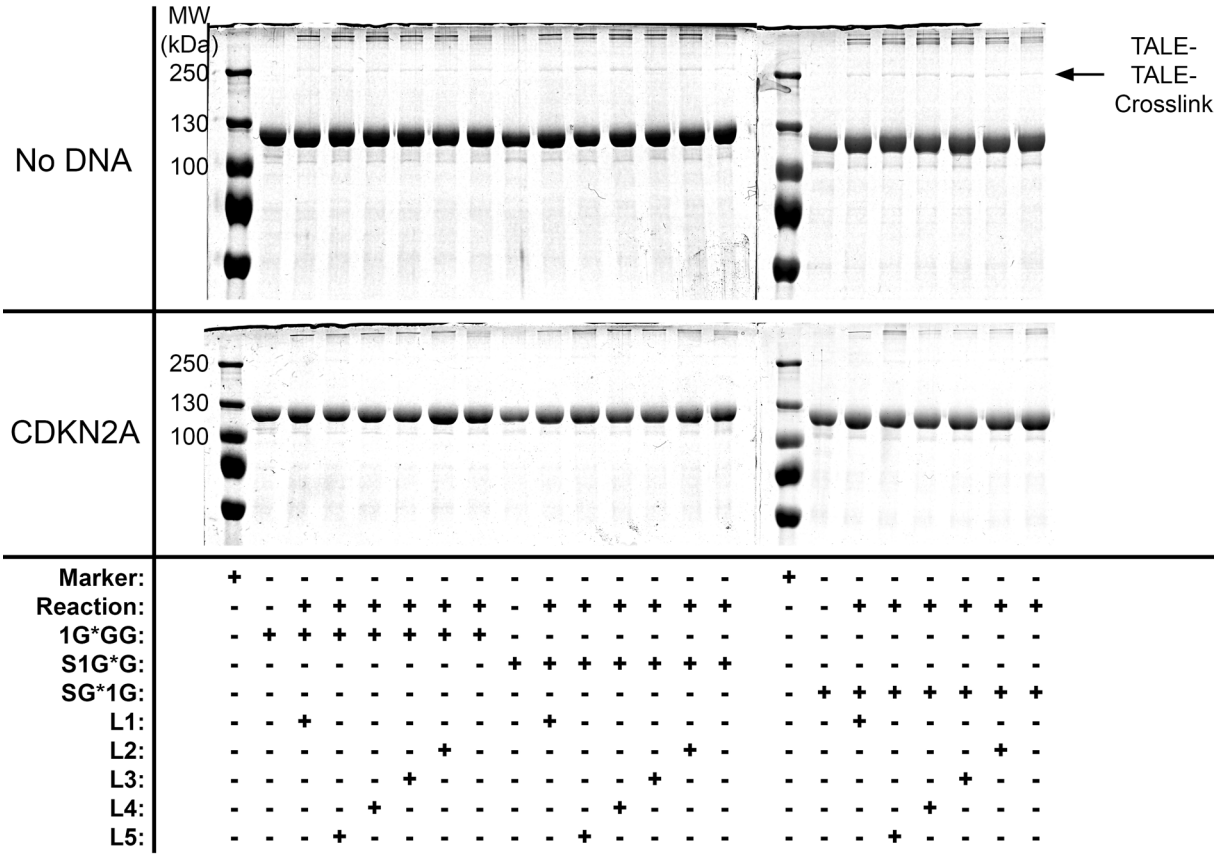


Figure 34: Test for protein-protein crosslinking of BRCA1b_TALEs and off-target crosslinking to CDKN2A-DNA with 5fC as indicated.

These results indicate that prior to the denaturation for SDS-PAGE the TALE was still able to bind its target DNA sequence despite the bulkier 5fC-linker adduct and acidic pH value. Indeed, TALE binding under reaction conditions still resulted in high percentage band shifts in EMSA (**Fig. SI17**).

Most importantly, together these results demonstrate that the bands correspond to 5fC-specific, programmable and *p*AcF-dependent TALE-DNA crosslinking without significant cross reactivity to other sequences. Furthermore, according to the several tested hydroxylamines in chapter 3.1.2, steric clash would not be expected between

the molecular structure of the linear linker and the size-reduced repeat. In the end, structural analysis of the covalently connected TALE-DNA complexes would be required to explain the different crosslinking profiles of the TALEs.

To quantify the crosslinking yields for the different TALE-linker combinations, the previous experiment was repeated with ^{32}P-labeled oligonucleotides.

Denaturing SDS-PAGE revealed that crosslinking yields were overall slightly lower for TALEa_S**1***GG except for **L5**, while TALEa_**1**G*GG and TALEa_SG***1**G displayed similar crosslinking efficiency (**Fig. 35**). Expectedly, no adduct formation was visible for samples without linker or without DNA (**Fig. 35; Fig. SI18**).

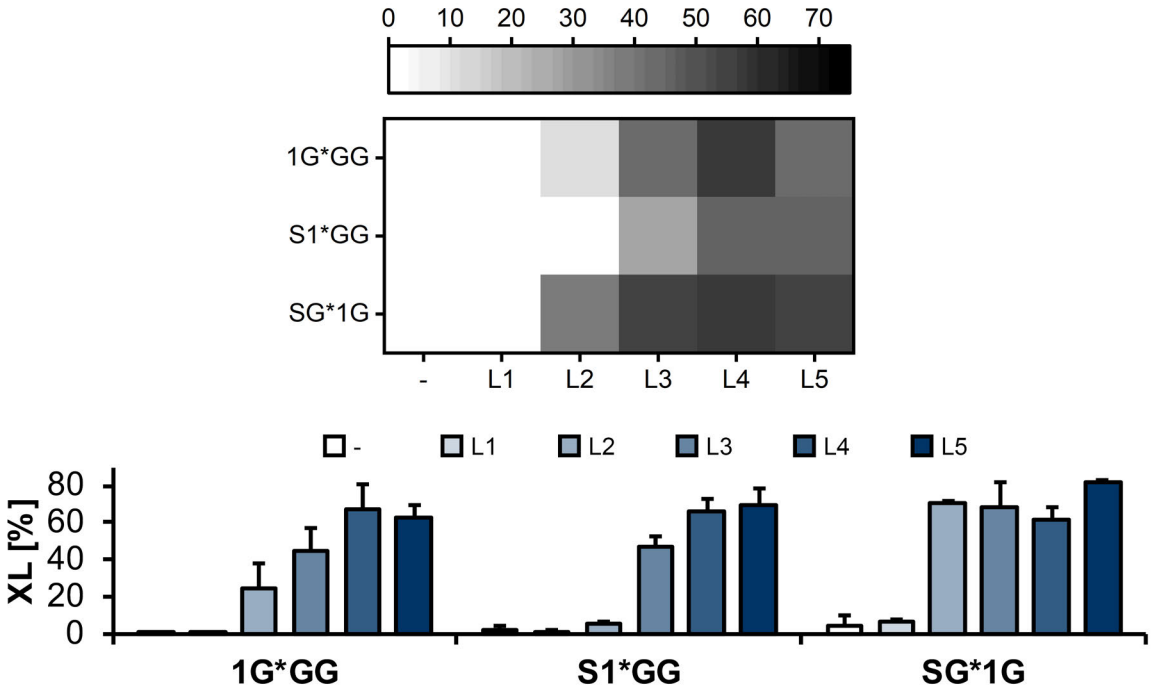


Figure 35: Quantification of crosslinking reactions with indicated TALE-linker combinations by SDS-PAGE analysis using ^{32}P-labeled DNA. XL: Crosslink formation

Readouts revealed a low amount (~7 %) crosslinking between **L1** and TALEa_SG***1**G whereas no crosslinking with **L1** could be observed for TALE_**1**G*GG and TALEa_S**1***GG (**Fig. 35**). Both TALEs required a linker with a minimal length of two ethylene glycol moieties (**L2**) (**Fig. 35**). From the heat maps it can be concluded that greater linker length yields higher crosslinking. While almost 40 % crosslinking could be observed for TALEa_SG***1**G and **L2**, maximal crosslinking for the three different TALEs was observed for **L3**-**L5** (**Fig. 35**). This supports the previous assumptions regarding positioning of **1** within the TALE repeat, linker length and flexibility.

To assess potential off-target crosslinking to 5fC-sites within the target DNA, TALEs targeting a triple-CpG sequence in the BRCA1 sequence context (BRCA1b, **Fig. 36a**) were constructed and expressed analog to BRCA1a_TALEs (**Fig. SI11b; Fig. SI12c**). Instead of the canonical C-selective repeat_SHDGG, universal repeat_SG*GG was positioned opposite the two Cs of the neighboring CpG-sites (TALEb_**1**G*GG, TALEb_S**1***GG, TALEb_SG***1**G) (**Fig. 36a**). TALE proteins were functional and displayed good and uniform binding to target DNA with 5fC at different positions (5fCpG1, 5fCpG2 or 5fCpG3) of the target strand or complementary strand (rev_5fCpG2) according to EMSA (**Fig. 36b**). The affinity of TALEb_1G*GG and TALEb_S1*GG with target DNA was comparable. Again, only TALEb_SG*1G bound weaker to the DNA (**Fig. 36b**). This result supports the hypothesis that the position of **1** in TALEa_SG***1**G and TALEb_SG***1**G prevents effective binding to DNA. Moreover, these results are in good accordance to the EMSA of TALEs for the BRCA1a target DNA (**Fig. 31**).

Additionally, no significant decrease in binding could be observed for incorporation of repeat_SG*GG modules. This indicates that incorporation of multiple size-reduced repeats with RVD G* does not significantly reduce TALE binding to the target DNA similar to TALE repeats with RVD N*.[322]

SDS-PAGE analysis of crosslinking experiments with DNAs bearing a single 5fC in any of the two off-target CpGs on the target strand revealed a certain degree of positional off-target crosslinking for **L3**-**L5** in all cases and for TALEb_SG***1**G also with **L2** (**Fig. 37**). This shows that more flexible and longer linkers have a higher positional off-target effect and that there is sufficient chemical space in the TALE-DNA complex for the linker to be guided to the 5-modification at other positions. For the on-target position, the crosslinking profile was similar to the BRCA1a-TALEs.

Interestingly, **L1** did not yield crosslinks to 5fCpG1 for all three TALEs but a strong band was only visible for TALEb_SG***1**G and 5fCpG3. In fact, strong bands can be observed for this TALE and 5fC-position with all five linkers.

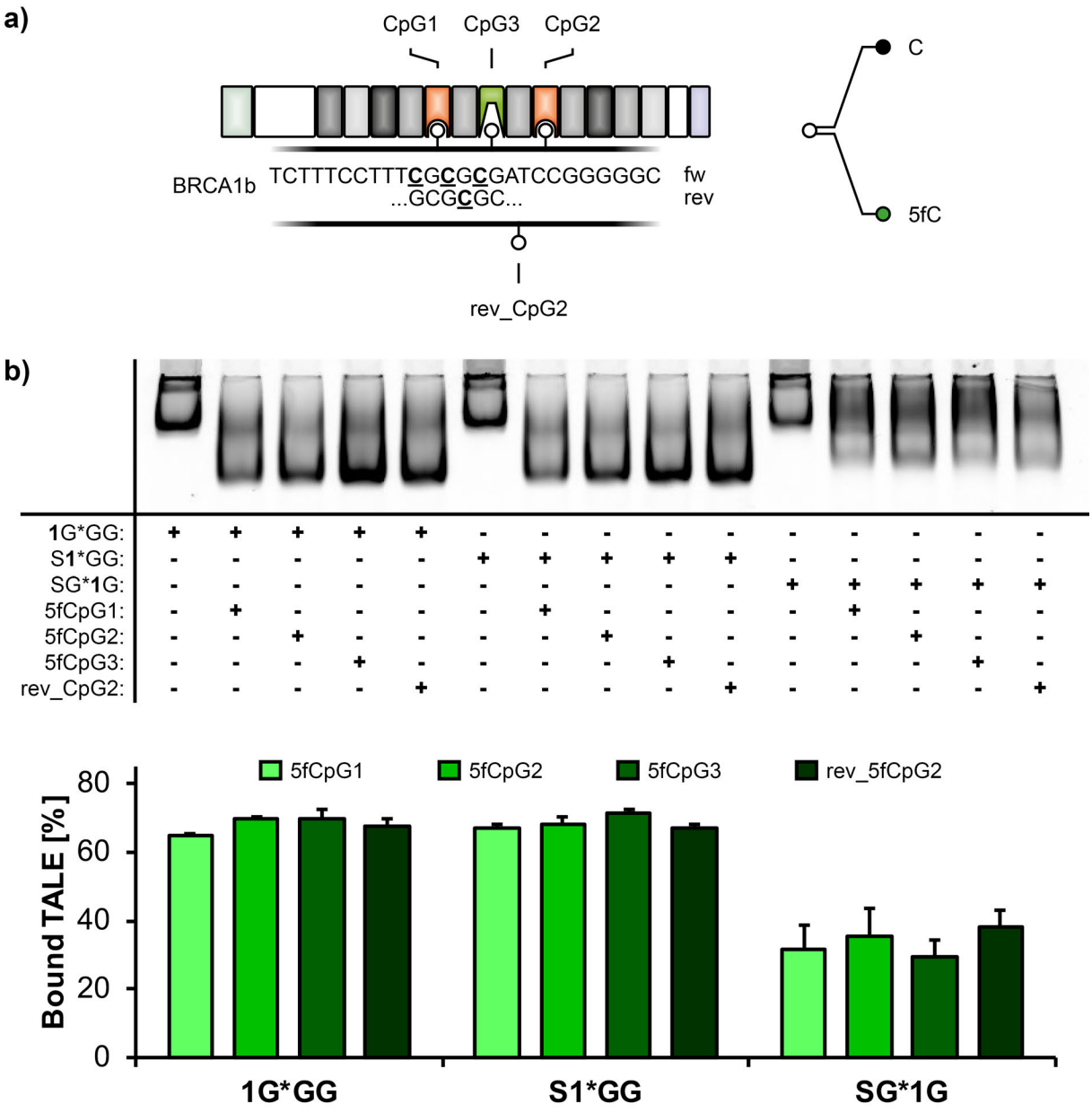


Figure 36: TALEs for positional resolution analysis of 5fC-crosslinking. a) TALE model showing the interaction of TALEs with repeats bearing **1** flanked by two universal repeat_SG*GG modules. BRCA1b target DNA bear a single 5fC at one of the indicated CpGs. b) EMSA for testing interaction of 0.75 pmol TALEs with 7.5 pmol BRCA1b target DNA as indicated.

Another highly resolved position is the combination of **L2** and TALEb_S**1***GG for rev_5fCpG2, the 5fC opposite of the on-target position but on the complementary strand.

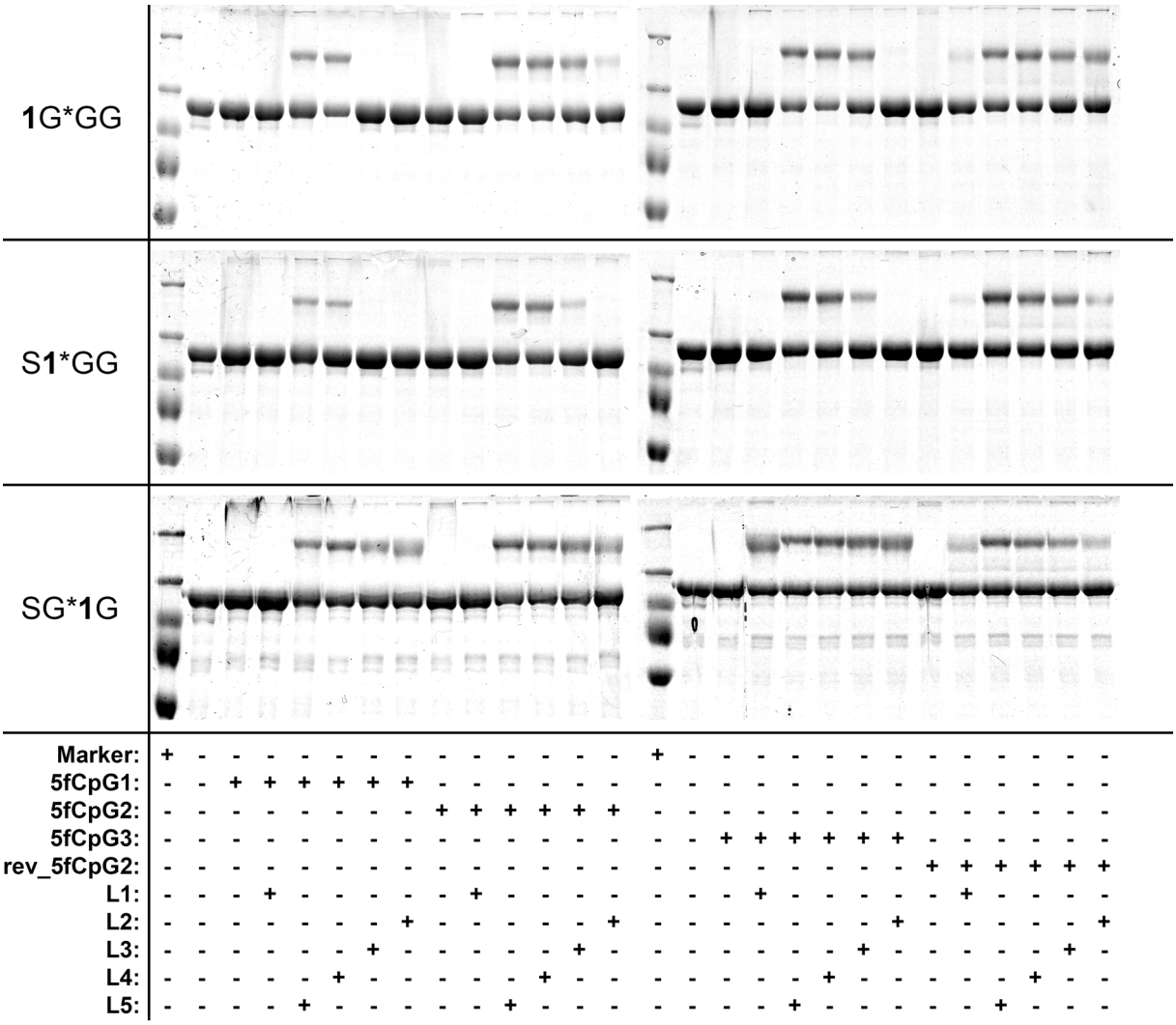


Figure 37: SDS-PAGE images of positional resolution of TALE-5fC-DNA crosslinking for BRCA1b.

Again, the experiment was repeated with radiolabeled DNA to quantify the off-target crosslinking (**Fig. 38, Fig. SI19**). On- and off-target crosslinking for linker L4 and L5 was equally high (between 40-70 %) for all positions and TALEs. This indicates that the free hydroxylamine group can be guided more freely from the DNA to different repeat positions through the TALE-DNA-complex. Selectivity for single CpG-sites could be observed for the target 5fCpG2 with the combination of TALEb_**1**G*GG and **L2** but with a yield of only ~20 % (**Fig. 38a, e**). **L3** was somewhat specific for CpG3 in combination with TALEb_S**1***GG (**Fig. 38b, f**). Furthermore, the 5fCpG3-selective combination of TALEb_SG***1**G and **L1** produced crosslinks with excellent yield of ~70 % (**Fig. 38c, g**). As **L1** is the shortest of the linkers and does not make good crosslinks with 5fCpG2 this raises questions about how **1** or the TALE is positioned in

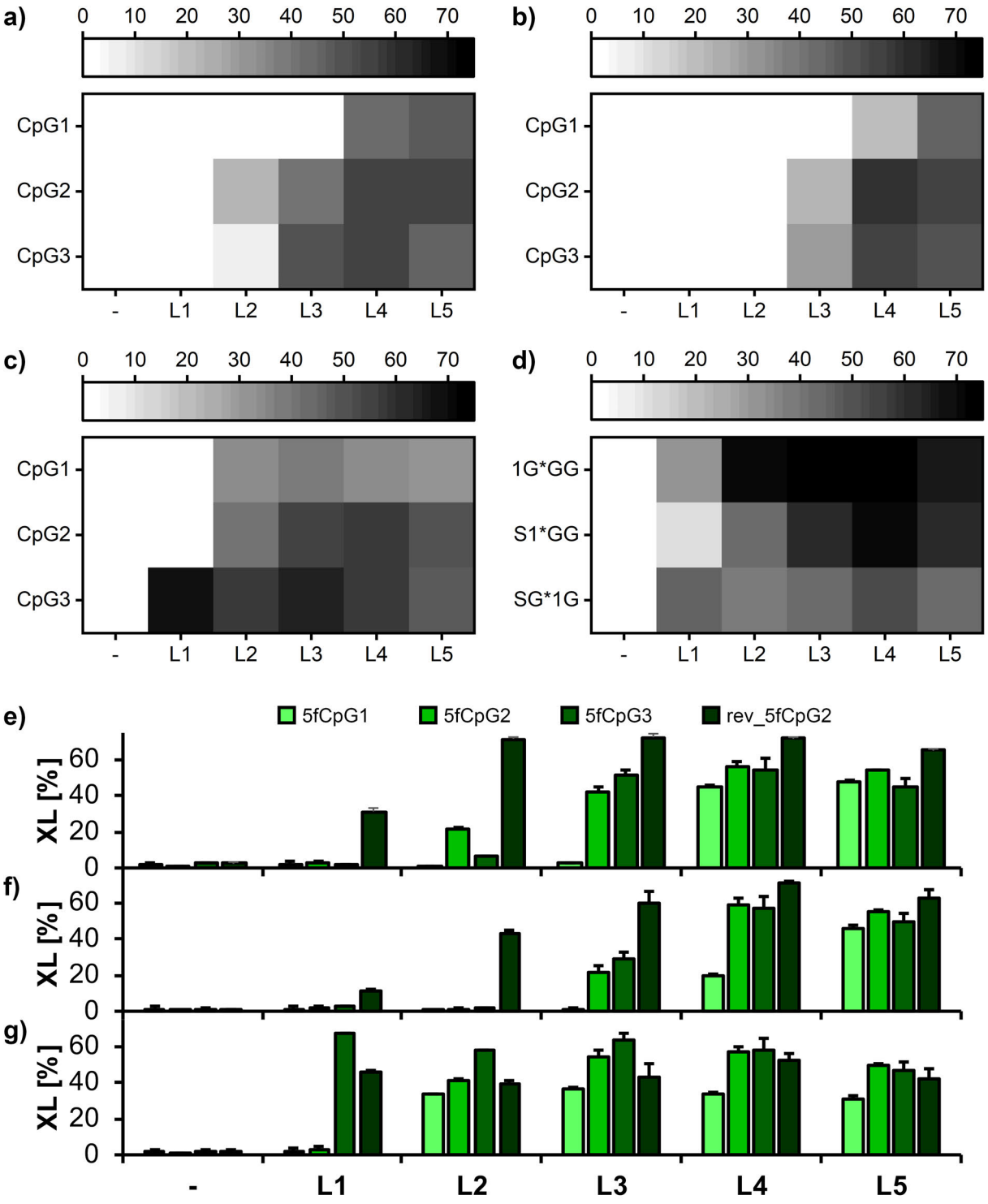


Figure 38: Quantification of positional resolution of TALE-5fC-DNA crosslinking for BRCA1b using ^{32}P-labeled DNA. Crosslinking profiles for **a), e)** TALEb_**1**G*GG**; b), f)** TALEb_S**1***GG**; c), g)** TALEb_SG***1**G**. d)** Crosslinking profile for the three TALEs to 5fC opposite the variable repeat as indicated on the complementary DNA strand.

these cases and how changes of the TALE and DNA structure play into this observation. For the 5fCpG on the complementary strand opposite the central CpG, considerable to high off-target crosslinking could be observed for all three TALEs in combination with the **L1**-**L5** (**Fig. 38d**). This gives more insight, how **1** could be positioned in the TALE repeat. Potentially, instead of facing away from 5fC it could rather extend past the nucleobase accommodated in the repeat and connect to the 5fC on the other strand.

Considering all 4 possible combinations, TALEb_**1**G*GG and TALEb_S**1***GG with **L2** yielded selective crosslinks with good yields of above 60 % and 40 %, respectively. Crosslinking of **1** to rev_5fCpG2 with **L1** could be observed for TALE_**1**G*GG and TALE_SG***1**G. These results indicate that there are several combinations of linker and TALE proteins, which enable resolved, position-specific crosslinking to 5fC-sites.

They also show that limitations to the targeting of TALE-linker combinations for CpG-rich target DNA exist. Currently it is unclear if the reactivity and selectivity profiles differ when multiple vicinal 5fC are present. In general, it needs to be considered as well that short synthetic oligonucleotides might react differently than natural, complex chromosomal DNA. Therefore, the observed effects, selectivities and crosslinking efficiencies might not apply to genomic DNA.

Another approach for programmable 5fC-selective crosslinking could be the use of TALE proteins bearing *p*AcF outside the CRD in the LR (**Fig. 39a**). Therefore, the canonical TALE structure could be completely conserved. Additionally, due to the more flexible and less constraint LR and C-terminal structure this could provide a scaffold with novel positional resolution for targeting a 5fCpG in the 3'-region of a BRCA1 target DNA (BRCA1c). TALEs consisting of 20 or 21 modules including the last repeat with **1** at aa position 12 were assembled and expressed analog to BRCA1a_TALEs (**Fig. SI11c; Fig. SI12d**). Like for the other TALEs in this study, EMSA revealed a strong, specific shift corresponding to TALE-DNA interaction only for the target site (**Fig. 39b**).

SDS-PAGE analysis showed that **L2**-**L5** was able to mediate crosslinks of LR_S**1***GG opposite of the target nucleobase with 5fC (LR_S**1***GG(20), **Fig. 40**). Interestingly, crosslinking of the five linker **L1-L5** with 5fC could be observed with the LR_S**1***GG targeting the nucleobase preceding 5fC (LR_S**1***GG(19), **Fig. 40**).

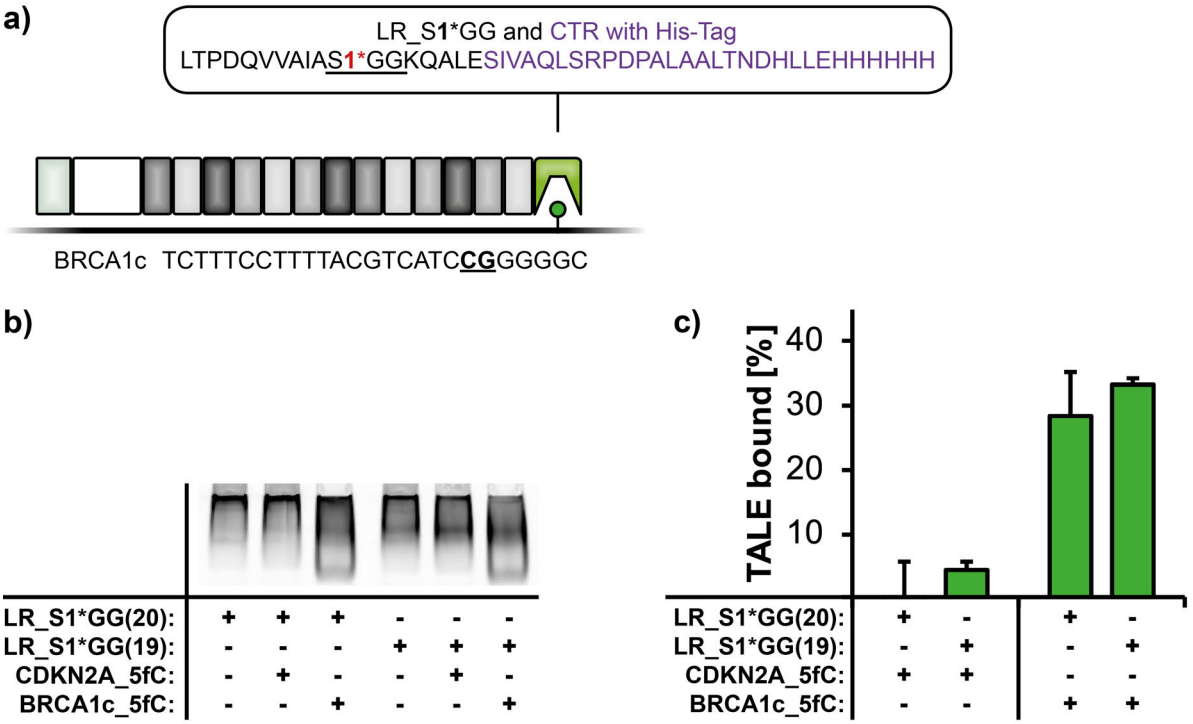


Figure 39: Programmable TALE-DNA crosslinking with engineered last repeats. a) Schematic TALE with 18 or 19 RVDs and last repeats with **1** at aa position 12 targeting base 19 (LR_S**1***GG(19)) or 20 (LR_S**1***GG(20)), respectively. The 5fC is opposite of position 20. **b)** Native PAGE image after EMSA of the 1.5 pmol TALEs with 7.5 pmol off-target CDKN2A DNA or 7.5 pmol target BRCA1c as indicated. **c)** Binding profiles of TALE-DNA interaction from b). Controls without DNA were subtracted from both samples.

Again, no crosslinking could be observed with off-target 5fC-DNA. This indicates that even though polarity effects of TALE-DNA interactions have been reported,[295] the binding is still specific between these sequences. However, bands corresponding to TALE-TALE (~220 kDa) crosslinking occurred in the absence of DNA (**Fig. 40**).

The band formation gives an indication that there is a positional resolution to these TALEs. Quantifying the crosslink formation of TALEs with LR_S**1***GG(19) and LR_S**1***GG(20) with ^{32}P-labeled DNA confirmed the high crosslinking yield with **L3**-**L5** for both TALEs comparable to the TALE_S**1***GG for BRCA1a and BRCA1b target DNA (**Fig. 41; Fig. SI18**). Same goes for the combination of **L2** and TALE_LR_S**1***GG(20), while the combination of **L2** and TALE_LR_S**1***GG(19) shows approximately 60 % crosslinking.

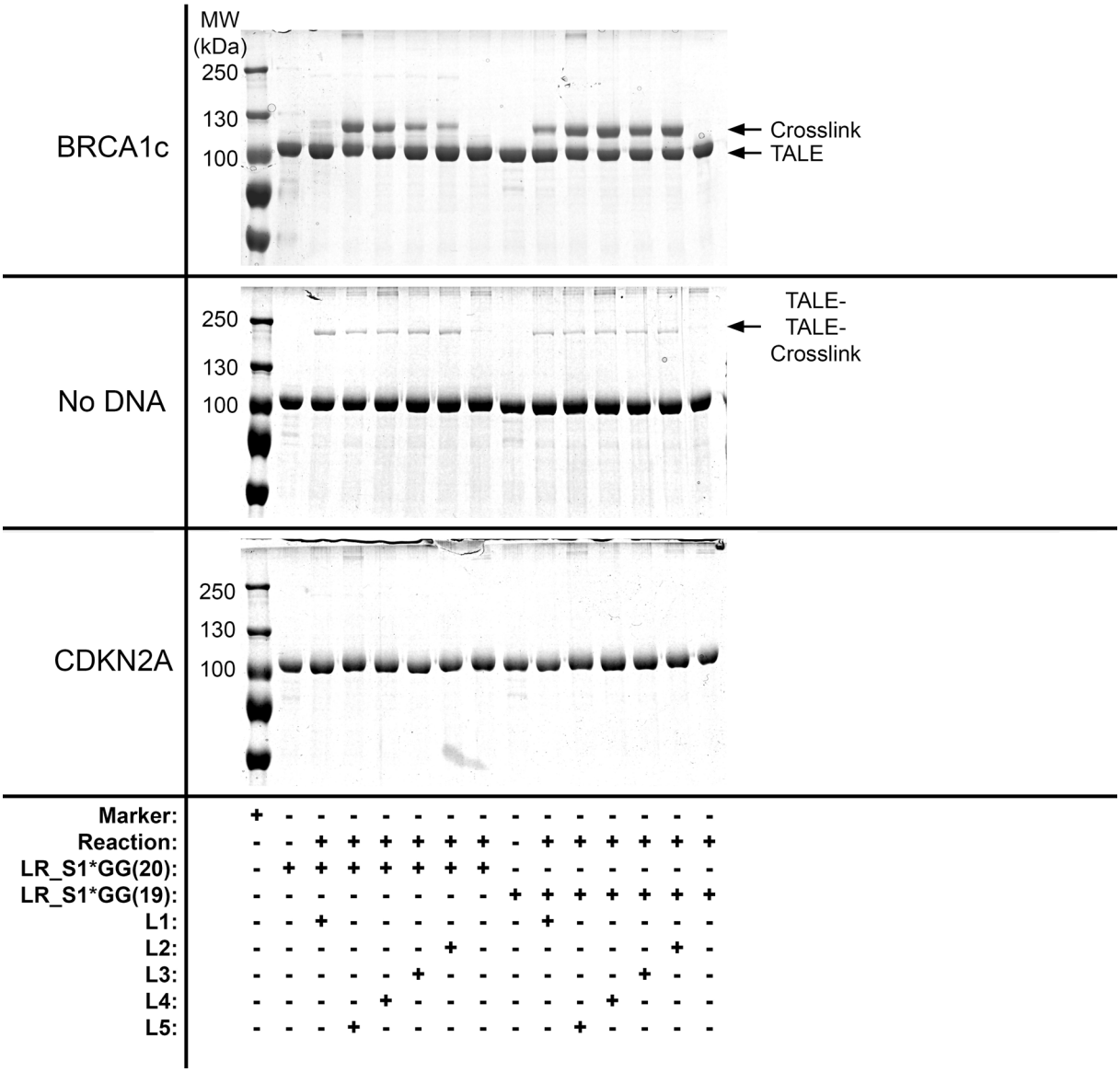


Figure 40: Test for protein-protein crosslinking of BRCA1c_TALEs and off-target crosslinking to CDKN2A-DNA with 5fC as indicated.

These results show that it is possible to design selective programmable probes for efficient crosslinking of target 5fC-DNA at different positions of the CRD. It is possible that the ncAA in the C-terminal domain is more accessible than in the CRD. If so, this could enable the crosslinking of 5fC-DNA with the TALE after formation of a non-covalent TALE-DNA complex. This would allow for programmable TALE-DNA binding *in vitro* and *in vivo*, subsequent crosslinking and enrichment of 5fC-sites. However, it remains to be shown if the use of these constructs can match the selectivity and specificity of constructs allowing accommodation of labeled 5fC for crosslinking within the TALE structure.

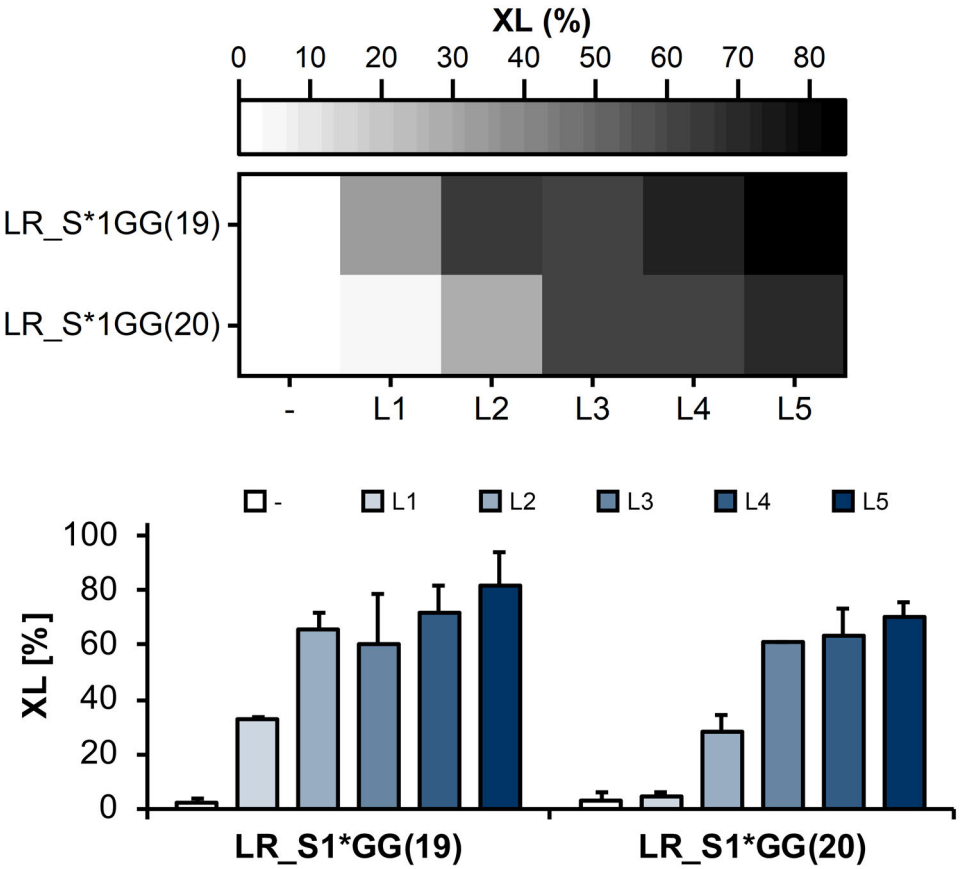


Figure 41: Quantification of crosslinking reaction yields with indicated TALE-linker combinations by SDS-PAGE analysis using ^{32}P-labeled DNA.

3.2.3. Detection and quantification of single 5fC-sites in genomic DNA

The next step was the detection and quantification of 5fC levels at single CpG in the background of mammalian genomic DNA. Considering both EMSA and SDS-PAGE analysis, TALEa_**1**G*GG was selected in combination with **L4** due to good 5fC-binding and high yields in crosslinking reactions. Additionally, **L2** was tested due to his selectivity in the context of the neighboring 5fCpG on the target DNA strand.

The enrichment assay based on magnetic bead-immobilized TALEs established in chapter 3.1.3 was slightly modified with more stringent washing to separate non-covalent bound target DNA from the crosslinked 5fC sequences (**Fig. 42a**).

This was necessary because EMSA data had shown that TALEa_**1**G*GG has high universal affinity to DNA targets with cytosine 5-modifications. Different denaturing methods for example with SDS or other ionic detergents, heat or chemical additives were tried, but in the end a buffer containing 2 M NaCl proved best compatible with the enrichment process and most effective to disrupt non-covalent or off-target TALE-DNA binding.[295] For enrichments, PCR targets with precisely defined 5fC levels at single positions were spiked into 5fC-free, fragmented and digested human gDNA (Encode NA18507) at a concentration of a single genome copy (*cf.* chapter 3.1.3 for details, **Fig. SI20**). Similarly, 5hmC as a structurally comparable off-target nucleobase was introduced at the BRCA1a target position via synthetic primers (**Fig. SI20**).

For enrichments, 1 µg DNA with 10 µmol linker **L2** or **L4** and 2 µmol *p*PDA was reacted as previously described and purified. Then, samples were incubated with 10 pmol TALEa_**1**G*GG and fresh *p*PDA. After 6 h the reaction was quenched with 18.5 µl Tris buffer (pH = 9.0), purified it via gel filtration and incubated it with Ni-NTA magnetic beads in enrichment buffer. The beads were washed with buffer containing high amount of NaCl to disrupt non-covalent or off-target TALE-DNA binding.[295] After another washing step with enrichment buffer, the beads were suspended in water and incubated at 95°C for 15 minutes. The beads were collected, and the supernatant was used for qPCRs, targeting a sequence adjacent to the BRCA1 TALE binding site.

For both crosslinking reactions, 5fC-DNA was selectively enriched from human genomic background DNA in comparison to 5hmC-DNA (**Fig. 42b**). According to the previous experiments, higher crosslink efficiency could be expected for reaction with **L4.** Indeed, crosslinking with **L4** led to approximately 3x fold higher enrichment of 5fC-DNA compared to reaction with **L2** (**Fig. 42b,**).

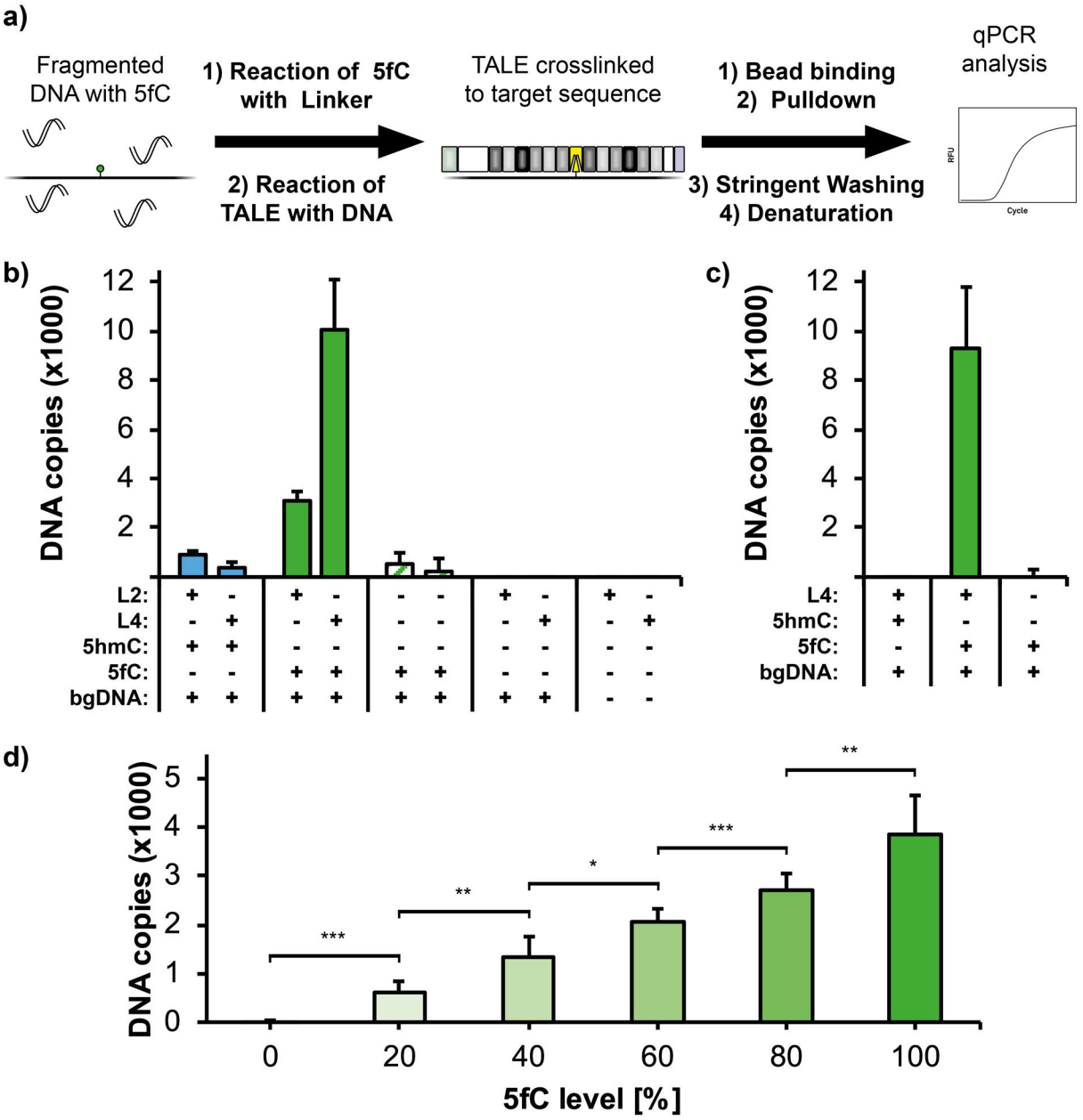


Figure 42: Covalent enrichment of 5fC-containing DNA sequences from mammalian genomic DNA backgrounds. a) Assay workflow. **b)** qPCR quantification of enrichments with TALE_S**1***GG and DNA with 100 % 5fC or 5hmC levels in the BRCA1 target CpG, including basic controls as indicated. A background control without TALE was subtracted from other enrichments. Enrichments were conducted in triplicates, each quantified with triplicate qPCRs. **c)** qPCR quantification of enrichments with TALE_S**1***GG performed as in b), but with background of mESC gDNA. A background control without linker was subtracted from other samples **d)** Enrichments as in b) with different 5fC levels. * = $p < 0.005$. ** = $p < 0.0005$. *** = $p < 0.00005$ from student-t-test. R2 from linear regression = 0.99.

Enriched copy numbers of 5hmC-containing DNA and 5fC-target in absence of linker were negligible in the experiments, confirming covalent enrichment as a robust and reliable method for 5fC detection in gDNA (**Fig. 42b**).

An experiment with non-amplified mESC gDNA extracted from J1 cells containing background off-target 5fC did not decrease enrichment of target sites from background gDNA after crosslinking with **L4** (**Fig. 42c**). This confirms the selectivity of user-defined TALE-DNA crosslinking from the previous experiments and demonstrate, that no significant amount of the enriched 5fC-target DNA is lost due to off-target 5fC.

Regarding the sensitivity and potential for quantification of 5fC in target DNA, the experiment was repeated with different 5fC levels obtained from a mixture of 5fC- and 5hmC-target DNA. Copy numbers of enriched 5fC-DNA for single levels differed significantly from each other (**Fig. 42d**). Moreover, an increase in 5fC-levels correlated proportionally with enriched copy numbers (**Fig. 42d**).

Taken together, this shows that the approach allows selective quantification of 5fC levels at a single, user-defined CpG in a human genome background by TALE crosslinking.

4. Summary and Outlook

Utilizing TALE proteins as programmable probes enabled sequence-specific targeting of single nucleobase positions. However, previous efforts to engineer new TALE selectivities did either not provide full selectivity for the target epigenetic 5-modified cytosine nucleobases or displayed decreased overall affinity. In this work, TALE-based approaches utilizing cytosine 5-modification-selective chemical conjugation reactions enabled complete, programmable decoding of ox5mCs in DNA and the direct capture and quantification of 5fC.

In the first project, size-reduced TALE repeats were designed. TALE repeat_SG*GG was found to universally bind to all eight human DNA nucleobases. Chemoselective modifications of 5hmC, 5fC and 5caC with substituents with distinct chemical properties via enzymatic glucosylation, oxime condensation or amide formation, respectively, was found to effectively block TALE binding and induce full selectivity for the modified nucleobase in the universal repeat_SG*GG. This enabled complete, programmable decoding of each ox5mC in target DNA at single user-defined positions with oligonucleotides as well as with a complex human gDNA background.

Additionally, PyAOP-catalyzed amide formation provides a new conjugation strategy for 5caC. This could be of use for other chemical modification-based approaches.

In the next project, the direct, programmable targeting of TALEs was advanced to covalent capture of user-defined target 5fC-DNA. First, size-reduced TALE repeats bearing the ncAA *p*AcF were engineered. These repeats were still able to bind to the five human cytosine nucleobases. Second, chemoselective crosslinking of repeats bearing *p*AcF to 5fC with different hydroxylamine-functionalized linker through oxime condensation was demonstrated. Positional resolution analysis revealed TALE repeat-linker combinations with single CpG resolution for target DNA-stands. However positional off-target crosslinking to the complementary DNA-strand was observed for all cases. Notably, the combination TALEb_**1**G*GG and **L2** was selective for the off-target 5fC on the complementary DNA strand. Third, covalent enrichment of target 5fC-DNA withTALEa_**1**G*GG and **L2** and **L4** demonstrate that the method can be used for robust and sensitive detection and quantification of 5fC-levels in gDNA.

For future applications, the design of TALEs for different target DNA and comparison with data from other highly sensitive and established approaches will help to develop both chemoselective blockage or crosslinking for the detection of naturally occurring

epigenetic nucleobases. Additionally, programmable covalent enrichment of DNA could enable to detect associated proteins.

However, *p*PDA-catalyzed oxime formation is incompatible with pH-sensitive proteins and live cells. Schiff base formation between 5fC and lysine residues of protein has been reported *in vitro* and *in vivo.* Encoding lysine or photocaged derivatives might be an approach for a biocompatible 5fC-selective crosslinking in live cells that could be spatially and temporally controlled.

Additionally, TALE repeats bearing amino acids with distinct chemical handles could extend covalent enrichment to the other ox5mCs.

Alternatively, the possibility to encode ncAA at the RVD of TALE repeats could add a new layer of functionality to TALEs as DNA-binding scaffolds. The introduction of residues with distinct chemical properties could provide new selectivity profiles of TALE repeats for epigenetic 5-modified cytosines.

Taken together, the two methods extend the application scope of TALEs as versatile probes for the programmable targeting of nucleobases beyond A, G, T and C. In comparison with non-programmable DNA-binding probes and tagging strategies alone, the use of TALEs in combination with chemoselective modification provides a user-defined sequence specificity and nucleotide resolution which overcomes previous limits of TALE-based approaches for the selective detection and analysis of the epigenetic cytosine nucleobases 5hmC, 5fC and 5caC in DNA *in vitro*.

5. Materials and Methods

5.1. Materials

5.1.1. Table 1: List of Services

Service	Product	Company
Synthesis	Oligonucleotides	Sigma Aldrich
Synthesis	Oligonucleotides bearing 5mC or ox5mCs	Metabion
Analysis	ESI-TOF mass spectrometry	Metabion
Analysis	Sanger Sequencing	GATC (now Eurofins Genomics)
Cell line	HEK293T	ATCC
Cell line	J1 mESC	ATCC
Plasmids	TALE toolbox	Addgene
Human genomic DNA	Male Yoruban individual, Encode sample NA18507	Coriell Institute

5.1.2. Table 2: List of Software and Tools

Name	Purpose	Developer/Distributor
BioDoc Analyze 2.1	Imaging of Agarose gels	Jena Biosciences
CFX Manager	Analysis and presentation of data from qPCR	Bio-Rad
Chemdraw 18.0	Chemical structure and reaction scheme design	PerkinElmer
Citavi 6	Citation	Swiss Academic Software
Excel	Data analysis and presentation	Microsoft
ExPASy	Bioinformatic tools (UniProt KB, BLAST, Compute pI/MW)	Swiss Institute of Bioinformatics
flexAnalysis 3.0	Analysis and presentation of data from mass spectrometry	Bruker
ImageJ	Picture editing and data analysis	National Institutes of Health
ImageQuant TL ver. 8.1.0.0	Analysis of data from EMSA and phosphor screens	GE-Healthcare
InDesign CC 2017	Figure design and formatting	Adobe
mMass	Analysis and presentation of data from mass spectrometry	Martin Strohalm
NanoDrop2000 Software 1.6.198	Data analysis of DNA concentration measurements	Thermo Fisher Scientific
OligoCalc	Oligonucleotide Properties Calculator	Northwestern University
Origin 2018	Data analysis and presentation	OriginLab
Pymol	Protein structure design	Schroedinger
Quantity One	Analysis of data from gel images and phosphor screens	Bio-Rad
Research MestReNova Lite CDE	Analysis and presentation of data from NMR spectrometry	Mestrelab

Spectra Manager Suite	Analysis of CD spectroscopy	Jasco
SnapGene	Analysis of data from Sanger sequencing, design of plasmid maps	GSL Biotech
Word	Data presentation and writing of manuscript	Microsoft

5.1.3. Table 3: List of Lab Equipment

Modell	Purpose	Manufacturer/Distributor
102-C	Converter for Digital Sonifiers	Branson
300TH	Ultrasonic cleaner	vwr
5810R/5424/5415R/5427R	Centrifuges	Eppendorf
Avance III 400	NMR spectrometry	Bruker
AX224	Analytical scale	Satorius
BAS-Cassette 2025	Phosphor screen cassettes	FUJIFILM Life Science
BAS-IP MS 2025 E	Storage phosphor screen	GE Healthcare
BIOER Thermocell	Cooling and heating block	biozym
Biophotometer Plus	Cuvette photometer	eppendorf
CanoScan 9000F	Scanner	Canon
CB 161	Magnetic stirrer	Stuart
CFX384 Touch Real-Time PCR Detection System	qPCR	Bio-Rad
Concentrator plus	Centrifugal vacuum concentrator	Eppendorf
Digital Sonifier 450	Power and control unit for sonification	Branson
Easypet	Pipettor	eppendorf
ED8525.3J	Microwave	exquisit
Eporator	Electric transformation	eppendorf
EV233	Electrophoresis power supply	Consort
FiveEasy Standard	pH meter	Mettler Toledo
GFX 460	Contact thermometer	ebro
Ground steel MTP 384 target	Plate for spotting MALDI-TOF samples	Bruker
I26	Incubator shaker	New Brunswick Scientific
Infinite M1000	Plate reader	Tecan

J-715	CD spectropolarimeter	Jasco
JB Aqua 12 PW	Water bath	Grant
kuroGEL MiniPlus 10	Agarose gel electrophoresis system	vwr
Laborota 4000-efficient	Rotary evaporator	Heidolph
LA-DC58K	Conversion lens adapter	Canon
LE410	3-in-1 pH glass electrode	Mettler Toledo
LQT Orbitrap Fournier Transformation mass spectrometer	High resolution mass spectrometry	Thermo Fisher Scientific
MagRack 6	Magnetic stand	GE Healthcare
MGD-4534	Gel dryer	vwr
Mini-Protean Tetra System	Gel electrophoresis system	Bio-Rad
Modell 100-800	Incubator	memmert
MR Hei Standard	Magnetic stirrer and heater	Heidolph
MyCycler	Thermocycler	Bio-Rad
MZ 2 NT	Membrane pump	vacubrand
MZ2C NT+AK+EK	Vacuum system	vacubrand
Nanodrop 2000	Spectrometer for concentration determination of DNA	Thermo Fisher Scientific
PCR workstation pro	PCR workstation	peqlab/vwr
PM400	Scale	Mettler
PowerPac basic	Power supply for PAGE	Bio-Rad
PowerShot G10	Camera for agarose gel documentation	Canon
Premium GGU 1500	Freezer	Liebherr
ProfiLine GG 4010	Freezer	Liebherr
ProfiLine FKU 1800	Fridge	Liebherr
RCT classic	Magnetic stirrer and heater	IKA

Research plus 1000/100/10/2.5 µl	Pipettes	eppendorf
RZ-6	Rotary vane pump	vacubrand
SimpliAmp	Thermocycler	applied biosystems
Sorvall Lynx 6000	Ultracentrifuge	Thermo Scientific
SSE-1	Sonifier sound enclosure	Branson
Steel chamber 18/10	Agarose gel staining bath	Bochem
Thermomixer compact/comfort	Thermomixer	eppendorf
Thermostat plus	Heating block	eppendorf
TPersonal	Thermocycler	Biometra
Tube Revolver	Tube Revolver	Thermo Scientific
Typhon FLA 9500	Imaging of native gels and radioactive screens	GE Healthcare
U725-G	Ultra-low temperature freezer	New Brunswick Scientific
ultrafleXtreme MALDI TOF/TOF instrument	MALDI-TOF mass spectrometry	Bruker
Unimax 1010	Plate shaker	Heidolph
UV star	UV Chamber for agarose gel documentation	Biometra
VL-6.LC	365 nm/254 nm UV lamp	Vilber Lourmat
VortexGenie2	Vortexer	Scientific Industries

5.1.4. Table 4: List of Disposables and Glass Ware

Product	Manufacturer/Distributor
384 well Lightcycler plate PP	Sarstedt
96 deep well plates	Sarstedt
96 fritted deep well plates	Thermo Fisher Scientific
96 well plates transparent	Carl Roth
Adhesive clear qPCR seals	Biozym
Amicon Ultra MWCO 3 kDa	Millipore
Amicon Ultra-0.5 ml Centrifugal Filters	Merck Millipore
Amicon Ultra-15, PLQK Ultracel-PL Membrane, 50 kDa	Merck Millipore
C18 Ziptips	Millipore
Cap, snap ring, 11 mm, blue, soft	vwr
Cuvettes	Sarstedt
Disposable glass Pasteur pipettes, sealed point, pre-marked, 230 mm	vwr
Electroporation cuvettes 1 mm	Carl Roth
epT.I.P.S. LoRetention 0.1 -10 µl	eppendorf
Erlenmayer flasks (2 l, 1 l, 500 ml, 100 ml, 25 ml)	vwr
Conical plastic tubes, 15 ml	Sarstedt
Conical plastic tubes, 50 ml	Sarstedt
Filter discs, grade 3hw, 125 mm	Sartorius
Filter discs, grade 3hw, 45 mm	Sartorius
Filter tips, low retention (2-100µl)	vwr
Folded qualitative filter paper, 301	vwr
Folded qualitative filter paper, 305	vwr
Glass beads	VWR
Glass test tubes, boro 5.1	vwr
Glass vial, snap ring, 1.5	vwr
Gloves Nitrile	VWR
illustra MicroSpin G-25 columns	GE Healthcare

MµLTIGUARD 100-1000 µl low retention filter tips	Sorenson
Microplate 384 well, PS, F-Bottom, small volume, black	Greiner
Multiply Pro 0.2 mL reaction tube	Sarstedt
NORM-JECT single-use syringes (20 ml, 5 ml)	Henke-Sass, Wolf
Nunc sealing tape, breathable, sterile	Thermo Fisher Scientific
Omnifix single-use syringes (50 ml, 20 ml, 10 ml)	B.Braun
Omnifix-F single-use syringes (1 ml)	B.Braun
Parafilm PM-996	Bemis
Pierce C18 tips, 10 µl	Thermo Fisher Scientific
Pierce centrifuge columns, 0.8 ml	Thermo Fisher Scientific
Pipette tips (10 µL, 200 µL, 1000 µL) Sarstedt	Sarstedt
Polypropylene columns, 1 ml	Qiagen
Polypropylene columns, 5 ml	Qiagen
Protein LoBind Tubes	eppendorf
Reaction tubes 1.5 mL	Sarstedt
Reaction tubes 2.0 mL	Sarstedt
Round bottom flask with standard ground joint (250 ml, 100 ml)	Duran
Scalpel	B.Braun
Schott flasks (1000 ml, 500 ml, 250 ml, 100 ml, 50 ml)	Duran
Schott flasks (1000 ml, 500 ml, 250 ml, 100 ml, 50 ml)	vwr
Serological pipette 10 mL	Sarstedt
Serological pipette 25 mL	Sarstedt
Slide-A-Lyzer Dialysis Cassette MWCO 10 kDa	Thermo Fisher Scientific

Slide-A-Lyzer MINI Dialysis Unit MWCO 10 kDa	Thermo Fisher Scientific
Sterican single-use needles 0.80 x 120 mm, 21 G x 4 ¾''	B.Braun
Sterican single-use needles 0.90 x 40 mm, 20 G x 1 ½''	B.Braun
Sterican single-use needles 0.90 x 70 mm, 20 G x 2 ¾''	B.Braun
Sterican single-use needles 1.20 x 40 mm, 18 G x 1 ½''	B.Braun
Syringe filter 0.2 µM	Sarstedt
UV cuvettes	Sarstedt
ZelluTrans dialysis membrane T1 MWCO 3.5 kDa	Carl Roth

5.1.5. Table 5 List of Consumables

Name	Use	Manufacturer/Distributor
10 % FBS	Supplement for DMEM	PAN Biotech
2-Log DNA Ladder (0.1-10 kb)	Standard DNA ladder for agarose gels	NEB
Bright-Glo	Luciferase assay master mix	Promega
DMEM	Medium for HEK293T cells	PAN Biotech
DPBS	Wash buffer for Luciferase assay	Pan Biotech
Dynabeads His-Tag Isolation and Pulldown	Magnetic Ni-NTA beads	Invitrogen
EmbryoMAX ES cell qualified FBS	Supplement for stem cell culture	Merck Milipore
Gelatine solution	Coating of cell culture dishes	Pan Biotech
GlutaMAX	Supplement for stem cell culture	Thermo Fisher Scientific
HisPur Ni-NTA Resin	High-capacity, high-performance Ni-IMAC resin	Thermo Fisher Scientific
KnockOut DMEM	Medium for Medium for culturing J1 mESC	Thermo Fisher Scientific
LB agar (Lennox)	Solid medium for *E.coli*	Carl Roth
LB-medium (Lennox)	Liquid medium for *E.coli*	Carl Roth
Lipofectamin 2000	Transfection reagent	Thermo Fisher Scientific
Low Molecular Weight DNA Ladder (25 bp to 766 bp)	DNA ladder for high percentage agarose gels	NEB
Mouse recombinant LIF	Supplement for stem cell culture	StemCell Technologies
Ni-NTA Magnetic Agarose Beads	Nickel-charged magnetic agarose beads	Qiagen

Non-essential amino acids	Supplement for stem cell culture	Thermo Fisher Scientific
Opti MEM	Medium for HEK293T cells	Gibco
PageRuler Plus Prestained Protein Ladder (10 to 250 kDa)	Protein ladder for SDS-PAGE	Thermo Fisher Scientific
Penicillin/Streptavidin solution	Antibiotics for stem cell culture	Pan Biotech
Unstained Protein Ladder, Broad Range (2-212 kDa)	Protein ladder for SDS-PAGE	NEB

5.1.6. Table 6: List of Commercial Kits and Master Mixes

Name	Type	Manufacturer/Distributor
GeneJET Gel extraction	Purification of PCR products and DNA from agarose gels	Thermo Fisher Scientific
GeneJET PCR purification	Purification of PCR products and DNA from reaction mixes	Thermo Fisher Scientific
GeneJET Plasmid Miniprep Kit	Plasmid DNA purification	Thermo Fisher Scientific
GoTaq qPCR Master Mix	qPCR Master Mix	Promega
illustra Ready-To-Go GenomiPhi V3	Whole genome amplification kit	GE Healthcare
Monarch PCR & DNA Cleanup Kit	Purification of PCR products and DNA from reaction mixes	NEB
NucleoSpin Plasmid Easy Pure	Plasmid DNA purification	Macherey Nagel
NucleoSpin Gel and PCR Clean-up	Purification of PCR products and DNA from reaction mixes and agarose gels	Macherey Nagel
Nucleotide Removal Kit	Clean-up of DNA labeling reactions	Qiagen
Pierce BCA Protein Assay Kit	Protein concentration determination	Thermo Fisher Scientific
QIAamp DNA Mini Kit	Purification of WGA DNA or gDNA	Qiagen
REPLI-g Mini Kit	Whole genome amplification kit	Qiagen

5.1.7. Table 7: List of Enzymes

Name	Distributor
AbaSI	NEB
BanII	NEB
BsaI	NEB
BsmBI	NEB
EarI	NEB
FatI	NEB
FokI	NEB
PlasmidSafe nuclease	Biozym
Proteinase K	Roche
KF(exo)$^-$	NEB
Q5 polymerase	NEB
RNase A	Thermo Fisher Scientific
Lysozyme	Sigma Aldrich
XhoI	NEB
NcoI	NEB
SacI	NEB
T4 ligase	Thermo Fisher Scientific
Pfu DNA Polymerase	Promega
T4 Phage β-glucosyltransferase (T4 BGT)	NEB
T4 Polynucleotide kinase (T4 PNK)	Thermo Fisher Scientific
Taq polymerase	NEB

5.1.8. Table 8: List of Chemicals

Name	CAS	Distributor
(7-Azabenzotriazol-1-yloxy)tri-pyrrolidinophosphonium hexa-fluorophosphate (PyAOP)	156311-83-0	Acros Organics
[ɣ-^{32}P]-ATP (370MBq) in 50 mM Tricine-buffer (pH 7.6)	-	Hartmann Analytic
1,4-Dithiothreitol (DTT)	3483-12-3	Carl Roth
2-(N-morpholino)ethanesulfonic acid (MES)	145224-94-8	Fisher Scientific
2-Mercaptoethanol	60-24-2	Merck
2-Propanol	67-63-0	Fisher Scientific
3-Hydroxypicolinic acid	874-24-8	Sigma Aldrich
5-Brom-4-chlor-3-indoxyl-β-D-galactopyranosid (X-Gal)	7240-90-6	Thermo Fisher Scientific
Acetic Acid	64-19-7	Carl Roth
Acetonitrile	75-05-8	Sigma Aldrich
Adenosintriphosphat, 10 mM solution	56-65-5	NEB
Agarose	9012-36-6	Biozym
Albumin, Bovine Serum (BSA)	9048-46-8	Cell Signaling Technology
Alexa Fluor 488	-	Thermo Fisher Scientific
Ammonium acetate	631-61-8	Sigma Aldrich
Ammonium peroxodisulfate	7727-54-0	Carl Roth
Benzylamine	100-46-9	Sigma Aldrich
Boric acid	10043-35-3	Carl Roth
Bromophenol blue	115-39-9	Sigma Aldrich
Calcium chloride	10043-52-4	Fisher Scientific
Carbenicillin disodium salt	4800-94-6	Carl Roth
Chloramphenicol	56-75-7	Carl Roth
Chloroform-d	865-49-6	Sigma Aldrich
Coomassie Brilliant Blue G250	6104-58-1	Carl Roth
Diammonium hydrogen citrate	3012-65-5	Sigma Aldrich

Dichlormethan	75-09-2	Acros Organics
Diethyl ether	60-29-7	Sigma Aldrich
Diethylene glycol	111-46-6	abcr
Diisopropyl azodicarboxylate (DIAD)	2446-83-5	abcr
Dimethyl sulfoxide (DMSO)	67-68-5	Carl Roth
dNTP solution mix, 10 mM	-	NEB
Ethanol 96 %	64-17-5	Fisher Scientific
Ethanol, absolute	64-17-5	Sigma Aldrich
Ethidium bromide	1239-45-8	Sigma Aldrich
Ethylamin	75-04-7	Alfa Aesar
Ethylene glycol	107-21-1	abcr
Ethylenediaminetetraacetate (EDTA)	60-00-4	Carl Roth
Formamide	75-12-7	Acros Organics
Glycerole	56-81-5	Carl Roth
Glycine	56-40-6	Carl Roth
HEPES	7365-45-9	Carl Roth
Hexaethylene glycol	2615-15-8	abcr
Hydrazine hydrate	7803-57-8	Fisher Chemical
Hydrochloric acid (1 M)	7647-01-0	vwr
Hydrochloric acid (37 %)	7647-01-0	vwr
Imidazole	288-32-4	abcr
Isopropyl-β-D-thiogalactopyranosid	367-93-1	Roth
Kanamycinsulfate	25389-94-0	Carl Roth
L(+)-Arabinose	5328-37-0	Carl Roth
Magnesium chloride hexadydrate	7791-18-6	Acros Organics
Magnesium sulfate heptahydrate	10034-99-8	Merck
MES sodium salt	71119-23-8	Acros Organics
Methanol	67-56-1	Sigma Aldrich
N-(3-Dimethylaminopropyl)-N'-ethylcarbodiimidhydrochlorid (EDC)	25952-53-8	Sigma Aldrich
N,N,N',N'-Tetramethylene-ethylenediamine (TEMED)	110-18-9	Carl Roth
N-Hydroxyphtalimide	524-38-9	Acros Organics

O-(tert-Butyl)hydroxylamine	39684-28-1	Acros Organics
O-(Tetrahydro-2H-pyran-2-yl)hydroxylamine	6723-30-4	Acros Organics
O-Benzylhydroxylamin hydrochloride	2687-43-6	Alfa Aesar
O-Ethylhydroxylamine hydrochloride	3332-29-4	Sigma Aldrich
para-Acetylphenylalanine (*p*AcF)	122555-04-8	Chem Impex
para-Phenylenediamine (*p*PDA)	7447-40-7	Acros Organics
Phenylmethylsulfonyl fluoride	329-98-6	Carl Roth
Phenylmethylsulfonylfluorid (PMSF)	329-98-6	Carl Roth
Potassium chloride	7447-40-7	Carl Roth
Roti-Phenol/Chloroform/Isoamylalkohol	-	Carl Roth
Rotiphorese Gel 40 (37.5:1)	-	Carl Roth
Rotiphorese sequencing gel concentrate	-	Carl Roth
Silica gel 60M 0.04-0.063 nm	112926-00-8	Macherey Nagel
Sodium acetate	127-09-3	Merck
Sodium chloride	7647-14-5	Carl Roth
Sodium dihydrogen phosphate monohydrate	10049-21-5	Merck
Sodium dodecyl sulfate (SDS)	151-21-3	Carl Roth
Sodium hydroxide	1310-73-2	Fisher Scientific
Sodium lauroyl sarcosinate	137-16-6	AppliChem
Sodium phosphate dibasic dihydrate	10028-24-7	Sigma Aldrich
Spermine tetrahydrochloride	306-67-2	Alfa Aesar
tert-Butylamin	75-64-9	Acros Organics
Tetracycline hydrochloride	64-75-5	Carl Roth
Tetraethylene glycol	112-60-7	abcr
Tetrahydropyran-2-ylmethylamine	6628-83-7	Maybridge
Triethylene glycol	112-27-6	abcr
Trifluoroacetic acid (TFA)	76-05-1	Sigma Aldrich
Triphenylphosphine	603-35-0	Alfa Aesar
Tris(hydroxymethyl)aminomethane	77-86-1	Sigma Aldrich
Triton-X	9002-93-1	Fluka
Trizma base (Tris)	77-86-1	Carl Roth
Tween20	9005-64-5	Fisher Scientific
Urea	57-13-6	Carl Roth

Uridine Diphosphate Glucose solution, 2 mM	133-89-1	NEB
Water, nuclease-free	7732-18-5	Qiagen
Xylencyanol	2650-17-1	Carl Roth
Yeast extract	8013-01-2	Carl Roth

5.1.9. Table 9: List of Buffers and Stock Solutions

Name	Components
4xPBS	548 mM NaCl, 43 mM KCl, 69 mM Na_2HPO_4, 3.2 g/l KH_2PO_4, pH = 8
Amide reaction buffer	50 mM MES, pH = 8
Cam stock	34 mg/ml chloramphenicol in ethanol
Carb stock	50 mg/ml carbenicillin disodium salt in 1:1 ethanol:water
CD spectroscopy buffer	20 mM Tris-HCl, 50 mM NaCl, pH = 7.5
CutSmart buffer	50 mM KOAc, 20 mM Tris-acetate, 10 mM $Mg(OAc)_2$**,** 100 µg/ml BSA**,** pH = 7.9
Enrichment buffer (EB) 1	150 mM NaCl, 30 mM Tris-HCl, 5 mM $MgCl_2$, 0.05 % Tween20, 0.5 mg/ml BSA, pH = 7.9
Enrichment buffer (EB) 2	500 mM NaCl, 30 mM Tris-HCl, 5 mM $MgCl_2$, 100 µM spermine, 0,005 % Tween20, 0.5 mg/ml BSA, pH = 7.9
Enrichment high-salt wash buffer (EWB)	2 M NaCl, 30 mM Tris-HCl, 5 mM $MgCl_2$, 10 µM spermine, 0,005 % Tween20, 0.5 mg/ml BSA, pH = 7.9
HEK293T lysis buffer	100 mM NaH_2PO_4, 0.2 % Triton X-100
Kan stock	50 mg/ml kanamycin sulfate in water
KF (exo^-) storage buffer	25 mM Tris-HCl, 1 mM DTT, 0.1 mM EDTA, 50 % glycerol, pH = 7.4
LB Medium	10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl, pH = 7
mESC lysis buffer	100 mM Tris, 5 mM EDTA, 200 mM NaCl, 0.2 % SDS, 400 µg/ml proteinase K, 200 µg/ml RNase A, pH = 5.5
NEB1	10 mM bis-Tris-propane-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH = 7
NEB2	50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH = 7.9
NEB3	100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH = 7.9

NEB4	50 mM KOAc, 20 mM Tris-acetate, 10 mM $Mg(OAc)_2$, 1 mM DTT, pH = 7.9
Oxime reaction buffer	50 mM $NaHPO_4/Na_2PO_4$, pH = 6.0
Purple loading dye	2.5 % Ficoll-400, 10 mM EDTA, 3.33 mM Tris-HCl, 0.08 % SDS, 0.02 % Dye 1, 0.001 % Dye 2, pH = 8
Qiagen lysis buffer	50 mM NaH_2PO_4 monohydrate, 300 mM NaCl, pH = 8
Qiagen lysis buffer 20	Qiagen Lysis buffer + 20 mM imidazole
Qiagen lysis buffer 50	Qiagen Lysis buffer + 50 mM imidazole
Qiagen lysis buffer 500	Qiagen Lysis buffer + 500 mM imidazole
SDS-PAGE running buffer	3.03 g/l Tris base, 14.40 g/l glycine, 1.00 g/l SDS
SDS-PAGE sample buffer	50 mM Tris-HCl (pH = 6.8), 2 % SDS, 10 % glycerol, 0.02 % bromophenol blue, 1 % β-mercaptoethanol
SDS-PAGE staining solution	10 % v/v acetic acid, 50 % v/v methanol, 40 % v/v water, 0.1 % w/v Comassie brilliant blue
SDS-PAGE destaining solution	10 % v/v acetic acid, 20 % v/v ethanol, 70 % v/v water
SOB^{++}	20 g/l tryptone, 5 g/l yeast extract, 0.5 g/l NaCl, 0.186 g/l KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$
SOC medium	0.58 g/l NaCl, 2.03 g/l $MgCl_2$ hexahydrate, 2.46 g/l $MgSO_4$ heptahydrate, 5 g/l yeast extract, 20 g/l tryptone, 1 M glucose, pH = 7.5
Spec stock	100 mg/ml spectinomycin in water
T4 DNA ligase buffer	50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP, 10 mM DTT, pH = 7.5
T4 PNK buffer A	350 mM Tris-HCl (pH = 7.6), 50 mM $MgCl_2$, 500 mM KCl, 5 mM 2-mercaptoethanol
TAE buffer	40 mM Tris-acetate, 2.5 EDTA, pH = 7.8
TAL storage buffer	200 M NaCl, 20 mM Tris, 5 % glycerol, pH = 7.5

TALE Binding Buffer (TBB) 1	20 mM Tris-HCl (pH = 8.0), 50 mM NaCl, 5 mM $MgCl_2$, 5 % Glycerol
TALE-DNA-Crosslinking (TDC) Buffer	20 mM Tris-HCl, 10 mM NaCl, 1.25 mM $MgCl_2$, 5 % glycerol, pH = 6
Taq lysis buffer	10 mM Tris-HCl, 300 mM NaCl, 2.5 mM $MgCl_2$, 0.1 % Triton X-100, pH = 9
TBE Buffer	10.78 g/l Tris base, 5.50 g/l boric acid, 0.585 g/l EDTA, pH = 8.3
TE buffer	100 mM Tris-HCl, 10 mM EDTA, pH = 7.4
Tet stock	12.5 mg/ml tetracycline in ethanol

5.1.10. Table 10: List of Oligonucleotides

Name	Sequence	Description
o7	CTGCAGTTTCAAACGCTAAATTGCCTGATG	Sequencing primer pEVOL *p*AcF
o41	AGTTTAAACGGTCTCCAGC	Sequencing primer pEVOL *p*AcF
o42	ATTGCAGAGATCATGTAGG	Sequencing primer pEVOL *p*AcF
o83	AGTTCCCTACTCTCGCATG	Sequencing primer pEVOL *p*AcF
o247	GGGTTATGCTAGTTATTGCTCAG	Sequencing primer pET TALE expression vector
o315	CGTAGAGGATCGAGATC	Sequencing primer pET TALE expression vector
o410	GAGTAACAGCGGTAGAG	Sequencing primer pET TALE expression vector
o734	TTGATGCCTGGCAGTTCCCT	Sequencing primer pFUS vector
o735	CGAACCGAACAGGCTTATGT	Sequencing primer pFUS vector
o814	AATGCGCAAACCAACCC	Sequencing primer pMV vector
o890	AGATATGATTGCGGCCCTG	Sequencing primer pET TALE expression vector
o1152	GGATGTGGAAACGGAAGAGCTCACGGTGGAAGAAGCCACTGAAGATGCTGTAACGCTGGGGCTGAAGAGTGGGAATCCA	Reference sequence_1297_rev
o1408	AGAAGCGCGATCACATG	Sequencing primer pET TALE expression vector
o1411	TCAATATCACTGTGTGGC	Sequencing primer pLR vector
o1422	TGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTTCCACCGTGAGCTCTTCXGTTTCCACATCC (X = 5caC)	Reference sequence_1297_C → 5caC

o1423	TGGATTCCCACTCTTCAGCCCCAGCGTTA CAGCATCTTCAGTGGCTTCTTCCACCGTG AGCTCTTCXGTTTCCACATCC (X = 5fC)	Reference sequence_1297_C → 5fC
o1494	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCGGCGGCAA GCAAGCGCTCGAAACGGTGCAGCGG	pHD2_SHDGG → pG*2_SG*GG
o1495	TTTTCTCGAGGGTCTCGGTCCTGGCACA GCACCGGCAACAGCCGCTGCACCGTTTC GAGCGCTTGCTTGCCGCCGCCGCT	pHD2_SHDGG → pG*2_SG*GG
o1496	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCGGCAAGCA AGCGCTCGAAACGGTGCAGCGGCTG	pHD2_SHDGG → pG**2_S***G
o1497	TTTTCTCGAGGGTCTCGGTCCTGGCACA GCACCGGCAACAGCCGCTGCACCGTTTC GAGCGCTTGCTTGCCGCCGCTGGC	pHD2_SHDGG → pG**2_S***G
o1498	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCAAGCAAGC GCTCGAAACGGTGCAGCGGCTGTTG	pHD2_SHDGG → pG***2_S***G
o1499	TTTTCTCGAGGGTCTCGGTCCTGGCACA GCACCGGCAACAGCCGCTGCACCGTTTC GAGCGCTTGCTTGCCGCTGGCGAT	pHD2_SHDGG → pG***2_S***G
o1500	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCAAGCAAGCGCTC GAAACGGTGCAGCGGCTGTTGCCG	pHD2_SHDGG → pS****2_ S****
o1501	TTTTCTCGAGGGTCTCGGTCCTGGCACA GCACCGGCAACAGCCGCTGCACCGTTTC GAGCGCTTGCTTGCTGGCGATAGC	pHD2_SHDGG → pS****2_ S****
o1502	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAATAAGCAAGCGCTC GAAACGGTGCAGCGGCTGTTGCCG	pHD2_SHDGG → pN****2_ N****
o1503	TTTTCTCGAGGGTCTCGGTCCTGGCACA GCACCGGCAACAGCCGCTGCACCGTTTC GAGCGCTTGCTTATTGGCGATAGC	pHD2_SHDGG → pN****2_ N****

o1504	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCGGCGGCAA GCAAG	pHD5_SHDGG → pG*5_SG*GG
o1505	TTTTCTCGAGGGTCTCCTGCACCGTTTCG AGCGCTTGCTTGCCGCCGCCGCTGGCGA TAG	pHD5_SHDGG → pG*5_SG*GG
o1506	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCGGCAAGCA AGCGC	pHD5_SHDGG → pG**5_S***G
o1507	TTTTCTCGAGGGTCTCCTGCACCGTTTCG AGCGCTTGCTTGCCGCCGCTGGCGATAG CCA	pHD5_SHDGG → pG**5_S***G
o1508	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCAAGCAAGC GCTCG	pHD5_SHDGG → pG***5_S***G
o1509	TTTTCTCGAGGGTCTCCTGCACCGTTTCG AGCGCTTGCTTGCCGCTGGCGATAGCCA CCA	pHD5_SHDGG → pG***5_S***G
o1510	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCAAGCAAGCGCTC GAAA	pHD5_SHDGG → pS****5_ S****
o1511	TTTTCTCGAGGGTCTCCTGCACCGTTTCG AGCGCTTGCTTGCTGGCGATAGCCACCA CTT	pHD5_SHDGG → pS****5_ S****
o1512	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAATAAGCAAGCGCTC GAAA	pHD5_SHDGG → pN****5_ N****
o1513	TTTTCTCGAGGGTCTCCTGCACCGTTTCG AGCGCTTGCTTATTGGCGATAGCCACCA CTT	pHD5_SHDGG → pN****5_ N****
o1516	CTTCCTCTTCCGTCTCTTTCCTTTTACGTC ATCCGGGGGCAGACT	BRCA1a
o1518	CTTCCTCTTCCGTCTCTTTCCTTTTAXGTC ATCCGGGGGCAGACT (X = 5mC)	BRCA1a_C13 → 5mC

o1520	CTTCCTCTTCXGTCTCTTTCCTTTTACGTC ATCCGGGGGCAGACT (X = 5hmC)	Reference sequence_BRCA1_C → 5hmC
o1521	CTTCCTCTTCCGTCTCTTTCCTTTTAXGTC ATCCGGGGGCAGACT (X = 5hmC)	BRCA1a_C13 → 5hmC
o1524	CTTCCTCTTCCGTCTCTTTCCTTTTAXGTC ATCCGGGGGCAGACT (X = 5fC)	BRCA1a_C13 → 5fC
o1525	CTTCCTCTTCCGTCTCTTTCCTTTTACGTC ATCXGGGGGCAGACT (X = 5fC)	BRCA1c_C13 → 5fC
o1527	CTTCCTCTTCCGTCTCTTTCCTTTTAXGTC ATCCGGGGGCAGACT (X = 5caC)	BRCA1a_C13 → 5caC
o1529	AGTCTGCCCCCGGATGACGTAAAAGGAA AGAGACGGAAGAGGAAG	BRCA1a_C13_rev
o1570	CAATCCAGAGCCCCG	Sequencing primer BRCA1a spike-in
o1577	GGACTCTACTACCTTTACC	Sequencing primer BRCA1a spike-in
o1591	GGCCAGCCAGTCAGCCGAAGGCTCCATG CTGCTCCCCGCCGCCGGC	CDKN2A
o1592	GGCCAGCCAGTCAGCXGAAGGCTCCATG CTGCTCCCCGCCGCCGGC (X = 5mC)	CDKN2A _C5 → 5mC
o1593	GGCCAGCCAGTCAGCXGAAGGCTCCATG CTGCTCCCCGCCGCCGGC (X = 5hmC)	CDKN2A _C5 → 5hmC
o1594	GGCCAGCCAGTCAGCXGAAGGCTCCATG CTGCTCCCCGCCGCCGGC (X = 5fC)	CDKN2A _C5 → 5fC
o1595	GGCCAGCCAGTCAGCXGAAGGCTCCATG CTGCTCCCCGCCGCCGGC (X = 5caC)	CDKN2A _C5 → 5caC
o1596	AGGGCGGTGTGGGGGGCAGG TGGGGAGGAGCCCAGTCCTC CTTCCT	CDKN2A_PCR primer for spike-in amplicons
o1601	GTCAAAGAATACCCATCTGTCAGCTTCGG AAATCCACTCTCCCAC	BRCA1a_PCR primer for spike-in amplicons
o1607	AGGAGCCCAGTCCTC	CDKN2A_qPCR primer

o1617	GCCGGCGGCGGGGAGCAGCATGGAGCC TTCGGCTGACTGGCTGGCC	CDKN2A_rev
o1750	TTTTCCATGGCCTGACTCCGGACCAAGT GGTGGCTATCGCCAGCGGCGGCGGCAA GCAAGCGCTCGAAACGGTGCAGCGG	pHD4_SHDGG → pG*4_SG*GG
o1751	TTTTCTCGAGGGTCTCCAACAGCCGCTG CACCGTTTCGAGCGCTTGCTT	pHD4_SHDGG → pG*4_SG*GG
o1902	GGCCAGCCAGTCAGCAGAAGGCTCCATG CTGCTCCCCGCCGCCGGC	CDKN2A _C5 → A
o1903	GGCCAGCCAGTCAGCTGAAGGCTCCATG CTGCTCCCCGCCGCCGGC	CDKN2A _C5 → T
o1904	GGCCAGCCAGTCAGCGGAAGGCTCCATG CTGCTCCCCGCCGCCGGC	CDKN2A _C5 → G
o2016	TTTTGTCGACTCAGCCGAAGGCTCCATG CTGCTCCCACTAGTTTTT	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6
o2017	AAAAACTAGTGGGAGCAGCATGGAGCCT TCGGCTGAGTCGACAAAA	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6
o2018	TTTTGTCGACTCAGCAGAAGGCTCCATGC TGCTCCCACTAGTTTTT	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6 → A6
o2019	AAAAACTAGTGGGAGCAGCATGGAGCCT TCTGCTGAGTCGACAAAA	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6 → A6
o2020	TTTTGTCGACTCAGCGGAAGGCTCCATG CTGCTCCCACTAGTTTTT	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6 → G6
o2021	AAAAACTAGTGGGAGCAGCATGGAGCCT TCCGCTGAGTCGACAAAA	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK

		oN2A_TALE binding_C6 → G6
o2022	TTTTGTCGACTCAGCTGAAGGCTCCATGC TGCTCCCACTAGTTTTT	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6 → T6
o2023	AAAAACTAGTGGGAGCAGCATGGAGCCT TCAGCTGAGTCGACAAAA	pSF_MinCMV_Fluc → pSF_MinCMV_Fluc_CDK N2A_TALE binding_C6 → T6
o2087	GCCGGCGGCGGGGAGCAGCATGGAGCC TTCTGCTGACTGGCTGGCC	CDKN2A _C5 → A_rev
o2088	GCCGGCGGCGGGGAGCAGCATGGAGCC TTCAGCTGACTGGCTGGCC	CDKN2A _C5 → T_rev
o2089	GCCGGCGGCGGGGAGCAGCATGGAGCC TTCCGCTGACTGGCTGGCC	CDKN2A _C5 → G_rev
o2129	TGAGTCGGAGTCTCATTCTGTC	CDKN2A_PCR primer for spike-in amplicons
o2130	CTCACAAAACAGGAGTAGGGAG	Sequencing primer CDKN2A spike-in
o2131	CCTCTTCTTTGCAGGATTCCTC	BRCA1a_PCR primer for spike-in amplicons
o2132	CACTCTTGTGCTGACTTACCAG	BRCA1a_PCR primer for spike-in amplicons
o2135	GAAGCCATACTTTCCCTATG	CDKN2A_PCR primer for spike-in amplicons
o2136	CCTTTCATTCCGCAACGC	BRCA1a_PCR primer for spike-in amplicons
o2165	XGCCGGCGGCGGGGAGCAGCATGGAGC CTTCGGCTGACTGGCTGGCC (X = 5'-Cy5)	CDKN2A_rev_Cy5
o2313	CTCTTTCTTCCTCCGGTGCTG	CDKN2A_qPCR primer
o2485	CAAGTGGTGGCTATCGCCAGCGGCTAGG GCAAGCAAGCG	pG*2_SG*GG → pG*12_SG*1G

o2486	CCGTTTCGAGCGCTTGCTTGCCCTAGCC GCTGGCGATAG	pG*2_SG*GG → pG*12_SG*1G
o2487	CGGACCAAGTGGTGGCTATCGCCTAGGG CGGCGGCAAGC	pG*2_SG*GG → p1G*2_1G*GG
o2488	TCGAGCGCTTGCTTGCCGCCGCCCTAGG CGATAGCCACC	pG*2_SG*GG → p1G*2_1G*GG
o2489	CAAGTGGTGGCTATCGCCAGCTAGGGCG GCAAGCAAGCG	pG*2_SG*GG → p1*2_S1*GG
o2490	CCGTTTCGAGCGCTTGCTTGCCGCCCTA GCTGGCGATAG	p1*2_SG*GG → p1*2_S1*GG
o2551	AGTCTGCCCCCGGATCGCGCGAAAGGAA AGAGACGGAAGAGGAAG	BRCA1b_rev
o2552	CTTCCTCTTCCGTCTCTTTCCTTTXGCGC GATCCGGGGGCAGACT (X = 5fC)	BRCA1b_C11 → 5fC
o2553	CTTCCTCTTCCGTCTCTTTCCTTTCGXGC GATCCGGGGGCAGACT (X = 5fC)	BRCA1b_C13 → 5fC
o2554	CTTCCTCTTCCGTCTCTTTCCTTTCGCGX GATCCGGGGGCAGACT (X = 5fC)	BRCA1b_C15 → 5fC
o2555	AGTCTGCCCCCGGATCGXGCGAAAGGAA AGAGACGGAAGAGGAAG (X = 5fC)	BRCA1b_C13 → 5fC_rev
o2556	CTTCCTCTTCCGTCTCTTTCCTTTCGCGC GATCCGGGGGCAGACT	BRCA1b
o2901	TTTTTTTTTTTCGTTTTTTTTTTT	Reference sequence_RefSeq_rev
o2910	XAGTCTGCCCCCGGATGACGTAAAAGGA AAGAGACGGAAGAGGAAG (X = 5'-Cy5)	BRCA1a_rev_Cy5
o2939	TCTGAGAGGCTGCTGCTTAG	BRCA1a _qPCR primer
o2943	GGAAGTCTCAGCGAGCTCAC	BRCA1a _qPCR primer
o2963	CGGACCAAGTGGTGGCTATCGCCAGCGG CGGCGGGAGAC	pHD10_SHDGG → pG*10_SG*GG
o2964	GTGGCTCGAGGGTCTCCCGCCGCCGCT GGCGATAGCCAC	pHD10_SHDGG → pG*10_SG*GG

o3116	YAAAAAAAAAAAXGAAAAAAAAAAA (Y = 5’-FAM; X = 5fC)	Reference sequence_RefSeq_5fC12

5.1.11. Table 11: List of Plasmids

Name	Description	Antibiotic Resistance
pDaS71	pEVOL vector for expression of *M. jannaschii* tyrosyl-tRNA synthetase/tRNA pair for incorporation of *p*AcF via amber suppression	Cam
pAnI521	pET TALE expression vector with N-terminal GFP-domain, shortened AvrBs3-type TALE N-terminus and C-terminal His-tag	Carb
pMaG635	pHD2_SHDGG → SG*GG module vector with H12G and D13*	Tet
pMaG636	pHD2_SHDGG → SG**G module vector with H12G, D13* and G14*	Tet
pMaG637	pHD2_SHDGG → SG*** module vector with H12G, D13*, G14* and G15*	Tet
pMaG638	pHD2_SHDGG → S**** module vector with H12*, D13*, G14* and G15*	Tet
pMaG639	pHD2_SHDGG → N**** module vector with S11N, H12*, D13*, G14* and G15*	Tet
pMaG640	pHD5_SHDGG → SG*GG module vector with H12G and D13*	Tet
pMaG641	pHD5_SHDGG → SG**G module vector with H12G, D13* and G14*	Tet
pMaG642	pHD5_SHDGG → SG*** module vector with H12G, D13*, G14* and G15*	Tet
pMaG643	pHD5_SHDGG → S**** module vector with H12*, D13*, G14* and G15*	Tet
pMaG644	pHD5_SHDGG → N**** module vector with S11N, H12*, D13*, G14* and G15*	Tet
pMaG659	pET_GFP_BRCA1a_TALE_wt_6His	Carb
pMaG660	pET_GFP_BRCA1a_TALE_SG*GG_6His	Carb
pMaG661	pET_GFP_BRCA1a_TALE_SG**G_6His	Carb
pMaG662	pET_GFP_BRCA1a_TALE_SG***_6His	Carb
pMaG663	pET_GFP_BRCA1a_TALE_S****_6His	Carb

pMaG664	pET_GFP_BRCA1a_TALE_N****_6His	Carb
pMaG665	pET_GFP_CDKN2A_TALE_wt_6His	Carb
pMaG666	pET_GFP_CDKN2A_TALE_SG*GG_6His	Carb
pMaG667	pET_GFP_CDKN2A_TALE_SG**G_6His	Carb
pMaG668	pET_GFP_CDKN2A_TALE_SG***_6His	Carb
pMaG669	pET_GFP_CDKN2A_TALE_S****_6His	Carb
pMaG670	pET_GFP_CDKN2A_TALE_N****_6His	Carb
pMaG1017	pHD4 → SG*GG module vector with H12G and D13*	Tet
pMaG1184	pMaG635_SG*GG → SG*1G with G14*p*AcF	Tet
pMaG1185	pMaG635_SG*GG → 1G*GG with S11*p*AcF	Tet
pMaG1186	pMaG635_SG*GG → S1*GG with G12*p*AcF	Tet
pMaG1187	pMaG640_SG*GG → SG*1G with G14*p*AcF	Tet
pMaG1188	pMaG640_SG*GG → 1G*GG with S11*p*AcF	Tet
pMaG1189	pMaG640_SG*GG → S1*GG with G12*p*AcF	Tet
pMaG1190	pLR_NG_SNGGG→S1*GG with N12*p*AcF and G13*	Tet
pMaG1201	pET_GFP_BRCA1a_TALE_1G*GG_6His	Carb
pMaG1202	pET_GFP_BRCA1a_TALE_S1*GG_6His	Carb
pMaG1203	pET_GFP_BRCA1a_TALE_SG*1G_6His	Carb
pMaG1208	pET_GFP_BRCA1c_TALE_LR_S1*GG(20)_6His	Carb
pMaG1209	pET_GFP_BRCA1c_TALE_LR_S1*GG(19)_6His	Carb
pMaG1269	pET_GFP_BRCA1b_TALE_1G*GG_6His	Carb
pMaG1270	pET_GFP_BRCA1b_TALE_S1*GG_6His	Carb
pMaG1271	pET_GFP_BRCA1b_TALE_SG*1G_6His	Carb
pMaG1535	pHD10_SHDGG → SG*GG module vector with H12G and D13*	Tet

5.1.11. Table 12: List of TALE Proteins

Name	RVD Sequence (non-canonical repeats in brackets)
CDKN2A_TALE_wt	HD NI NN HD HD NN NI NI NN NN HD NG HD HD NI NG NN HD NG NN HD NG HD HD HD
CDKN2A_TALE_SG*GG	HD NI NN HD (SG*GG) NN NI NI NN NN HD NG HD HD NI NG NN HD NG NN HD NG HD HD HD
CDKN2A_TALE_SG**G	HD NI NN HD (SG**G)NN NI NI NN NN HD NG HD HD NI NG NN HD NG NN HD NG HD HD HD
CDKN2A_TALE_SG***	HD NI NN HD (SG***) NN NI NI NN NN HD NG HD HD NI NG NN HD NG NN HD NG HD HD HD
CDKN2A_TALE_S****	HD NI NN HD (S****) NN NI NI NN NN HD NG HD HD NI NG NN HD NG NN HD NG HD HD HD
CDKN2A_TALE_N****	HD NI NN HD (N****) NN NI NI NN NN HD NG HD HD NI NG NN HD NG NN HD NG HD HD HD
BRCA1a_TALE_wt	HD NG NG NG HD HD NG NG NG NG NI HD NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_SG*GG	HD NG NG NG HD HD NG NG NG NG NI (SG*GG) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_SG**G	HD NG NG NG HD HD NG NG NG NG NI (SG**G) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_SG***	HD NG NG NG HD HD NG NG NG NG NI (SG***) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_S****	HD NG NG NG HD HD NG NG NG NG NI (S****) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_N****	HD NG NG NG HD HD NG NG NG NG NI (N****) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_**1**G*GG	HD NG NG NG HD HD NG NG NG NG NI (**1**G*GG) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_S**1**GG	HD NG NG NG HD HD NG NG NG NG NI (S**1**GG) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD
BRCA1a_TALE_SG**1**G	HD NG NG NG HD HD NG NG NG NG NI (SG**1**G) NN NG HD NI NG HD HD NN NN NN NN NN LR_HD

BRCA1b_TALE_**1**G*GG	HD NG NG NG HD HD NG NG NG (SG*GG) NN (**1**G*GG) NN (SG*GG) NN NI NG HD HD NN NN NN NN NN LR_HD
BRCA1b_TALE_S**1***GG	HD NG NG NG HD HD NG NG NG (SG*GG) NN (S**1***GG) NN (SG*GG) NN NI NG HD HD NN NN NN NN NN LR_HD
BRCA1b_TALE_SG***1**G	HD NG NG NG HD HD NG NG NG (SG*GG) NN (SG***1**G) NN (SG*GG) NN NI NG HD HD NN NN NN NN NN LR_HD
BRCA1c_ TALE_LR_S**1***GG(20)	HD NG NG NG HD HD NG NG NG NG NI HD NN NG HD NI NG HD LR_S**1***GG
BRCA1c_ TALE_LR_S**1***GG(19)	HD NG NG NG HD HD NG NG NG NG NI HD NN NG HD NI NG LR_S**1***GG

5.1.12. Table 13: List of Strains/Cell Lines

Type	Host
GH371	*E. coli*
BL21 DE3 Gold	*E. coli*
HEK293T	*H. sapiens*
J1 mESC	*M. musculus*

5.2. Methods[e]

5.2.1. Biological/Biochemical Methods

5.2.1.1. Plate Culture

Bacterial swabs were streaked onto LB-Agar supplemented with antibiotic in plastic petri dish, if necessary. Alternatively, bacterial suspension (either water or SOC medium) were pipetted onto LB-Agar supplemented with antibiotic in plastic petri dishes and spread out using glass beads close to the flame. Plates were incubated at 37°C and grown overnight in an incubator.

5.2.1.2. Liquid Overnight Culture

Single bacterial colonies were transferred to LB medium supplemented with antibiotics and incubated at 37 °C whilst shaking at 180 rpm overnight in glass Erlenmeyer flasks or conical plastic tubes.

5.2.1.3. Preparation of Chemical-competent GH317 *E.coli* bacteria

From the strain stock, bacteria were streaked onto LB-Agar plates and incubated at 37°C overnight. A single colony of *E. coli* GH317 strain was transferred to 15 ml LB medium and incubated overnight at 37 °C shaking at 180 rpm (rounds per minute). 8 ml of bacterial overnight culture were added to 800 ml of LB medium and incubated at 37 °C whilst shaking at 180 rpm until an OD600 of 0.5 was reached. The culture was first cooled on ice and subsequently centrifuged at 4 °C for 10 minutes at 4000 rpm. The supernatant was promptly discarded. The pellet was resuspended in 80 ml of precooled, sterile 100 mM $MgCl_2$ on ice. The suspension was again centrifuged (4 °C, 15 min, 4000 rpm), the pellet recovered, kept on ice and resuspended in 80 ml of ice-cold, sterile 50 mM $CaCl_2$. After another round of centrifugation, the pellet was again resuspended in 4 ml ice-cold, sterile 50 mM $CaCl_2$ supplemented with 15 % glycerol. Finally, the suspension was aliquoted into fractions of 50 µl on ice and immediately frozen in liquid nitrogen. Frozen cells were stored at -80 °C.

5.2.1.4. Preparation of Chemical-competent BL21 *E.coli* bacteria

From the strain stock, bacteria were streaked onto LB-Agar plates and incubated at 37°C overnight. A single colony of *E. coli* BL21 strain was transferred to 15 ml LB

[e] The protocols of this section have been (partly) published in Gieß et al. (2018) and Gieß et al. (2019).[369,370] Reprinted (adapted) with permission from ref.[369] and ref.[370] Copyright 2019 American Chemical Society

medium and incubated overnight at 37 °C shaking at 180 rpm. 800 ml of SOB++ medium was inoculated with 5 ml of bacterial BL21 overnight culture and incubated at room temperature (~25-30 °C) whilst shaking till an OD_{600} = 0.5 was reached. The culture was chilled on ice for 10 min. and then kept on ice as much as possible during the procedure. The culture was centrifuged at 4 °C for 10 minutes at 4000 rpm. Then, the supernatant was discarded, and the pellet was resuspended in 200 ml cold TB-Buffer and incubated on ice for 10 min. After 10 minutes of centrifugation at 4 °C for and 4000 rpm the supernatant was discarded, and the pellet was resuspended in 30 ml of cold TB-Buffer. Next, 2 ml DMSO was added, and the suspension incubated for 10 minutes on ice. Aliquots of 200 µl were prepared on ice and immediately frozen in liquid nitrogen. Frozen cells were stored at -80 °C.

5.2.1.5. Preparation of Electro-competent GH371 or BL21 *E.coli*

From LB-Agar plates single colonies of an *E. coli* strain were inoculated into 25 ml LB-overnight culture, if necessary supplied with appropriate antibiotics. The suspension was incubated at 37 °C and shaking at 180 rpm. Next, 5 ml of overnight culture was transferred to 500 ml of LB medium and incubated at 37 °C whilst shaking at 200 rpm until an OD_{600} = 0.4 was reached. The culture was chilled on ice and kept cool as much as possible during the procedure. The culture was centrifuged at 4 °C for 15 minutes at 3500 g. The supernatant was discarded, and the pellet resuspended in 20 ml ice cold sterile MilliQ-Water. The suspension was centrifuged again at 4 °C for 15 minutes at 3500 g. This was repeated once with 10 ml ice cold sterile MilliQ-Water. The supernatant was discarded, and the pellet resuspended in 7 ml ice cold sterile 10 % $glycerole_{aq}$. The suspension was centrifuged again at 4 °C for 20 minutes at 4000 rpm. The supernatant was discarded, and the pellet resuspended in 2.5 ml ice cold sterile 10 % $glycerole_{aq}$. Finally, the suspension was aliquoted in volumes of 25 µl, frozen in liquid nitrogen and stored at -80 °C.

5.2.1.6. Transformation by Heat Shock

Depending on purity and concentration of the plasmid solution up to 5 µl (usually 0.1 µl if purified and concentrated plasmid at >100 ng/µl had been used and 2.5 µl if crude plasmid solutions were used) were very carefully mixed with 25-50 µl bacterial suspension (GH317 *E.coli* for plasmid amplification) or (BL21 *E.coli* for protein expression) in a 1.5 ml micro centrifuge tube and cooled on ice for 30 minutes.

After a heat shock at 42 °C for maximal 45 s, the tube was immediately mixed with 1 ml tempered SOC medium (37 °C). Afterwards the suspension was incubated for 1 h at 37 °C whilst shaking at 600 rpm. Finally, up to 500 µl of the suspension (usually below 50 µl if purified and concentrated plasmid had been used and 100 µl if crude plasmid solutions were used) was plated onto LB-Agar supplemented with antibiotics in plastic petri dishes using glass beads and incubated at 37 °C overnight.

5.2.1.7. Transformation by Electroporation

Depending on purity and concentration of the plasmid solution, 1µl of 10 – 100 ng of DNA plasmids were resuspended in a 25 µl cell aliquot on ice. The cell suspension was transferred to a pre-cooled electroporation cuvette and transformed with 1800 V in an electroporator. Cells were resuspended with 1 ml pre-warmed SOC medium (37 °C) and incubated in a Thermomixer at 37 °C, 1400 rpm shaking for 1 h. 10-500 µl of the suspension was plated on LB-agar supplemented with the appropriate antibiotics in plastic petri dishes and incubated overnight at 37 °C.

5.2.1.8. Cassette Mutagenesis

4 µM complementary oligonucleotide pairs bearing the mutation were hybridized in 100 µl 1x NEB2 at 95 °C for 10 minutes. If necessary, the oligonucleotides were extended with 5 Units (Un) Klenow fragment (3'→5' exo-) and 3 µl of 10 mM dNTPs for 2 h at 37 °C and purified with the PCR purification kit according to the manufacturer's protocol.

Then, 60 µl of paired oligonucleotides (>150 ng/µl) were digested in 1x CutSmart buffer with 15 Un of NcoI and XhoI. The reaction was incubated for 1 h at 37 °C before stopping the reaction at 80 °C for 20 minutes. The reaction was purified with the PCR purification kit. Similarly, 60 µl of pMV (>100 ng/µl) were digested and purified.

Digested vectors and paired oligonucleotides were ligated in 10 µl 1x T4 DNA ligase reaction buffer using 400 U T4 DNA ligase for 4 hours at 16 °C. The reaction was stopped by heating the mixture to 65 °C for 10 minutes.

5 µl of the crude reaction mixture was transformed into GH317 and plated onto LB-Agar supplemented with tetracycline and incubated overnight. Single colonies were picked, inoculated in 5 ml LB-Medium supplemented with tetracycline and grown overnight. Plasmids were isolated, purified, checked for correct sequence and stored at -20 °C.

5.2.1.9. Site-directed Mutagenesis (SDM) by PCR

Primer containing the amber stop codon targeting the site of mutation were designed according to the guidelines of standard QuickChange protocols with a short overlap, moderate melting temperature (Tm) and GC content, if possible.

The PCR reactions were performed as indicated below. Next, samples were digested without any further purification with 10 Un DpnI at 37 °C for 1 h. 10 µl of this reaction mixture were transformed into 50 µl chemically competent *E.coli* GH371 cells. After growing single colonies in liquid overnight cultures, plasmids were isolated, purified, checked for correct sequence and stored at -20 °C.

Table 14: SDM-PCR reagent concentrations in 25 µl

Reagent	Volume
Template (100 ng/µl)	1 µl
Primer, forward (10 µM)	0.5 µl
Primer, reverse(10 µM)	0.5 µl
Pfu buffer (10x)	2.5 µl
dNTPs (10mM each)	0.5 µl
Sterile water	19.5 µl
Pfu polymerase (2.5 Un/µl)	0.5 µl

Table 15: SDM-PCR reaction times and temperatures

T	t	Cycles
95 °C	60 s	1
95 °C	30 s	18
55 °C	30 s	
68 °C	600 s	
68 °C	600 s	1

5.2.1.10. Agarose Gel Electrophoresis

First, Agarose was dissolved in 0.5x TBE Buffer. The solution was heated until boiling, then kept at room temperature whilst stirring. When the solution had cooled down but was still liquid, it was poured into the gel casting chamber and a comb was added for sample well creation. After solidification, the gel was transferred to the running chamber filled with 0.5x TBE buffer. Then, 12 µl of samples with loading dye were

pipetted into the sample wells and the electrophoresis was performed at 5 V/cm for 90 minutes. Afterwards the gel was stained with ethidium bromide and DNA bands were visualized by UV light.

5.2.1.11. Colony PCR

Single colonies were inoculated in 20 µl of sterile water and kept on ice. The PCR reaction was pipetted on ice, with the polymerase being added last. Colony PCR was performed as indicated below. Product formation was analyzed via agarose gel electrophoresis with a DNA ladder as a reference.

Table 16: Colony PCR reagent concentrations in 25 µl

Reagent	Volume
Bacterial colony suspension	1 µl
Primer, forward (10 µM)	0.25 µl
Primer, reverse(10 µM)	0.25 µl
Thermo Pol buffer (10x)	2.5 µl
dNTPs (10mM each)	0.625 µl
Sterile water	20.2 µl
Taq polymerase (5 Un/µl)	0.2 µl

Table 17: Colony PCR reaction times and temperatures

T	t	Cycles
95 °C	120 s	1
95 °C	30 s	35
55 °C	30 s	
68 °C	180 s	
68 °C	300 s	1

5.2.1.12. TALE Assembly via Golden Gate Reaction

pFUS vectors were assembled according to the target sequence for CDKN2A, BRCA1a, BRCA1b and BRCA1c as indicated below.

Canonical and mutated pMV for the respective CDKN2A, BRCA1a and BRCA1b sequence position 1-10 were assembled into pFUSA30A whereas pMV for BRCA1c sequence position 1-10 were assembled into pFUSA. CDKN2A, BRCA1a and BRCA1b

sequence position 11-20 were assembled into pFUSA30B. CDKN2A, BRCA1a and BRCA1b sequence position 21-24 were assembled into pFUSB4 whereas pMV for BRCA1c sequence position 11-17 or sequence position 11-18 into pFUSB7 and pFUSB8, respectively.

After Golden Gate 1 reaction, all non-circular plasmid DNA was removed by adding 5 nmol ATP and 5 Un Plasmid-safe Nuclease and incubating the mixture for 1h at 37°C.

10 µl of Golden Gate reaction mixes were directly transformed into GH317 *E.coli* and plated onto LB-Agar plates supplemented with spectinomycine and 50 µg/ml x-Gal for blue-white screening.

Four white colonies from the previous step were used to perform a colony PCR to check for correct insertion of pMV fragments into pFUS vectors with agarose gel electrophoresis.

One correct colony per pFUS vector was used for overnight culture. Plasmid were isolated, purified and sequenced.

Table 18: Golden Gate 1 reagent concentrations in 25 µl

Reagent	Volume
Each pMV (150 ng/µl)	1 µl
pFUS vector (75 ng/µl)	1 µl
T4 ligase buffer (10x)	2.5 µl
BsaI (10 Un/µl)	1 µl
T4 DNA ligase (400 Un/µl)	1 µl
Sterile water	to 25 µl

Table 19: Golden Gate 1 reaction times and temperatures

T	t	Cycles
37 °C	300 s	15
16 °C	600 s	
50 °C	300 s	1
80 °C	300 s	1

Next, pFUS vector for CDKN2A, BRCA1a, BRCA1b and BRCA1c were assembled with their specific original or mutated pLR vectors into pAnI521.

After Golden Gate 2 reaction, 10 µl of reaction mixes were directly transformed into GH317 *E.coli* and plated onto LB-Agar plates supplemented with carbenicillin and 50 µg/ml x-Gal for blue-white screening.

Six white colonies were picked for colony PCR to check for correct insertion of pFUS and pLR fragments into the pET TEV with agarose gel electrophoresis.

Two colonies with the correct insert length were inoculated in 5 ml LB-Medium supplemented with carbenicillin and incubated at 37 °C overnight. Plasmids were isolated, purified and sequenced.

Table 20: Golden Gate 2 reagent concentrations in 25 µl

Reagent	Volume
Each assembled pFUS vector (150 ng/µl)	1 µl
pLR vector (150 ng/µl)	1 µl
T4 ligase buffer (10x)	2.5 µl
BsmBI (10 Un/µl)	1 µl
T4 DNA ligase (400 Un/µl)	1 µl
Sterile water	to 25 µl

Table 21: Golden Gate 2 reaction times and temperatures

T	t	Cycles
37 °C	600 s	10
16 °C	900 s	
37 °C	900 s	1
80 °C	300 s	1

5.2.1.13. SDS-PAGE

Protein samples were mixed with SDS loading buffer and were denatured for 5 minutes at 95 °C. The samples were allowed to cool to room temperature and were loaded on an SDS polyacrylamide gel with protein marker as a reference. SDS gels were run for 1 h 15 minutes at 120 V. After electrophoresis, the gels were washed with water and stained in Comassie solution for 1 h on a shaker. The gels were developed first in destaining solution for 2 h, then in a diluted aqueous destaining solution overnight. Gel images were scanned on the CanoScan 9000F.

5.2.1.14. TALE Expression

For expression of standard TALE proteins, single clones of E. coli BL21(DE3) Gold transformed with a TALE expression plasmid were grown in LB media supplemented with 50 µg/ml carbenicillin at 37 °C overnight. These cultures were diluted 50-fold into LB medium supplemented with the same antibiotic and incubated at 37 °C and 200 rpm shaking until an OD_{600} of 0.4 was reached. IPTG was added to a final concentration of 0.2 mM and the cultures were further incubated under the same conditions. Then, cells were pelleted at 4 °C for 20 minutes at 4000 rpm, the supernatant discarded and if necessary frozen overnight.

The pellets were lysed in lysis-buffer (10 mM Tris-HCl, 300 mM NaCl, 2.5 mM $MgCl_2$, 0.1 % Triton X-100, pH = 9) containing 1 mM PMSF and 50 µg/ml by shaking at room temperature at 1400 rpm for 30 minutes. The suspension was pelleted by centrifugation and the supernatant was collected and extracted with Ni-NTA. Ni-NTA beads were washed two times with 4x PBS-buffer (0.55 M NaCl, 43 mM KCl, 69 mM $Na_2HPO_4 \cdot 2\ H_2O$, 24 mM KH_2PO_4, pH = 8), four times with wash buffer (50 mM $NaH_2PO_4 \cdot H_2O$, 300 mM NaCl, pH = 8) containing 20 mM imidazole and once with wash buffer containing 50 mM imidazole. The protein was eluted three times or until all protein was visibly eluted from the Ni-NTA with wash buffer containing 500 mM imidazole. Pooled elution fractions were added to a dialysis tube and dialyzed against TALE storage buffer (20 mM Tris-HCl, 200 mM NaCl, 10 % glycerol, 1 mM DTT, pH = 7.5). For CD-spectroscopy, TALEs were dialyzed against a different TALE storage buffer (20 mM Tris-HCl, 50 mM NaCl, pH = 7.5). Purity of the TALE protein was analyzed with SDS-PAGE. Concentrations were quantified with a BCA assay according to the manufacturer's protocol. The proteins were snap-frozen and stored in aliquots at −80 °C.

For Expression of TALEs bearing *p*AcF, TALE plasmids were co-transformed with plasmid pDaS71 (encoding for *p*AcF tRNA/RS; under control of araBAD promotor) into *E. coli* BL21 DE3 gold cells and grown at 37 °C on carbenicillin/chloramphenicol LB agar plates. Single colonies were inoculated into 5 ml LB medium with carbenicillin/chloramphenicol and grown at 37 °C overnight. Overnight culture was added to 250 mL LB medium with carbenicillin/chloramphenicol containing 1 mM of *p*AcF. Cultures were grown at 37°C whilst shaking at 180 rpm and until an OD_{600} of 0.4 was reached. Then, expression of the tRNA synthetase/tRNA pair was induced by

adding arabinose to a final concentration of 0.02 %. After an hour, TALE expression was induced by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.2 mM. The expression was continued overnight. Then, cells were pelleted at 4 °C for 20 minutes at 4000 rpm and the supernatant discarded.

The pellet was re-suspended in lysis-buffer (10 mM Tris-HCl, 300 mM NaCl, 2,5 mM MgCl2, 5 % DMSO, 0.2 % sodium lauroyl sarcosinate, 0.1 % Triton X-100, pH = 9) containing 1 mM PMSF and 50 µg/mL lysozyme. The suspension was lysed by repeated sonication on ice (2 runs of 3 minutes; 20 % amplitude; 4s on, 2s off), letting samples cool on ice for 1 minute between runs. Cell debris was pelleted by ultracentrifugation and the supernatant was collected and extracted with Ni-NTA. Ni-NTA beads were washed in 4 steps with 25 times the bead volume with wash buffer (50 mM NaH_2PO_4·H2O, 300 mM NaCl, pH = 8) containing 5 mM, 10 mM, 20 mM imidazole and 50 mM imidazole. The protein was eluted three times or until all protein was visibly eluted from the Ni-NTA with wash buffer containing 500 mM imidazole. Pooled elution fractions were added to a dialysis tube and dialyzed against TALE storage buffer (20 mM Tris-HCl, 200 mM NaCl, 10 % glycerol, 1 mM DTT, pH = 7,5). Purity of the TALE protein was analyzed with SDS-PAGE. Concentrations were determined with a BCA assay according to manufacturer's instructions. The proteins were snap-frozen and stored in aliquots at −80 °C.

5.2.1.15. Circular Dichroism Spectroscopy

Circular Dichroism (CD) spectra were recorded on a J-715 spectropolarimeter between 190 nm and 260 nm with a scanning speed of 50 nm/min and a band-width of 1 nm in a cylindrical quartz cell with a path length of 0.01 cm. The spectra represent the average of ten scans for 350 µl of samples with a concentration of 1 mg/ml, which were corrected for the baseline of a buffer control and normalized for the molecular concentration of the sample in the Spectra Manager Suite. Obtained values were plotted in OriginPro.

5.2.1.16. Electromobility Shift Assay

Oligonucleotide pairs for TALE binding (see SI) were hybridized in TALE binding buffer (20 mM Tris, 50 mM NaCl, 5 mM $MgCl_2$ and 5 % v/v glycerol, pH = 8) by incubating at 95 °C for 5 minutes, followed by 30 minutes at room temperature. Then, defined amounts of DNA duplex were incubated with the respective TALE protein in TALE binding buffer with a final volume of 10 µl. The reaction was kept at room temperature

for 60 minutes and then at 4 °C for 30 minutes. The native polyacrylamide gel (0.5 x TAE buffer, 8 % Rotiphorese gel 40, 0.1 % APS and 0.01 % TEMED) was pre-run at 4 °C for 30 minutes in a Mini Protean vertical electrophoresis cell at 70 V. The gel was loaded with 7,5 µl sample and then run for 90 minutes at 4 °C and 70 V. GFP fluorescence of the gels was recorded with a Typhoon FLA-9500 laser scanner and analyzed using the software ImageJ or ImageQuant TL 8.1.

5.2.1.17. Structure Modelling

The modelling workflow was performed using MOE 2016.08[378] and Chimera.[379] As starting point, the crystal structure PDB 4GJR (chains A, I and J) was used. The base pairs of the DNA strand were mutated to the corresponding base pairs of the used DNA sequence using Chimera. After that, the respective cytosine modifications were added to the DNA in MOE, resulting in five different DNA models (C, 5mC, 5hmC, 5fC, 5caC). After that, one homology model was created for each combination of protein variant and C modification using the homology modelling module of MOE. Here, the option "include selected atoms as environment for induced fit option" was enabled to transfer the modified DNA to the homology model and to consider it in the modelling process. 25 main chain models with 3 sidechain samples each were created and refined using the "Fine" setting with an RMS gradient of 0.1. Apart from that, standard settings were used with the default force field Amber10:EHT.[380,381]

For analyzing the impact of further modifications at the 5fC and 5caC base, the respective modifications were introduced using the "Build Molecule" functionality in MOE. To get reasonable structures of these modifications that do not affect binding, the modified base and residues within a 4.5 Å radius were minimized using MOE default settings.

5.2.1.18. Luciferase Assay

HEK293T cells were maintained in DMEM media supplemented with 1 % penicillin/streptomycin, 10 % FBS and 1 % L-glutamine. 1.6 x 104 cells were cultured in a 96 well plate overnight prior transfection. Opti MEM and Lipofectamin 2000 were mixed according to the manufacturer's protocol. 25 ng of the luciferase reporter plasmid (encoding the TALE binding site and a minCMV promoter upstream of a firefly luciferase gene) and 175 ng of the TALE plasmid (encoding the TALE-VP64 fusion constructs; for plasmid maps, see the SI Fig. 13). Transfection mix was added to each plasmid pair and incubated for 20 minutes at RT. The solution was added to wells of

the 96 well plate and incubated at 37 °C and 5 % CO_2 for 48 hr. Each well was then washed with 20 µl of DPBS. 40 µl of lysis buffer (100 mM NaH_2PO_4 and 0.2 % Triton X-100) was added to each well and mixed vigorously. The plate was then incubated on ice for 20 min. After incubation, 20 µL of the lysis solution from each well was combined with each 90 µL of Bright-Glo in a second 96 well plate. The plate was quickly spun down and the luminescence was immediately measured on a TECAN M1000 plate reader (wavelength 380-600 nM). Ratio of luminescence from each sample to that of a sample transfected with TALE-VP64 and the luciferase reporter plasmid bearing the CDKN2A_TALE binding_C6 was plotted as relative luminescence. The error bars represent standard errors from duplicate experiments.

5.2.1.19. Phenol/Chloroform Extraction and Ethanol Precipitation of DNA

An equal volume of a 25:24:1 mixture of phenol/chloroform/isoamylalcohol (pH = 7.5-8.0) was added to at least 100 µl of sample. The mixture was vortexed vigorously for 15 s at max speed. Then, the dispersion was shortly spun down. The upper aqueous phase was transferred to a fresh microcentrifuge tube. The extraction was repeated once. The extracted aqueous phase without any organic phase or precipitate from the interphase was mixed thoroughly with 0.1x fold of the sample volume 3M sodium acetate (pH = 5.2). Then, 3x fold of ice-cold absolute ethanol was added to the sample. The solution was mixed by inverting and kept at -20 °C overnight. The next day, the solution was centrifuged at full speed for 30 minutes at 4 °C to pellet the precipitated nucleic acid. The solution was carefully removed to not disturb the pellet. Next, the pellet was washed with 100 µl of ice-cold 70 % $ethanol_{aq}$ and the sample centrifuged for 5 minutes at full speed at 4 °C. The supernatant was carefully removed and the washing step repeated again. Finally, the pellet was dried in a vacuum concentrator and then taken up in nuclease-free water or sterile TE-buffer.

5.2.1.20. Enzymatic Glucosylation

500 pmol of dsDNA containing 5hmC was incubated for 18 h at 37 °C with 30 units T4 BGT and 800 nmol UDP-glucose in 50 µl of NEB4 (50 mM KOAc, 20 mM Tris-HOAc, 10 mM $Mg(OAc)_2$, 1 mM DTT, pH = 7.9). The reaction was stopped by incubation at 65 °C for 10 minutes. The DNA was purified using standard phenol-chloroform-extraction and ethanol precipitation. Remaining contaminants were removed by using the Nucleotide Removal Kit according to the manufacturer's protocol. The concentration and purity of the DNA solution was determined with the Nanodrop 2000.

5.2.1.21. Preparation of Oligonucleotides for MALDI-TOF MS

Oligonucleotides were hybridized in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH = 7.4) by incubating at 95 °C for 5 minutes, followed by 30 minutes at room temperature. The dsDNA was then digested for 1 h with SacI in order to improve the resolution of the DNA fragments in mass spectrometry. After phenol-chloroform-extraction and ethanol precipitation, the DNA was dissolved in nuclease-free water and used for subsequent treatment. After the reaction, the sample was again purified as described and dried. The pellet was dissolved in 10 µl of an aqueous solution of 0.1 % TFA and 1 % Acetonitrile. The sample was aspirated into an equilibrated C18 tip and rinsed with 10 µl of an aqueous solution with 0.1 % TFA and 5 % acetonitrile according to the manufacturer's instructions. The sample was eluted by slowly aspirating 2 µl of a solution of 50 mg/mL 3-hydroxypicolinic acid in 50 % $acetonitrile_{aq}$ / 0.1 % TFA_{aq} and 10 mg/mL diammonium hydrogen citrate and slowly dispensing the complete volume onto a position on the ground steel MTP 384 target. Spectra were recorded in linear positive mode on a Bruker ultrafleXtreme MALDI.TOF/TOF system and analyzed using flexControl and mMass.

5.2.1.22. Crosslinking Efficiency Analysis

Oligonucleotides were 5´-^{32}P-end labeled using γ-^{32}P-ATP and T4 polynucleotide kinase and purified using a G-25 gel filtration column as per manufacturer's instructions. 0.5 pmol of the labeled oligonucleotide was diluted into 100 pmol of the same, but unlabeled oligonucleotide. The mix was hybridized with 125 pmol of a fully reverse, complementary oligonucleotide in 1 x TE buffer at 95 °C for 5 minutes. After 30 minutes at room temperature, the DNA was cross-linked to 100 pmol of TALE protein as described above. After the second reaction step, 15 µl of 4x SDS-PAGE buffer were added to the sample and the mixture was incubated at 95 °C for 15 minutes. After letting the sample cool to ambient temperature, they were loaded on a denaturing 12 % SDS-polyacrylamide gel and run in 1 x running buffer (250 mM Tris-HCl (pH = 8.3), 1.918 M glycine, 1 w/v % SDS) at 120 V until the loading dye front had migrated to the bottom. Gels were carefully dried and incubated on a storage phosphor screen. Data on the screen was read out using the Typhoon FLA-9500 laser scanner and analyzed using the software ImageQuant TL 8.1.

5.2.1.23. Preparation of Spike-In DNA

To introduce single site cytosine modifications in the CDKN2A-TALE binding site, 421 bp long spike-ins containing either 5hmC, 5fC or 5caC at position 5 of the TALE binding site were obtained by PCR on amplicons derived from HeLa cell gDNA. First, amplicons from original HeLa cell gDNA were produced with primer pair o2129/o2130 via Template PCR. This fragment was used without further purification as template for Pre-Spike-In PCR with primer pair o2129/o2135. The resulting PCR product was purified using the GeneJet PCR Purification Kit according to manufacturer's instructions and subsequently used as template for introducing a single 5hmC, 5fC, or 5caC at position 6 within the target TALE binding sequence through primer pair o1593 (for 5hmC), o1594 (for 5fC) or o1595 (for 5caC) and o1596 for the Spike-in PCR.

To introduce single site cytosine modifications in the BRCA1-TALE binding site, 421 bp long spike-ins containing either 5hmC, 5fC or 5-caC at position 5 of the TALE binding site were obtained by PCR on amplicons derived from HeLa cell DNA. First, amplicons from original HeLa cell gDNA were produced with primer pair o2131/o2132 via template PCR. This fragment was used for Pre-Spike-In PCR without further purification as template for primer pair o2132/o2136. The resulting PCR product was purified using the Monarch PCR & DNA Cleanup kit according to manufacturer's instructions and subsequently used as template for introducing a single 5hmC or 5fC at position 13 within the target TALE binding sequence through primer pair o1521 (for 5hmC) or o1524 (for 5fC) and o1529for the Spike-in PCR. All amplicons were purified by agarose gel electrophoresis and the sequence was verified by Sanger sequencing.

Table 22: PCR reagent concentrations for the preparation of spike-in DNA in 50 µl

	Template PCR	**Pre-Spike-In PCR**	**Spike-In PCR**
Reagent	**Volume**	**Volume**	**Volume**
Template	1 µl (~100 ng/µl)		1 µl (<10 ng/µl)
Primer, forward (10 µM)	2.5 µl		5 µl
Primer, reverse (10 µM)	2.5 µl		5 µl
Q5 buffer (5x)	10 µl		10 µl
dNTPs (10 mM)	1 µl		1 µl
DMSO	1.5 µl		1.5 µl
Water, nuclease-free	31 µl		26 µl
Q5 polymerase (2 Un/µl)	0.5 µl		0.5 µl

Table 23: PCR settings for the preparation of spike-in DNA

Template PCR			Pre-Spike-In PCR			Spike-In PCR		
T	**t**	**Cycles**	**T**	**t**	**Cycles**	**T**	**t**	**Cycles**
98 °C	30 s	1	98 °C	120 s	1	98 °C	120 s	1
98 °C	30 s	35	98 °C	30 s	35	98 °C	20 s	20
66 °C	30 s		66 °C	60 s		68 °C	30 s	
72 °C	300 s		72 °C	600 s		72 °C	120 s	
72 °C	600 s	1	72 °C	120 s	1	98 °C	20 s	40
						72 °C	180 s	
						72 °C	600 s	1

5.2.1.24. Purification of DNA via Agarose Gel Electrophoresis

3 % agarose gels were prepared as described above, but with big sample wells. The full sample was loaded, run and stained as described above. The gel was illuminated with UV light as shortly as possible and position of single bands corresponding to the PCR amplicon were marked. Gel fragments were cut out, dissolved and the DNA recovered via commercial kits as per manufacturer's instructions. The concentration and purify was determined on a Nanodrop 2000. Low purity could be improved via standard phenol/chloroform extraction and ethanol precipitation. A higher amount of DNA could be obtained by using the purified sample in another round of PCR with the respective primer and purifying the amplicons with a PCR purification kit as per manufacturer's instruction.

5.2.1.25. Preparation of Genomic DNA (for chapter 3.1.3)

Human gDNA (male Yoruban individual, Encode sample NA18507) was obtained from the Coriell Institute. For erasing epigenetic cytosine modifications, gDNA samples were whole genome amplified and purified using the QIAamp DNA mini kit according to manufacturer's instructions. gDNA samples were sheared by sonication (20 % power, 8 cycles of 30 s on / 30 s off), resulting in a broad fragment size distribution around 500 bp.

Respective amount of amplicon were spiked into non-methylated WGA gDNA that had been digested with FatI, BanII and EarI, all cutting within the binding sites of qPCR primers, and purified by the Monarch PCR & DNA cleanup kit according to the manufacturer's instructions.

5.2.1.26. Quantitative PCR (qPCR)

Quantitative PCR on the CFX384 Touch real-time PCR detection system was performed with the 2x GoTaq master mix. First, 4µl of (diluted) sample was pipetted in a PCR workstation into a 384 well plate. Then, 5 µl of master mix was added to 0.5 µl of each primer and mixed carefully. This mixture was pipetted to the sample into the 384 well plate without mixing. The plate was tightly sealed with clear qPCR seals. The plate was centrifuged at room temperature for 1 minute at 2500 rpm and transferred to the qPCR cycler.

Table 24: qPCR reaction times and temperatures

T	t	Cycles
95 °C	180 s	1
95 °C	15 s	45
65 °C	60 s	
95 °C	30 s	1
65 °C (0.5°/cycle)	5 s	60

5.2.1.27. Enrichment (for chapter 3.1.3)

Before each enrichment, Ni-NTA magnetic agarose beads were freshly washed/equilibrated with enrichment buffer (150 mM NaCl, 30 mM Tris-HCl, 5 mM MgCl2, 0.05 % Tween20, 0.5 mg/ml BSA, pH = 7.9) at room temperature using a magnetic stand with keeping the overall volume constant. To prepare the beads for enrichment, 5 µl of the bead suspension were added to 25 µl of 200 nM TALE protein. After addition of 20 µl enrichment buffer, the suspension was incubated in a thermomixer at room temperature for 30 minutes whilst shaking at 600 rpm.

Then, 500 ng of sheared gDNA (with the respective amount of spike-in) were added and the volume of the mixture was adjusted to 500 µl with enrichment buffer. The samples were first incubated in a thermo-mixer at room temperature, and then cooled to 4 °C for 30 minutes while shaking at 600 rpm, each. The tubes were placed on the magnetic stand for 2 minutes and the liquid was carefully removed without disturbing the beads. The beads were washed once with 500 µl of pre-chilled enrichment buffer by slowly pipetting up and down once. The liquid was completely removed and 500 µl of nuclease-free water was added. The tube was shaken at 1400 rpm at 95 °C for 15 minutes. The tubes were cooled down to room temperature, spun shortly (2-3 s) and

placed on the stand. The supernatant was collected in a fresh tube, dried in a vacuum concentrator and re-suspended with 20 µl of DNA-, nuclease-free water. qPCRs were conducted with 5 µl GoTaq qPCR Master Mix, 0.5 µl of each 10 µM primer and 4 µl of the sample as template in the CFX384 Touch real-time PCR detection system. Copy number quantification for enriched gDNA was done by linear regression from serial dilutions of the respective gDNA with each spike-in DNA.

5.2.1.28. Preparation of Genomic DNA (for chapter 3.2.3)

Human gDNA (male Yoruban individual, Encode entry NA18507) was purchased from the Coriell Institute. For erasing epigenetic cytosine modifications, gDNA samples were whole genome amplified using the illustra Ready-To-Go GenomiPhi V3 DNA Amplification kit and purified using the Monarch PCR & DNA Cleanup kit according to the manufacturer's instructions.

Genomic DNA from mouse embryonic stem cells (mESC) was isolated from J1 mESC (ATCC SCRC-1010). J1 cells were cultured on gelatin-coated plates using KnockOut DMEM supplemented with 10 % of EmbryoMAX ES cell qualified FBS, 103 U/ml of mouse recombinant LIF, 2 mM of GlutaMAX, 50 µM 2-mercaptoethanol, 1x Pen/Strep and 1x non-essential amino acids. Cells were split every 2-3 days using 0.25 % Trypsin/EDTA. Samples were collected by trypsinization, washed with DPBS once and harvested by centrifugation for 10 minutes at 400x g. Isolation of genomic DNA from mESC was adapted from Booth et al..[78] J1 cells were lysed overnight at 55 °C in 200 µl lysis buffer (100 mM Tris, pH 5.5, 5 mM EDTA, 200 mM NaCl, 0.2 % SDS, 400 µg/ml proteinase K, 200 µg/ml RNase A). After standard phenol-chloroform extraction, the DNA was precipitated in ice-cold 70 % ethanol and washed two times using ice-cold 70 % ethanol. The DNA was dried for 15 minutes in a vacuum concentrator and dissolved in nuclease free water by incubating the solution for 2 h at 65 °C and overnight at 4 °C.

gDNA samples were sheared by sonication (8 cycles of 30 s on / 30 s off), resulting in a broad fragment size distribution.

Respective amount of amplicon was spiked into original, non-methylated WGA DNA that was previously digested with FokI and SacI and purified by the Monarch PCR & DNA Cleanup kit according to the manufacturer's instructions.

5.2.1.29. Enrichment (for chapter 3.2.3)

Per sample, 10 µl Dynabeads His-Tag Isolation and Pulldown were equilibrated in buffer EB2 at room temperature using the MagRack 6 magnetic stand with keeping the overall volume constant and added to the sample. The suspension was rotated at 15 rpm for 30 minutes at 4°C in a tube revolver/rotator.

The tubes were placed on the magnetic stand for 2 minutes and the liquid was carefully removed without disturbing the beads. The beads were then incubated for 5 minutes with 500 µl of pre-chilled buffer EWB (2 M NaCl, 30 mM Tris-HCl, 5 mM $MgCl_2$, 10 µM Spermine, 0,005 % Tween20, 0.5 mg/mL BSA, pH 7.9). After removing the supernatant, the beads were washed twice with 500 µl of pre-chilled buffer EB1 (150 mM NaCl, 30 mM Tris-HCl, 5 mM $MgCl_2$, 0,005 % Tween20, 0.5 mg/ml BSA, pH = 7.9) by slowly pipetting up and down twice while keeping the tubes in the magnetic stand the whole time. Between each washing step the beads were let to rest for 2 minutes on the stand. After taking off the supernatant in the last washing step, the beads were shortly spun down (1000x g, 5 s), collected again at the magnet and any remaining buffer was completely removed.

The beads were taken up in 50 µl DNA-, nuclease-free water and the suspension in the tube was shaken at 1400 rpm at 95 °C for 13 minutes, then shortly spun down (1000x g, 5 s) and incubated for 2 minutes at 95°C. The tubes were then immediately placed on the stand, the beads collected on the magnet and the supernatant was transferred to a fresh tube. qPCRs were conducted with 5 µl GoTaq qPCR Master Mix, 0.5 µl of primer pair o2139/o2143 (10 µM each) and 4 µl of the sample as template in the CFX384 Touch real-time PCR detection system. Copy number quantification for enriched gDNA was done by linear regression from serial dilutions of the respective gDNA with each spike-in DNA.

5.2.2. Chemical Methods

5.2.2.1. Oxime Formation

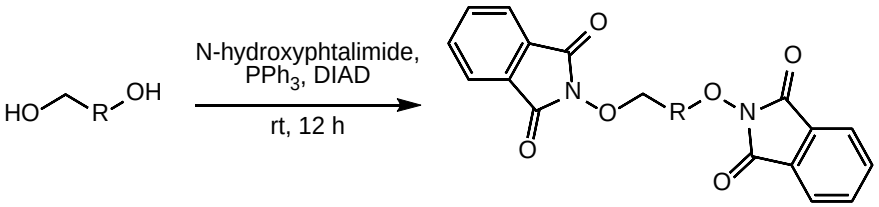
In 20 µl of reaction buffer (50 mM $NaHPO_4/Na_2HPO_4$, pH = 6.0), 200 pmol DNA containing 5fC, 2 µmol 1,4-benzenediamine (pPDA) and 2 µmol of the hydroxylamine were combined and mixed thoroughly. The sample was shaken and incubated in the dark at room temperature overnight. To remove excessive reagents, the DNA was precipitated through standard ethanol precipitation, pelleted, dried and taken up in water. Remaining contaminants were removed by using the Nucleotide Removal Kit according to the manufacturer's protocol. The concentration of the DNA solution was

determined with the Nanodrop 2000. The resulting UV spectrum, 260/280 nm and 260/230 nm ratio values confirmed the purity of the sample.
For oxime ligation of the amplicons spiked into bgDNA, the protocol was carried out as described above but with 0.2 µmol pPDA. Furthermore, the DNA was purified with the Monarch PCR & DNA cleanup kit (NEB). The concentration of the gDNA samples was quantified by qPCR and interpolation on a calibration curve of the untreated gDNA with spike-in DNA. The concentration was adjusted to 200 ng/µl.

5.2.2.2. Amide Formation

200 pmol dsDNA containing 5caC in 72 µl of 50 mM MES buffer (pH = 8.0) was treated with 5 µl of 0.6 M EDC-HCl_{aq} and incubated for 15 minutes at room temperature and shaking at 800 rpm. Then, 16 µl of a 1 M $NaCl_{aq}$ solution were added and the solution was adjusted to 120 µl with nuclease-free water.
The aqueous mixture was added carefully to a 30 µl solution containing 3 µmol PyAOP and 2.75 µmol of amine in DMSO. The mixture was shaken at 800 rpm and incubated for 24 h at room temperature. The reaction was quenched by addition of 20 µl of 3 M NH_4OAc (pH = 5.2) and incubation at room temperature for 30 minutes. The DNA was subsequently precipitated overnight by addition of 620 µl ice-cold absolute ethanol. The DNA was pelleted, dried and taken up in water. Remaining contaminants were removed by using the Nucleotide Removal Kit according to the manufacturer's protocol. The concentration of the DNA solution was determined with the Nanodrop 2000. The resulting UV spectrum, 260/280 nm and 260/230 nm ratio values confirmed the purity of the sample.
For 5caC-bearing amplicons spiked into bgDNA, 96 µl of a pre-activated solution containing 3 µmol EDC-HCl and 1.2 µg of gDNA in 50 mM MES, 125 mM NaCl (pH = 8) was added carefully to 4 µl of 0.3 µmol PyAOP and 0.275 µmol benzylamine in DMSO. The reaction mixture was incubated in a thermomixer for 24 h at room temperature and shaking 800 rpm. After 24 h, the reaction was quenched by adding 15 µl of 0.5 M ammonium acetate, purified with the Monarch PCR & DNA cleanup kit. The concentration of the gDNA samples was quantified by qPCR and interpolation on a calibration curve of the untreated gDNA with spike-in DNA. The concentration was adjusted to 200 ng/µl.

5.2.2.3. Linker Synthesis (modified from Holder, Francis 2007[377])



1a-e **2a-e**

A flame-dried, argon purged 100 mL round-bottom flask equipped with a magnetic stir bar was charged with N-hydroxyphthalimide (2.2 equiv.), triphenylphosphine (2.1 equiv.) and **1a-e** (1.1g, 1 equiv.). The flask was evacuated and purged with argon. The solids where dissolved in anhydrous dichloromethane (25 mL). After 15 minutes, the mixture was treated in small portions (~0.25 mL) with DIAD (2.2 equiv.), allowing the orange color to fade between each addition. After 12 h of stirring, the mixture was concentrated *in vacuo* and diluted with 100 mL of diethyl ether and stirred vigorously at 4 °C for 30 minutes. The white precipitate was collected via filtration and was recrystallized two times from diethyl ether and dried *in vacuo* to give **2a-e** used for hydrazinolysis without further purification.

2,2'-(ethane-1,2-diylbis(oxy))bis(isoindoline-1,3-dione) (2a): The product was obtained as an off-white solid (3.91 g, 10.97 mmol, 62 %).

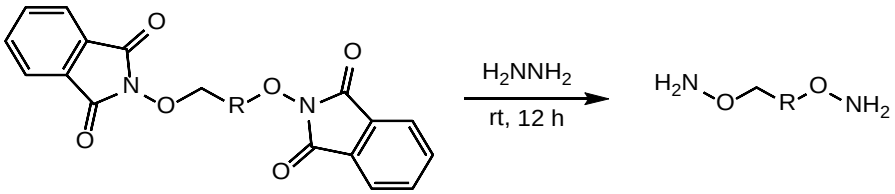
2,2'-((oxybis(ethane-2,1-diyl))bis(oxy))bis(isoindoline-1,3-dione) (2b): The product was obtained as an off-white solid (2.7 g, 6.82 mmol, 65 %).

2,2'-(((ethane-1,2-diylbis(oxy))bis(ethane-2,1-diyl))bis(oxy))bis(isoindoline-1,3-dione) (2c): The product was obtained as an off-white solid (2.6 g, 5.9 mmol, 79 %).

2,2'-((((oxybis(ethane-2,1-diyl))bis(oxy))bis(ethane-2,1-diyl))bis(oxy))bis(isoindoline-1,3-dione): The product was obtained as an off-white solid **(2d):** (1.54 g, 3.22 mmol, 62.5 %).

2,2'-((3,6,9,12,15-pentaoxaheptadecane-1,17-diyl)bis(oxy))bis(isoindoline-1,3-dione) (2e): The product was obtained as an off-white solid (1.81 g, 3.16 mmol, 89 %).

Synthesis of linkers L1-5 (3a-e) (modified procedure from Holder & Francis[1])



2a-e **3a-e**

A 100 mL round-bottom flask was charged with a solution of **2a-e** in anhydrous dichloromethane. The solution was treated in one portion with 2.2 eqv. of hydrazine hydrate and stirred vigorously at room temperature for 12 h. The resulting pasty off-white solid was removed by filtration. The clear eluent was diluted into diethyl ether and stirred vigorously for 1 h at 4°C.The resulting solid was removed by filtration. The filtrate was dried *in vacuo* and the resulting residue was taken up in chloroform. The solution was filtered again, and the filtrate was dried *in vacuo* yielding the desired product **3a-e** as clear oil.

L1: O, O'-(ethane-1,2-diyl)bis(hydroxylamine) (3a): The product was obtained as colorless oil (0.82 g, 8.9 mmol, 96 %).

^{1}H-NMR (400 MHz, $CDCl_3$): δ 5.51 (4H, s), 3.85 (4H, m),

^{13}C-NMR (101 MHz, $CDCl_3$): δ 73.88

LCMS (m/z): $[M+H]^+$ (calculated for $C_2H_8N_2O_2$) = 93.06585 u; $[M+H]^+$ (observed) = 93.06535.

L2: O, O'-(oxybis(ethane-2,1-diyl))bis(hydroxylamine) (3b): The product was obtained as yellowish oil (0.81 g, 5.95 mmol, 60 %).

^{1}H-NMR (400 MHz, $CDCl_3$): δ 5.47 (4H, br s), 3.85-3.82 (4H, m), 3.68-3.65 (4H, m)

^{13}C-NMR (101 MHz, $CDCl_3$): δ 74.87, 69.88

LCMS (m/z): $[M+H]^+$ (calculated for $C_4H_{12}N_2O_3$) = 137.09207 u; $[M+H]^+$ (observed) = 137.09184 u.

L3: O,O'-((ethane-1,2-diylbis(oxy))bis(ethane-2,1-diyl))bis(hydroxylamine) (3c): The product was obtained as colorless oil (0.6 g, 3.33 mmol, 56 %).

^{1}H-NMR (400 MHz, $CDCl_3$): δ 5.50 (4H, br s), 3.84-3.82 (4H, m), 3.69-3.66 (8H, m)

^{13}C-NMR (101 MHz, $CDCl_3$): δ 74.95, 70.66, 69.75

LCMS (m/z): $[M+H]^+$ (calculated for $C_6H_{16}N_2O_4$) = 181.11828 u; $[M+H]^+$ (observed) = 181.11807 u.

L4: O,O'-(((oxybis(ethane-2,1-diyl))bis(oxy))bis(ethane-2,1-diyl))bis(hydroxylamine) (3d): The product was obtained as colorless oil (0.58 g, 2.59 mmol, 81 %).

^{1}H-NMR (400 MHz, $CDCl_3$): δ 5.52 (4H, br s), 3.84-3.82 (4H, m), 3.69-3.64 (12H, m)

^{13}C-NMR (101 MHz, $CDCl_3$): δ 74.94, 70.73, 70.69, 69.77

LCMS (m/z): $[M+H]^+$ (calculated for $C_8H_{20}N_2O_5$) = 225.14450 u; $[M+H]^+$ (observed) = 225.14451 u.

L5: O,O'-(3,6,9,12,15-pentaoxaheptadecane-1,17-diyl)bis(hydroxylamine) (3e): The product was obtained as colorless oil (1.58 g, 5.06 mmol, 80 %).

^{1}H-NMR (400 MHz, $CDCl_3$): δ 5.51 (4H, br s), 3.84-3.82 (4H, m), 3.72-3.65 (20H, m)

^{13}C-NMR (101 MHz, $CDCl_3$): δ 74.94, 70.76, 70.75, 70.72, 70.70, 69.77

LCMS (m/z): $[M+H]^+$ (calculated for $C_{12}H_{28}N_2O_7$) = 313.19693 u; $[M+H]^+$ (observed) = 313.19700 u.

5.2.2.4. Crosslinking Reaction

In 20 µl of sodium phosphate buffer (20 mM, pH = 6.0) 100 pmol DNA containing 5fC, 2 µmol *para*-phenylenediamine (*p*PDA) (dissolved in 5 µl of 0.5 % acetic acid) and 10 µmol of Linker **L1-5** were combined and mixed thoroughly. The sample was shaken at 550 rpm and incubated at room temperature overnight. The next day, 100 pmol of TALE protein in 10 µl of TALE storage buffer was added to the reaction mix together with 5µl of freshly prepared *p*PDA-solution and 10 µl of 5x TDC buffer (20 mM Tris-HCl, 10 mM NaCl, 1.25 mM $MgCl_2$, 5 % glycerol, pH = 6). The reaction mix was filled up to 50 µl with nuclease-free water and the reaction was kept at room temperature for 6 h whilst shaking at 550 rpm. The reaction was stopped by quenching with 4x SDS-PAGE loading buffer (200 mM Tris-HCl (pH = 6.8), 8 w/v % SDS, 0.08 w/v % bromophenol blue, 40 v/v % glycerol, 1 v/v % 2-mercaptoethanol) or 100 mM Tris-HCl (pH = 9).

For 5fC-DNA spiked into a gDNA background, 1 µg of sheared, digested gDNA (either whole genome amplified from human gDNA or directly isolated from J1 mESC) with the respective amount of intact spike-in DNA was incubated in 20 µl of 20 mM phosphate buffer (pH = 6) overnight with 2 µmol of *p*PDA (dissolved in 5 µl 0.5 % acetic acid) and 10 µmol of linker at room temperature whilst shaking at 550 rpm. To remove excess of linker and oxidized catalyst, the reaction mix was purified using the Monarch PCR purification kit. The purified DNA was taken up in 50 µl of 1x TDC buffer (20 mM Tris-HCl, 10 mM NaCl, 1.25 mM $MgCl_2$, 5 % glycerol, pH = 6) and treated with 5 pmol

*p*PDA (dissolved in 5 µl 0.5 % acetic acid) and 10 pmol of TALE protein and incubated for 6 h at room temperature and shaking at 550 rpm.

To terminate the reaction, the sample was quenched with 18.5 µl of 100 mM Tris-HCl (pH 9) and purified via a G25-Sephadex column according to the manufacturer's instructions. The filtrate was buffered into 90 µl 1x EB2 (500 mM NaCl, 30 mM Tris-HCl, 5 mM $MgCl_2$, 100 µM Spermine, 0,005 % Tween20, 0.5 mg/mL BSA, pH = 7.9).

6. Supplementary Data[f]

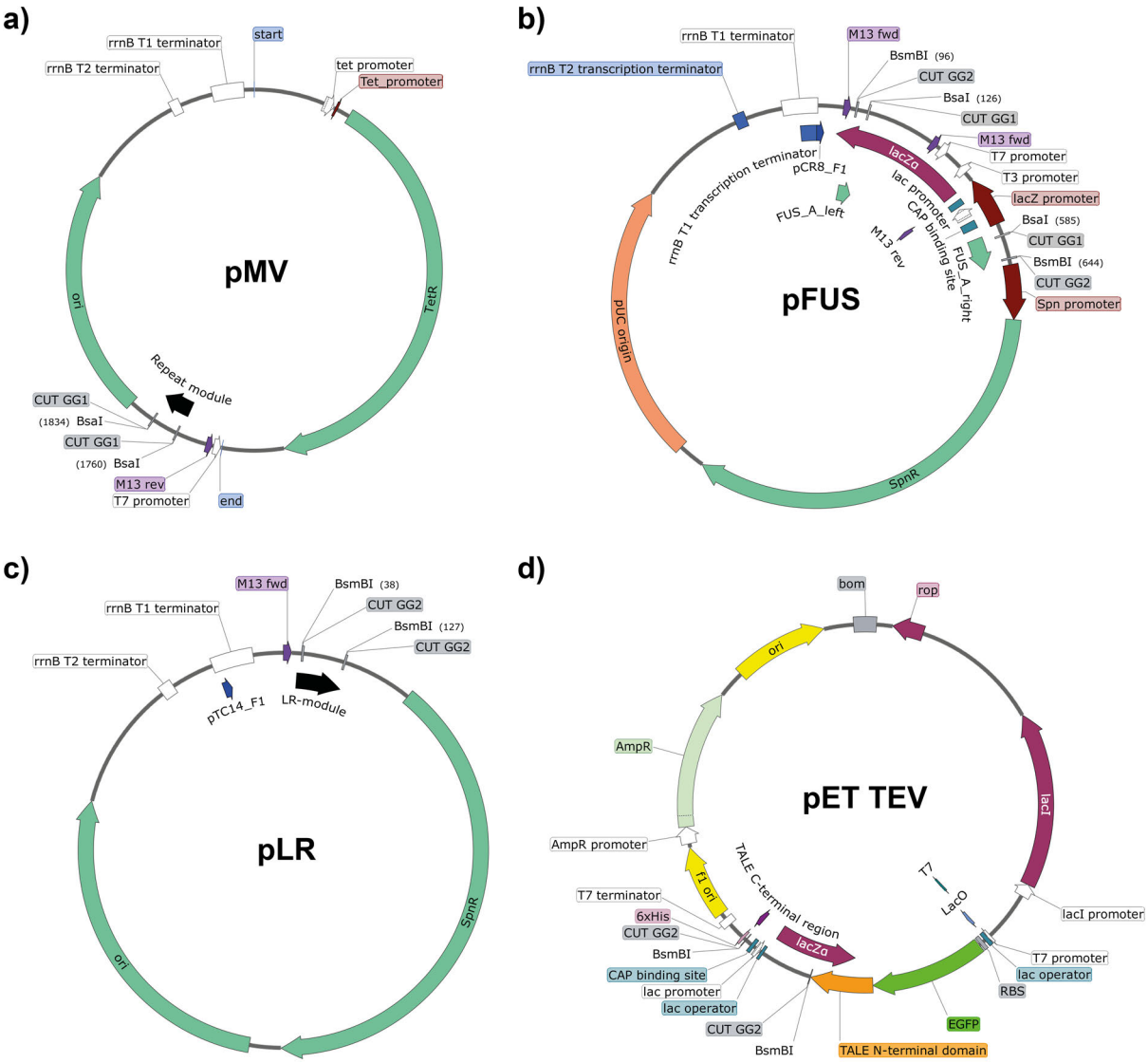


Figure SI1a: Module, array and backbone vectors for TALE assembly. a) Module vectors (MV) for single TALE repeats. **b)** pFUS (A, A30A, A30B and B1-B10) vectors for assembly of arrays of 10 (for A, A30A, A30B) or up to 10 repeat modules (for B) in the Golden Gate 1 reaction. **c)** Last repeat (LR) vector containing also part of the TALE C-terminal domain. **d)** TALE expression vector (TEV) with N-terminal GFP domain and C-terminal His-tag. Assembled array plasmids can be cloned into the expression vector in the Golden Gate 2 reaction.

[f] The contents of this section have been (partly) published in Gieß et al. (2018) and Gieß et al. (2019).[369,370] Reprinted (adapted) with permission from ref.[369] and ref.[370] Copyright 2019 American Chemical Society

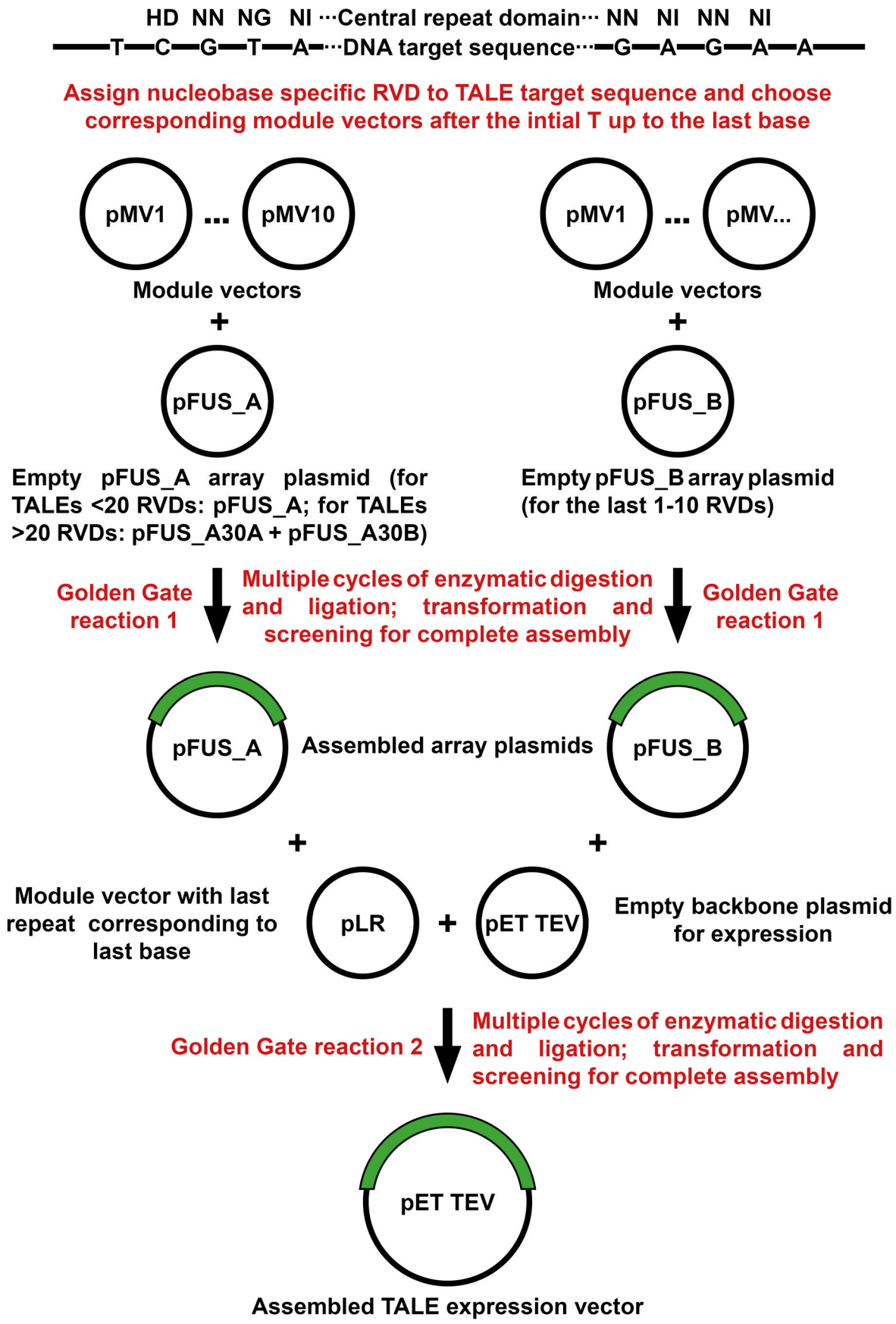


Figure SI1b: TALE assembly via two-step Golden Gate reaction.

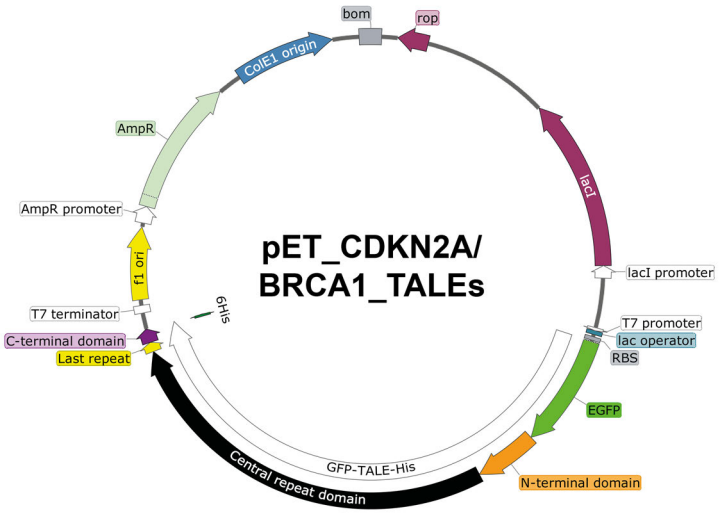


Figure SI1c: Exemplary pET vector for TALE expression.

CDKN2A_TALEs

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP
WPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKANFKIRH
NIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVT
AAGITLGMDELYKTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQH
PAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQL
DTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
NGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH
GLTPDQVVAIA**XXXXX**KQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIAS
NIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAI
ASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQAL
ETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLLEHHHHH
H

	XXXXX	MW(calc.) [Da]
TALE_SHDGG	SHDGG	131660.95
TALE_SG*GG	SG*GG	131465.78
TALE_SG**G	SG**G	131408.72
TALESG***	SG***	131351.67
TALE_S****	S****	131294.62
TALE_N****	N****	131321.64

Figure SI2a: Protein sequence and molecular mass of CDKN2A_TALEs. Sequence colored green marks the GFP domain; orange marks the N-terminal domain; Black marks the central repeat domain; Red marks the variable repeat (**XXXXX** marks the variable positions); Yellow marks the last repeat; Purple marks the C-terminal domain with His-tag. * represents the of the amino acid deletion.

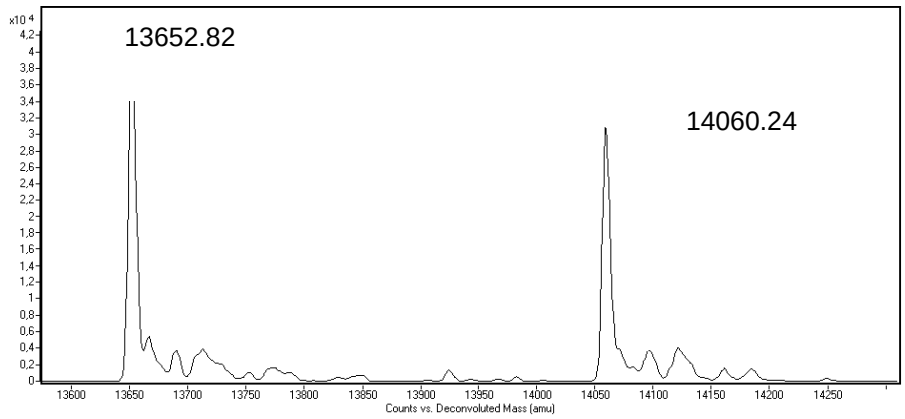
BRCA1_TALEs

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP
WPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKANFKIRH
NIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVT
AAGITLGMDELYKTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQH
PAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQL
DTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIAS
NGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALE
TVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVA
IASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLC
QDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIA**XXXXX**KQALETVQRLLPVLCQDHGLTPDQV
VAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLT
PDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQR
LLPVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLLEHHHH
HH

	XXXXX	MW(calc.) [Da]
TALE_SHDGG	SHDGG	131305.52
TALE_SG*GG	SG*GG	131110.34
TALE_SG**G	SG**G	131053.29
TALESG***	SG***	130996.23
TALE_S****	S****	130939.18
TALE_N****	N****	130966.21

Figure SI2b: Protein sequence and molecular mass of BRCA1_TALEs. Color code as in Fig. SI2a

a)



b)

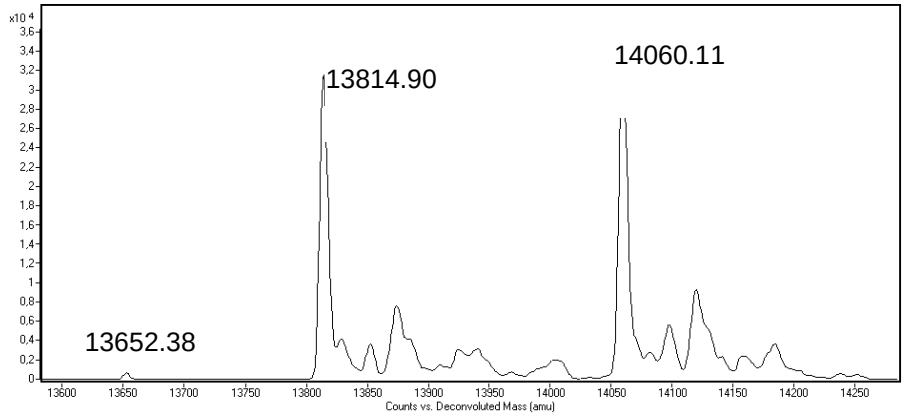


Figure SI3: ESI-MS spectra of dsDNA before and after glycosylation. a) ESI-TOF mass spectrum of oligonucleotide pair o1520/o1529 before glycosylation. **b)** ESI-TOF mass spectrum of oligonucleotide pair o1520/o1529 after glycosylation.

Table SI1: Theoretical and observed peak list of BRCA1_5hmC before and after glycosylation by ESI-TOF mass spectrometry.

	Theoretical Mass			Δm_{calc}	Observed Mass			Δm_{obs}
	starting	**+ CPD**	**converted**		**starting**	**+ CPD**	**converted**	
ODN-5hmC	13652.8	-	-	-	13652.82	-	-	-
		+ **1**	13815.0	162.2	13562.38	+ **1**	13814.90	162.52

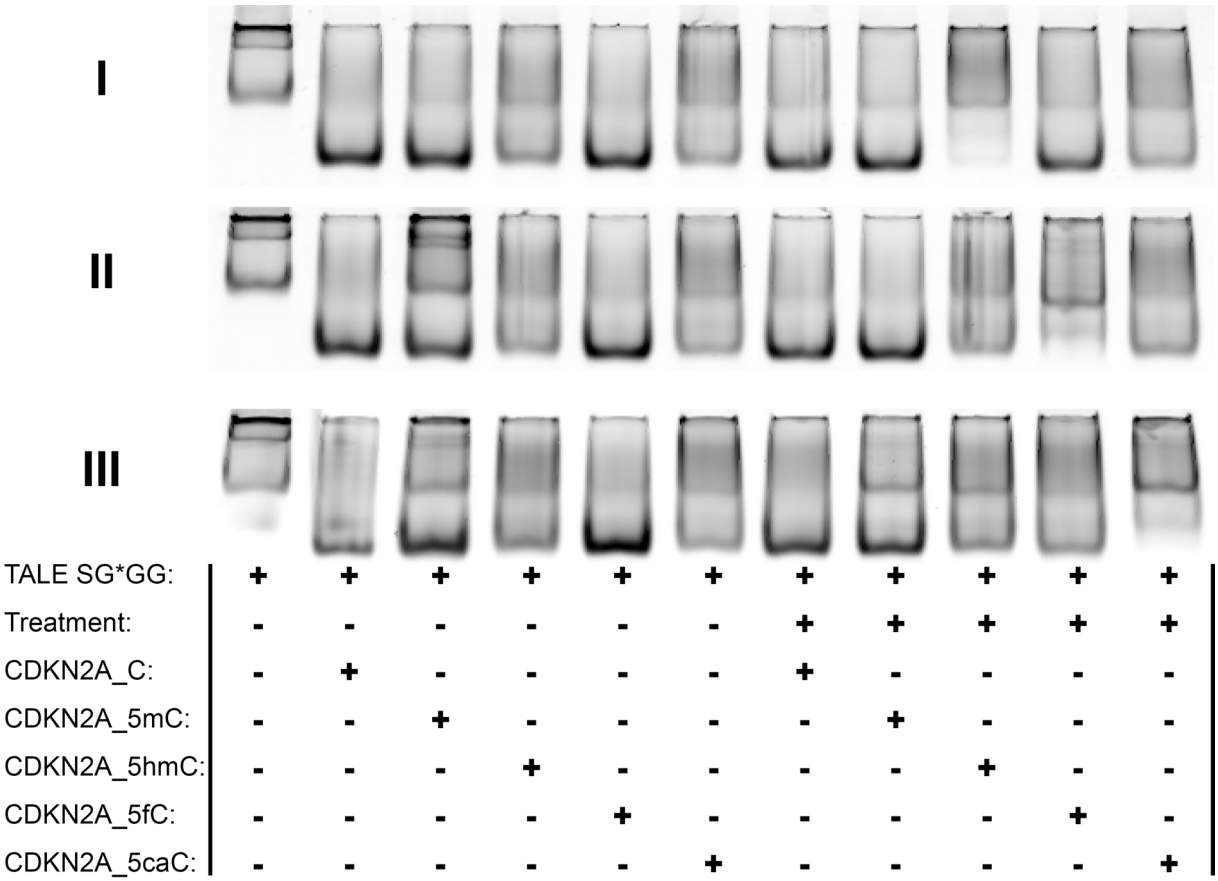


Figure SI4: Representative EMSA data for differential TALE interaction with their cognate targets before and after treatment as shown in Fig. 24d, Fig. 26a, and Fig. 28a. Treatment refers to the different reaction types of I: Glycosylation, II: Oxime formation, III: Amide formation.

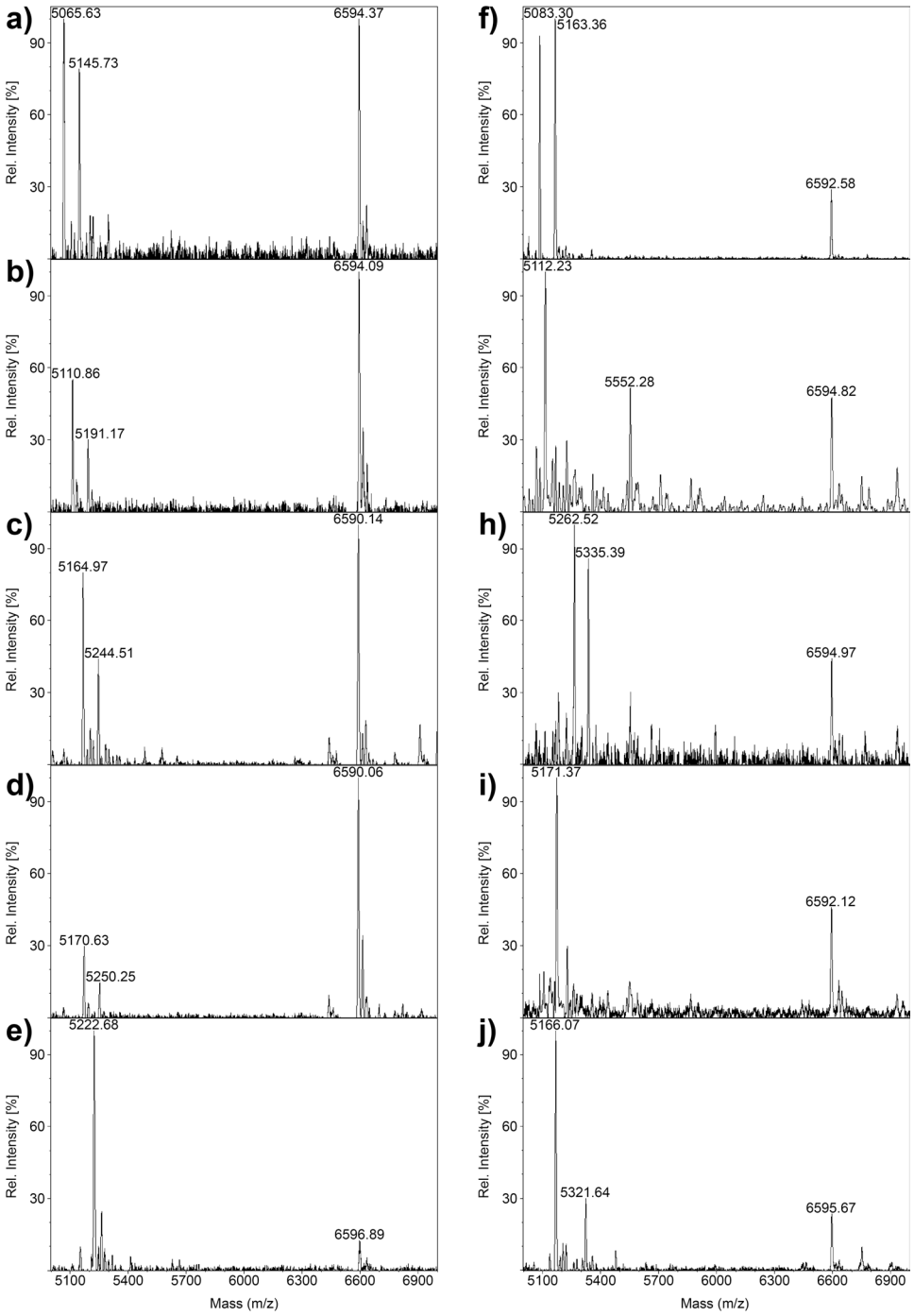


Figure SI5: MALDI-TOF MS spectra of dsDNAs used for 5fC conversion with hydroxylamines and 5caC conversion with amines. a) 5fC-ODN untreated**. b)** 5fC-ODN after reaction with O-ethylhydroxylamine **(3a)**. **c)** 5fC-ODN after reaction with O-(tetrahydro-2H-pyran-2-yl)hydroxylamine **(3b)**. **d)** 5fC-ODN after reaction with O-(benzyl)hydroxylamine **(3c)**. **e)** 5fC-ODN after reaction with O-(tert.-butyl)hydroxylamine **(3d)**. **f)** 5caC-ODN untreated. **g)** 5caC-ODN after reaction with ethylamine **(5a)**. **h)** 5caC-ODN after reaction with (tetrahydro-2H-pyran-2-yl)methylamine **(5b)**. **i)** 5caC-ODN after reaction with benzylamine **(5c)**. **j)** 5caC-ODN after reaction with tert.-butylamine **(5d)**.

Table SI2: Observed and calculated masses for 5fC- and 5caC-ODN from data of Fig. SI5. Asterisks indicate oligonucleotides with a mass difference of 79.9 compared to standard calculated mass, matching the weight of an additional phosphate group. These were observed in MS analyses of digested dsDNA oligonucleotides for both non-reacted and reacted DNA. Δmcalc and Δmobs correspond to masses of introduced oxime and amide substituents.

	Theoretical Mass			Δm_{calc}	Observed Mass			Δm_{obs}
	starting	**+ CPD**	**converted**		**starting**	**+ CPD**	**converted**	
ODN-5fC	5060.3 5140.3)*	**-**	-	-	5065.6 5145.7)*	**-**	-	-
		+ 2a	5103.4 5183.4)*	43.1	5065.4 5145.8)*	**+ 2a**	5110.9 5191.1)*	45.5 45.3)*
		+ 2b	5159.4 5139.4)*	99.1	5064.8 5145.3)*	**+ 2b**	5164.9 5244.5)*	100.1 99.2)*
		+ 2c	5165.4 5245.3)*	105.1	5064.7 5145.3)*	**+ 2c**	5170.6 5250.2)*	105.9 104.9)*
		+ 2d	5131.4 5212.4)*	71.1	5068.3 5151.2)*	**+ 2d**	5143.1 5222.7)*	74.8 71.5)*
ODN-5caC	5077.3 5157.3)*		-	-	5083.3 5163.4)*	**-**	-	-
		+ 4e	5094.4 5284.4)*	27.1	5084.5 5165.5)*	**+ 4e**	5112.2 5193.9	27.7 28.4)*
		+ 4f	5159.4 5139.4)*	97.2	5082.1 5164.1)*	**+ 4f**	5180.8 5262.5)*	98.6 98.3)*
		+ 4g	5165.4 5245.3)*	89.1	5082.7 5160.1)*	**+ 4g**	5171.4 5250.6)	88.7 90.5)*
		+ 4h	5131.4 5212.4)*	55.1	5083.1 5166.1)*	**+ 4h**	5135.8 5219.5)*	52.7 53.4)*

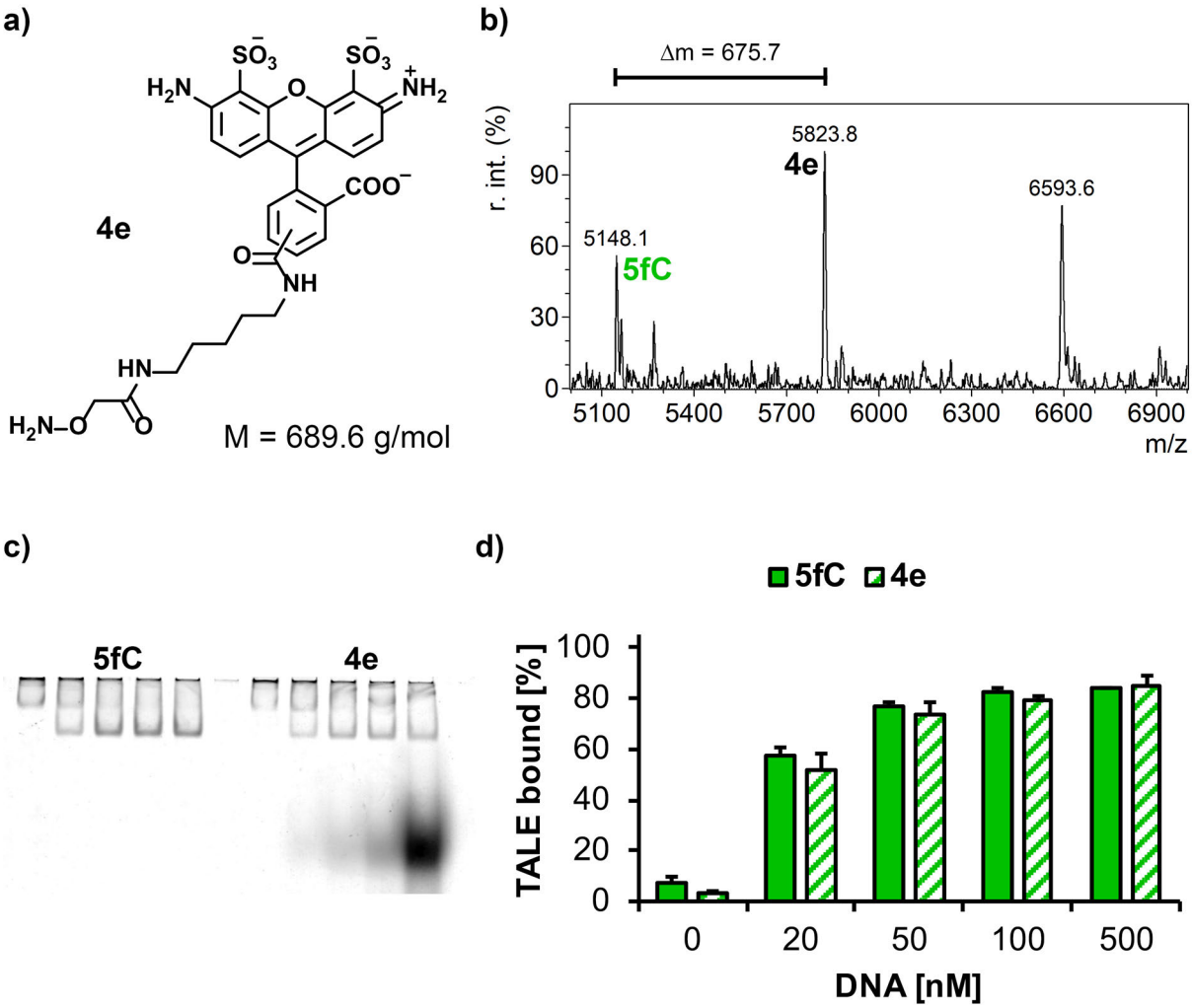


Figure SI6: Oxime formation of 5fC-DNA with Alexa Fluor 488 hydroxylamine. a) Chemical structure of Alexa Fluor 488 hydroxylamine. **b)** MALDI-TOF MS of reacted 5fC shows correct product formation of **6e**. **c)** Native PAGE image. Increasing fluorescent bands of the fluorescent probe shows presence of 5fC DNA at bands for unbound and excessive DNA. **d)** Comparison of shifted bands for unreacted and reacted 5fC.

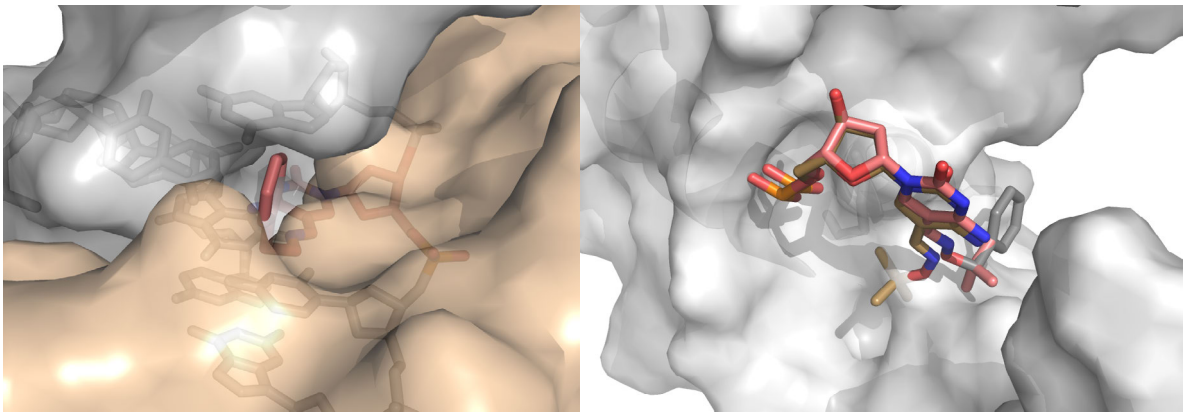


Figure SI7: Superimposed models of DNA containing benzyloxime (4c) or tert-butyloxime (4d) bound by TALE_SG*GG. a) Positioning of benzyl substituent of **4c** (red sticks) to evade steric clash with a minimized TALE repeat structure. (DNA surface grey, TALE surface light brown). **b)** Hypothesis for a steric clash of **4d** as a result of different orientation of benzyl substituent of **4c** (grey sticks) and tert-butyl substituent of **4d** (pale red sticks).

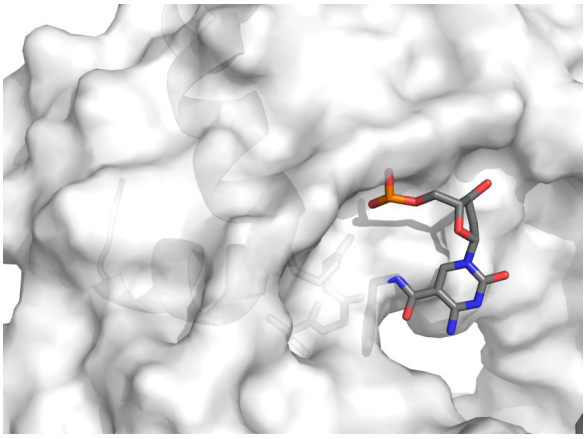


Figure SI8: Model of DNA containing benzylamide (6c) (grey sticks) clashing with non-minimized structure of repeat_SG*GG.

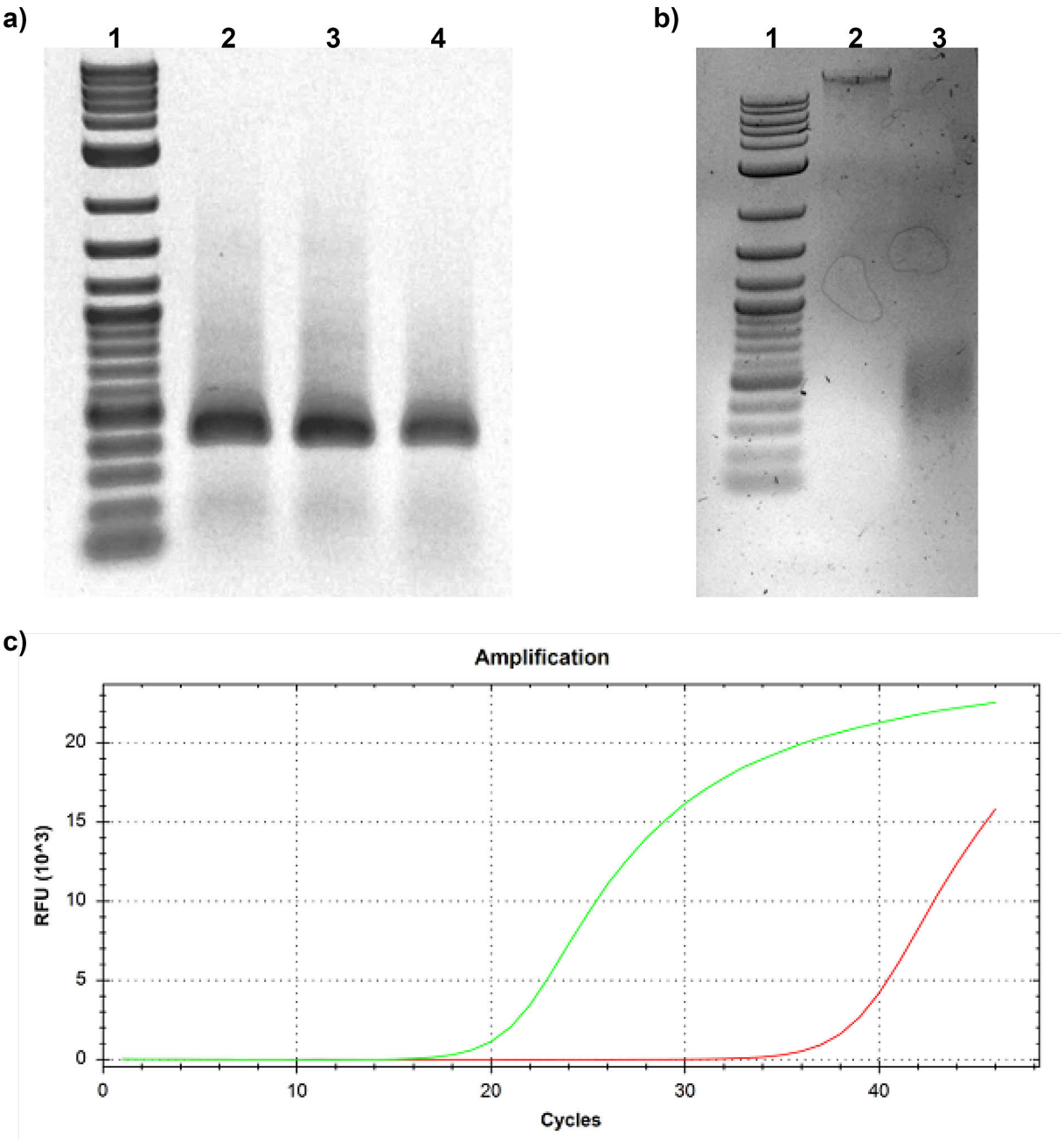
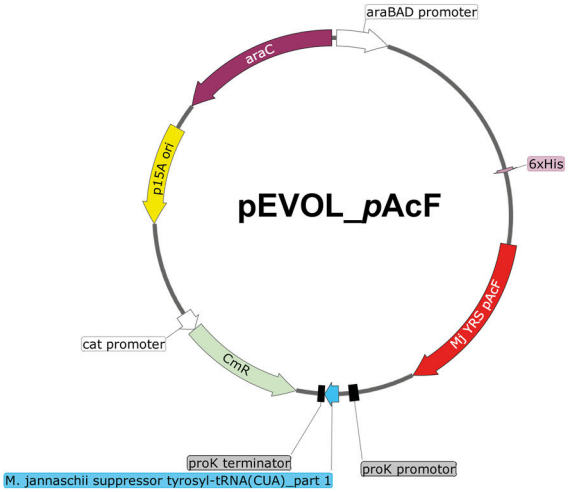
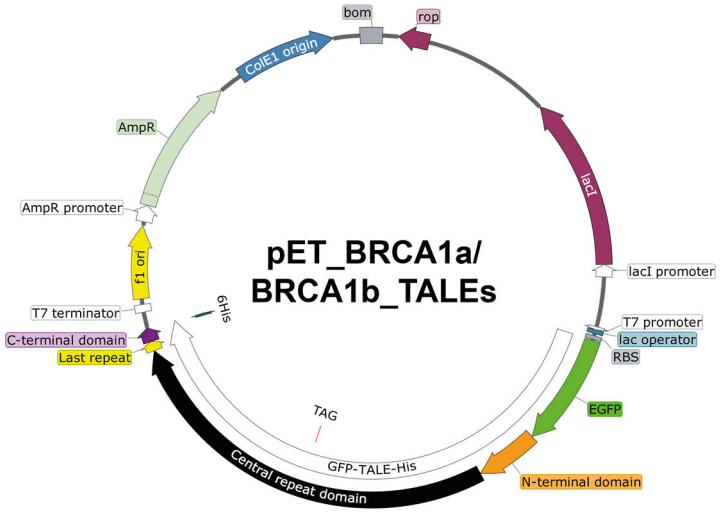


Figure SI9: Spike-ins and genomic DNA for enrichment. a) Agarose gel of 421bp_C6 → X spike-in PCR products. Lane 1: 2-log DNA-ladder, Lane 2: CDKN2A_421bp_TALE binding_C6 → 5hmC6, Lane 3: CDKN2A_421bp_TALE binding_C6 → 5fC6, Lane 4: CDKN2A_421bp_TALE binding_C6 → 5caC6. **b)** Agarose gel analysis of human whole genome amplified DNA randomly sheared by sonication. Lane 1: 2-log DNA Ladder, Lane 2: Whole genome amplified human DNA (100ng), Lane 3: Fragmented whole genome amplified human DNA (100ng). **c)** qPCR on human genomic DNA samples with and without spike-in DNA. red: 500 ng sheared, digested gDNA, green: 500 ng sheared, digested gDNA with 421 bp-spike-in

a)



d)



c)

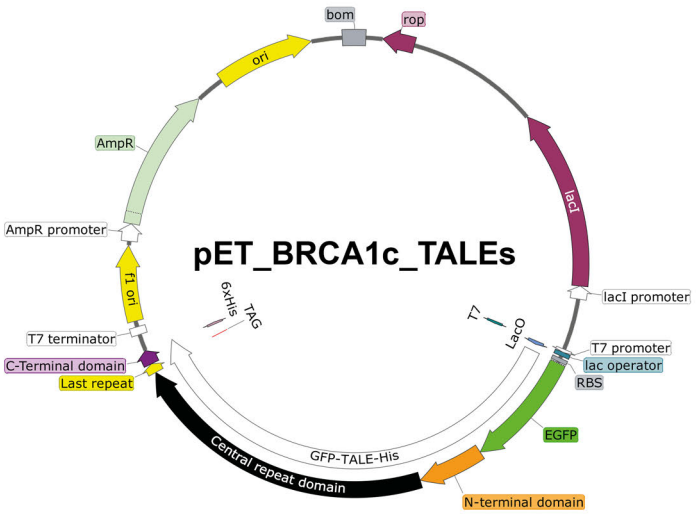


Figure SI10: Plasmid maps for vectors containing genes for a) *p*AcF tRNA/aaRS, for b) BRCA1a and BRCA1b or c) BRCA1c genes. TAG denotes the codon for the position of pAcF in the protein.

BRCA1a_TALEs

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP
WPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKANFKIRH
NIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVT
AAGITLGMDELYKTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQH
PAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQL
DTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIAS
NGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALE
TVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVA
IASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLC
QDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIA**XX***X**GKQALETVQRLLPVLCQDHGLTPDQVV
AIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQ
VVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLP
VLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGG
KQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRL
LPVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLLEHHHHH
H

	XX*X**	MW(calc.) [Da]
TALE_**1**G*GG	**1**G*GG	131305.52
TALE_S**1***GG	S**1***GG	131110.34
TALE_SG***1**G	SG***1**G	131053.29

Figure SI11a: Protein sequence and molecular mass of BRCA1a_TALEs. Color code as in Fig. SI2a. * represents the of the amino acid deletion. **1** represents the *para*-acetylphenylalanine (*p*AcF).

BRCA1b_TALEs

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP
WPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKANFKIRH
NIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVT
AAGITLGMDELYKTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQH
PAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQL
DTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIAS
NGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALE
TVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVA
IASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLC
QDHGLTPDQVVAIASG*GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIA**XX***X*GKQALETVQRLLPVLCQDHGLTPDQVV
AIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASG*GGKQALETVQRLLPVLC
QDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASHDGGKQALESIVAQLSRPDPALAALTNDHLLEHHHHH
H

	XX*X*	MW(calc.) [Da]
TALE_**1**G*GG	**1**G*GG	130961.19
TALE_S**1***GG	S**1***GG	130991.22
TALE_SG***1**G	SG***1**G	130991.22

Figure SI11b: Protein sequence and molecular mass of BRCA1b_TALEs. Color code as in Fig. SI2a. * represents the of the amino acid deletion. **1** represents the *para*-acetylphenylalanine (*p*AcF).

BRCA1c_TALEs

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVP
WPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAE
VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKANFKIRH
NIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVT
AAGITLGMDELYKTLGYSQQQQEKIKPKVRSTVAQHHEALVGHGFTHAHIVALSQH
PAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQL
DTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIAS
NGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALE
TVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVA
IASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLC
QDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQA
LETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG**(**LTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHG**)**LTPDQVVAIAS**1***GGKQALESIVAQL
SRPDPALAALTNDHLLEHHHHHH

	MW(calc.) [Da]
TALE_LR_S1*GG(20)	109810.09
TALE_LR_S**1***GG(19)	106199.00

Figure SI11c: Protein sequence and molecular mass of BRCA1c_TALEs. Color code as in Fig. SI2a. * represents the of the amino acid deletion. **1** represents the *para*-acetylphenylalanine (*p*AcF). The full sequence represents TALE_LR_S**1***GG(20). For TALE_LR_S**1***GG(19) only the sequence outside of round brackets **(…)** need to be considered.

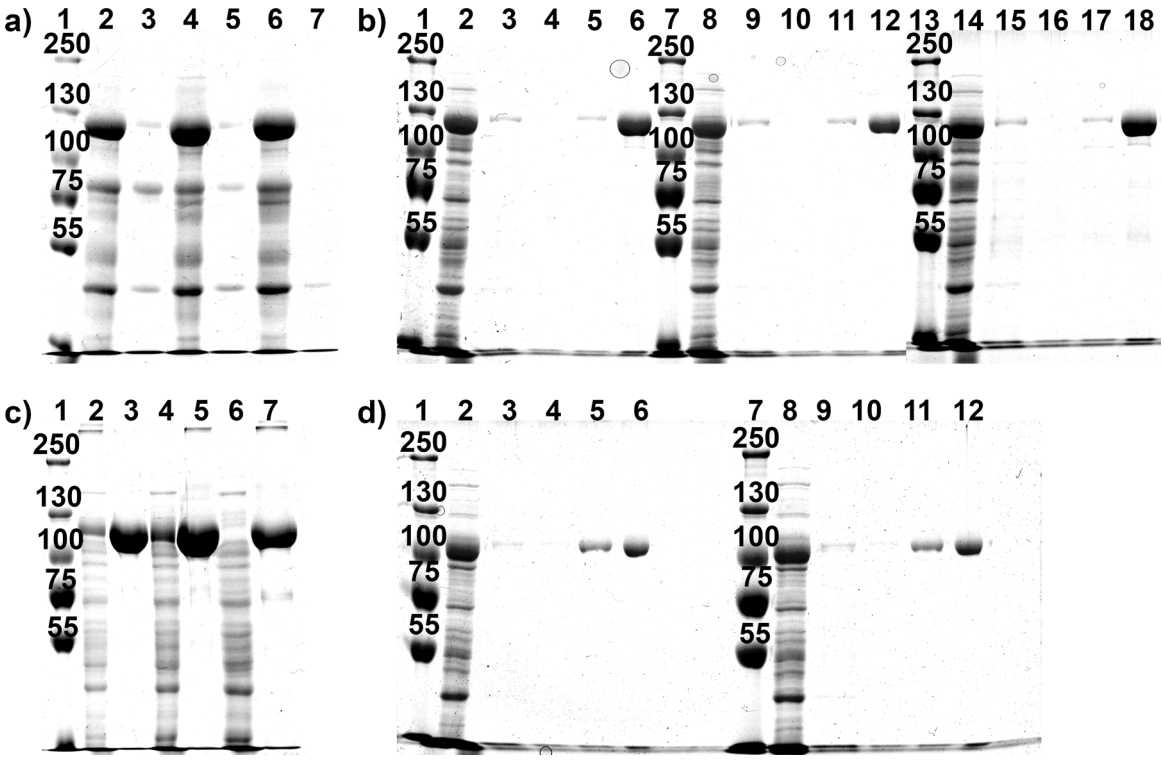
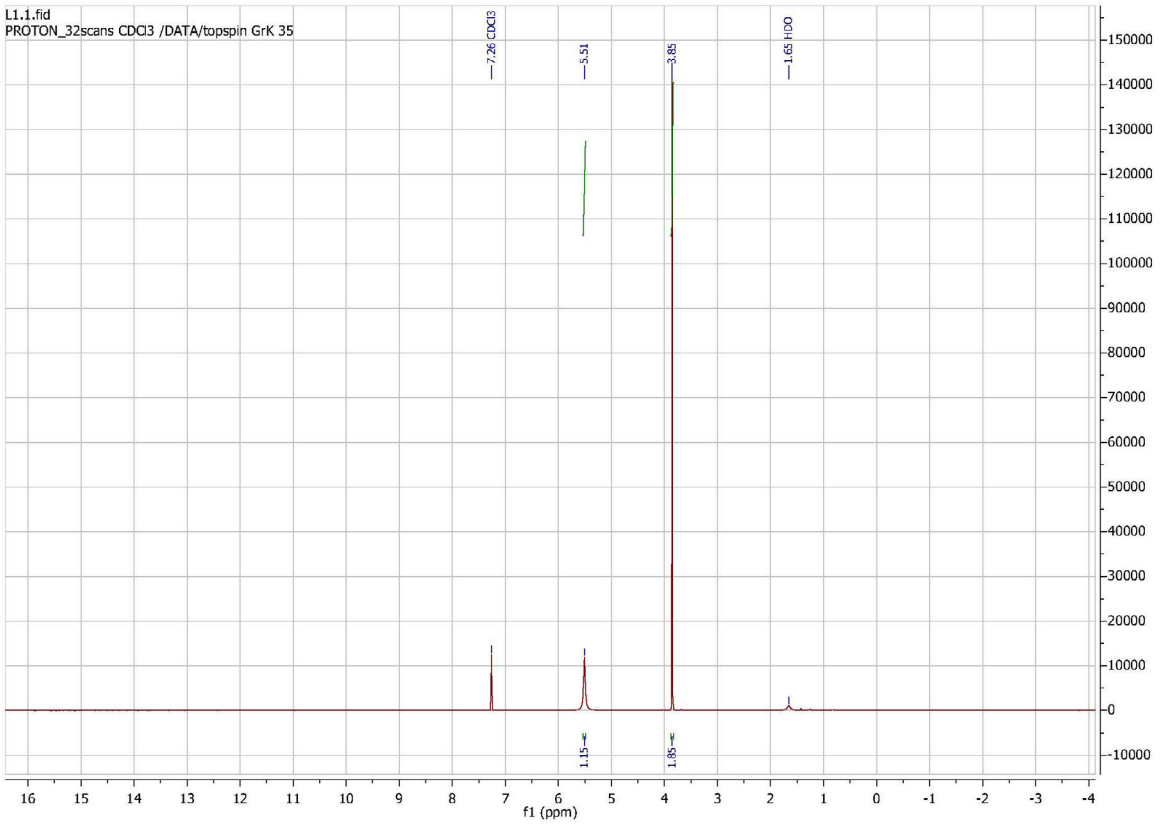
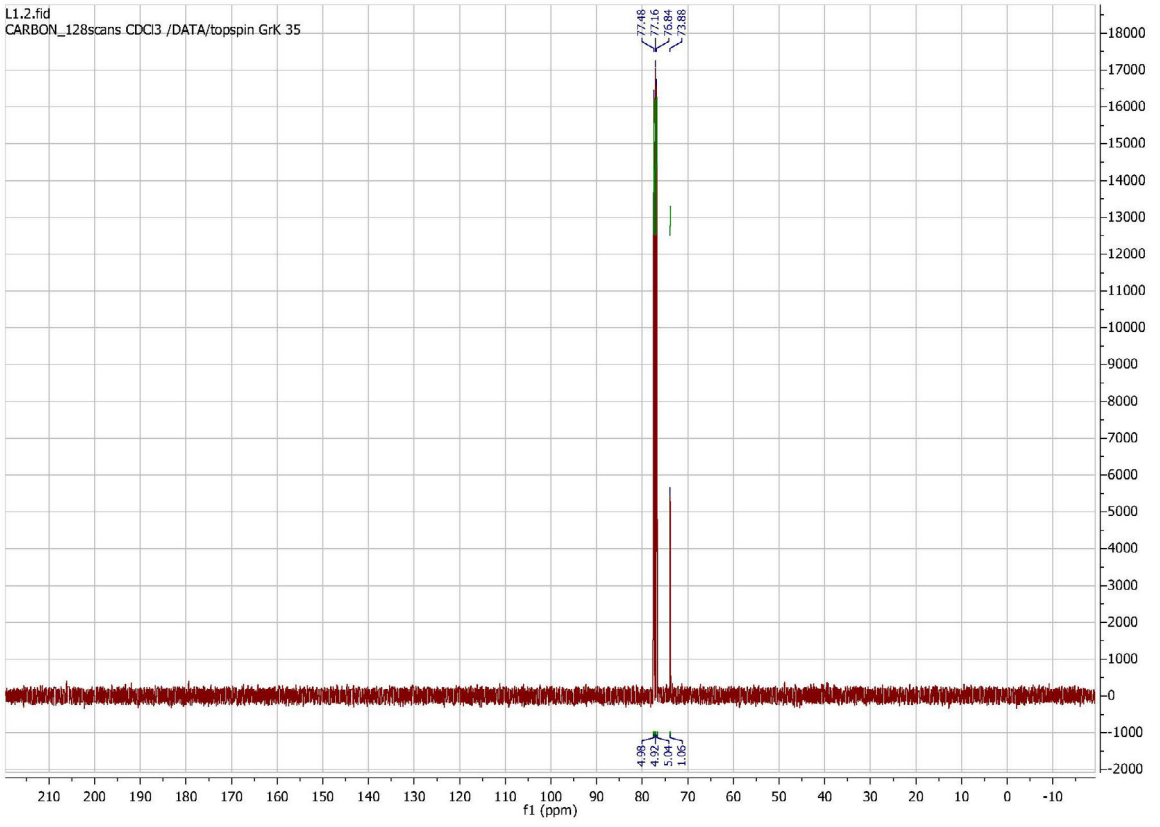
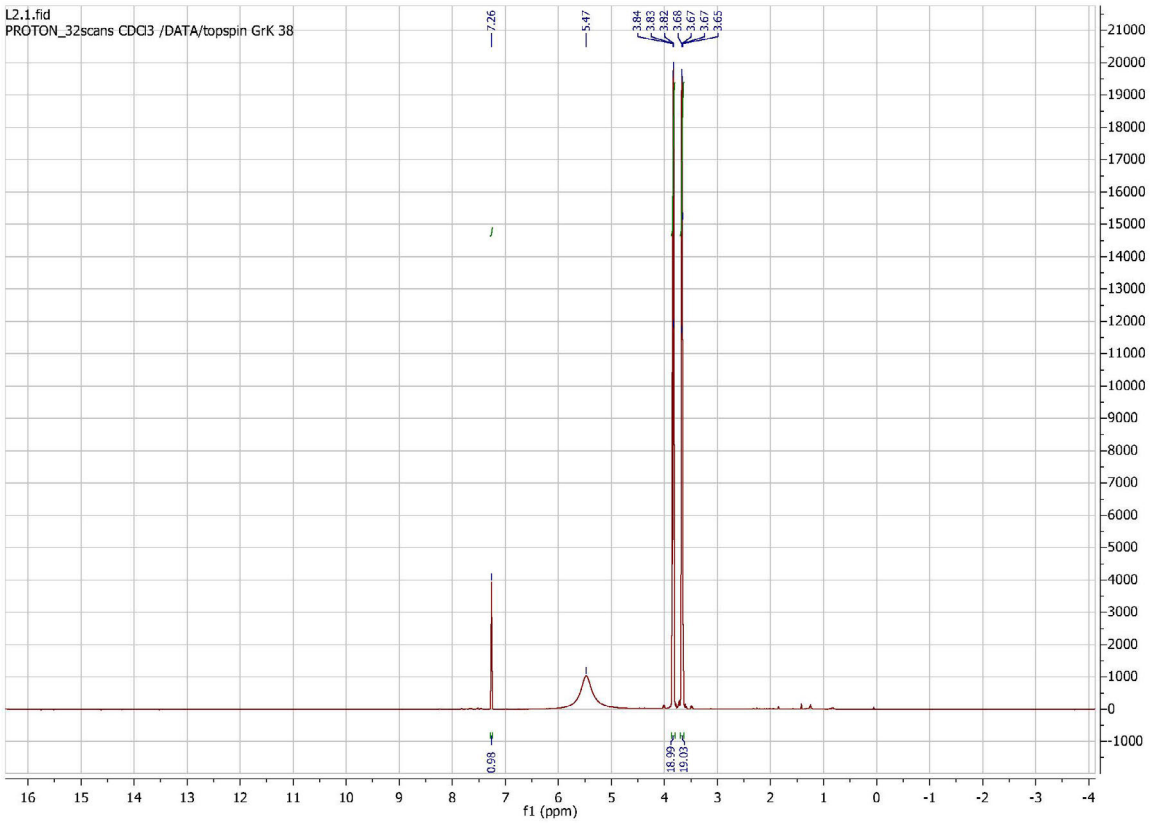


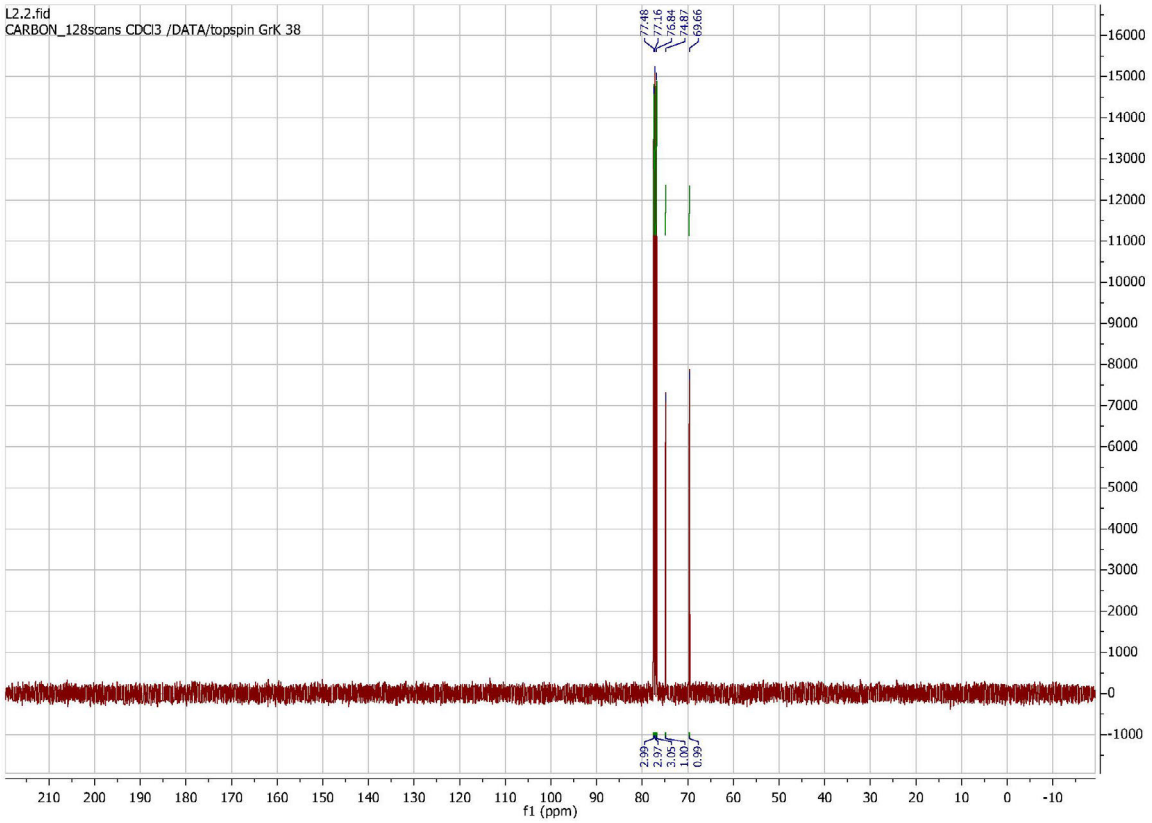
Figure SI12: SDS-PAGE images of TALE proteins. a) Full image of Fig. 30d). Lane 1: Protein ladder (bands correspond to MW in kDa). Lane 2: Isolated protein fraction from expression with *p*AcF of pET_BRCA1a_TALE_**1**G*GG and pEVOL_*p*AcF with IPTG and Arabinose. Lane 3: Isolated protein fraction from expression without *p*AcF of pET_BRCA1a_TALE_**1**G*GG and pEVOL_*p*AcF with IPTG and Arabinose. Lanes 4-7: As lane 2 and lane 3 but for TALE_S**1***GG (lanes 4-5) and for TALE_SG***1**G (lanes 6-7). **b)** SDS-PAGE image of the expression and purification of pET_BRCA1b_TALEs. Lane 1: Protein ladder (bands correspond to MW in kDa). Lane 2: Cell lysate after pET_BRCA1a_TALE_**1**G*GG expression. Lanes 3-5: Washing fractions of TALE_**1**G*GG purification with PBS (Lane 3), wash buffer with 20 mM imidazole (Lane 4), wash buffer with 50 mM imidazole (Lane 5). Lane 6: Isolated TALE_**1**G*GG protein. Lanes 7-18: As lanes 1-6 but for TALE_S**1***GG (lanes 7-12) and TALE_SG***1**G (lanes 13-18). **c)** SDS-PAGE image of the expression and purification of pET_BRCA1b_TALEs. Lane 1: Protein ladder (bands correspond to MW in kDa). Lane 2: Flowthrough after Ni-NTA purification of BRCA1b TALE_**1**G*GG. Lane 3: Purified TALE_**1**G*GG. Lanes 4-7: As lane 2 and lane 3 but for TALE_S**1***GG (lanes 4-5) and for TALE_SG***1**G (lanes 6-7). **d)** SDS-PAGE image of the expression of TALE_LR_S**1***GG(20) and TALE_LR_S**1***GG(19). Lanes 1-12: As lanes 1-6 in b) but for TALE_LR_S**1***GG(20) (lanes 2-6) and TALE_LR_S**1***GG(19) (lanes 8-12).

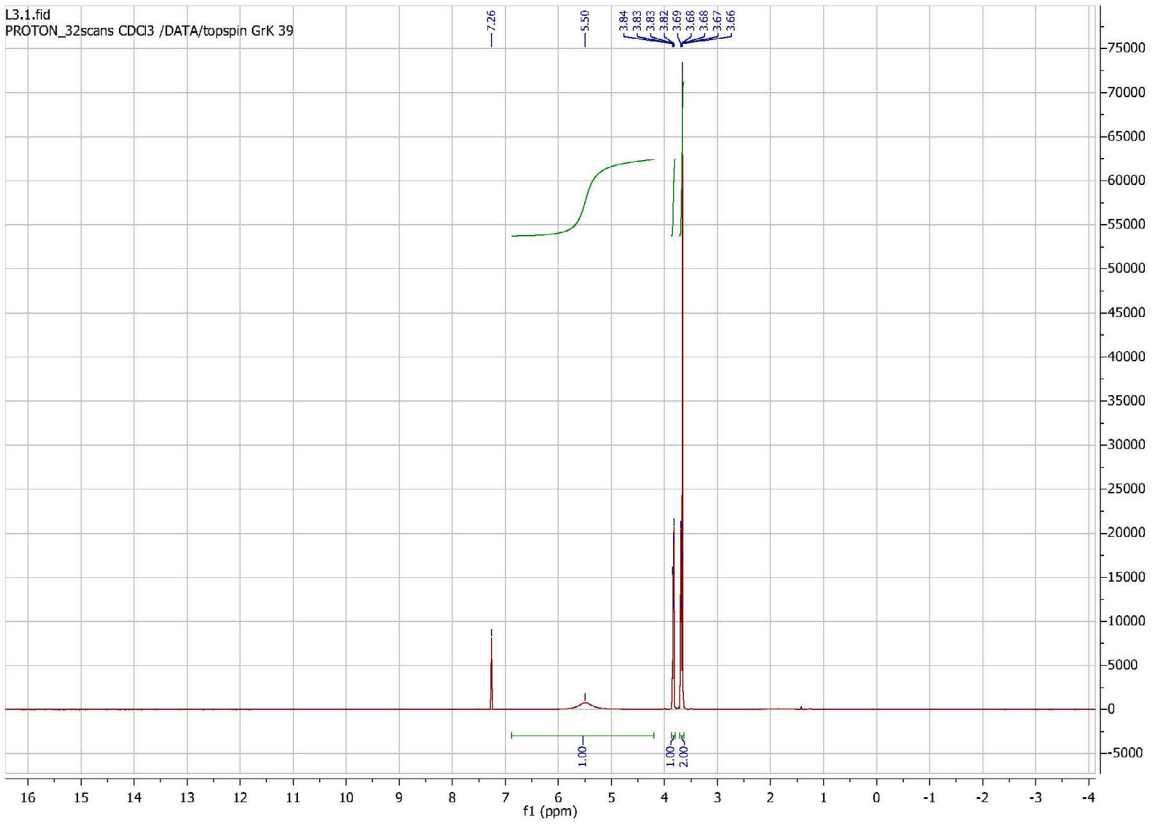
Figure SI13: ^{1}H and ^{13}C NMR spectra of L1-5

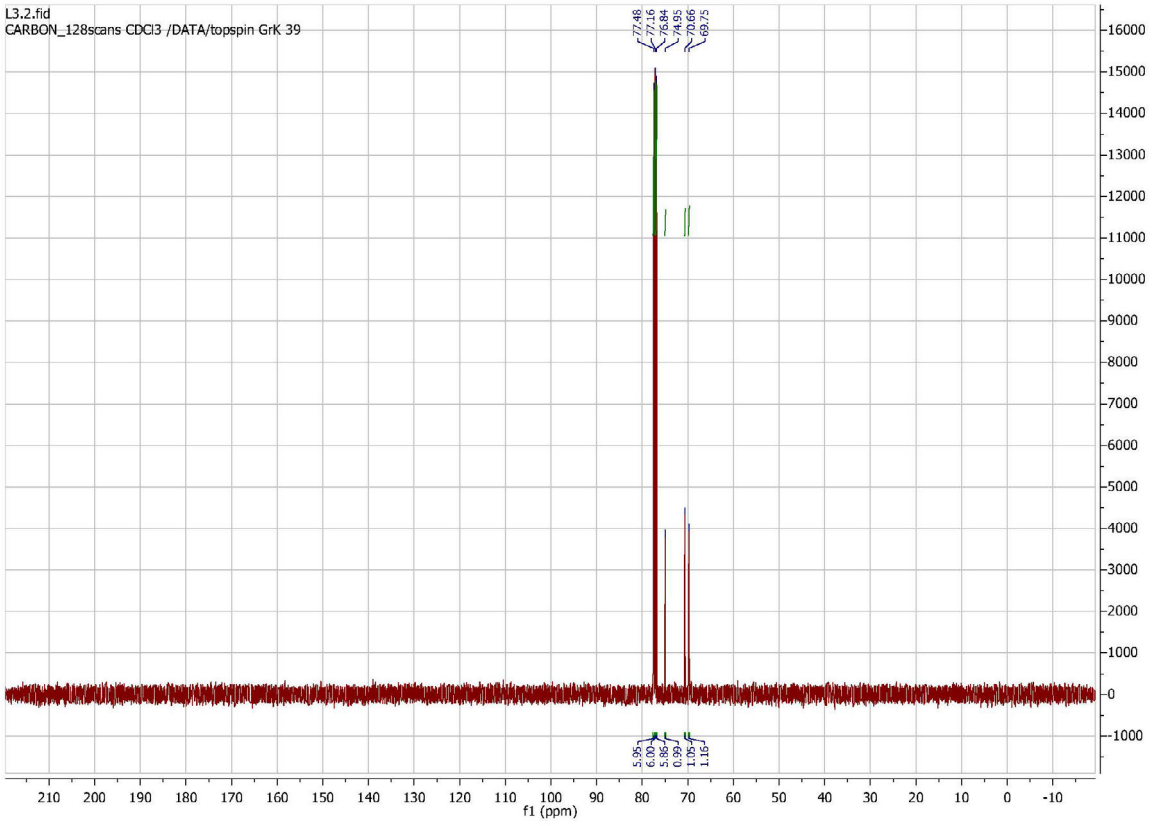


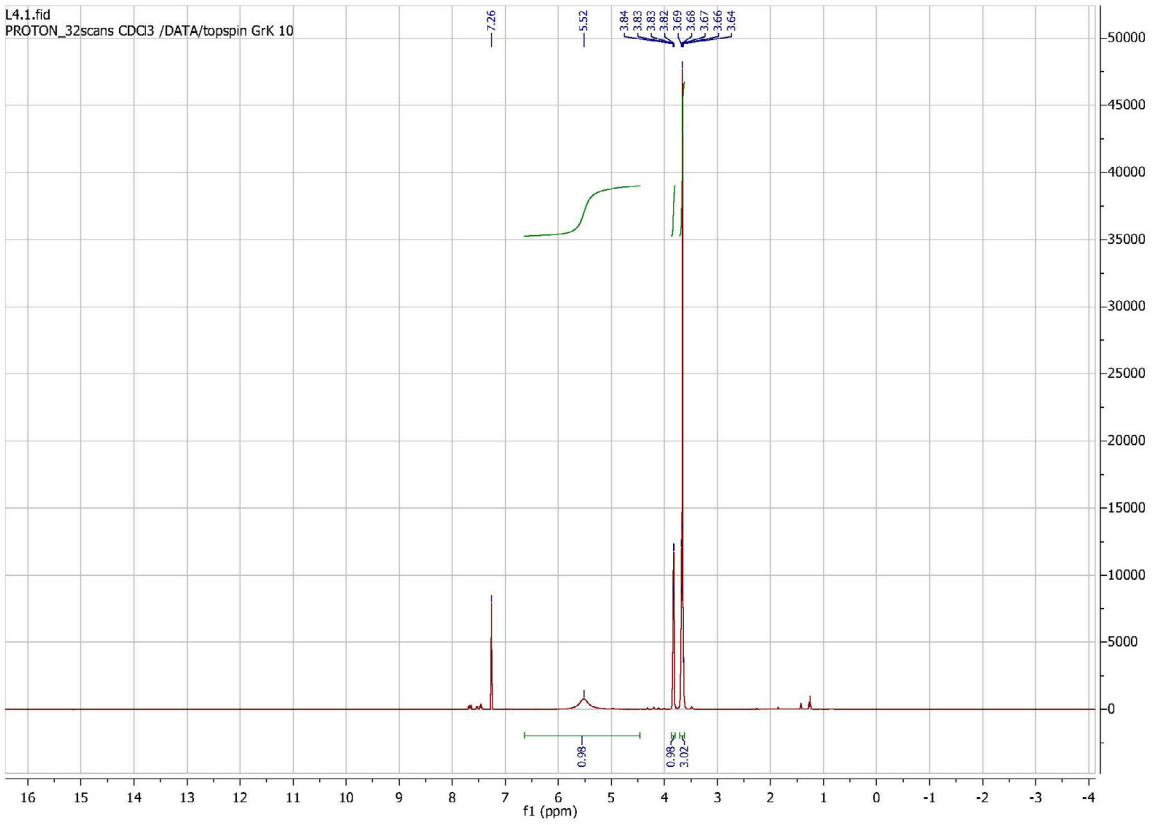


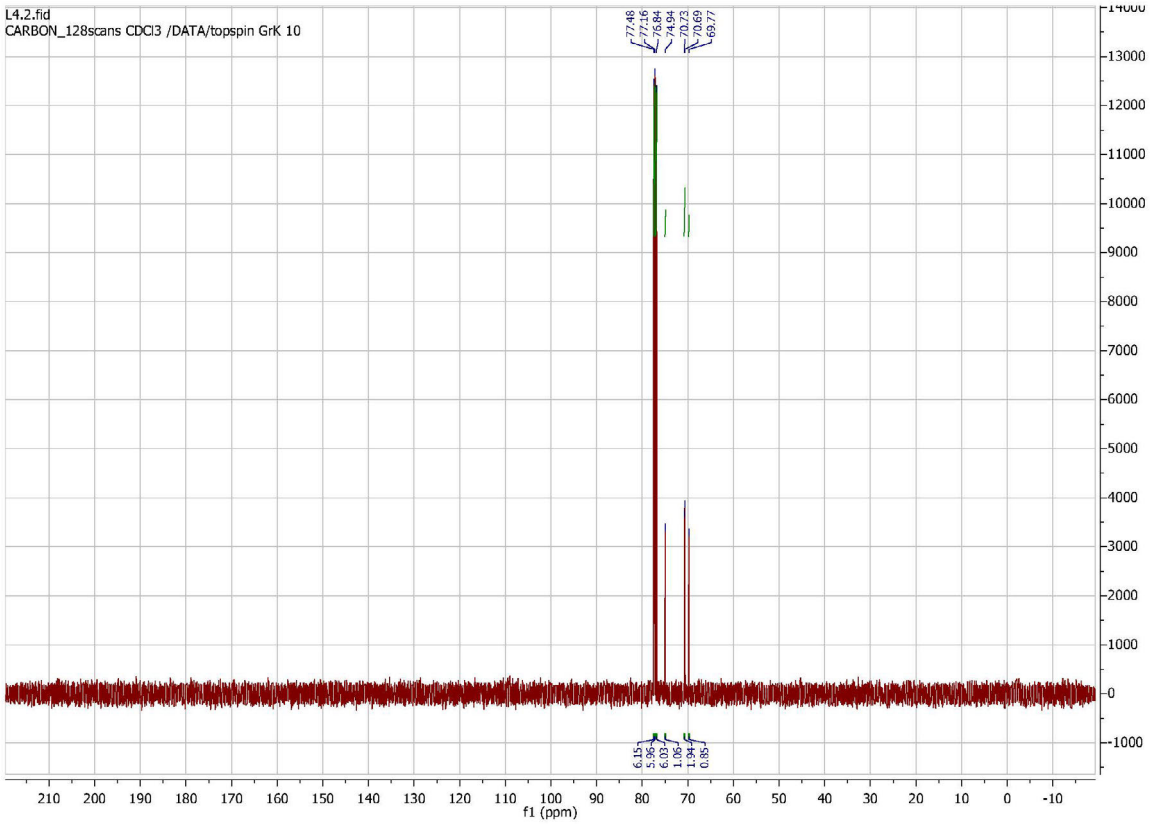

L2.1.fid
PROTON_32scans CDCl3 /DATA/topspin GrK 38
7.26
5.47
3.84
3.83
3.82
3.68
3.67
3.67
3.65
0.98
18.99
19.03
f1 (ppm)

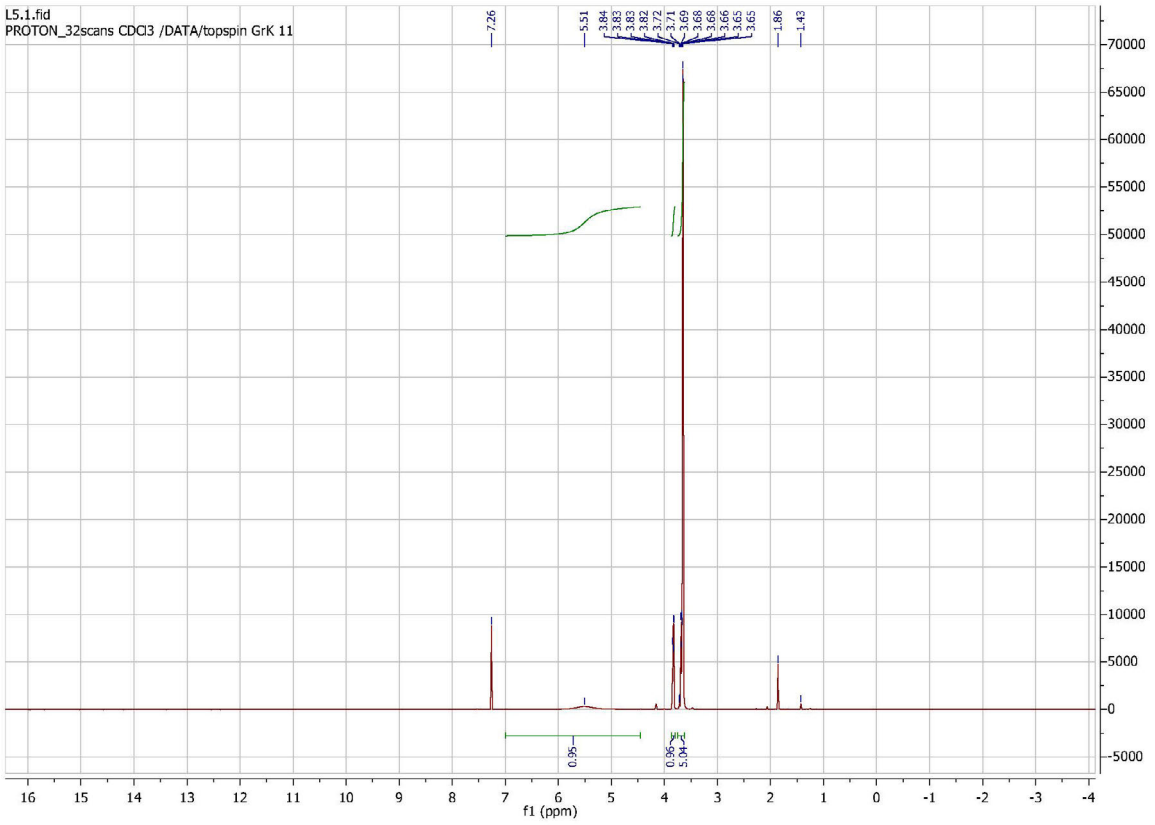

L2.2.fid
CARBON_128scans CDCl3 /DATA/topspin GrK 38
77.48
77.16
76.84
74.87
69.66
2.99
2.97
3.05
1.00
0.99
f1 (ppm)


L3.1.fid
PROTON_32scans CDCl3 /DATA/topspin GrK 39
7.26
5.50
3.84
3.83
3.83
3.82
3.69
3.68
3.68
3.67
3.66
1.00
1.00
2.00
f1 (ppm)


L3.2.fid
CARBON_128scans CDCl3 /DATA/topspin GrK 39
77.48
77.16
76.84
74.95
70.66
69.75
5.95
6.00
5.86
0.99
1.05
1.16
f1 (ppm)


L4.1.fid
PROTON_32scans CDCl3 /DATA/topspin GrK 10
7.26
5.52
3.84
3.83
3.83
3.82
3.69
3.68
3.67
3.66
3.64
0.98
0.99
3.02
f1 (ppm)


L4.2.fid
CARBON_128scans CDCl3 /DATA/topspin GrK 10
77.48
77.16
76.84
74.94
70.73
70.69
69.77
6.15
5.95
5.03
1.06
1.94
0.85
f1 (ppm)


L5.1.fid
PROTON_32scans CDCl3 /DATA/topspin GrK 11
7.26
5.51
3.84
3.83
3.83
3.82
3.72
3.71
3.69
3.68
3.68
3.66
3.65
3.65
1.86
1.43
0.95
0.96
5.04
f1 (ppm)

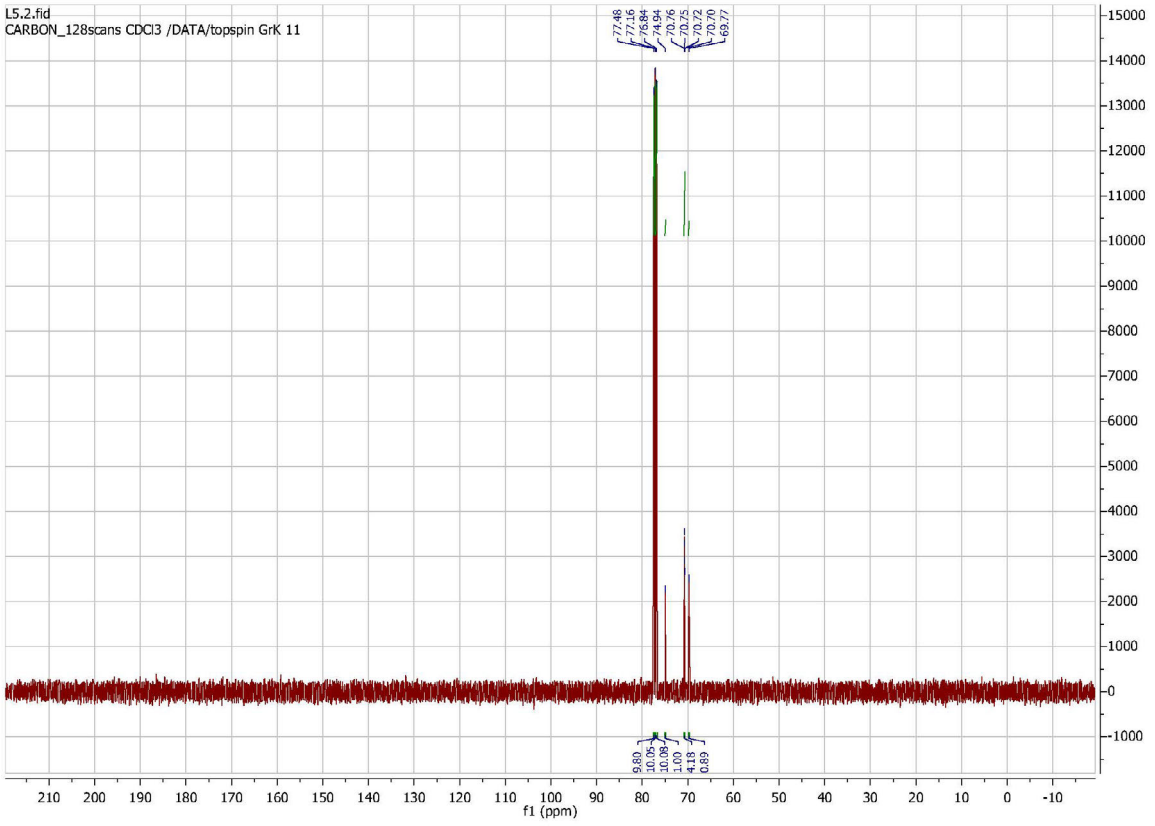
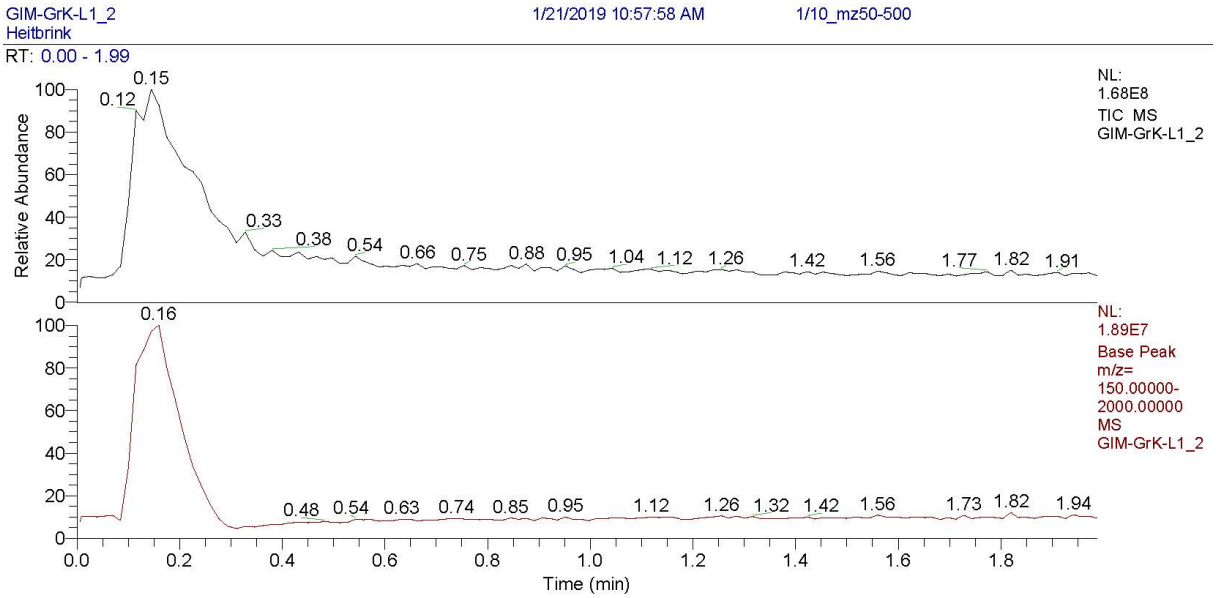
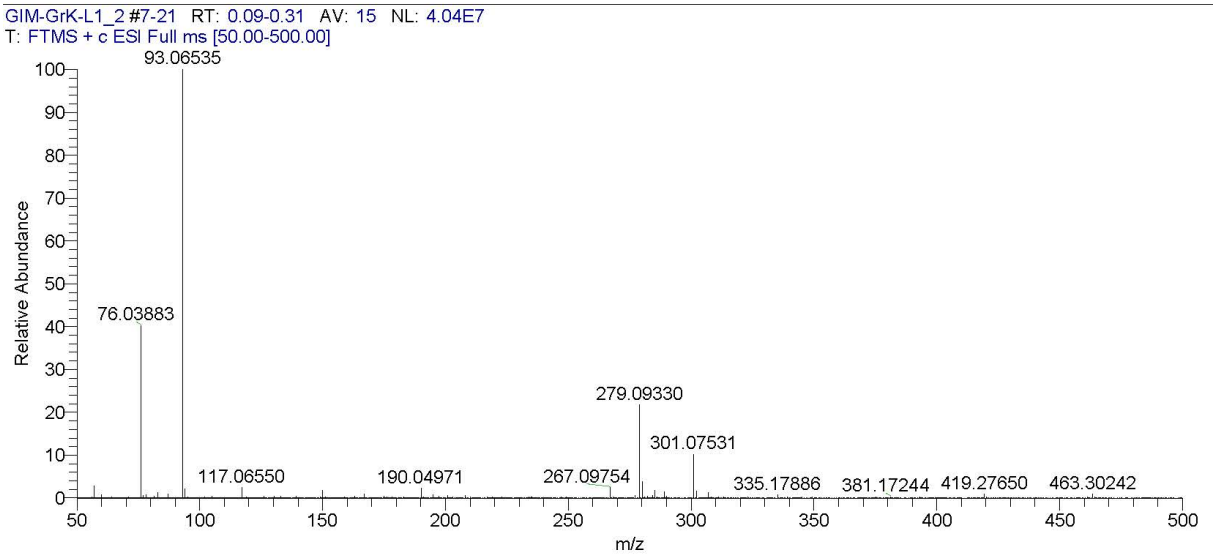
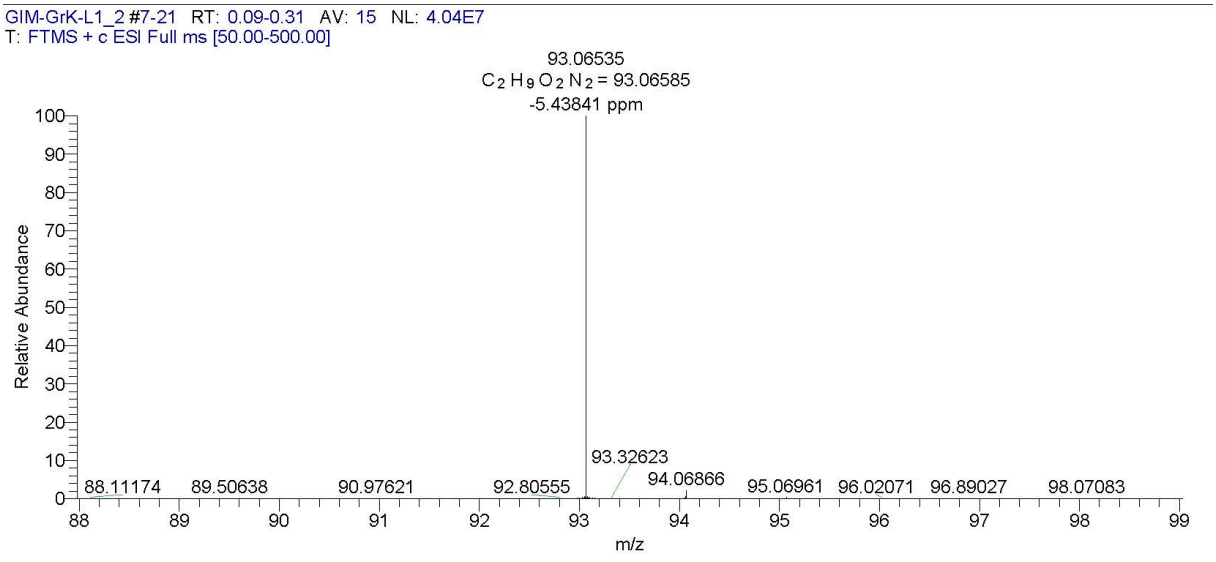
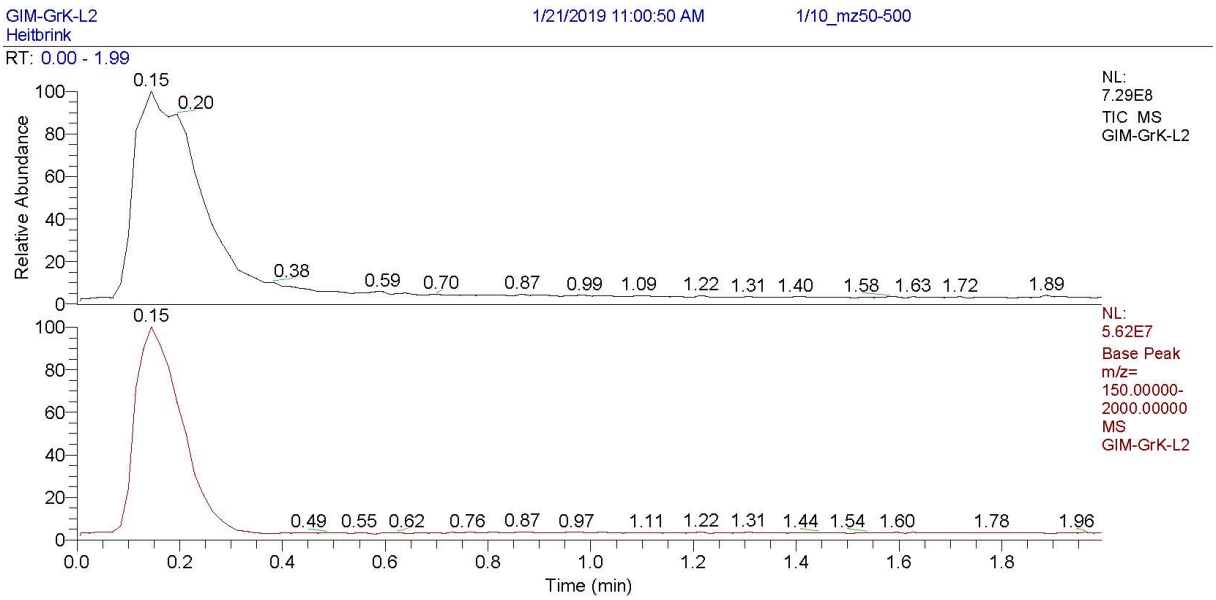

L5.2.fid
CARBON_128scans CDCl3 /DATA/topspin GrK 11
77.48
77.16
76.84
74.94
70.76
70.75
70.72
70.70
69.77
9.80
10.05
10.09
1.00
4.18
0.89
f1 (ppm)

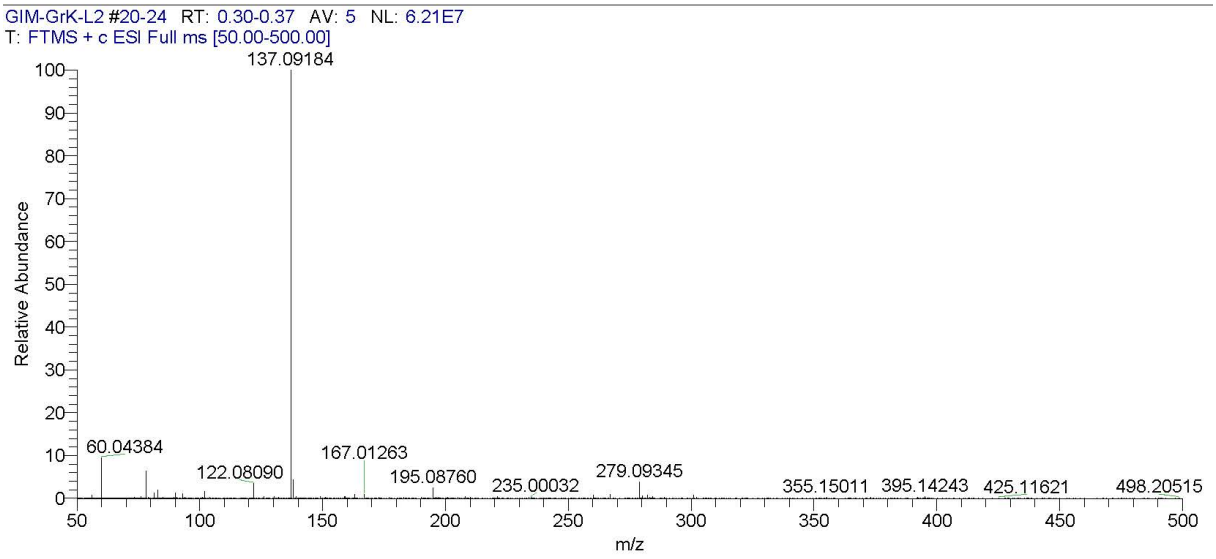
Figure SI14: High resolution mass spectra of L1-5

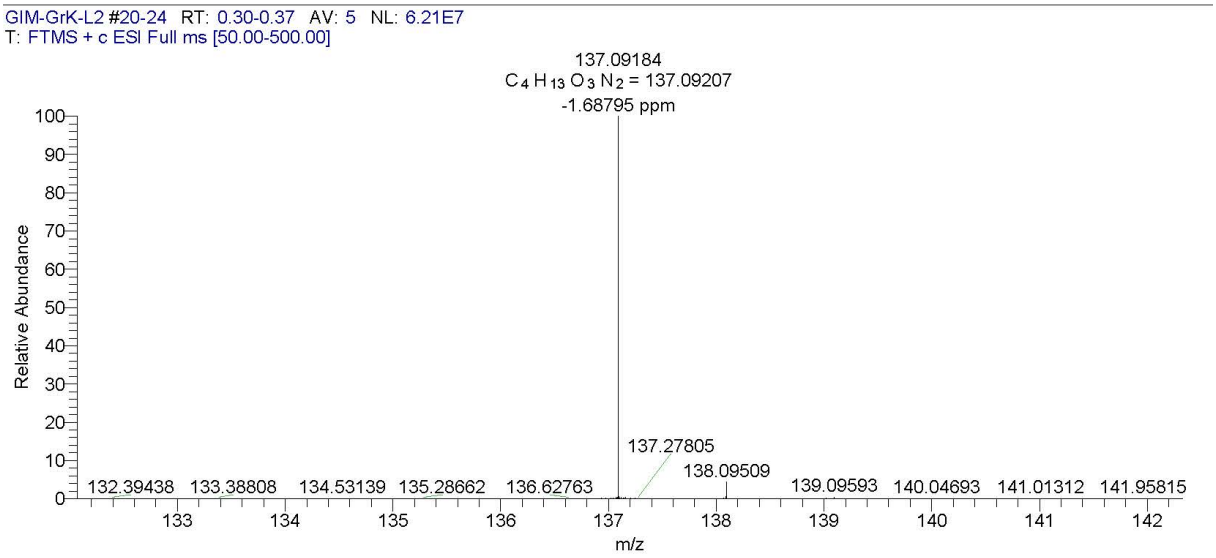


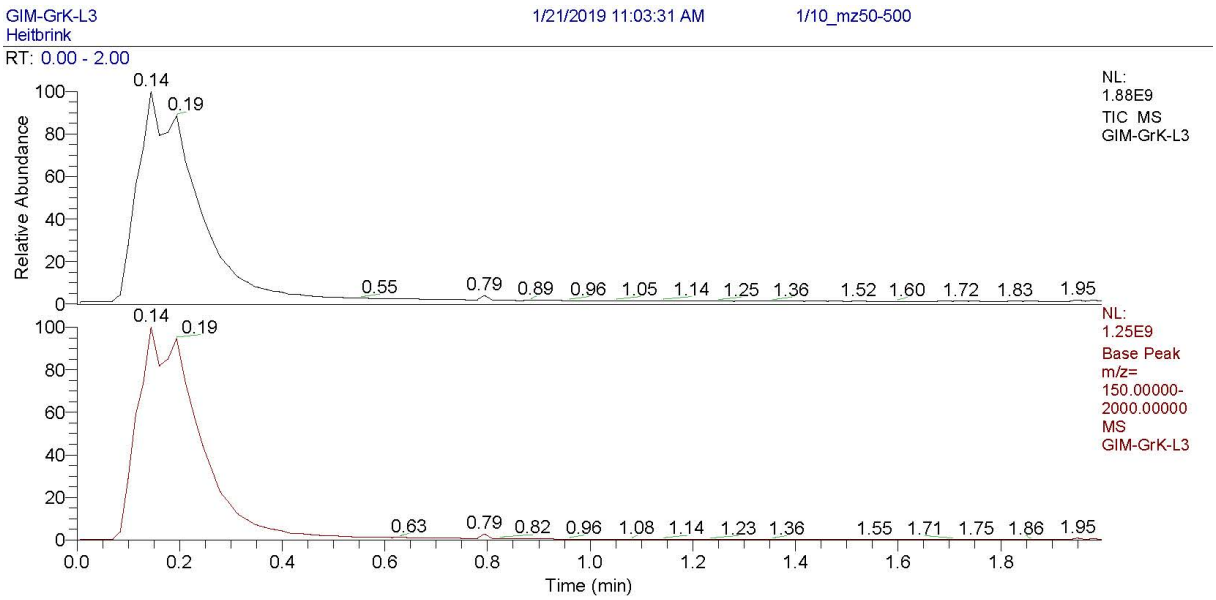


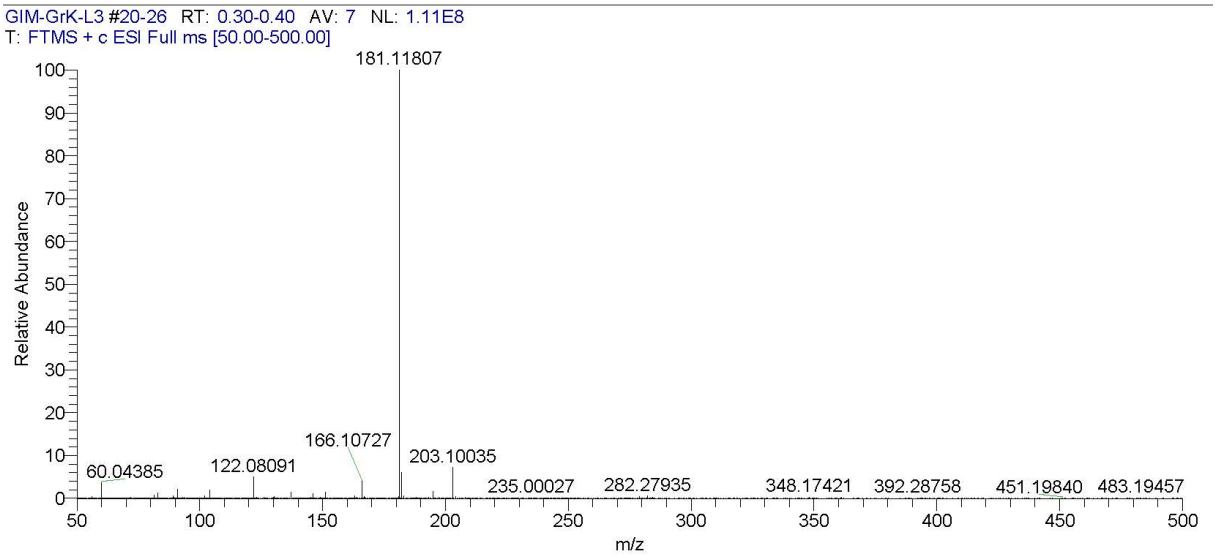


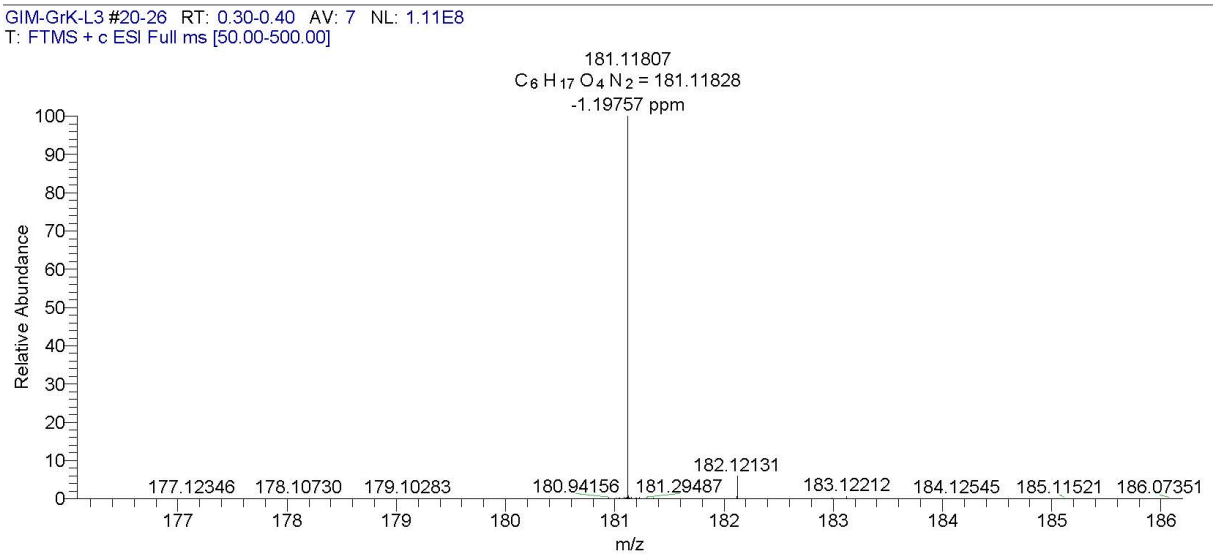

GIM-GrK-L2
Heitbrink
1/21/2019 11:00:50 AM
1/10_mz50-500
RT: 0.00 - 1.99
NL:
7.29E8
TIC MS
GIM-GrK-L2
NL:
5.62E7
Base Peak
m/z=
150.00000-
2000.00000
MS
GIM-GrK-L2
Relative Abundance
Time (min)


GIM-GrK-L2 #20-24 RT: 0.30-0.37 AV: 5 NL: 6.21E7
T: FTMS + c ESI Full ms [50.00-500.00]
137.09184
Relative Abundance
m/z


GIM-GrK-L2 #20-24 RT: 0.30-0.37 AV: 5 NL: 6.21E7
T: FTMS + c ESI Full ms [50.00-500.00]
137.09184
C4 H13 O3 N2 = 137.09207
-1.68795 ppm
Relative Abundance
m/z

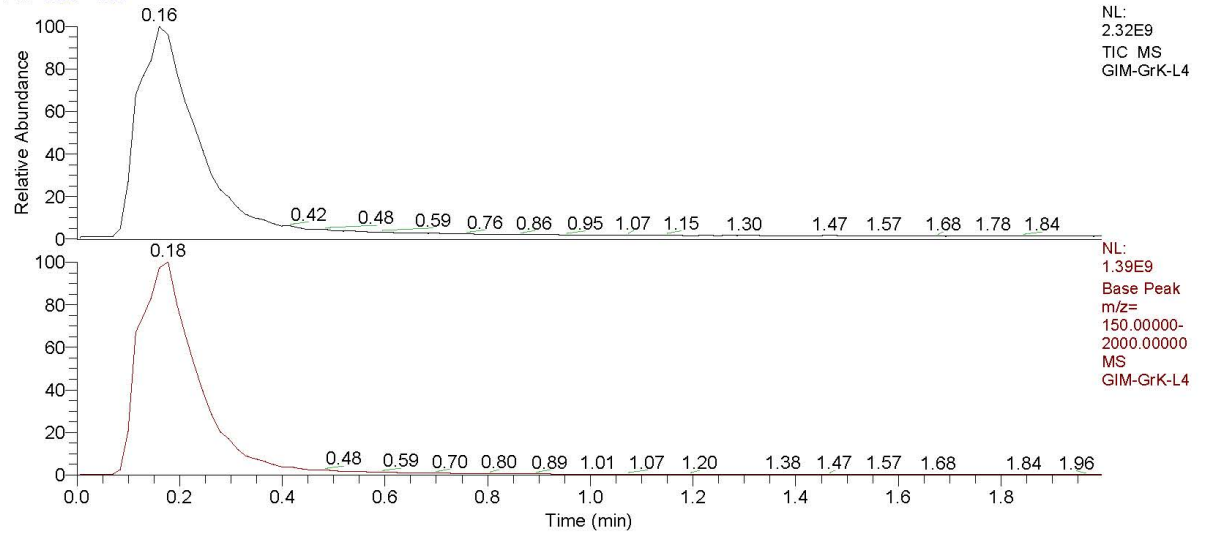

GIM-GrK-L3
Heitbrink
1/21/2019 11:03:31 AM
1/10_mz50-500
RT: 0.00 - 2.00
NL:
1.88E9
TIC MS
GIM-GrK-L3
NL:
1.25E9
Base Peak
m/z=
150.00000-
2000.00000
MS
GIM-GrK-L3
Relative Abundance
Time (min)


GIM-GrK-L3 #20-26 RT: 0.30-0.40 AV: 7 NL: 1.11E8
T: FTMS + c ESI Full ms [50.00-500.00]
181.11807
166.10727
203.10035
60.04385
122.08091
235.00027
282.27935
348.17421
392.28758
451.19840
483.19457
Relative Abundance
m/z

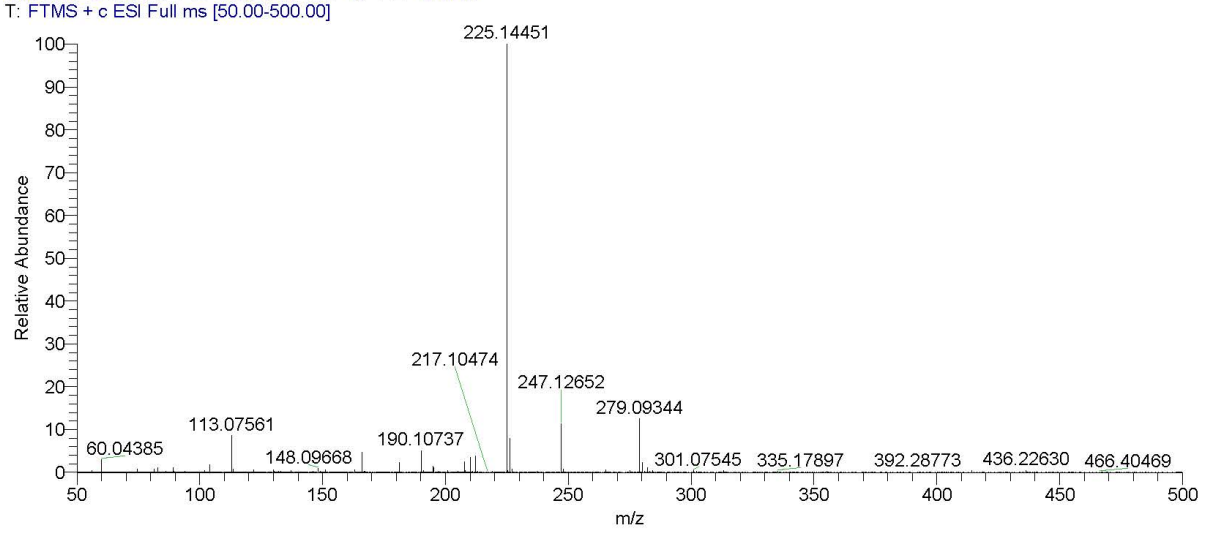

GIM-GrK-L3 #20-26 RT: 0.30-0.40 AV: 7 NL: 1.11E8
T: FTMS + c ESI Full ms [50.00-500.00]
181.11807
C6 H17 O4 N2 = 181.11828
-1.19757 ppm
182.12131
177.12346
178.10730
179.10283
180.94156
181.29487
183.12212
184.12545
185.11521
186.07351
Relative Abundance
m/z

GIM-GrK-L4 1/21/2019 11:06:11 AM 1/10_mz50-500
Heitbrink

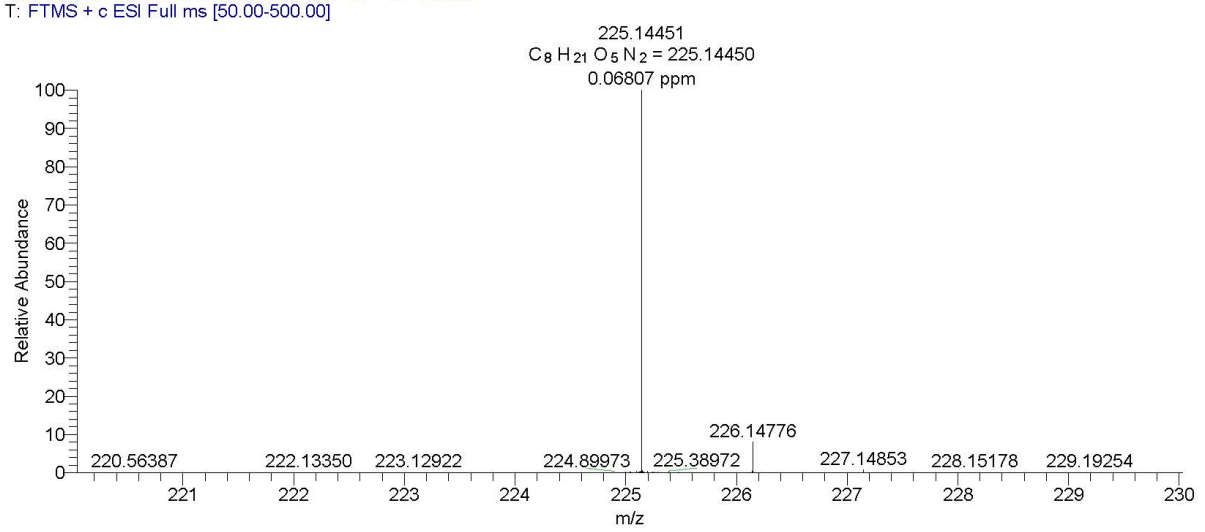
RT: 0.00 - 1.99



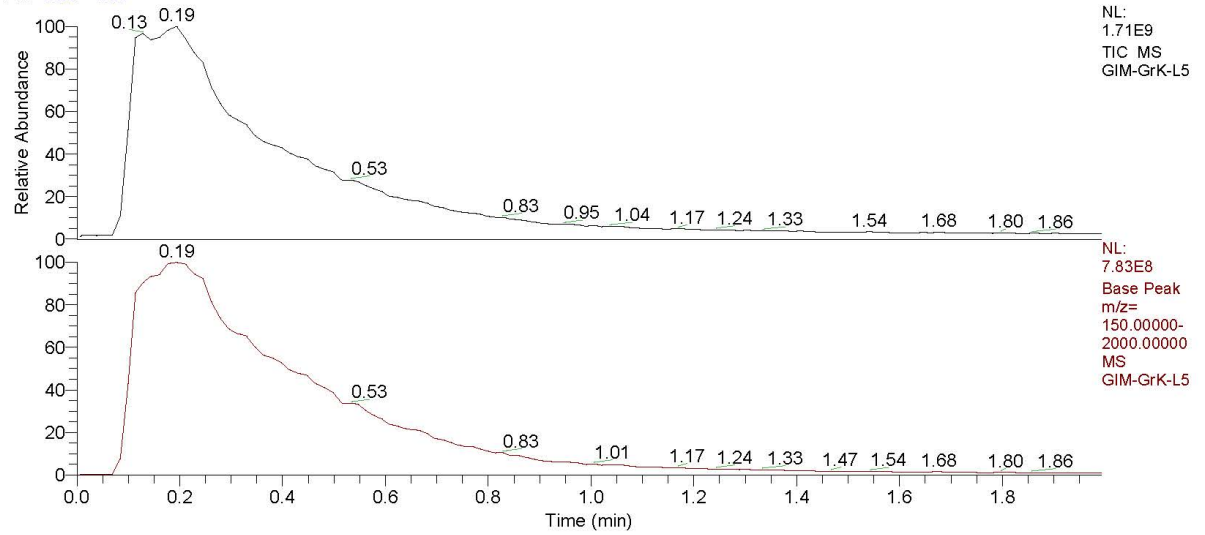
GIM-GrK-L4 #19-28 RT: 0.28-0.43 AV: 10 NL: 1.23E8
T: FTMS + c ESI Full ms [50.00-500.00]



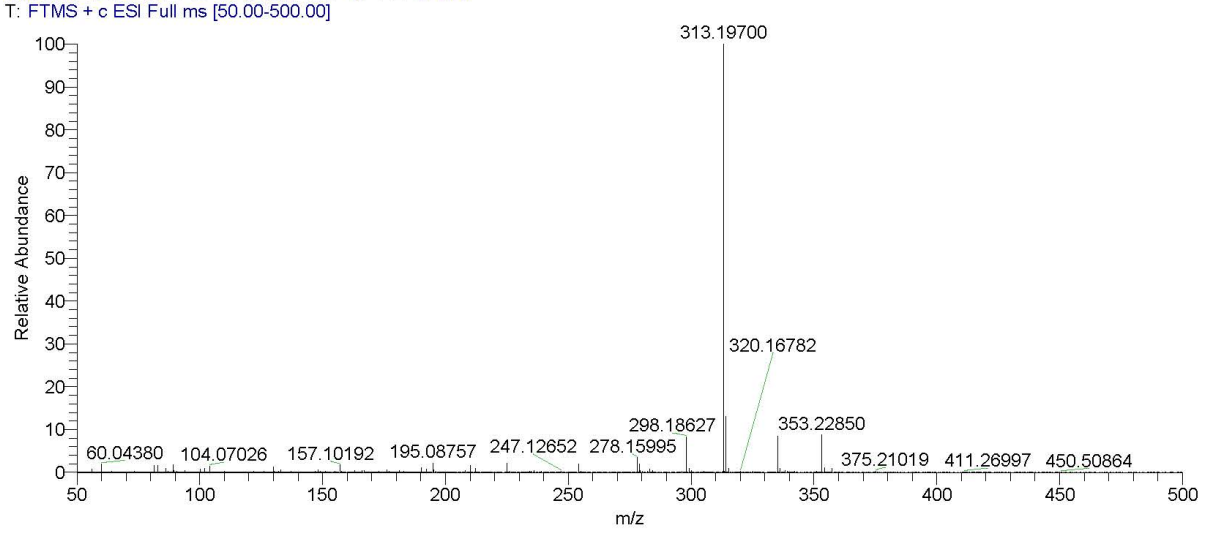
GIM-GrK-L4 #19-28 RT: 0.28-0.43 AV: 10 NL: 1.23E8
T: FTMS + c ESI Full ms [50.00-500.00]



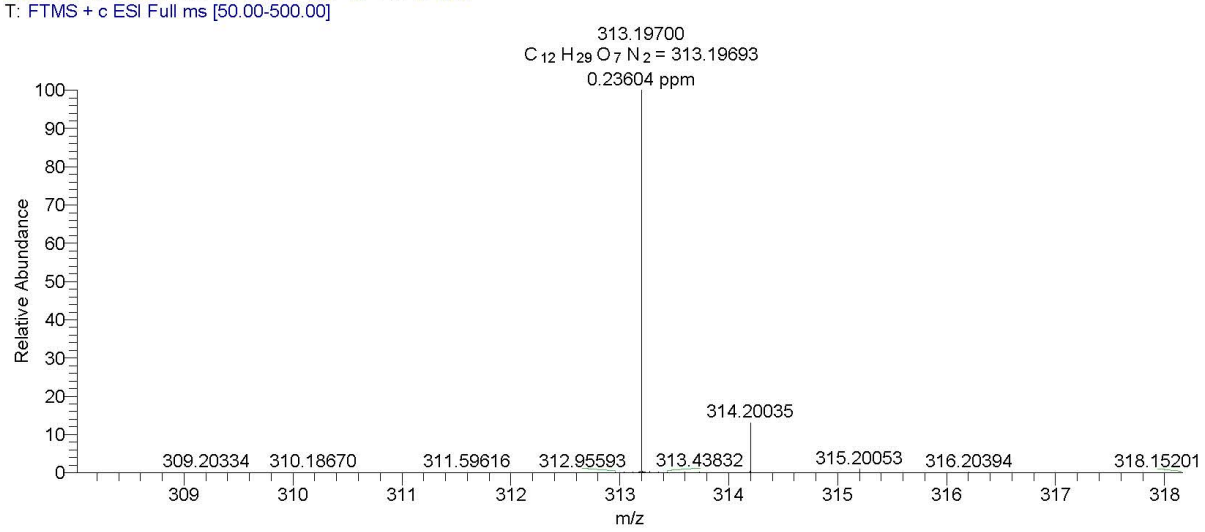
GIM-GrK-L5 1/21/2019 11:08:52 AM 1/10_mz50-500
Heitbrink
RT: 0.00 - 1.99



GIM-GrK-L5 #45-59 RT: 0.70-0.90 AV: 15 NL: 9.13E7
T: FTMS + c ESI Full ms [50.00-500.00]



GIM-GrK-L5 #45-59 RT: 0.70-0.90 AV: 15 NL: 9.13E7
T: FTMS + c ESI Full ms [50.00-500.00]



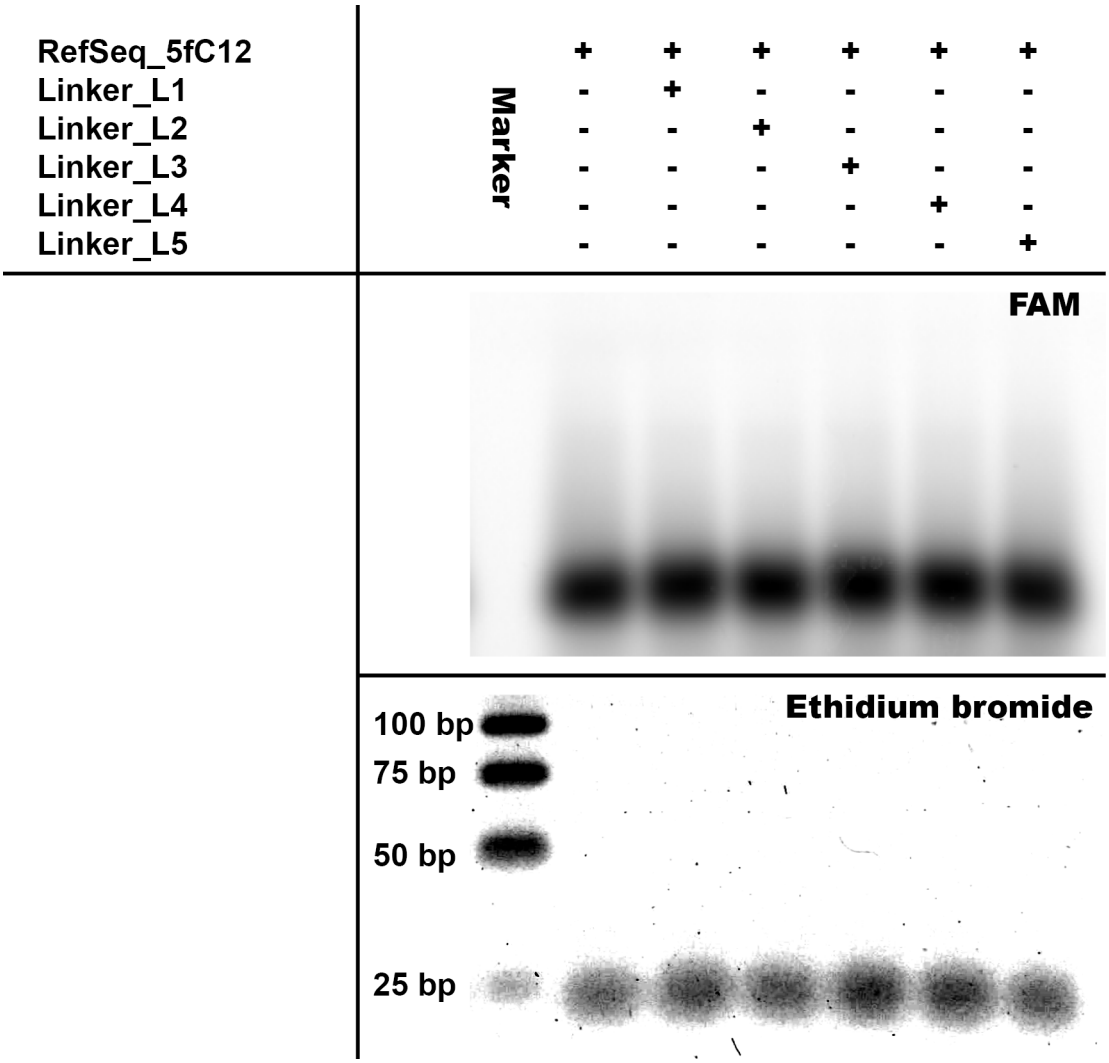
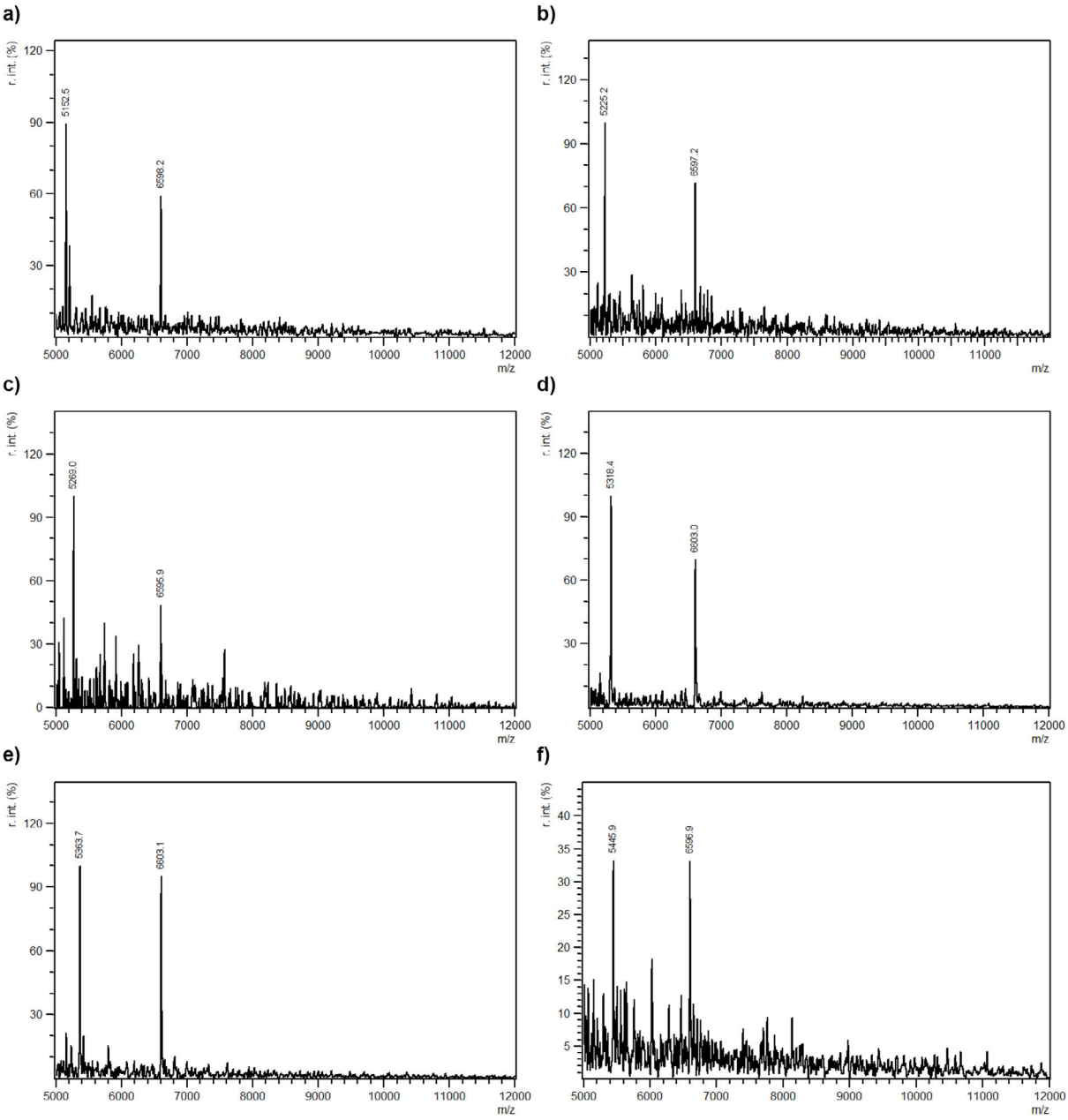


Figure SI15: 5 % agarose gels of 5'-FAM labeled RefSeq_5fC12 oligonucleotide after incubation under crosslinking conditions do not show crosslinking between 5fC-bearing oligonucleotides.



	-	+L1	+L2	+L3	+L4	+L5
obs. m/z [u]	5152.5	5225.2	5316.4	5269.0	5363.7	5445.9
calc. Δm/z [u]		75.1	119.1	163.1	207.1	295.2
obs. Δm/z [u]		72.7	116.5	163.9	211.2	293.4

Figure SI16: MALDI-TOF mass spectrometry analysis of reference oligonucleotides with 5fC after incubation under crosslinking conditions without or with linkers L1-5. Mass peaks of oligonucleotides with 5fC treated with linker L1-5 shift from the mass observed for the oligonucleotide missing a linker in the reaction to values corresponding with expected masses after oxime ligation (see table). Mass peaks of the complementary reverse strand appear at ~6600 u and do not shift in the presence of linker L1-5. Crosslinking between two oligonucleotides could not be observed for linker L1-5.

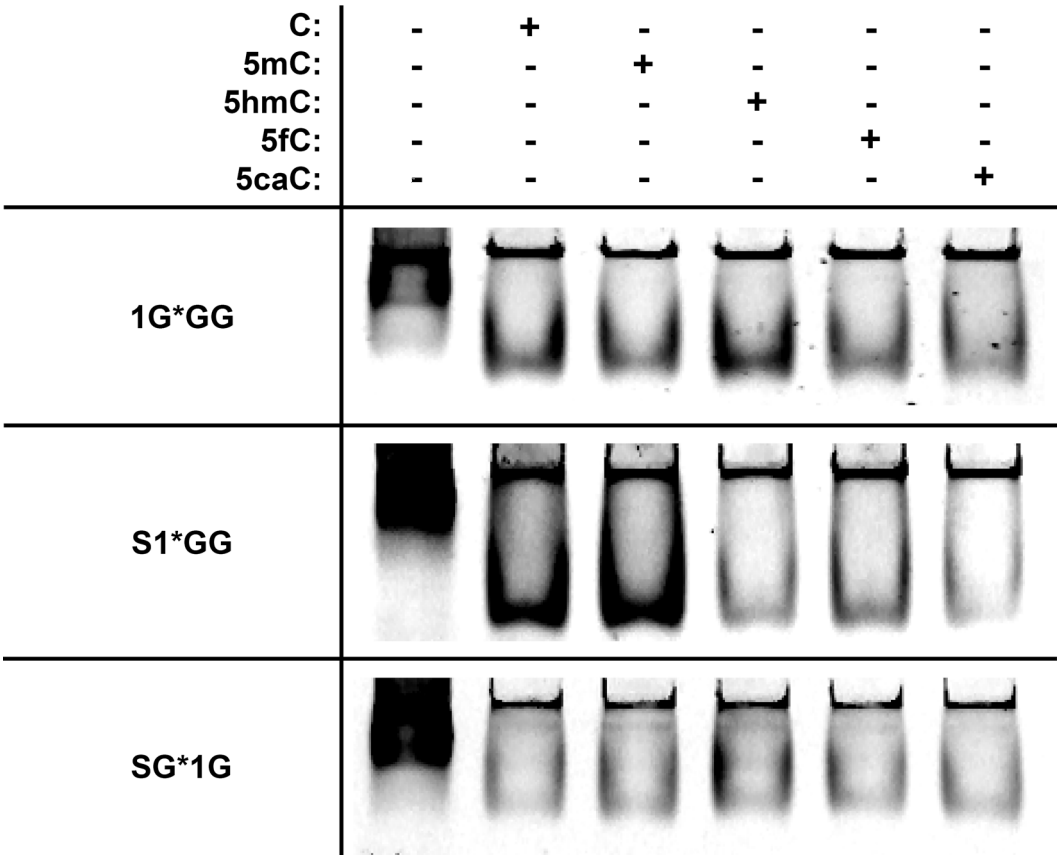


Figure SI17: EMSA assays of BRCA1a TALEs with pAcF and its target DNA under crosslinking reaction conditions. 100 pmol of oligonucleotides were buffered into phosphate buffer (pH = 6) and added to 100 pmol of TALE protein in a final volume of 50 µl of 1x TALE binding buffer BMaG1 (pH = 6). After incubation at room temperature and at 4 °C at 30 minutes each, 1 µl (2 pmol TALE/DNA complex) of the solution was buffered into 10 µl of TALE binding buffer (pH = 8) and subjected to native PAGE.

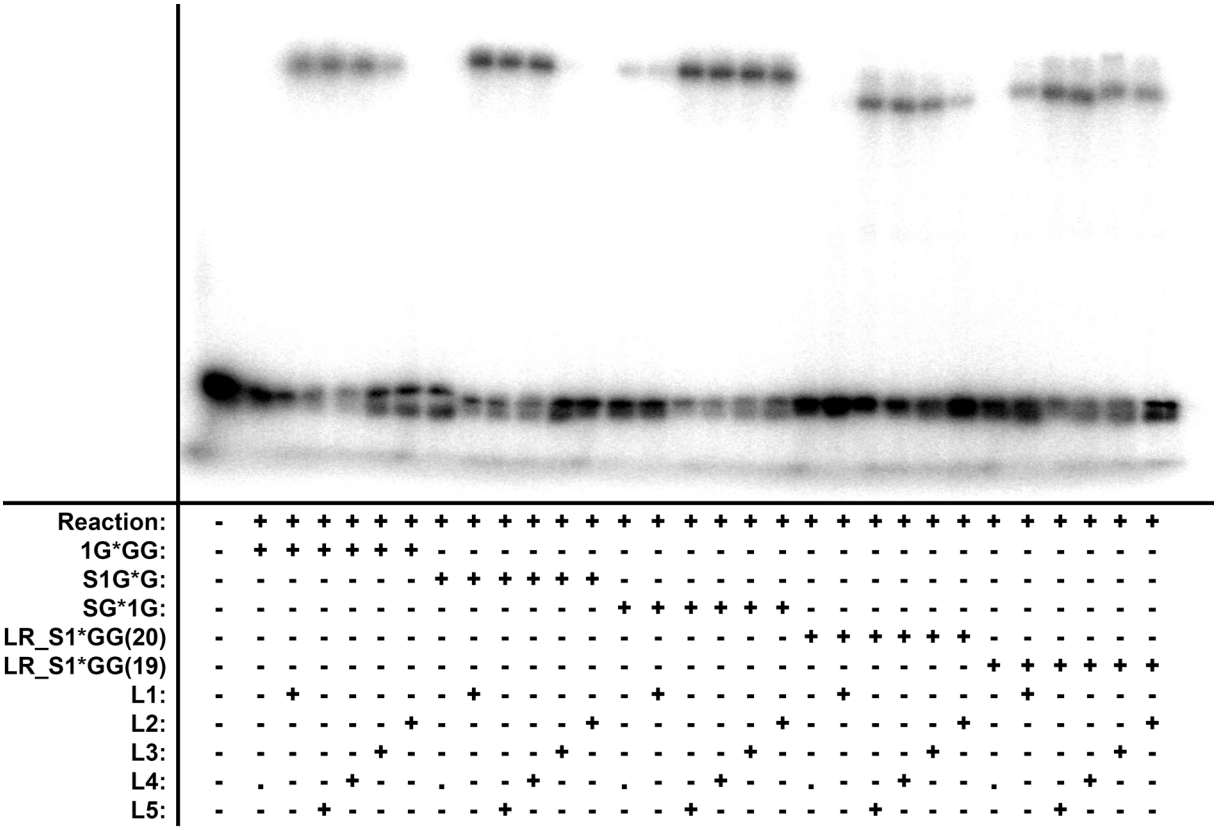


Figure SI18: SDS-PAGE gel image for TALE-DNA crosslinking as indicated from Fig. 35.

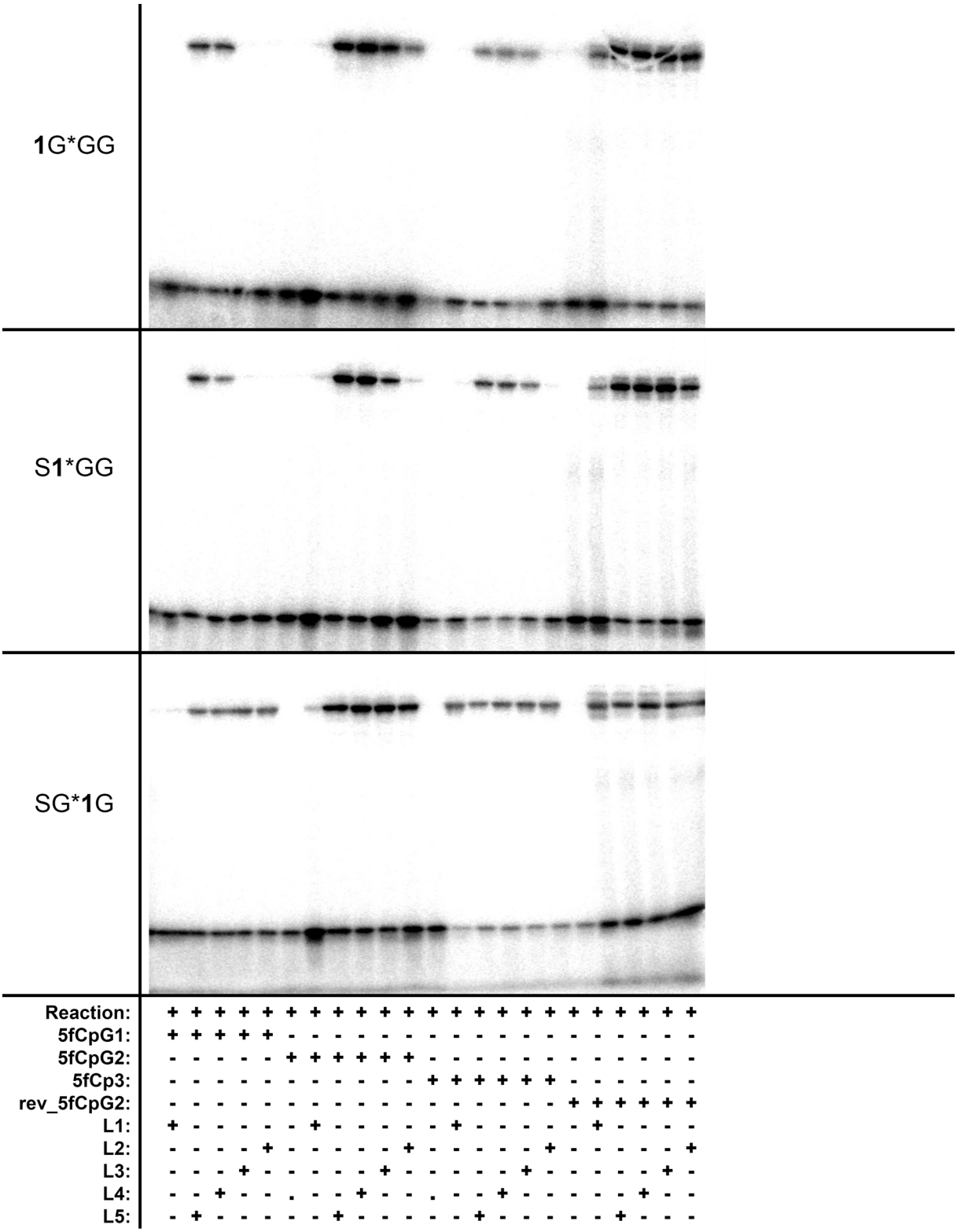


Figure SI19: SDS-PAGE gel images for positional resolutions of TALE-DNA crosslinking as indicated from Fig. 38.

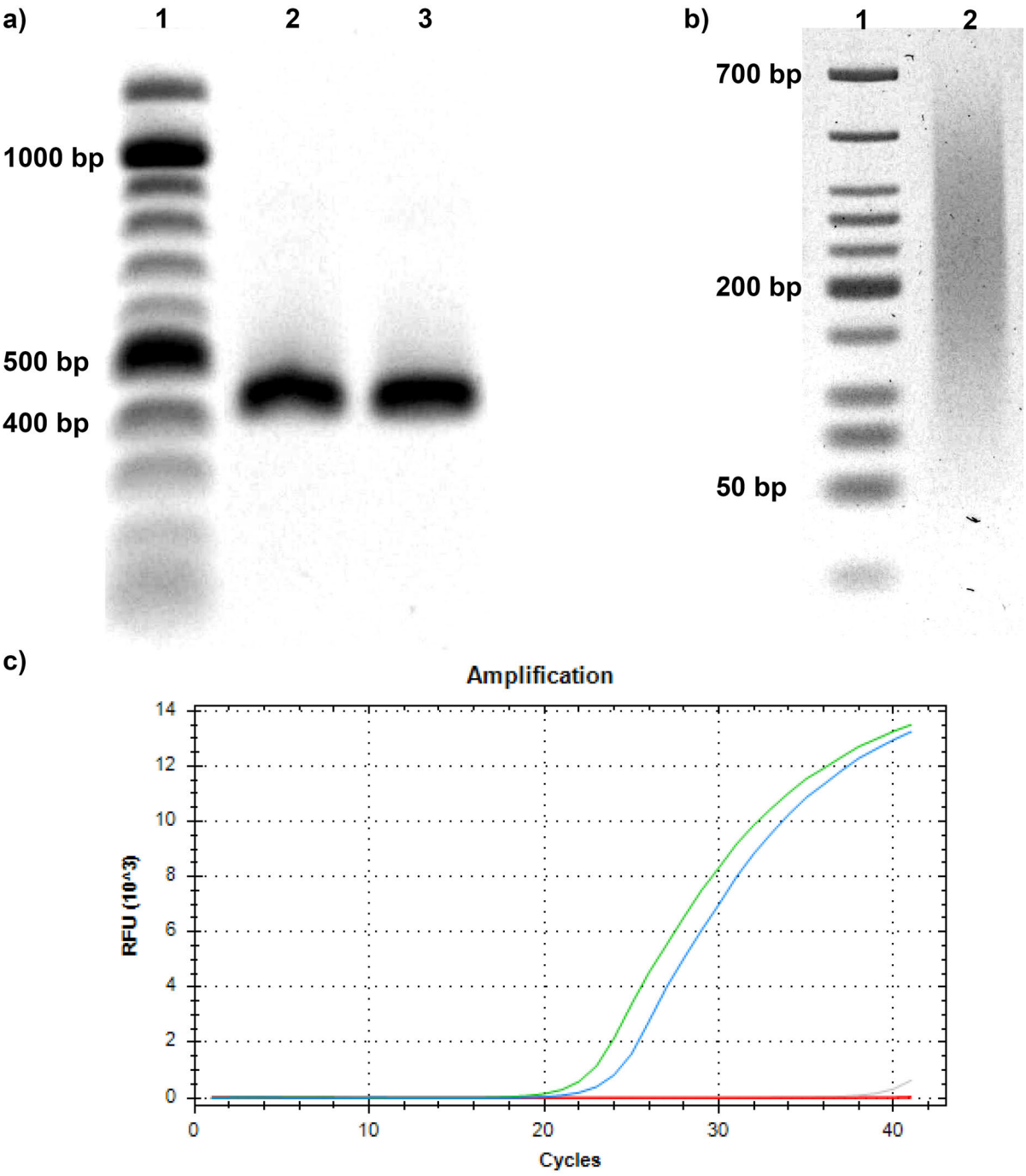


Figure SI20: Spike-ins and genomic DNA for enrichment. a) Agarose gel of 421bp_C13 → X spike-in PCR products. Lane 1: DNA-ladder, Lane 2: BRCA1a_421bp_TALE binding_C13 → 5hmC13, Lane 3: BRCA1a _421bp_TALE binding_C13 → 5fC13. **b)** Agarose gel analysis of human whole genome amplified DNA randomly sheared by sonication. Lane 1: DNA Ladder, Lane 2: Fragmented whole genome amplified human DNA (100ng). **c)** qPCR activity of 100 ng sheared WGA DNA (blue), 100ng background (bg) DNA (WGA DNA, sheared/digested) without (gray) and with BRCA1a_5fC13 spike-in (green).

7. Eidesstattliche Versicherung (Affidavit)

Gieß, Mario David
Name, Vorname
(Surname, first name)

191358
Matrikel-Nr.
(Enrolment number)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden, § 63 Abs. 5 Hochschulgesetz NRW.

Die Abgabe einer falschen Versicherung an Eides statt ist strafbar.

Wer vorsätzlich eine falsche Versicherung an Eides statt abgibt, kann mit einer Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft werden, § 156 StGB. Die fahrlässige Abgabe einer falschen Versicherung an Eides statt kann mit einer Freiheitsstrafe bis zu einem Jahr oder Geldstrafe bestraft werden, § 161 StGB.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:

Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offence can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offences of this type is the chancellor of the TU Dortmund University. In the case of multiple or other serious attempts at deception, the candidate can also be unenrolled, Section 63, paragraph 5 of the Universities Act of North Rhine-Westphalia.

The submission of a false affidavit is punishable.

Any person who intentionally submits a false affidavit can be punished with a prison sentence of up to three years or a fine, Section 156 of the Criminal Code. The negligent submission of a false affidavit can be punished with a prison sentence of up to one year or a fine, Section 161 of the Criminal Code.

I have taken note of the above official notification.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

Titel der Dissertation:
(Title of the thesis):

Chemoselective Conjugation Strategies for the Programmable Detection of Epigenetic Cytosine 5-Modifications with Transcription Activator-Like Effectors

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.
Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder der TU Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

I hereby swear that I have completed the present dissertation independently and without inadmissible external support. I have not used any sources or tools other than those indicated and have identified literal and analogous quotations.

The thesis in its current version or another version has not been presented to the TU Dortmund University or another university in connection with a state or academic examination.*

*Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the PhD thesis is the official and legally binding version.

Ort, Datum
(Place, date)

Unterschrift
(Signature)

8. References

(1) Pearce, M. *DNA: blueprint of life*; Wiley: New York, 1968.

(2) Doonan, S. *Nucleic Acids* 20; Royal Society of Chemistry: Cambridge, 2004.

(3) Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Stryer Biochemie*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013.

(4) Müller, S. *Nucleic Acids from A to Z*; Wiley-VCH: Hoboken, 2008.

(5) Travers, A.; Muskhelishvili, G. DNA structure and function. *The FEBS journal* **2015**, *282*, 2279–2295, DOI: 10.1111/febs.13307.

(6) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic acids research* **2006**, *34*, 564–574, DOI: 10.1093/nar/gkj454.

(7) Ussery, D. W. DNA Structure: A-, B- and Z-DNA Helix Families. *Encyclopedia of life sciences*; Wiley: Chichester, England, 2005; p 1.

(8) Cutter, A. R.; Hayes, J. J. A brief review of nucleosome structure. *FEBS letters* **2015**, *589*, 2914–2922, DOI: 10.1016/j.febslet.2015.05.016.

(9) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular biology of the cell*, Fourth edition; Garland Science Taylor & Francis Group: New York, NY, 2002.

(10) Lodish, H. *Molecular cell biology*, 4. ed., [Nachdr.]; Media connected; Freeman: New York, NY, 2002.

(11) Andrew Scott. Diagram of Chromatin / Nucleosome structure. https://www.oist.jp/news-center/photos/diagram-chromatin-nucleosome-structure (accessed May 28, 2019).

(12) Luger, K.; Dechassa, M. L.; Tremethick, D. J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature reviews. Molecular cell biology* **2012**, *13*, 436–447, DOI: 10.1038/nrm3382.

(13) Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **2007**, *128*, 787–800, DOI: 10.1016/j.cell.2007.01.028.

(14) Kornberg, R. D. The eukaryotic gene transcription machinery. *Biological chemistry* **2001**, *382*, 1103–1107, DOI: 10.1515/BC.2001.140.

(15) Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences* **2007**, *104*, 12955–12961, DOI: 10.1073/pnas.0704138104.

(16) Pang, Y. L. J.; Poruri, K.; Martinis, S. A. tRNA synthetase: tRNA aminoacylation and beyond. *Wiley interdisciplinary reviews. RNA* **2014**, *5*, 461–480, DOI: 10.1002/wrna.1224.

(17) Cooper, G. M. *The cell: A molecular approach*, 2. ed.; ASM Press: Washington, DC, 2000.

(18) Merrick, W. C.; Pavitt, G. D. Protein Synthesis Initiation in Eukaryotic Cells. *Cold Spring Harbor perspectives in biology* **2018**, *10,* DOI: 10.1101/cshperspect.a033092.

(19) Dever, T. E.; Dinman, J. D.; Green, R. Translation Elongation and Recoding in Eukaryotes. *Cold Spring Harbor perspectives in biology* **2018**, *10,* DOI: 10.1101/cshperspect.a032649.

(20) Hellen, C. U. T. Translation Termination and Ribosome Recycling in Eukaryotes. *Cold Spring Harbor perspectives in biology* **2018**, *10,* DOI: 10.1101/cshperspect.a032656.

(21) Bird, A. Perceptions of epigenetics. *Nature* **2007**, *447*, 396–398, DOI: 10.1038/nature05913.

(22) Bannister, A. J.; Kouzarides, T. Regulation of chromatin by histone modifications. *Cell research* **2011**, *21*, 381–395, DOI: 10.1038/cr.2011.22.

(23) Fu, Y.; He, C. Nucleic acid modifications with epigenetic significance. *Current opinion in chemical biology* **2012**, *16*, 516–524, DOI: 10.1016/j.cbpa.2012.10.002.

(24) Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature reviews. Genetics* **2018**, *19*, 81–92, DOI: 10.1038/nrg.2017.80.

(25) Ren, W.; Gao, L.; Song, J. Structural Basis of DNMT1 and DNMT3A-Mediated DNA Methylation. *Genes* **2018**, *9,* DOI: 10.3390/genes9120620.

(26) Bird, A. DNA methylation patterns and epigenetic memory. *Genes & development* **2002**, *16*, 6–21, DOI: 10.1101/gad.947102.

(27) Breiling, A.; Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & chromatin* **2015**, *8*, 24, DOI: 10.1186/s13072-015-0016-6.

(28) Spruijt, C. G.; Vermeulen, M. DNA methylation: old dog, new tricks? *Nature structural & molecular biology* **2014**, *21*, 949–954, DOI: 10.1038/nsmb.2910.

(29) Edwards, J. R.; Yarychkivska, O.; Boulard, M.; Bestor, T. H. DNA methylation and DNA methyltransferases. *Epigenetics & chromatin* **2017**, *10*, 23, DOI: 10.1186/s13072-017-0130-8.

(30) Kumar, S.; Chinnusamy, V.; Mohapatra, T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Frontiers in genetics* **2018**, *9*, 640, DOI: 10.3389/fgene.2018.00640.

(31) Robertson, K. D. DNA methylation and human disease. *Nature reviews. Genetics* **2005**, *6*, 597–610, DOI: 10.1038/nrg1655.

(32) Jang, H. S.; Shin, W. J.; Lee, J. E.; Do, J. T. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes* **2017**, *8,* DOI: 10.3390/genes8060148.

(33) Raiber, E.-A.; Hardisty, R.; van Delft, P.; Balasubramanian, S. Mapping and elucidating the function of modified bases in DNA. *Nat Rev Chem* **2017**, *1*, 69, DOI: 10.1038/s41570-017-0069.

(34) Luo, C.; Hajkova, P.; Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science (New York, N.Y.)* **2018**, *361*, 1336–1340, DOI: 10.1126/science.aat6806.

(35) Muñoz-López, Á.; Summerer, D. Recognition of Oxidized 5-Methylcytosine Derivatives in DNA by Natural and Engineered Protein Scaffolds. *Chemical record (New York, N.Y.)* **2018**, *18*, 105–116, DOI: 10.1002/tcr.201700088.

(36) Law, J. A.; Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* **2010**, *11*, 204–220, DOI: 10.1038/nrg2719.

(37) Lister, R.; Pelizzola, M.; Dowen, R. H.; Hawkins, R. D.; Hon, G.; Tonti-Filippini, J.; Nery, J. R.; Lee, L.; Ye, Z.; Ngo, Q.-M. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **2009**, *462*, 315–322, DOI: 10.1038/nature08514.

(38) Smith, Z. D.; Meissner, A. DNA methylation: roles in mammalian development. *Nature reviews. Genetics* **2013**, *14*, 204–220, DOI: 10.1038/nrg3354.

(39) Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **2012**, *13*, 484–492, DOI: 10.1038/nrg3230.

(40) Hardwick, J. S.; Lane, A. N.; Brown, T. Epigenetic Modifications of Cytosine: Biophysical Properties, Regulation, and Function in Mammalian DNA. *BioEssays : news and reviews in molecular, cellular and developmental biology* **2018**, *40,* DOI: 10.1002/bies.201700199.

(41) Wu, S. C.; Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nature reviews. Molecular cell biology* **2010**, *11*, 607–620, DOI: 10.1038/nrm2950.

(42) Kohli, R. M.; Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **2013**, *502*, 472–479, DOI: 10.1038/nature12750.

(43) Branco, M. R.; Ficz, G.; Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature reviews. Genetics* **2011**, *13*, 7–13, DOI: 10.1038/nrg3080.

(44) Wu, X.; Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nature reviews. Genetics* **2017**, *18*, 517–534, DOI: 10.1038/nrg.2017.33.

(45) Shi, D.-Q.; Ali, I.; Tang, J.; Yang, W.-C. New Insights into 5hmC DNA Modification: Generation, Distribution and Function. *Frontiers in genetics* **2017**, *8*, 100, DOI: 10.3389/fgene.2017.00100.

(46) Loenarz, C.; Schofield, C. J. Physiological and biochemical aspects of hydroxylations and demethylations catalyzed by human 2-oxoglutarate oxygenases. *Trends in biochemical sciences* **2011**, *36*, 7–18, DOI: 10.1016/j.tibs.2010.07.002.

(47) Wu, H.; Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes & development* **2011**, *25*, 2436–2452, DOI: 10.1101/gad.179184.111.

(48) Tahiliani, M.; Koh, K. P.; Shen, Y.; Pastor, W. A.; Bandukwala, H.; Brudno, Y.; Agarwal, S.; Iyer, L. M.; Liu, D. R.; Aravind, L. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)* **2009**, *324*, 930–935, DOI: 10.1126/science.1170116.

(49) Ito, S.; Shen, L.; Dai, Q.; Wu, S. C.; Collins, L. B.; Swenberg, J. A.; He, C.; Zhang, Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)* **2011**, *333*, 1300–1303, DOI: 10.1126/science.1210597.

(50) He, Y.-F.; Li, B.-Z.; Li, Z.; Liu, P.; Wang, Y.; Tang, Q.; Ding, J.; Jia, Y.; Chen, Z.; Li, L. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (New York, N.Y.)* **2011**, *333*, 1303–1307, DOI: 10.1126/science.1210944.

(51) Lu, X.; Zhao, B. S.; He, C. TET family proteins: oxidation activity, interacting molecules, and functions in diseases. *Chemical reviews* **2015**, *115*, 2225–2239, DOI: 10.1021/cr500470n.

(52) Maiti, A.; Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *The Journal of biological chemistry* **2011**, *286*, 35334–35338, DOI: 10.1074/jbc.C111.284620.

(53) Drohat, A. C.; Coey, C. T. Role of Base Excision "Repair" Enzymes in Erasing Epigenetic Marks from DNA. *Chemical reviews* **2016**, *116*, 12711–12729, DOI: 10.1021/acs.chemrev.6b00191.

(54) Kriukienė, E.; Liutkevičiūtė, Z.; Klimašauskas, S. 5-Hydroxymethylcytosine--the elusive epigenetic mark in mammalian DNA. *Chemical Society reviews* **2012**, *41*, 6916–6930, DOI: 10.1039/c2cs35104h.

(55) Iwan, K.; Rahimoff, R.; Kirchner, A.; Spada, F.; Schröder, A. S.; Kosmatchev, O.; Ferizaj, S.; Steinbacher, J.; Parsa, E.; Müller, M. *et al.* 5-Formylcytosine to cytosine conversion by C-C bond cleavage in vivo. *Nature chemical biology* **2018**, *14*, 72–78, DOI: 10.1038/nchembio.2531.

(56) Renciuk, D.; Blacque, O.; Vorlickova, M.; Spingler, B. Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic acids research* **2013**, *41*, 9891–9900, DOI: 10.1093/nar/gkt738.

(57) Lercher, L.; McDonough, M. A.; El-Sagheer, A. H.; Thalhammer, A.; Kriaucionis, S.; Brown, T.; Schofield, C. J. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chemical communications (Cambridge, England)* **2014**, *50*, 1794–1796, DOI: 10.1039/c3cc48151d.

(58) Hu, L.; Lu, J.; Cheng, J.; Rao, Q.; Li, Z.; Hou, H.; Lou, Z.; Zhang, L.; Li, W.; Gong, W. *et al.* Structural insight into substrate preference for TET-mediated oxidation. *Nature* **2015**, *527*, 118–122, DOI: 10.1038/nature15713.

(59) Szulik, M. W.; Pallan, P. S.; Nocek, B.; Voehler, M.; Banerjee, S.; Brooks, S.; Joachimiak, A.; Egli, M.; Eichman, B. F.; Stone, M. P. Differential stabilities and sequence-dependent base pair opening dynamics of Watson-Crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. *Biochemistry* **2015**, *54*, 1294–1305, DOI: 10.1021/bi501534x.

(60) Irrera, S.; Portalone, G. First X-ray diffraction and quantum chemical study of proton-acceptor and proton-donor forms of 5-carboxylcytosine, the last-discovered nucleobase. *Journal of Molecular Structure* **2013**, *1050*, 140–150, DOI: 10.1016/j.molstruc.2013.07.023.

(61) Ngo, T. T. M.; Yoo, J.; Dai, Q.; Zhang, Q.; He, C.; Aksimentiev, A.; Ha, T. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nature communications* **2016**, *7*, 10813, DOI: 10.1038/ncomms10813.

(62) Dai, Q.; Sanstead, P. J.; Peng, C. S.; Han, D.; He, C.; Tokmakoff, A. Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. *ACS chemical biology* **2016**, *11*, 470–477, DOI: 10.1021/acschembio.5b00762.

(63) Hardwick, J. S.; Ptchelkine, D.; El-Sagheer, A. H.; Tear, I.; Singleton, D.; Phillips, S. E. V.; Lane, A. N.; Brown, T. 5-Formylcytosine does not change the global structure of DNA. *Nature structural & molecular biology* **2017**, *24*, 544–552, DOI: 10.1038/nsmb.3411.

(64) Liu, S.; Wang, J.; Su, Y.; Guerrero, C.; Zeng, Y.; Mitra, D.; Brooks, P. J.; Fisher, D. E.; Song, H.; Wang, Y. Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. *Nucleic acids research* **2013**, *41*, 6421–6429, DOI: 10.1093/nar/gkt360.

(65) Kriaucionis, S.; Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (New York, N.Y.)* **2009**, *324*, 929–930, DOI: 10.1126/science.1169786.

(66) Globisch, D.; Münzel, M.; Müller, M.; Michalakis, S.; Wagner, M.; Koch, S.; Brückl, T.; Biel, M.; Carell, T. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one* **2010**, *5*, e15367, DOI: 10.1371/journal.pone.0015367.

(67) Ito, S.; D'Alessio, A. C.; Taranova, O. V.; Hong, K.; Sowers, L. C.; Zhang, Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **2010**, *466*, 1129–1133, DOI: 10.1038/nature09303.

(68) Pfeifer, G. P.; Xiong, W.; Hahn, M. A.; Jin, S.-G. The role of 5-hydroxymethylcytosine in human cancer. *Cell and tissue research* **2014**, *356*, 631–641, DOI: 10.1007/s00441-014-1896-7.

(69) Szulwach, K. E.; Li, X.; Li, Y.; Song, C.-X.; Wu, H.; Dai, Q.; Irier, H.; Upadhyay, A. K.; Gearing, M.; Levey, A. I. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature neuroscience* **2011**, *14*, 1607–1616, DOI: 10.1038/nn.2959.

(70) Chia, N.; Wang, L.; Lu, X.; Senut, M.-C.; Brenner, C. A.; Ruden, D. M. Hypothesis: Environmental regulation of 5-hydroxymethylcytosine by oxidative stress. *Epigenetics* **2011**, *6*, 853–856, DOI: 10.4161/epi.6.7.16461.

(71) Delatte, B.; Jeschke, J.; Defrance, M.; Bachman, M.; Creppe, C.; Calonne, E.; Bizet, M.; Deplus, R.; Marroquí, L.; Libin, M. *et al.* Genome-wide hydroxymethylcytosine pattern changes in response to oxidative stress. *Scientific reports* **2015**, *5*, 12714, DOI: 10.1038/srep12714.

(72) Shen, L.; Wu, H.; Diep, D.; Yamaguchi, S.; D'Alessio, A. C.; Fung, H.-L.; Zhang, K.; Zhang, Y. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **2013**, *153*, 692–706, DOI: 10.1016/j.cell.2013.04.002.

(73) Bachman, M.; Uribe-Lewis, S.; Yang, X.; Williams, M.; Murrell, A.; Balasubramanian, S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature chemistry* **2014**, *6*, 1049–1055, DOI: 10.1038/nchem.2064.

(74) Bachman, M.; Uribe-Lewis, S.; Yang, X.; Burgess, H. E.; Iurlaro, M.; Reik, W.; Murrell, A.; Balasubramanian, S. 5-Formylcytosine can be a stable DNA modification in mammals. *Nature chemical biology* **2015**, *11*, 555–557, DOI: 10.1038/nchembio.1848.

(75) Su, M.; Kirchner, A.; Stazzoni, S.; Müller, M.; Wagner, M.; Schröder, A.; Carell, T. 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angewandte Chemie (International ed. in English)* **2016**, *55*, 11797–11800, DOI: 10.1002/anie.201605994.

(76) Pfaffeneder, T.; Hackner, B.; Truss, M.; Münzel, M.; Müller, M.; Deiml, C. A.; Hagemeier, C.; Carell, T. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angewandte Chemie (International ed. in English)* **2011**, *50*, 7008–7012, DOI: 10.1002/anie.201103899.

(77) Song, C.-X.; Szulwach, K. E.; Dai, Q.; Fu, Y.; Mao, S.-Q.; Lin, L.; Street, C.; Li, Y.; Poidevin, M.; Wu, H. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **2013**, *153*, 678–691, DOI: 10.1016/j.cell.2013.04.001.

(78) Booth, M. J.; Marsico, G.; Bachman, M.; Beraldi, D.; Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nature chemistry* **2014**, *6*, 435–440, DOI: 10.1038/nchem.1893.

(79) Iurlaro, M.; McInroy, G. R.; Burgess, H. E.; Dean, W.; Raiber, E.-A.; Bachman, M.; Beraldi, D.; Balasubramanian, S.; Reik, W. In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome biology* **2016**, *17*, 141, DOI: 10.1186/s13059-016-1001-5.

(80) Zhu, C.; Gao, Y.; Guo, H.; Xia, B.; Song, J.; Wu, X.; Zeng, H.; Kee, K.; Tang, F.; Yi, C. Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell stem cell* **2017**, *20*, 720-731.e5, DOI: 10.1016/j.stem.2017.02.013.

(81) Li, F.; Zhang, Y.; Bai, J.; Greenberg, M. M.; Xi, Z.; Zhou, C. 5-Formylcytosine Yields DNA-Protein Cross-Links in Nucleosome Core Particles. *Journal of the American Chemical Society* **2017**, *139*, 10617–10620, DOI: 10.1021/jacs.7b05495.

(82) Ji, S.; Shao, H.; Han, Q.; Seiler, C. L.; Tretyakova, N. Y. Reversible DNA-Protein Cross-Linking at Epigenetic DNA Marks. *Angewandte Chemie (International ed. in English)* **2017**, *56*, 14130–14134, DOI: 10.1002/anie.201708286.

(83) Ji, S.; Fu, I.; Naldiga, S.; Shao, H.; Basu, A. K.; Broyde, S.; Tretyakova, N. Y. 5-Formylcytosine mediated DNA-protein cross-links block DNA replication and induce mutations in human cells. *Nucleic acids research* **2018**, *46*, 6455–6469, DOI: 10.1093/nar/gky444.

(84) Raiber, E.-A.; Portella, G.; Martínez Cuesta, S.; Hardisty, R.; Murat, P.; Li, Z.; Iurlaro, M.; Dean, W.; Spindel, J.; Beraldi, D. *et al.* 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nature chemistry* **2018**, *10*, 1258–1266, DOI: 10.1038/s41557-018-0149-x.

(85) Hofer, A.; Liu, Z. J.; Balasubramanian, S. Detection, Structure and Function of Modified DNA Bases. *J. Am. Chem. Soc.* **2019**, *141*, 6420–6429, DOI: 10.1021/jacs.9b01915.

(86) Raiber, E.-A.; Beraldi, D.; Ficz, G.; Burgess, H. E.; Branco, M. R.; Murat, P.; Oxley, D.; Booth, M. J.; Reik, W.; Balasubramanian, S. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome biology* **2012**, *13*, R69, DOI: 10.1186/gb-2012-13-8-r69.

(87) Wheldon, L. M.; Abakir, A.; Ferjentsik, Z.; Dudnakova, T.; Strohbuecker, S.; Christie, D.; Dai, N.; Guan, S.; Foster, J. M.; Corrêa, I. R. *et al.* Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. *Cell reports* **2014**, *7*, 1353–1361, DOI: 10.1016/j.celrep.2014.05.003.

(88) Eleftheriou, M.; Pascual, A. J.; Wheldon, L. M.; Perry, C.; Abakir, A.; Arora, A.; Johnson, A. D.; Auer, D. T.; Ellis, I. O.; Madhusudan, S. *et al.* 5-Carboxylcytosine levels are elevated in human breast cancers and gliomas. *Clinical epigenetics* **2015**, *7*, 88, DOI: 10.1186/s13148-015-0117-x.

(89) Lewis, L. C.; Lo, P. C. K.; Foster, J. M.; Dai, N.; Corrêa, I. R.; Durczak, P. M.; Duncan, G.; Ramsawhook, A.; Aithal, G. P.; Denning, C. *et al.* Dynamics of 5-carboxylcytosine during hepatic differentiation: Potential general role for active demethylation by DNA repair in lineage specification. *Epigenetics* **2017**, *12*, 277–286, DOI: 10.1080/15592294.2017.1292189.

(90) Wang, L.; Zhou, Y.; Xu, L.; Xiao, R.; Lu, X.; Chen, L.; Chong, J.; Li, H.; He, C.; Fu, X.-D. *et al.* Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature* **2015**, *523*, 621–625, DOI: 10.1038/nature14482.

(91) Wang, D.; Hashimoto, H.; Zhang, X.; Barwick, B. G.; Lonial, S.; Boise, L. H.; Vertino, P. M.; Cheng, X. MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic acids research* **2017**, *45*, 2396–2407, DOI: 10.1093/nar/gkw1184.

(92) Nanan, K.; Sturgill, D.; Prigge, M.; Thenoz, M.; Dillman, A.; Mandler, M.; Oberdoerffer, S. *TET-catalyzed 5-carboxylcytosine promotes CTCF binding to suboptimal sequences genome-wide*, 2018.

(93) Spruijt, C. G.; Gnerlich, F.; Smits, A. H.; Pfaffeneder, T.; Jansen, P. W. T. C.; Bauer, C.; Münzel, M.; Wagner, M.; Müller, M.; Khan, F. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **2013**, *152*, 1146–1159, DOI: 10.1016/j.cell.2013.02.004.

(94) Iurlaro, M.; Ficz, G.; Oxley, D.; Raiber, E.-A.; Bachman, M.; Booth, M. J.; Andrews, S.; Balasubramanian, S.; Reik, W. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome biology* **2013**, *14*, R119, DOI: 10.1186/gb-2013-14-10-r119.

(95) Chen, Y.; Hong, T.; Wang, S.; Mo, J.; Tian, T.; Zhou, X. Epigenetic modification of nucleic acids: from basic studies to medical applications. *Chemical Society reviews* **2017**, *46*, 2844–2872, DOI: 10.1039/c6cs00599c.

(96) Shen, L.; Zhang, Y. Enzymatic analysis of Tet proteins: key enzymes in the metabolism of DNA methylation. *Methods in enzymology* **2012**, *512*, 93–105, DOI: 10.1016/B978-0-12-391940-3.00005-6.

(97) Reddy, M.V.; Gupta, R. C.; Randerath, K. 32P-base analysis of DNA. *Analytical Biochemistry* **1981**, *117*, 271–279, DOI: 10.1016/0003-2697(81)90722-3.

(98) Yuan, B.-F.; Feng, Y.-Q. Recent advances in the analysis of 5-methylcytosine and its oxidation products. *TrAC Trends in Analytical Chemistry* **2014**, *54*, 24–35, DOI: 10.1016/j.trac.2013.11.002.

(99) Fraga, M. F.; Uriol, E.; Diego, L. B.; Berdasco, M.; Esteller, M.; Cañal, M. J.; Rodríguez, R. High-performance capillary electrophoretic method for the quantification of 5-methyl 2'-deoxycytidine in genomic DNA: Application to plant, animal and human cancer tissues. *Electrophoresis* **2002**, *23*, 1677, DOI: 10.1002/1522-2683(200206)23:11<1677:AID-ELPS1677>3.0.CO;2-Z.

(100) Stach, D.; Schmitz, O. J.; Stilgenbauer, S.; Benner, A.; Döhner, H.; Wiessler, M.; Lyko, F. Capillary electrophoretic analysis of genomic DNA methylation levels. *Nucl Acids Res* **2003**, *31*, e2.

(101) Kato, D.; Sekioka, N.; Ueda, A.; Kurita, R.; Hirono, S.; Suzuki, K.; Niwa, O. A nanocarbon film electrode as a platform for exploring DNA methylation. *J. Am. Chem. Soc.* **2008**, *130*, 3716–3717, DOI: 10.1021/ja710536p.

(102) Bhattacharjee, R.; Moriam, S.; Umer, M.; Nguyen, N.-T.; Shiddiky, M. J. A. DNA methylation detection: recent developments in bisulfite free electrochemical and optical approaches. *The Analyst* **2018**, *143*, 4802–4818, DOI: 10.1039/c8an01348a.

(103) Barhoumi, A.; Halas, N. J. Detecting Chemically Modified DNA Bases Using Surface Enhanced Raman Spectroscopy. *The journal of physical chemistry letters* **2011**, *2*, 3118–3123, DOI: 10.1021/jz201423b.

(104) Guerrini, L.; Krpetić, Ž.; van Lierop, D.; Alvarez-Puebla, R. A.; Graham, D. Direct surface-enhanced Raman scattering analysis of DNA duplexes. *Angewandte Chemie (International ed. in English)* **2015**, *54*, 1144–1148, DOI: 10.1002/anie.201408558.

(105) Morla-Folch, J.; Xie, H.-n.; Gisbert-Quilis, P.; Gómez-de Pedro, S.; Pazos-Perez, N.; Alvarez-Puebla, R. A.; Guerrini, L. Ultrasensitive Direct Quantification of Nucleobase Modifications in DNA by Surface-Enhanced Raman Scattering: The Case of Cytosine. *Angewandte Chemie (International ed. in English)* **2015**, *54*, 13650–13654, DOI: 10.1002/anie.201507682.

(106) Razin, A.; Sedat, J. Analysis of 5-methylcytosine in DNA. *Analytical Biochemistry* **1977**, *77*, 370–377, DOI: 10.1016/0003-2697(77)90250-0.

(107) Breter, H.-J.; Seibert, G.; Zahn, R. K. The use of high-pressure liquid cation-exchange chromatography for determination of the 5-methylcytosine content of DNA. *Journal of Chromatography A* **1976**, *118*, 242–249, DOI: 10.1016/S0021-9673(00)81215-7.

(108) Torres, A. L.; Barrientos, E. Y.; Wrobel, K.; Wrobel, K. Selective derivatization of cytosine and methylcytosine moieties with 2-bromoacetophenone for submicrogram DNA methylation analysis by reversed phase HPLC with spectrofluorimetric detection. *Analytical chemistry* **2011**, *83*, 7999–8005, DOI: 10.1021/ac2020799.

(109) Traube, F. R.; Schiffers, S.; Iwan, K.; Kellner, S.; Spada, F.; Müller, M.; Carell, T. Isotope-dilution mass spectrometry for exact quantification of noncanonical DNA nucleosides. *Nature protocols* **2019**, *14*, 283–312, DOI: 10.1038/s41596-018-0094-6.

(110) Tang, Y.; Chu, J.-M.; Huang, W.; Xiong, J.; Xing, X.-W.; Zhou, X.; Feng, Y.-Q.; Yuan, B.-F. Hydrophilic material for the selective enrichment of 5-hydroxymethylcytosine and its liquid chromatography-tandem mass spectrometry detection. *Analytical chemistry* **2013**, *85*, 6129–6135, DOI: 10.1021/ac4010869.

(111) Zhang, H.-Y.; Xiong, J.; Qi, B.-L.; Feng, Y.-Q.; Yuan, B.-F. The existence of 5-hydroxymethylcytosine and 5-formylcytosine in both DNA and RNA in mammals. *Chemical communications (Cambridge, England)* **2016**, *52*, 737–740, DOI: 10.1039/c5cc07354e.

(112) Tang, Y.; Xiong, J.; Jiang, H.-P.; Zheng, S.-J.; Feng, Y.-Q.; Yuan, B.-F. Determination of oxidation products of 5-methylcytosine in plants by chemical derivatization coupled with liquid chromatography/tandem mass spectrometry analysis. *Analytical chemistry* **2014**, *86*, 7764–7772, DOI: 10.1021/ac5016886.

(113) Jiang, H.-P.; Liu, T.; Guo, N.; Yu, L.; Yuan, B.-F.; Feng, Y.-Q. Determination of formylated DNA and RNA by chemical labeling combined with mass spectrometry analysis. *Analytica chimica acta* **2017**, *981*, 1–10, DOI: 10.1016/j.aca.2017.06.009.

(114) Tang, Y.; Zheng, S.-J.; Qi, C.-B.; Feng, Y.-Q.; Yuan, B.-F. Sensitive and simultaneous determination of 5-methylcytosine and its oxidation products in genomic DNA by chemical derivatization coupled with liquid chromatography-tandem mass spectrometry analysis. *Analytical chemistry* **2015**, *87*, 3445–3452, DOI: 10.1021/ac504786r.

(115) Guo, M.; Li, X.; Zhang, L.; Liu, D.; Du, W.; Yin, D.; Lyu, N.; Zhao, G.; Guo, C.; Tang, D. Accurate quantification of 5-Methylcytosine, 5-Hydroxymethylcytosine, 5-Formylcytosine, and 5-Carboxylcytosine in genomic DNA from breast cancer by chemical derivatization coupled with ultra performance liquid chromatography- electrospray quadrupole time of flight mass spectrometry analysis. *Oncotarget* **2017**, *8*, 91248–91257, DOI: 10.18632/oncotarget.20093.

(116) Chowdhury, B.; Cho, I.-H.; Irudayaraj, J. Technical advances in global DNA methylation analysis in human cancers. *Journal of biological engineering* **2017**, *11*, 10, DOI: 10.1186/s13036-017-0052-9.

(117) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences* **1996**, *93*, 13770–13773, DOI: 10.1073/pnas.93.24.13770.

(118) Stoddart, D.; Heron, A. J.; Mikhailova, E.; Maglia, G.; Bayley, H. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 7702–7707, DOI: 10.1073/pnas.0901054106.

(119) Derrington, I. M.; Butler, T. Z.; Collins, M. D.; Manrao, E.; Pavlenok, M.; Niederweis, M.; Gundlach, J. H. Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences of the United States of America* **2010**, *107*, 16060–16065, DOI: 10.1073/pnas.1001831107.

(120) Wang, Y.; Yang, Q.; Wang, Z. The evolution of nanopore sequencing. *Frontiers in genetics* **2014**, *5*, 449, DOI: 10.3389/fgene.2014.00449.

(121) Rhee, M.; Burns, M. A. Nanopore sequencing technology: research trends and applications. *Trends in biotechnology* **2006**, *24*, 580–586, DOI: 10.1016/j.tibtech.2006.10.005.

(122) Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)* **2009**, *323*, 133–138, DOI: 10.1126/science.1162986.

(123) Korlach, J.; Bjornson, K. P.; Chaudhuri, B. P.; Cicero, R. L.; Flusberg, B. A.; Gray, J. J.; Holden, D.; Saxena, R.; Wegener, J.; Turner, S. W. Real-Time DNA Sequencing from Single Polymerase Molecules. In *Single molecule methods*; Walter, N. G., Ed.; Methods in Enzymology; Academic: London, 2010; pp 431–455.

(124) Plongthongkum, N.; Diep, D. H.; Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature reviews. Genetics* **2014**, *15*, 647–661, DOI: 10.1038/nrg3772.

(125) Ardui, S.; Ameur, A.; Vermeesch, J. R.; Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucl Acids Res* **2018**, *46*, 2159–2168, DOI: 10.1093/nar/gky066.

(126) Korlach, J.; Turner, S. W. Going beyond five bases in DNA sequencing. *Current opinion in structural biology* **2012**, *22*, 251–261, DOI: 10.1016/j.sbi.2012.04.002.

(127) Ross, M. G.; Russ, C.; Costello, M.; Hollinger, A.; Lennon, N. J.; Hegarty, R.; Nusbaum, C.; Jaffe, D. B. Characterizing and measuring bias in sequence data. *Genome biology* **2013**, *14*, R51, DOI: 10.1186/gb-2013-14-5-r51.

(128) van Dijk, E. L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends in genetics : TIG* **2018**, *34*, 666–681, DOI: 10.1016/j.tig.2018.05.008.

(129) Yang, Y.; Sebra, R.; Pullman, B. S.; Qiao, W.; Peter, I.; Desnick, R. J.; Geyer, C. R.; DeCoteau, J. F.; Scott, S. A. Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). *BMC genomics* **2015**, *16*, 350, DOI: 10.1186/s12864-015-1572-7.

(130) Song, C.-X.; Clark, T. A.; Lu, X.-Y.; Kislyuk, A.; Dai, Q.; Turner, S. W.; He, C.; Korlach, J. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nature methods* **2011**, *9*, 75–77, DOI: 10.1038/nmeth.1779.

(131) Clark, T. A.; Lu, X.; Luong, K.; Dai, Q.; Boitano, M.; Turner, S. W.; He, C.; Korlach, J. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC biology* **2013**, *11*, 4, DOI: 10.1186/1741-7007-11-4.

(132) Li, W.-W.; Gong, L.; Bayley, H. Single-molecule detection of 5-hydroxymethylcytosine in DNA through chemical modification and nanopore analysis. *Angewandte Chemie (International ed. in English)* **2013**, *52*, 4350–4355, DOI: 10.1002/anie.201300413.

(133) Zeng, T.; Liu, L.; Li, T.; Li, Y.; Gao, J.; Zhao, Y.; Wu, H.-C. Detection of 5-methylcytosine and 5-hydroxymethylcytosine in DNA via host-guest interactions inside α-hemolysin nanopores. *Chemical science* **2015**, *6*, 5628–5634, DOI: 10.1039/c5sc01436k.

(134) Shapiro, R.; Braverman, B.; Louis, J. B.; Servis, R. E. Nucleic acid reactivity and conformation. II. Reaction of cytosine and uracil with sodium bisulfite. *The Journal of biological chemistry* **1973**, *248*, 4060–4064.

(135) Shapiro, R.; DeFate, V.; Welcher, M. Deamination cytosine derivatives by bisulfite. Mechanism of the reaction. *J. Am. Chem. Soc.* **1974**, *96*, 906–912, DOI: 10.1021/ja00810a043.

(136) Grunau, C.; Clark, S. J.; Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic acids research* **2001**, *29*, E65-5.

(137) Tanaka, K.; Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorganic & medicinal chemistry letters* **2007**, *17*, 1912–1915, DOI: 10.1016/j.bmcl.2007.01.040.

(138) Wang, R. Y.; Gehrke, C. W.; Ehrlich, M. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic acids research* **1980**, *8*, 4777–4790.

(139) El-Maarri, O.; Herbiniaux, U.; Walter, J.; Oldenburg, J. A rapid, quantitative, non-radioactive bisulfite-SNuPE- IP RP HPLC assay for methylation analysis at specific CpG sites. *Nucl Acids Res* **2002**, *30*, e25.

(140) Ehrich, M.; Nelson, M. R.; Stanssens, P.; Zabeau, M.; Liloglou, T.; Xinarianos, G.; Cantor, C. R.; Field, J. K.; van den Boom, D. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proceedings of the National Academy of Sciences* **2005**, *102*, 15785–15790, DOI: 10.1073/pnas.0507816102.

(141) Bartošík, M.; Fojta, M.; Paleček, E. Electrochemical detection of 5-methylcytosine in bisulfite-treated DNA. *Electrochimica Acta* **2012**, *78*, 75–81, DOI: 10.1016/j.electacta.2012.05.115.

(142) Wang, P.; Wu, H.; Dai, Z.; Zou, X. Picomolar level profiling of the methylation status of the p53 tumor suppressor gene by a label-free electrochemical biosensor. *Chemical communications (Cambridge, England)* **2012**, *48*, 10754–10756, DOI: 10.1039/c2cc35615e.

(143) Xiong, Z.; Laird, P. W. COBRA: a sensitive and quantitative DNA methylation assay. *Nucl Acids Res* **1997**, *25*, 2532–2534, DOI: 10.1093/nar/25.12.2532.

(144) Herman, J. G.; Graff, J. R.; Myohanen, S.; Nelkin, B. d.; Baylin, S. B. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proceedings of the National Academy of Sciences* **1996**, *93*, 9821–9826, DOI: 10.1073/pnas.93.18.9821.

(145) Bibikova, M.; Lin, Z.; Zhou, L.; Chudin, E.; Garcia, E. W.; Wu, B.; Doucet, D.; Thomas, N. J.; Wang, Y.; Vollmer, E. *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome research* **2006**, *16*, 383–393, DOI: 10.1101/gr.4410706.

(146) Eads, C. A. MethyLight: a high-throughput assay to measure DNA methylation. *Nucl Acids Res* **2000**, *28*, 32e-0, DOI: 10.1093/nar/28.8.e32.

(147) Feng, F.; Wang, H.; Han, L.; Wang, S. Fluorescent conjugated polyelectrolyte as an indicator for convenient detection of DNA methylation. *J. Am. Chem. Soc.* **2008**, *130*, 11338–11343, DOI: 10.1021/ja8011963.

(148) Hu, J.; Zhang, C.-Y. Single base extension reaction-based surface enhanced Raman spectroscopy for DNA methylation assay. *Biosensors & bioelectronics* **2012**, *31*, 451–457, DOI: 10.1016/j.bios.2011.11.014.

(149) Wang, Y.; Wee, E. J. H.; Trau, M. Highly sensitive DNA methylation analysis at CpG resolution by surface-enhanced Raman scattering via ligase chain reaction. *Chemical communications (Cambridge, England)* **2015**, *51*, 10953–10956, DOI: 10.1039/c5cc03921e.

(150) Booth, M. J.; Raiber, E.-A.; Balasubramanian, S. Chemical methods for decoding cytosine modifications in DNA. *Chemical reviews* **2015**, *115*, 2240–2254, DOI: 10.1021/cr5002904.

(151) Berney, M.; McGouran, J. F. Methods for detection of cytosine and thymine modifications in DNA. *Nat Rev Chem* **2018**, *2*, 332–348, DOI: 10.1038/s41570-018-0044-4.

(152) Huang, Y.; Pastor, W. A.; Shen, Y.; Tahiliani, M.; Liu, D. R.; Rao, A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one* **2010**, *5*, e8888, DOI: 10.1371/journal.pone.0008888.

(153) Booth, M. J.; Branco, M. R.; Ficz, G.; Oxley, D.; Krueger, F.; Reik, W.; Balasubramanian, S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **2012**, *336*, 934–937, DOI: 10.1126/science.1220671.

(154) Kubik, G.; Summerer, D. Deciphering Epigenetic Cytosine Modifications by Direct Molecular Recognition. *ACS chemical biology* **2015**, *10*, 1580–1589, DOI: 10.1021/acschembio.5b00158.

(155) Josse, J.; Kornberg, A. Glucosylation of deoxyribonucleic acid. III. alpha- and beta-Glucosyl transferases from T4-infected Escherichia coli. *The Journal of biological chemistry* **1962**, *237*, 1968–1976.

(156) Tomaschewski, J.; Gram, H.; Crabb, J. W.; Rüger, W. T4-induced α- and β-glucosyltransferase: cloning of the genes and a comparison of their products based on sequencing data. *Nucl Acids Res* **1985**, *13*, 7551–7566, DOI: 10.1093/nar/13.21.7551.

(157) Yu, M.; Hon, G. C.; Szulwach, K. E.; Song, C.-X.; Zhang, L.; Kim, A.; Li, X.; Dai, Q.; Shen, Y.; Park, B. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **2012**, *149*, 1368–1380, DOI: 10.1016/j.cell.2012.04.027.

(158) Lu, X.; Song, C.-X.; Szulwach, K.; Wang, Z.; Weidenbacher, P.; Jin, P.; He, C. Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *Journal of the American Chemical Society* **2013**, *135*, 9315–9317, DOI: 10.1021/ja4044856.

(159) Wu, H.; Wu, X.; Shen, L.; Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nature biotechnology* **2014**, *32*, 1231–1240, DOI: 10.1038/nbt.3073.

(160) Meissner, A.; Gnirke, A.; Bell, G. W.; Ramsahoye, B.; Lander, E. S.; Jaenisch, R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research* **2005**, *33*, 5868–5877, DOI: 10.1093/nar/gki901.

(161) Guo, H.; Zhu, P.; Guo, F.; Li, X.; Wu, X.; Fan, X.; Wen, L.; Tang, F. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nature protocols* **2015**, *10*, 645–659, DOI: 10.1038/nprot.2015.039.

(162) Tierling, S.; Schmitt, B.; Walter, J. Comprehensive Evaluation of Commercial Bisulfite-Based DNA Methylation Kits and Development of an Alternative Protocol With Improved Conversion Performance. *Genetics & epigenetics* **2018**, *10*, 1179237X18766097, DOI: 10.1177/1179237X18766097.

(163) Münzel, M.; Lercher, L.; Müller, M.; Carell, T. Chemical discrimination between dC and 5MedC via their hydroxylamine adducts. *Nucleic acids research* **2010**, *38*, e192, DOI: 10.1093/nar/gkq724.

(164) Okamoto, A.; Tainaka, K.; Kamei, T. Sequence-selective osmium oxidation of DNA: efficient distinction between 5-methylcytosine and cytosine. *Organic & biomolecular chemistry* **2006**, *4*, 1638–1640, DOI: 10.1039/b600401f.

(165) Tanaka, K.; Tainaka, K.; Kamei, T.; Okamoto, A. Direct labeling of 5-methylcytosine and its applications. *J. Am. Chem. Soc.* **2007**, *129*, 5612–5620, DOI: 10.1021/ja068660c.

(166) Bareyt, S.; Carell, T. Selective detection of 5-methylcytosine sites in DNA. *Angewandte Chemie (International ed. in English)* **2008**, *47*, 181–184, DOI: 10.1002/anie.200702159.

(167) Xu, Y.; Niu, C.; Xiao, X.; Zhu, W.; Dai, Z.; Zou, X. Chemical-oxidation cleavage triggered isothermal exponential amplification reaction for attomole gene-specific methylation analysis. *Analytical chemistry* **2015**, *87*, 2945–2951, DOI: 10.1021/ac5044785.

(168) Tainaka, K.; Tanaka, K.; Okamoto, A. Development of bipyridine-modified nucleobase for methylcytosine-selective crosslink reaction. *Nucleic acids symposium series (2004)* **2006**, 129–130, DOI: 10.1093/nass/nrl064.

(169) Tanaka, K.; Tainaka, K.; Umemoto, T.; Nomura, A.; Okamoto, A. An osmium-DNA interstrand complex: application to facile DNA methylation analysis. *J. Am. Chem. Soc.* **2007**, *129*, 14511–14517, DOI: 10.1021/ja076140r.

(170) Tanabe, K.; Yamada, H.; Nishimoto, S.-i. One-Electron Photooxidation and Site-Selective Strand Cleavage at 5-Methylcytosine in DNA by Sensitization with 2-Methyl-1,4-naphthoquinone-Tethered Oligonucleotides. *J. Am. Chem. Soc.* **2007**, *129*, 8034–8040, DOI: 10.1021/ja071369s.

(171) Ogino, M.; Taya, Y.; Fujimoto, K. Highly selective detection of 5-methylcytosine using photochemical ligation. *Chemical communications (Cambridge, England)* **2008**, 5996–5998, DOI: 10.1039/b813677g.

(172) Ogino, M.; Taya, Y.; Fujimoto, K. Detection of methylcytosine by DNA photoligation via hydrophobic interaction of the alkyl group. *Organic & biomolecular chemistry* **2009**, *7*, 3163, DOI: 10.1039/b904941j.

(173) Sakamoto, T.; Ami, T.; Fujimoto, K. 5-Methylcytosine Selective Photoligation Using Photoresponsive Oligonucleotides Containing Various 5-Vinyl-2′-deoxyuridines Having an Aromatic Group. *Chem. Lett.* **2012**, *41*, 47–49, DOI: 10.1246/cl.2012.47.

(174) Fujimo, K.; Konishi-Hiratsuka, K.; Sakamoto, T. Quick, Selective and Reversible Photocrosslinking Reaction between 5-Methylcytosine and 3-Cyanovinylcarbazole in DNA Double Strand. *International journal of molecular sciences* **2013**, *14*, 5765–5774, DOI: 10.3390/ijms14035765.

(175) Yamayoshi, A.; Matsuyama, Y.; Kushida, M.; Kobori, A.; Murakami, A. Novel photodynamic effect of a psoralen-conjugated oligonucleotide for the discrimination of the methylation of cytosine in DNA. *Photochemistry and photobiology* **2014**, *90*, 716–722, DOI: 10.1111/php.12232.

(176) Dalhoff, C.; Lukinavicius, G.; Klimasăuskas, S.; Weinhold, E. Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. *Nature chemical biology* **2006**, *2*, 31–32, DOI: 10.1038/nchembio754.

(177) Kriukienė, E.; Labrie, V.; Khare, T.; Urbanavičiūtė, G.; Lapinaitė, A.; Koncevičius, K.; Li, D.; Wang, T.; Pai, S.; Ptak, C. *et al.* DNA unmethylome profiling by covalent capture of CpG sites. *Nature communications* **2013**, *4*, 2190, DOI: 10.1038/ncomms3190.

(178) Staševskij, Z.; Gibas, P.; Gordevičius, J.; Kriukienė, E.; Klimašauskas, S. Tethered Oligonucleotide-Primed Sequencing, TOP-Seq: A High-Resolution Economical Approach for DNA Epigenome Profiling. *Molecular cell* **2017**, *65*, 554-564.e6, DOI: 10.1016/j.molcel.2016.12.012.

(179) Liu, Y.; Siejka-Zielińska, P.; Velikova, G.; Bi, Y.; Yuan, F.; Tomkova, M.; Bai, C.; Chen, L.; Schuster-Böckler, B.; Song, C.-X. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nature biotechnology* **2019**, *37*, 424–429, DOI: 10.1038/s41587-019-0041-2.

(180) Matsushita, T.; Moriyama, Y.; Nagae, G.; Aburatani, H.; Okamoto, A. DNA-friendly Cu(ii)/TEMPO-catalyzed 5-hydroxymethylcytosine-specific oxidation. *Chemical communications (Cambridge, England)* **2017**, *53*, 5756–5759, DOI: 10.1039/c7cc02814h.

(181) Hu, J.; Chen, Y.; Xu, X.; Wu, F.; Xing, X.; Xu, Z.; Xu, J.; Weng, X.; Zhou, X. Discrimination between 5-hydroxymethylcytosine and 5-methylcytosine in DNA by selective chemical labeling. *Bioorganic & medicinal chemistry letters* **2014**, *24*, 294–297, DOI: 10.1016/j.bmcl.2013.11.017.

(182) Ma, S.; Sun, H.; Li, Y.; Qi, H.; Zheng, J. Discrimination between 5-Hydroxymethylcytosine and 5-Methylcytosine in DNA via Selective Electrogenerated Chemiluminescence (ECL) Labeling. *Analytical chemistry* **2016**, *88*, 9934–9940, DOI: 10.1021/acs.analchem.6b01265.

(183) Wang, H.; Liu, M.; Bai, W.; Sun, H.; Li, Y.; Deng, H. A convenient electrogenerated chemiluminescence biosensing method for selective detection of 5-hydroxymethylcytosine in genomic DNA. *Sensors and Actuators B: Chemical* **2019**, *284*, 236–242, DOI: 10.1016/j.snb.2018.12.132.

(184) Hong, T.; Wang, T.; Guo, P.; Xing, X.; Ding, F.; Chen, Y.; Wu, J.; Ma, J.; Wu, F.; Zhou, X. Fluorescent strategy based on cationic conjugated polymer fluorescence resonance energy transfer for the quantification of 5-(hydroxymethyl)cytosine in genomic DNA. *Analytical chemistry* **2013**, *85*, 10797–10802, DOI: 10.1021/ac4020676.

(185) Wang, Y.; Zhang, X.; Wu, F.; Chen, Z.; Zhou, X. Bisulfite-free, single base-resolution analysis of 5-hydroxymethylcytosine in genomic DNA by chemical-mediated mismatch. *Chemical science* **2019**, *10*, 447–452, DOI: 10.1039/C8SC04272A.

(186) Zeng, H.; He, B.; Xia, B.; Bai, D.; Lu, X.; Cai, J.; Chen, L.; Zhou, A.; Zhu, C.; Meng, H. *et al.* Bisulfite-Free, Nanoscale Analysis of 5-Hydroxymethylcytosine at Single Base Resolution. *J. Am. Chem. Soc.* **2018**, *140*, 13190–13194, DOI: 10.1021/jacs.8b08297.

(187) Szwagierczak, A.; Bultmann, S.; Schmidt, C. S.; Spada, F.; Leonhardt, H. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic acids research* **2010**, *38*, e181, DOI: 10.1093/nar/gkq684.

(188) Schutsky, E. K.; DeNizio, J. E.; Hu, P.; Liu, M. Y.; Nabel, C. S.; Fabyanic, E. B.; Hwang, Y.; Bushman, F. D.; Wu, H.; Kohli, R. M. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nature biotechnology* **2018,** DOI: 10.1038/nbt.4204.

(189) Wu, X.; Li, Z.; Chen, X.-X.; Fossey, J. S.; James, T. D.; Jiang, Y.-B. Selective sensing of saccharides using simple boronic acids and their aggregates. *Chemical Society reviews* **2013**, *42*, 8032–8048, DOI: 10.1039/c3cs60148j.

(190) Zhao, C.; Wang, H.; Zhao, B.; Li, C.; Yin, R.; Song, M.; Liu, B.; Liu, Z.; Jiang, G. Boronic acid-mediated polymerase chain reaction for gene- and fragment-specific detection of 5-hydroxymethylcytosine. *Nucleic acids research* **2014**, *42*, e81, DOI: 10.1093/nar/gku216.

(191) Zhou, Y.; Yang, Z.; Li, X.; Wang, Y.; Yin, H.; Ai, S. Electrochemical biosensor for detection of DNA hydroxymethylation based on glycosylation and alkaline phosphatase catalytic signal amplification. *Electrochimica Acta* **2015**, *174*, 647–652, DOI: 10.1016/j.electacta.2015.06.043.

(192) Wei, Y.; Sun, H.; Li, J.; Zhang, Y.; Li, Y.; Lin, J.; Wang, T.; Zhou, M. Electrogenerated chemiluminescence biosensing method for highly sensitive detection of DNA hydroxymethylation: Combining glycosylation with Ru(phen) 3 2+ -assembled graphene oxide. *Journal of Electroanalytical Chemistry* **2017**, *795*, 123–129, DOI: 10.1016/j.jelechem.2017.04.056.

(193) Zhang, Y.; Li, Y.; Wei, Y.; Sun, H.; Wang, H. A sensitive signal-off electrogenerated chemiluminescence biosensing method for the discrimination of DNA hydroxymethylation based on glycosylation modification and signal quenching from ferroceneboronic acid. *Talanta* **2017**, *170*, 546–551, DOI: 10.1016/j.talanta.2017.04.051.

(194) Chen, H.-Y.; Wei, J.-R.; Pan, J.-X.; Zhang, W.; Dang, F.-Q.; Zhang, Z.-Q.; Zhang, J. Spectroscopic quantification of 5-hydroxymethylcytosine in genomic DNA using boric acid-functionalized nano-microsphere fluorescent probes. *Biosensors & bioelectronics* **2017**, *91*, 328–333, DOI: 10.1016/j.bios.2016.12.039.

(195) Sui, C.; Wang, T.; Zhou, Y.; Yin, H.; Meng, X.; Zhang, S.; Waterhouse, G. I. N.; Xu, Q.; Zhuge, Y.; Ai, S. Photoelectrochemical biosensor for hydroxymethylated DNA detection and T4-β-glucosyltransferase activity assay based on WS2 nanosheets and carbon dots. *Biosensors & bioelectronics* **2019**, *127*, 38–44, DOI: 10.1016/j.bios.2018.11.054.

(196) Pastor, W. A.; Pape, U. J.; Huang, Y.; Henderson, H. R.; Lister, R.; Ko, M.; McLoughlin, E. M.; Brudno, Y.; Mahapatra, S.; Kapranov, P. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **2011**, *473*, 394–397, DOI: 10.1038/nature10102.

(197) Shahal, T.; Green, O.; Hananel, U.; Michaeli, Y.; Shabat, D.; Ebenstein, Y. Simple and cost-effective fluorescent labeling of 5-hydroxymethylcytosine. *Methods and applications in fluorescence* **2016**, *4*, 44003, DOI: 10.1088/2050-6120/4/4/044003.

(198) Chen, S.; Dou, Y.; Zhao, Z.; Li, F.; Su, J.; Fan, C.; Song, S. High-Sensitivity and High-Efficiency Detection of DNA Hydroxymethylation in Genomic DNA by Multiplexing Electrochemical Biosensing. *Analytical chemistry* **2016**, *88*, 3476–3480, DOI: 10.1021/acs.analchem.6b00230.

(199) Song, C.-X.; Szulwach, K. E.; Fu, Y.; Dai, Q.; Yi, C.; Li, X.; Li, Y.; Chen, C.-H.; Zhang, W.; Jian, X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature biotechnology* **2011**, *29*, 68–72, DOI: 10.1038/nbt.1732.

(200) Song, C.-X.; He, C. Bioorthogonal labeling of 5-hydroxymethylcytosine in genomic DNA and diazirine-based DNA photo-cross-linking probes. *Accounts of chemical research* **2011**, *44*, 709–717, DOI: 10.1021/ar2000502.

(201) Szulwach, K. E.; Li, X.; Li, Y.; Song, C.-X.; Han, J. W.; Kim, S.; Namburi, S.; Hermetz, K.; Kim, J. J.; Rudd, M. K. *et al.* Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS genetics* **2011**, *7*, e1002154, DOI: 10.1371/journal.pgen.1002154.

(202) Sérandour, A. A.; Avner, S.; Mahé, E. A.; Madigou, T.; Guibert, S.; Weber, M.; Salbert, G. Single-CpG resolution mapping of 5-hydroxymethylcytosine by chemical labeling and exonuclease digestion identifies evolutionarily unconserved CpGs as TET targets. *Genome biology* **2016**, *17*, 56, DOI: 10.1186/s13059-016-0919-y.

(203) Han, D.; Lu, X.; Shih, A. H.; Nie, J.; You, Q.; Xu, M. M.; Melnick, A. M.; Levine, R. L.; He, C. A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Molecular cell* **2016**, *63*, 711–719, DOI: 10.1016/j.molcel.2016.06.028.

(204) Michaeli, Y.; Shahal, T.; Torchinsky, D.; Grunwald, A.; Hoch, R.; Ebenstein, Y. Optical detection of epigenetic marks: sensitive quantification and direct imaging of individual hydroxymethylcytosine bases. *Chemical communications (Cambridge, England)* **2013**, *49*, 8599–8601, DOI: 10.1039/c3cc42543f.

(205) Shahal, T.; Gilat, N.; Michaeli, Y.; Redy-Keisar, O.; Shabat, D.; Ebenstein, Y. Spectroscopic quantification of 5-hydroxymethylcytosine in genomic DNA. *Analytical chemistry* **2014**, *86*, 8231–8237, DOI: 10.1021/ac501609d.

(206) Song, C.-X.; Diao, J.; Brunger, A. T.; Quake, S. R. Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, 4338–4343, DOI: 10.1073/pnas.1600223113.

(207) Liutkevičiūtė, Z.; Kriukienė, E.; Grigaitytė, I.; Masevičius, V.; Klimašauskas, S. Methyltransferase-Directed Derivatization of 5-Hydroxymethylcytosine in DNA. *Angew. Chem.* **2011**, *123*, 2138–2141, DOI: 10.1002/ange.201007169.

(208) Yang, Z.; Shi, Y.; Liao, W.; Yin, H.; Ai, S. A novel signal-on photoelectrochemical biosensor for detection of 5-hydroxymethylcytosine based on in situ electron donor producing strategy and all wavelengths of light irradiation. *Sensors and Actuators B: Chemical* **2016**, *223*, 621–625, DOI: 10.1016/j.snb.2015.09.159.

(209) Jiang, W.; Lu, Y.; Wang, H.; Wang, M.; Yin, H. Amperometric biosensor for 5-hydroxymethylcytosine based on enzymatic catalysis and using spherical poly(acrylic acid) brushes. *Microchim Acta* **2017**, *184*, 3789–3796, DOI: 10.1007/s00604-017-2401-2.

(210) Wu, Q.; Amrutkar, S. M.; Shao, F. Sulfinate Based Selective Labeling of 5-Hydroxymethylcytosine: Application to Biotin Pull Down Assay. *Bioconjugate chemistry* **2018**, *29*, 245–249, DOI: 10.1021/acs.bioconjchem.7b00826.

(211) Cui, L.; Hu, J.; Wang, M.; Li, C.-C.; Zhang, C.-Y. Label-Free and Immobilization-Free Electrochemical Magnetobiosensor for Sensitive Detection of 5-Hydroxymethylcytosine in Genomic DNA. *Analytical chemistry* **2019**, *91*, 1232–1236, DOI: 10.1021/acs.analchem.8b04663.

(212) Okamoto, A.; Sugizaki, K.; Nakamura, A.; Yanagisawa, H.; Ikeda, S. 5-Hydroxymethylcytosine-selective oxidation with peroxotungstate. *Chemical communications (Cambridge, England)* **2011**, *47*, 11231–11233, DOI: 10.1039/c1cc14782j.

(213) Hayashi, G.; Koyama, K.; Shiota, H.; Kamio, A.; Umeda, T.; Nagae, G.; Aburatani, H.; Okamoto, A. Base-Resolution Analysis of 5-Hydroxymethylcytosine by One-Pot Bisulfite-Free Chemical Conversion with Peroxotungstate. *J. Am. Chem. Soc.* **2016**, *138*, 14178–14181, DOI: 10.1021/jacs.6b06428.

(214) Wang, Y.; Liu, C.; Yang, W.; Zou, G.; Zhang, X.; Wu, F.; Yu, S.; Luo, X.; Zhou, X. Naphthalimide derivatives as multifunctional molecules for detecting 5-formylpyrimidine by both PAGE analysis and dot-blot assays. *Chemical communications (Cambridge, England)* **2018**, *54*, 1497–1500, DOI: 10.1039/c7cc08715b.

(215) Guo, P.; Yan, S.; Hu, J.; Xing, X.; Wang, C.; Xu, X.; Qiu, X.; Ma, W.; Lu, C.; Weng, X. *et al.* Selective detection of 5-formyl-2'-deoxycytidine in DNA using a fluorogenic hydroxylamine reagent. *Organic letters* **2013**, *15*, 3266–3269, DOI: 10.1021/ol401290d.

(216) Xu, L.; Chen, Y.-C.; Chong, J.; Fin, A.; McCoy, L. S.; Xu, J.; Zhang, C.; Wang, D. Pyrene-based quantitative detection of the 5-formylcytosine loci symmetry in the CpG duplex content during TET-dependent demethylation. *Angewandte Chemie (International ed. in English)* **2014**, *53*, 11223–11227, DOI: 10.1002/anie.201406220.

(217) Xu, L.; Chen, Y.-C.; Nakajima, S.; Chong, J.; Wang, L.; Lan, L.; Zhang, C.; Wang, D. A Chemical Probe Targets DNA 5-Formylcytosine Sites and Inhibits TDG Excision, Polymerases Bypass, and Gene Expression. *Chemical science* **2014**, *5*, 567–574, DOI: 10.1039/C3SC51849C.

(218) Wang, S.-R.; Wang, J.-Q.; Fu, B.-S.; Chen, K.; Xiong, W.; Wei, L.; Qing, G.; Tian, T.; Zhou, X. Supramolecular Coordination-Directed Reversible Regulation of Protein Activities at Epigenetic DNA Marks. *J. Am. Chem. Soc.* **2018**, *140*, 15842–15849, DOI: 10.1021/jacs.8b09113.

(219) Wang, S.-R.; Song, Y.-Y.; Wei, L.; Liu, C.-X.; Fu, B.-S.; Wang, J.-Q.; Yang, X.-R.; Liu, Y.-N.; Liu, S.-M.; Tian, T. *et al.* Cucurbit7uril-Driven Host-Guest Chemistry for Reversible Intervention of 5-Formylcytosine-Targeted Biochemical Reactions. *J. Am. Chem. Soc.* **2017**, *139*, 16903–16912, DOI: 10.1021/jacs.7b09635.

(220) Dietzsch, J.; Feineis, D.; Höbartner, C. Chemoselective labeling and site-specific mapping of 5-formylcytosine as a cellular nucleic acid modification. *FEBS letters* **2018**, *592*, 2032–2047, DOI: 10.1002/1873-3468.13058.

(221) Hardisty, R. E.; Kawasaki, F.; Sahakyan, A. B.; Balasubramanian, S. Selective Chemical Labeling of Natural T Modifications in DNA. *J. Am. Chem. Soc.* **2015**, *137*, 9270–9272, DOI: 10.1021/jacs.5b03730.

(222) Wang, Y.; Zhang, X.; Zou, G.; Peng, S.; Liu, C.; Zhou, X. Detection and Application of 5-Formylcytosine and 5-Formyluracil in DNA. *Accounts of chemical research* **2019**, *52*, 1016–1024, DOI: 10.1021/acs.accounts.8b00543.

(223) Samanta, B.; Seikowski, J.; Höbartner, C. Fluorogenic Labeling of 5-Formylpyrimidine Nucleotides in DNA and RNA. *Angewandte Chemie (International ed. in English)* **2016**, *55*, 1912–1916, DOI: 10.1002/anie.201508893.

(224) Hu, J.; Xing, X.; Xu, X.; Wu, F.; Guo, P.; Yan, S.; Xu, Z.; Xu, J.; Weng, X.; Zhou, X. Selective chemical labelling of 5-formylcytosine in DNA by fluorescent dyes. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2013**, *19*, 5836–5840, DOI: 10.1002/chem.201300082.

(225) Layer, R. W. The Chemistry of Imines. *Chem. Rev.* **1963**, *63*, 489–510, DOI: 10.1021/cr60225a003.

(226) Kölmel, D. K.; Kool, E. T. Oximes and Hydrazones in Bioconjugation: Mechanism and Catalysis. *Chemical reviews* **2017**, *117*, 10358–10376, DOI: 10.1021/acs.chemrev.7b00090.

(227) Xia, B.; Han, D.; Lu, X.; Sun, Z.; Zhou, A.; Yin, Q.; Zeng, H.; Liu, M.; Jiang, X.; Xie, W. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nature methods* **2015**, *12*, 1047–1050, DOI: 10.1038/nmeth.3569.

(228) Liu, C.; Wang, Y.; Yang, W.; Wu, F.; Zeng, W.; Chen, Z.; Huang, J.; Zou, G.; Zhang, X.; Wang, S. *et al.* Fluorogenic labeling and single-base resolution analysis of 5-formylcytosine in DNA. *Chemical science* **2017**, *8*, 7443–7447, DOI: 10.1039/c7sc03685j.

(229) Wang, Y.; Liu, C.; Zhang, X.; Yang, W.; Wu, F.; Zou, G.; Weng, X.; Zhou, X. Gene specific-loci quantitative and single-base resolution analysis of 5-formylcytosine by compound-mediated polymerase chain reaction. *Chemical science* **2018**, *9*, 3723–3728, DOI: 10.1039/c8sc00493e.

(230) *Bioconjugate Techniques;* Hermanson, G. T., Ed., [3rd edition]; Elsevier: Amsterdam, Boston, Heidelberg, op. 2013.

(231) Haugland, R. P. *The molecular probes handbook: A guide to fluorescent probes and labeling technologies*, 10. ed.; Invitogen Corp: Carlsbad, California, 2005.

(232) Song, C.-X.; Yi, C.; He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature biotechnology* **2012**, *30*, 1107–1116, DOI: 10.1038/nbt.2398.

(233) Haffner, M. C.; Chaux, A.; Meeker, A. K.; Esopi, D. M.; Gerber, J.; Pellakuru, L. G.; Toubaji, A.; Argani, P.; Iacobuzio-Donahue, C.; Nelson, W. G. *et al.* Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget* **2011**, *2*, 627–637, DOI: 10.18632/oncotarget.316.

(234) Inoue, A.; Shen, L.; Dai, Q.; He, C.; Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell research* **2011**, *21*, 1670–1676, DOI: 10.1038/cr.2011.189.

(235) Mohn, F.; Weber, M.; Schübeler, D.; Roloff, T.-C. Methylated DNA immunoprecipitation (MeDIP). *Methods in molecular biology (Clifton, N.J.)* **2009**, *507*, 55–64, DOI: 10.1007/978-1-59745-522-0_5.

(236) Nestor, C. E.; Meehan, R. R. Hydroxymethylated DNA immunoprecipitation (hmeDIP). *Methods in molecular biology (Clifton, N.J.)* **2014**, *1094*, 259–267, DOI: 10.1007/978-1-62703-706-8_20.

(237) Ficz, G.; Branco, M. R.; Seisenberger, S.; Santos, F.; Krueger, F.; Hore, T. A.; Marques, C. J.; Andrews, S.; Reik, W. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **2011**, *473*, 398–402, DOI: 10.1038/nature10008.

(238) Nair, S. S.; Coolen, M. W.; Stirzaker, C.; Song, J. Z.; Statham, A. L.; Strbenac, D.; Robinson, M. D.; Clark, S. J. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* **2011**, *6*, 34–44, DOI: 10.4161/epi.6.1.13313.

(239) Ko, M.; Huang, Y.; Jankowska, A. M.; Pape, U. J.; Tahiliani, M.; Bandukwala, H. S.; An, J.; Lamperti, E. D.; Koh, K. P.; Ganetzky, R. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **2010**, *468*, 839–843, DOI: 10.1038/nature09586.

(240) Lentini, A.; Lagerwall, C.; Vikingsson, S.; Mjoseng, H. K.; Douvlataniotis, K.; Vogt, H.; Green, H.; Meehan, R. R.; Benson, M.; Nestor, C. E. A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nature methods* **2018**, *15*, 499–504, DOI: 10.1038/s41592-018-0038-7.

(241) Aschenbrenner, J.; Drum, M.; Topal, H.; Wieland, M.; Marx, A. Direct sensing of 5-methylcytosine by polymerase chain reaction. *Angewandte Chemie (International ed. in English)* **2014**, *53*, 8154–8158, DOI: 10.1002/anie.201403745.

(242) Roberts, R. J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic acids research* **2010**, *38*, D234-6, DOI: 10.1093/nar/gkp874.

(243) Kinney, S. M.; Chin, H. G.; Vaisvila, R.; Bitinaite, J.; Zheng, Y.; Estève, P.-O.; Feng, S.; Stroud, H.; Jacobsen, S. E.; Pradhan, S. Tissue-specific distribution and dynamic changes of 5-hydroxymethylcytosine in mammalian genomes. *The Journal of biological chemistry* **2011**, *286*, 24685–24693, DOI: 10.1074/jbc.M110.217083.

(244) Wang, H.; Guan, S.; Quimby, A.; Cohen-Karni, D.; Pradhan, S.; Wilson, G.; Roberts, R. J.; Zhu, Z.; Zheng, Y. Comparative characterization of the PvuRts1I family of restriction enzymes and their application in mapping genomic 5-hydroxymethylcytosine. *Nucleic acids research* **2011**, *39*, 9294–9305, DOI: 10.1093/nar/gkr607.

(245) Sun, Z.; Terragni, J.; Jolyon, T.; Borgaro, J. G.; Liu, Y.; Yu, L.; Guan, S.; Wang, H.; Sun, D.; Cheng, X. *et al.* High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell reports* **2013**, *3*, 567–576, DOI: 10.1016/j.celrep.2013.01.001.

(246) Sun, Z.; Dai, N.; Borgaro, J. G.; Quimby, A.; Sun, D.; Corrêa, I. R.; Zheng, Y.; Zhu, Z.; Guan, S. A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Molecular cell* **2015**, *57*, 750–761, DOI: 10.1016/j.molcel.2014.12.035.

(247) Cross, S. H.; Charlton, J. A.; Nan, X.; Bird, A. P. Purification of CpG islands using a methylated DNA binding column. *Nature genetics* **1994**, *6*, 236–244, DOI: 10.1038/ng0394-236.

(248) Rauch, T.; Pfeifer, G. P. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Laboratory investigation; a journal of technical methods and pathology* **2005**, *85*, 1172–1180, DOI: 10.1038/labinvest.3700311.

(249) Yamagata, K. DNA methylation profiling using live-cell imaging. *Methods (San Diego, Calif.)* **2010**, *52*, 259–266, DOI: 10.1016/j.ymeth.2010.04.008.

(250) Hori, Y.; Otomura, N.; Nishida, A.; Nishiura, M.; Umeno, M.; Suetake, I.; Kikuchi, K. Synthetic-Molecule/Protein Hybrid Probe with Fluorogenic Switch for Live-Cell Imaging of DNA Methylation. *J. Am. Chem. Soc.* **2018**, *140*, 1686–1690, DOI: 10.1021/jacs.7b09713.

(251) Cipriany, B. R.; Zhao, R.; Murphy, P. J.; Levy, S. L.; Tan, C. P.; Craighead, H. G.; Soloway, P. D. Single molecule epigenetic analysis in a nanofluidic channel. *Analytical chemistry* **2010**, *82*, 2480–2487, DOI: 10.1021/ac9028642.

(252) Cipriany, B. R.; Murphy, P. J.; Hagarman, J. A.; Cerf, A.; Latulippe, D.; Levy, S. L.; Benítez, J. J.; Tan, C. P.; Topolancik, J.; Soloway, P. D. *et al.* Real-time analysis and selection of methylated DNA by fluorescence-activated single molecule sorting in a nanofluidic channel. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109*, 8477–8482, DOI: 10.1073/pnas.1117549109.

(253) Robertson, A. B.; Dahl, J. A.; Vågbø, C. B.; Tripathi, P.; Krokan, H. E.; Klungland, A. A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic acids research* **2011**, *39*, e55, DOI: 10.1093/nar/gkr051.

(254) Serre, D.; Lee, B. H.; Ting, A. H. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic acids research* **2010**, *38*, 391–399, DOI: 10.1093/nar/gkp992.

(255) Li, N.; Ye, M.; Li, Y.; Yan, Z.; Butcher, L. M.; Sun, J.; Han, X.; Chen, Q.; Zhang, X.; Wang, J. Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods (San Diego, Calif.)* **2010**, *52*, 203–212, DOI: 10.1016/j.ymeth.2010.04.009.

(256) Brinkman, A. B.; Simmer, F.; Ma, K.; Kaan, A.; Zhu, J.; Stunnenberg, H. G. Whole-genome DNA methylation profiling using MethylCap-seq. *Methods (San Diego, Calif.)* **2010**, *52*, 232–236, DOI: 10.1016/j.ymeth.2010.06.012.

(257) Robertson, A. B.; Dahl, J. A.; Ougland, R.; Klungland, A. Pull-down of 5-hydroxymethylcytosine DNA using JBP1-coated magnetic beads. *Nature protocols* **2012**, *7*, 340–350, DOI: 10.1038/nprot.2011.443.

(258) Cui, L.; Chung, T. H.; Tan, D.; Sun, X.; Jia, X.-Y. JBP1-seq: a fast and efficient method for genome-wide profiling of 5hmC. *Genomics* **2014**, *104*, 368–375, DOI: 10.1016/j.ygeno.2014.08.023.

(259) Hashimoto, H.; Zhang, X.; Cheng, X. Selective excision of 5-carboxylcytosine by a thymine DNA glycosylase mutant. *Journal of molecular biology* **2013**, *425*, 971–976, DOI: 10.1016/j.jmb.2013.01.013.

(260) Wolfe, S. A.; Nekludova, L.; Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure* **2000**, *29*, 183–212, DOI: 10.1146/annurev.biophys.29.1.183.

(261) Bhakta, M. S.; Segal, D. J. The generation of zinc finger proteins by modular assembly. *Methods in molecular biology (Clifton, N.J.)* **2010**, *649*, 3–30, DOI: 10.1007/978-1-60761-753-2_1.

(262) Hudson, N. O.; Buck-Koehntop, B. A. Zinc Finger Readers of Methylated DNA. *Molecules (Basel, Switzerland)* **2018**, *23,* DOI: 10.3390/molecules23102555.

(263) Stains, C. I.; Furman, J. L.; Segal, D. J.; Ghosh, I. Site-specific detection of DNA methylation utilizing mCpG-SEER. *J. Am. Chem. Soc.* **2006**, *128*, 9761–9765, DOI: 10.1021/ja060681j.

(264) Lungu, C.; Pinter, S.; Broche, J.; Rathert, P.; Jeltsch, A. Modular fluorescence complementation sensors for live cell detection of epigenetic signals at endogenous genomic sites. *Nature communications* **2017**, *8,* DOI: 10.1038/s41467-017-00457-z.

(265) Nomura, A.; Okamoto, A. Phosphopeptides designed for 5-methylcytosine recognition. *Biochemistry* **2011**, *50*, 3376–3385, DOI: 10.1021/bi102053d.

(266) Nomura, A.; Sugizaki, K.; Yanagisawa, H.; Okamoto, A. Discrimination between 5-hydroxymethylcytosine and 5-methylcytosine by a chemically designed peptide. *Chemical communications (Cambridge, England)* **2011**, *47*, 8277–8279, DOI: 10.1039/c1cc12131f.

(267) Kubik, G.; Summerer, D. TALEored Epigenetics: A DNA-Binding Scaffold for Programmable Epigenome Editing and Analysis. *Chembiochem : a European journal of chemical biology* **2016**, *17*, 975–980, DOI: 10.1002/cbic.201600072.

(268) Waryah, C. B.; Moses, C.; Arooj, M.; Blancafort, P. Zinc Fingers, TALEs, and CRISPR Systems: A Comparison of Tools for Epigenome Editing. *Methods in molecular biology (Clifton, N.J.)* **2018**, *1767*, 19–63, DOI: 10.1007/978-1-4939-7774-1_2.

(269) Boch, J.; Bonas, U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annual review of phytopathology* **2010**, *48*, 419–436, DOI: 10.1146/annurev-phyto-080508-081936.

(270) Büttner, D.; Gürlebeck, D.; Noël, L. D.; Bonas, U. HpaB from Xanthomonas campestris pv. vesicatoria acts as an exit control protein in type III-dependent protein secretion. *Molecular microbiology* **2004**, *54*, 755–768, DOI: 10.1111/j.1365-2958.2004.04302.x.

(271) Yang, Y.; Gabriel, D. W. Xanthomonas Avirulence/Pathogenicity Gene Family Encodes Functional Plant Nuclear Targeting Signals. *MPMI* **1995**, *8*, 627, DOI: 10.1094/MPMI-8-0627.

(272) Szurek, B.; Marois, E.; Bonas, U.; van den Ackerveken, G. Eukaryotic features of the Xanthomonas type III effector AvrBs3: Protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. *The Plant Journal* **2001**, *26*, 523–534, DOI: 10.1046/j.0960-7412.2001.01046.x.

(273) Kay, S.; Bonas, U. How Xanthomonas type III effectors manipulate the host plant. *Current opinion in microbiology* **2009**, *12*, 37–43, DOI: 10.1016/j.mib.2008.12.006.

(274) Bonas, U.; Stall, R. E.; Staskawicz, B. Genetic and structural characterization of the avirulence gene avrBs3 from Xanthomonas campestris pv. vesicatoria. *Molecular & general genetics : MGG* **1989**, *218*, 127–136.

(275) Boch, J.; Scholze, H.; Schornack, S.; Landgraf, A.; Hahn, S.; Kay, S.; Lahaye, T.; Nickstadt, A.; Bonas, U. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)* **2009**, *326*, 1509–1512, DOI: 10.1126/science.1178811.

(276) Rinaldi, F. C.; Doyle, L. A.; Stoddard, B. L.; Bogdanove, A. J. The effect of increasing numbers of repeats on TAL effector DNA binding specificity. *Nucleic acids research* **2017**, *45*, 6960–6970, DOI: 10.1093/nar/gkx342.

(277) Bogdanove, A. J.; Voytas, D. F. TAL effectors: customizable proteins for DNA targeting. *Science (New York, N.Y.)* **2011**, *333*, 1843–1846, DOI: 10.1126/science.1204094.

(278) Moscou, M. J.; Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science (New York, N.Y.)* **2009**, *326*, 1501, DOI: 10.1126/science.1178817.

(279) Deng, D.; Yan, C. Y.; Pan, X. J.; Wang, J. W.; Yan, N.; Shi, Y. G. *Crystal structure of the DNA-bound dHax3, a TAL effector, at 1.85 angstrom*, 2012.

(280) Mak, A. N.-S.; Bradley, P.; Cernadas, R. A.; Bogdanove, A. J.; Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science (New York, N.Y.)* **2012**, *335*, 716–719, DOI: 10.1126/science.1216211.

(281) Deng, D.; Yan, C.; Pan, X.; Mahfouz, M.; Wang, J.; Zhu, J.-K.; Shi, Y.; Yan, N. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science (New York, N.Y.)* **2012**, *335*, 720–723, DOI: 10.1126/science.1215670.

(282) Blainey, P. C.; Luo, G.; Kou, S. C.; Mangel, W. F.; Verdine, G. L.; Bagchi, B.; Xie, X. S. Nonspecifically bound proteins spin while diffusing along DNA. *Nature structural & molecular biology* **2009**, *16*, 1224–1229, DOI: 10.1038/nsmb.1716.

(283) Cuculis, L.; Abil, Z.; Zhao, H.; Schroeder, C. M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nature communications* **2015**, *6*, 7277, DOI: 10.1038/ncomms8277.

(284) Cuculis, L.; Abil, Z.; Zhao, H.; Schroeder, C. M. TALE proteins search DNA using a rotationally decoupled mechanism. *Nature chemical biology* **2016**, *12*, 831–837, DOI: 10.1038/nchembio.2152.

(285) Gao, H.; Wu, X.; Chai, J.; Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell research* **2012**, *22*, 1716–1720, DOI: 10.1038/cr.2012.156.

(286) Murakami, M. T.; Sforça, M. L.; Neves, J. L.; Paiva, J. H.; Domingues, M. N.; Pereira, A. L. A.; Zeri, A. C. d. M.; Benedetti, C. E. The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction. *Proteins* **2010**, *78*, 3386–3395, DOI: 10.1002/prot.22846.

(287) Tochio, N.; Umehara, K.; Uewaki, J.-I.; Flechsig, H.; Kondo, M.; Dewa, T.; Sakuma, T.; Yamamoto, T.; Saitoh, T.; Togashi, Y. *et al.* Non-RVD mutations that enhance the dynamics of the TAL repeat array along the superhelical axis improve TALEN genome editing efficacy. *Scientific reports* **2016**, *6*, 37887, DOI: 10.1038/srep37887.

(288) Schreiber, T.; Sorgatz, A.; List, F.; Blüher, D.; Thieme, S.; Wilmanns, M.; Bonas, U. Refined requirements for protein regions important for activity of the TALE AvrBs3. *PloS one* **2015**, *10*, e0120214, DOI: 10.1371/journal.pone.0120214.

(289) Schreiber, T.; Bonas, U. Repeat 1 of TAL effectors affects target specificity for the base at position zero. *Nucleic acids research* **2014**, *42*, 7160–7169, DOI: 10.1093/nar/gku341.

(290) Lamb, B. M.; Mercer, A. C.; Barbas, C. F. Directed evolution of the TALE N-terminal domain for recognition of all 5' bases. *Nucleic acids research* **2013**, *41*, 9779–9785, DOI: 10.1093/nar/gkt754.

(291) Mak, A. N.-S.; Bradley, P.; Bogdanove, A. J.; Stoddard, B. L. TAL effectors: function, structure, engineering and applications. *Current opinion in structural biology* **2013**, *23*, 93–99, DOI: 10.1016/j.sbi.2012.11.001.

(292) Miller, J. C.; Tan, S.; Qiao, G.; Barlow, K. A.; Wang, J.; Xia, D. F.; Meng, X.; Paschon, D. E.; Leung, E.; Hinkley, S. J. *et al.* A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **2011**, *29*, 143–148, DOI: 10.1038/nbt.1755.

(293) Zheng, C.-K.; Wang, C.-L.; Zhang, X.-P.; Wang, F.-J.; Qin, T.-F.; Zhao, K.-J. The last half-repeat of transcription activator-like effector (TALE) is dispensable and thereby TALE-based technology can be simplified. *Molecular plant pathology* **2014**, *15*, 690–697, DOI: 10.1111/mpp.12125.

(294) Wan, H.; Chang, S.; Hu, J.-p.; Tian, X.-h.; Wang, M.-H. Potential Role of the Last Half Repeat in TAL Effectors Revealed by a Molecular Simulation Study. *BioMed research international* **2016**, *2016*, 8036450, DOI: 10.1155/2016/8036450.

(295) Meckler, J. F.; Bhakta, M. S.; Kim, M.-S.; Ovadia, R.; Habrian, C. H.; Zykovich, A.; Yu, A.; Lockwood, S. H.; Morbitzer, R.; Elsäesser, J. *et al.* Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic acids research* **2013**, *41*, 4118–4128, DOI: 10.1093/nar/gkt085.

(296) Wan, H.; Hu, J.-p.; Li, K.-s.; Tian, X.-h.; Chang, S. Molecular dynamics simulations of DNA-free and DNA-bound TAL effectors. *PloS one* **2013**, *8*, e76045, DOI: 10.1371/journal.pone.0076045.

(297) Rogers, J. M.; Barrera, L. A.; Reyon, D.; Sander, J. D.; Kellis, M.; Joung, J. K.; Bulyk, M. L. Context influences on TALE-DNA binding revealed by quantitative profiling. *Nature communications* **2015**, *6*, 7440, DOI: 10.1038/ncomms8440.

(298) Wicky, B. I. M.; Stenta, M.; Dal Peraro, M. TAL effectors specificity stems from negative discrimination. *PloS one* **2013**, *8*, e80261, DOI: 10.1371/journal.pone.0080261.

(299) Deng, D.; Yan, C.; Wu, J.; Pan, X.; Yan, N. Revisiting the TALE repeat. *Protein & cell* **2014**, *5*, 297–306, DOI: 10.1007/s13238-014-0035-2.

(300) Le Cong; Zhou, R.; Kuo, Y.-C.; Cunniff, M.; Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **2012**, *3*, 968, DOI: 10.1038/ncomms1962.

(301) Streubel, J.; Blücher, C.; Landgraf, A.; Boch, J. TAL effector RVD specificities and efficiencies. *Nature biotechnology* **2012**, *30*, 593–595, DOI: 10.1038/nbt.2304.

(302) Yang, J.; Zhang, Y.; Yuan, P.; Zhou, Y.; Cai, C.; Ren, Q.; Wen, D.; Chu, C.; Qi, H.; Wei, W. Complete decoding of TAL effectors for DNA recognition. *Cell research* **2014**, *24*, 628–631, DOI: 10.1038/cr.2014.19.

(303) Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V.; Bogdanove, A. J.; Voytas, D. F. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **2011**, *39*, e82, DOI: 10.1093/nar/gkr218.

(304) Reyon, D.; Tsai, S. Q.; Khayter, C.; Foden, J. A.; Sander, J. D.; Joung, J. K. FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **2012**, *30*, 460–465, DOI: 10.1038/nbt.2170.

(305) Schmid-Burgk, J. L.; Schmidt, T.; Kaiser, V.; Höning, K.; Hornung, V. A ligation-independent cloning technique for high-throughput assembly of transcription activator–like effector genes. *Nature biotechnology* **2013**, *31*, 76–81, DOI: 10.1038/nbt.2460.

(306) Gogolok, S.; Garcia-Diaz, C.; Pollard, S. M. STAR: a simple TAL effector assembly reaction using isothermal assembly. *Scientific reports* **2016**, *6*, 33209, DOI: 10.1038/srep33209.

(307) Joung, J. K.; Sander, J. D. TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology* **2013**, *14*, 49–55, DOI: 10.1038/nrm3486.

(308) Zhang, F.; Le Cong; Lodato, S.; Kosuri, S.; Church, G. M.; Arlotta, P. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **2011**, *29*, 149–153, DOI: 10.1038/nbt.1775.

(309) Geiβler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S.-E.; Boch, J. Transcriptional Activators of Human Genes with Programmable DNA-Specificity. *PloS one* **2011**, *6*, e19509, DOI: 10.1371/journal.pone.0019509.

(310) Mahfouz, M. M.; Li, L.; Piatek, M.; Fang, X.; Mansour, H.; Bangarusamy, D. K.; Zhu, J.-K. Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant molecular biology* **2012**, *78*, 311–321, DOI: 10.1007/s11103-011-9866-x.

(311) Bernstein, D. L.; Le Lay, J. E.; Ruano, E. G.; Kaestner, K. H. TALE-mediated epigenetic suppression of CDKN2A increases replication in human fibroblasts. *The Journal of clinical investigation* **2015**, *125*, 1998–2006, DOI: 10.1172/JCI77321.

(312) Maeder, M. L.; Angstman, J. F.; Richardson, M. E.; Linder, S. J.; Cascio, V. M.; Tsai, S. Q.; Ho, Q. H.; Sander, J. D.; Reyon, D.; Bernstein, B. E. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature biotechnology* **2013**, *31*, 1137–1142, DOI: 10.1038/nbt.2726.

(313) Ma, H.; Reyes-Gutierrez, P.; Pederson, T. Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, 21048–21053, DOI: 10.1073/pnas.1319097110.

(314) Ma, Y.; Wang, M.; Li, W.; Zhang, Z.; Zhang, X.; Tan, T.; Zhang, X.-E.; Cui, Z. Live cell imaging of single genomic loci with quantum dot-labeled TALEs. *Nature communications* **2017**, *8,* DOI: 10.1038/ncomms15318.

(315) Yuan, K.; Shermoen, A. W.; O'Farrell, P. H. Illuminating DNA replication during Drosophila development using TALE-lights. *Current biology : CB* **2014**, *24*, R144-5, DOI: 10.1016/j.cub.2014.01.023.

(316) Miller, J. C.; Zhang, L.; Xia, D. F.; Campo, J. J.; Ankoudinova, I. V.; Guschin, D. Y.; Babiarz, J. E.; Meng, X.; Hinkley, S. J.; Lam, S. C. *et al.* Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nature methods* **2015**, *12*, 465–471, DOI: 10.1038/nmeth.3330.

(317) Juillerat, A.; Pessereau, C.; Dubois, G.; Guyot, V.; Maréchal, A.; Valton, J.; Daboussi, F.; Poirot, L.; Duclert, A.; Duchateau, P. Optimized tuning of TALEN specificity using non-conventional RVDs. *Scientific reports* **2015**, *5,* DOI: 10.1038/srep08150.

(318) Valton, J.; Dupuy, A.; Daboussi, F.; Thomas, S.; Maréchal, A.; Macmaster, R.; Melliand, K.; Juillerat, A.; Duchateau, P. Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *The Journal of biological chemistry* **2012**, *287*, 38427–38432, DOI: 10.1074/jbc.C112.408864.

(319) Bultmann, S.; Morbitzer, R.; Schmidt, C. S.; Thanisch, K.; Spada, F.; Elsaesser, J.; Lahaye, T.; Leonhardt, H. Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic acids research* **2012**, *40*, 5368–5377, DOI: 10.1093/nar/gks199.

(320) Kubik, G.; Schmidt, M. J.; Penner, J. E.; Summerer, D. Programmable and highly resolved in vitro detection of 5-methylcytosine by TALEs. *Angewandte Chemie (International ed. in English)* **2014**, *53*, 6002–6006, DOI: 10.1002/anie.201400436.

(321) Deng, D.; Yin, P.; Yan, C.; Pan, X.; Gong, X.; Qi, S.; Xie, T.; Mahfouz, M.; Zhu, J.-K.; Yan, N. *et al.* Recognition of methylated DNA by TAL effectors. *Cell research* **2012**, *22*, 1502–1504, DOI: 10.1038/cr.2012.127.

(322) Kubik, G.; Summerer, D. Achieving single-nucleotide resolution of 5-methylcytosine detection with TALEs. *Chembiochem : a European journal of chemical biology* **2015**, *16*, 228–231, DOI: 10.1002/cbic.201402408.

(323) Kubik, G.; Batke, S.; Summerer, D. Programmable sensors of 5-hydroxymethylcytosine. *Journal of the American Chemical Society* **2015**, *137*, 2–5, DOI: 10.1021/ja506022t.

(324) Rathi, P.; Maurer, S.; Kubik, G.; Summerer, D. Isolation of Human Genomic DNA Sequences with Expanded Nucleobase Selectivity. *Journal of the American Chemical Society* **2016**, *138*, 9910–9918, DOI: 10.1021/jacs.6b04807.

(325) Maurer, S.; Giess, M.; Koch, O.; Summerer, D. Interrogating Key Positions of Size-Reduced TALE Repeats Reveals a Programmable Sensor of 5-Carboxylcytosine. *ACS chemical biology* **2016**, *11*, 3294–3299, DOI: 10.1021/acschembio.6b00627.

(326) Maurer, S.; Buchmuller, B.; Ehrt, C.; Jasper, J.; Koch, O.; Summerer, D. Overcoming conservation in TALE-DNA interactions: a minimal repeat scaffold enables selective recognition of an oxidized 5-methylcytosine. *Chemical science* **2018**, *9*, 7247–7252, DOI: 10.1039/c8sc01958d.

(327) Zhang, Y.; Liu, L.; Guo, S.; Song, J.; Zhu, C.; Yue, Z.; Wei, W.; Yi, C. Deciphering TAL effectors for 5-methylcytosine and 5-hydroxymethylcytosine recognition. *Nature communications* **2017**, *8*, 901, DOI: 10.1038/s41467-017-00860-6.

(328) Tsuji, S.; Futaki, S.; Imanishi, M. Sequence-specific recognition of methylated DNA by an engineered transcription activator-like effector protein. *Chemical communications (Cambridge, England)* **2016**, *52*, 14238–14241, DOI: 10.1039/c6cc06824c.

(329) Rathi, P.; Witte, A.; Summerer, D. Engineering DNA Backbone Interactions Results in TALE Scaffolds with Enhanced 5-Methylcytosine Selectivity. *Scientific reports* **2017**, *7*, 15067, DOI: 10.1038/s41598-017-15361-1.

(330) Liu, C. C.; Schultz, P. G. Adding new chemistries to the genetic code. *Annual review of biochemistry* **2010**, *79*, 413–444, DOI: 10.1146/annurev.biochem.052308.105824.

(331) Noren, C.; Anthony-Cahill, S.; Griffith, M.; Schultz, P. A general method for site-specific incorporation of unnatural amino acids into proteins. *Science* **1989**, *244*, 182–188, DOI: 10.1126/science.2649980.

(332) Bain, J. d.; Diala, E. S.; Glabe, C. G.; Dix, T. A.; Chamberlin, A. R. Biosynthetic site-specific incorporation of a non-natural amino acid into a polypeptide. *J. Am. Chem. Soc.* **1989**, *111*, 8013–8014, DOI: 10.1021/ja00202a052.

(333) Wang, L.; Schultz, P. G. Expanding the genetic code. *Angewandte Chemie (International ed. in English)* **2004**, *44*, 34–66, DOI: 10.1002/anie.200460627.

(334) Bossi, L.; Roth, J. R. The influence of codon context on genetic code translation. *Nature* **1980**, *286*, 123–127.

(335) Nakamura, Y.; Gojobori, T.; Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic acids research* **2000**, *28*, 292.

(336) Garen, A.; Siddiqi, O. Suppression of mutations in the alkaline phosphatase structural cistron of E. coli. *Proceedings of the National Academy of Sciences of the United States of America* **1962**, *48*, 1121–1127.

(337) Benzer, S.; Champe, S. P. A change from nonsense to sense in the genetic code. *Proceedings of the National Academy of Sciences* **1962**, *48*, 1114–1121.

(338) Wakasugi, K.; Quinn, C. L.; Tao, N.; Schimmel, P. Genetic code in evolution: switching species-specific aminoacylation with a peptide transplant. *The EMBO Journal* **1998**, *17*, 297–305, DOI: 10.1093/emboj/17.1.297.

(339) Kwok, Y.; Wong, J. T. Evolutionary relationship between Halobacterium cutirubrum and eukaryotes determined by use of aminoacyl-tRNA synthetases as phylogenetic probes. *Canadian journal of biochemistry* **1980**, *58*, 213–218.

(340) Doctor, B. P.; Mudd, A. J. Species Specificity of Amino Acid Acceptor Ribonucleic Acid and Aminoacyl Soluble Ribonucleic Acid Synthetases. *The Journal of biological chemistry* **1963**, *238*, 3677–3681.

(341) Liu, D. R.; Schultz, P. G. Progress toward the evolution of an organism with an expanded genetic code. *Proceedings of the National Academy of Sciences* **1999**, *96*, 4780–4785, DOI: 10.1073/pnas.96.9.4780.

(342) Melançon, C. E.; Schultz, P. G. One plasmid selection system for the rapid evolution of aminoacyl-tRNA synthetases. *Bioorganic & medicinal chemistry letters* **2009**, *19*, 3845–3847, DOI: 10.1016/j.bmcl.2009.04.007.

(343) Wang, L.; Schultz, P. G. A general approach for the generation of orthogonal tRNAs. *Chemistry & biology* **2001**, *8*, 883–890, DOI: 10.1016/S1074-5521(01)00063-1.

(344) Santoro, S. W.; Wang, L.; Herberich, B.; King, D. S.; Schultz, P. G. An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nature biotechnology* **2002**, *20*, 1044–1048, DOI: 10.1038/nbt742.

(345) Bryson, D. I.; Fan, C.; Guo, L.-T.; Miller, C.; Söll, D.; Liu, D. R. Continuous directed evolution of aminoacyl-tRNA synthetases. *Nature chemical biology* **2017**, *13*, 1253–1260, DOI: 10.1038/nchembio.2474.

(346) Liu, D. R.; Magliery, T. J.; Schultz, P. G. Characterization of an 'orthogonal' suppressor tRNA derived from E. coli tRNA2Gln. *Chemistry & biology* **1997**, *4*, 685–691, DOI: 10.1016/S1074-5521(97)90224-6.

(347) Wang, L.; Magliery, T. J.; Liu, D. R.; Schultz, P. G. A New Functional Suppressor tRNA/Aminoacyl−tRNA Synthetase Pair for the in Vivo Incorporation of Unnatural Amino Acids into Proteins. *Journal of the American Chemical Society* **2000**, *122*, 5010–5011, DOI: 10.1021/ja000595y.

(348) Chin, J. W. Expanding and reprogramming the genetic code. *Nature* **2017**, *550*, 53–60, DOI: 10.1038/nature24031.

(349) Davis, L.; Chin, J. W. Designer proteins: applications of genetic code expansion in cell biology. *Nature reviews. Molecular cell biology* **2012**, *13*, 168–182, DOI: 10.1038/nrm3286.

(350) Schmidt, M. J.; Summerer, D. Genetic code expansion as a tool to study regulatory processes of transcription. *Frontiers in chemistry* **2014**, *2*, 7, DOI: 10.3389/fchem.2014.00007.

(351) Xie, J.; Wang, L.; Wu, N.; Brock, A.; Spraggon, G.; Schultz, P. G. The site-specific incorporation of p-iodo-L-phenylalanine into proteins for structure determination. *Nature biotechnology* **2004**, *22*, 1297–1301, DOI: 10.1038/nbt1013.

(352) Wang, L.; Xie, J.; Deniz, A. A.; Schultz, P. G. Unnatural amino acid mutagenesis of green fluorescent protein. *The Journal of organic chemistry* **2003**, *68*, 174–176, DOI: 10.1021/jo026570u.

(353) Schultz, K. C.; Supekova, L.; Ryu, Y.; Xie, J.; Perera, R.; Schultz, P. G. A genetically encoded infrared probe. *J. Am. Chem. Soc.* **2006**, *128*, 13984–13985, DOI: 10.1021/ja0636690.

(354) Cellitti, S. E.; Jones, D. H.; Lagpacan, L.; Hao, X.; Zhang, Q.; Hu, H.; Brittain, S. M.; Brinker, A.; Caldwell, J.; Bursulaya, B. *et al.* In vivo incorporation of unnatural amino acids to probe structure, dynamics, and ligand binding in a large protein by nuclear magnetic resonance spectroscopy. *Journal of the American Chemical Society* **2008**, *130*, 9268–9281, DOI: 10.1021/ja801602q.

(355) Schmidt, M. J.; Borbas, J.; Drescher, M.; Summerer, D. A genetically encoded spin label for electron paramagnetic resonance distance measurements. *Journal of the American Chemical Society* **2014**, *136*, 1238–1241, DOI: 10.1021/ja411535q.

(356) Summerer, D.; Chen, S.; Wu, N.; Deiters, A.; Chin, J. W.; Schultz, P. G. A genetically encoded fluorescent amino acid. *Proceedings of the National Academy of Sciences* **2006**, *103*, 9785–9789, DOI: 10.1073/pnas.0603965103.

(357) Wang, J.; Xie, J.; Schultz, P. G. A genetically encoded fluorescent amino acid. *J. Am. Chem. Soc.* **2006**, *128*, 8738–8739, DOI: 10.1021/ja062666k.

(358) Lee, H. S.; Guo, J.; Lemke, E. A.; Dimla, R. D.; Schultz, P. G. Genetic incorporation of a small, environmentally sensitive, fluorescent probe into proteins in Saccharomyces cerevisiae. *J. Am. Chem. Soc.* **2009**, *131*, 12921–12923, DOI: 10.1021/ja904896s.

(359) Young, D. D.; Deiters, A. Photochemical control of biological processes. *Organic & biomolecular chemistry* **2007**, *5*, 999–1005, DOI: 10.1039/b616410m.

(360) Courtney, T.; Deiters, A. Recent advances in the optical control of protein function through genetic code expansion. *Current opinion in chemical biology* **2018**, *46*, 99–107, DOI: 10.1016/j.cbpa.2018.07.011.

(361) Lee, H. S.; Dimla, R. D.; Schultz, P. G. Protein-DNA photo-crosslinking with a genetically encoded benzophenone-containing amino acid. *Bioorganic & medicinal chemistry letters* **2009**, *19*, 5222–5224, DOI: 10.1016/j.bmcl.2009.07.011.

(362) Chin, J. W.; Martin, A. B.; King, D. S.; Wang, L.; Schultz, P. G. Addition of a photocrosslinking amino acid to the genetic code of Escherichiacoli. *Proceedings of the National Academy of Sciences* **2002**, *99*, 11020–11024, DOI: 10.1073/pnas.172226299.

(363) Xiao, H.; Schultz, P. G. At the Interface of Chemical and Biological Synthesis: An Expanded Genetic Code. *Cold Spring Harbor perspectives in biology* **2016**, *8,* DOI: 10.1101/cshperspect.a023945.

(364) Kim, C. H.; Axup, J. Y.; Schultz, P. G. Protein conjugation with genetically encoded unnatural amino acids. *Current opinion in chemical biology* **2013**, *17*, 412–419, DOI: 10.1016/j.cbpa.2013.04.017.

(365) Wang, L.; Zhang, Z.; Brock, A.; Schultz, P. G. Addition of the keto functional group to the genetic code of Escherichia coli. *Proceedings of the National Academy of Sciences* **2003**, *100*, 56–61, DOI: 10.1073/pnas.0234824100.

(366) Datta, D.; Wang, P.; Carrico, I. S.; Mayo, S. L.; Tirrell, D. A. A designed phenylalanyl-tRNA synthetase variant allows efficient in vivo incorporation of aryl ketone functionality into proteins. *J. Am. Chem. Soc.* **2002**, *124*, 5652–5653.

(367) Zhang, Z.; Smith, B. A. C.; Wang, L.; Brock, A.; Cho, C.; Schultz, P. G. A new strategy for the site-specific modification of proteins in vivo. *Biochemistry* **2003**, *42*, 6735–6746, DOI: 10.1021/bi0300231.

(368) Spicer, C. D.; Davis, B. G. Selective chemical protein modification. *Nature communications* **2014**, *5*, 4740, DOI: 10.1038/ncomms5740.

(369) Gieß, M.; Witte, A.; Jasper, J.; Koch, O.; Summerer, D. Complete, Programmable Decoding of Oxidized 5-Methylcytosine Nucleobases in DNA by Chemoselective Blockage of Universal Transcription-Activator-Like Effector Repeats. *Journal of the American Chemical Society* **2018,** DOI: 10.1021/jacs.8b02909.

(370) Gieß, M.; Munoz-Lopez, A.; Buchmuller, B.; Kubik, G.; Summerer, D. Programmable Protein-DNA Crosslinking for the Direct Capture and Quantification of 5-Formylcytosine. *J. Am. Chem. Soc.* **2019**, 9453–9457.

(371) Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols* **2006**, *1*, 2876–2890, DOI: 10.1038/nprot.2006.202.

(372) Flade, S.; Jasper, J.; Gieß, M.; Juhasz, M.; Dankers, A.; Kubik, G.; Koch, O.; Weinhold, E.; Summerer, D. The N6-Position of Adenine Is a Blind Spot for TAL-Effectors That Enables Effective Binding of Methylated and Fluorophore-Labeled DNA. *ACS chemical biology* **2017**, *12*, 1719–1725, DOI: 10.1021/acschembio.7b00324.

(373) Rappoport, Z.; Liebman, J. F. *The chemistry of hydroxylamines, oximes and hydroxamic acids*; Patai series; Wiley: Chichester, England, 2009.

(374) Shibutani, T.; Ito, S.; Toda, M.; Kanao, R.; Collins, L. B.; Shibata, M.; Urabe, M.; Koseki, H.; Masuda, Y.; Swenberg, J. A. *et al.* Guanine- 5-carboxylcytosine base pairs mimic mismatches during DNA replication. *Scientific reports* **2014**, *4*, 5220, DOI: 10.1038/srep05220.

(375) Steigenberger, B.; Schiesser, S.; Hackner, B.; Brandmayr, C.; Laube, S. K.; Steinbacher, J.; Pfaffeneder, T.; Carell, T. Synthesis of 5-hydroxymethyl-, 5-formyl-, and 5-carboxycytidine-triphosphates and their incorporation into oligonucleotides by polymerase chain reaction. *Organic letters* **2013**, *15*, 366–369, DOI: 10.1021/ol3033219.

(376) Baumann, H.; Knapp, S.; Lundbäck, T.; Ladenstein, R.; Härd, T. Solution structure and DNA-binding properties of a thermostable protein from the archaeon Sulfolobus solfataricus. *Nature structural biology* **1994**, *1*, 808–819.

(377) Holder, P. G.; Francis, M. B. Integration of a Self-Assembling Protein Scaffold with Water-Soluble Single-Walled Carbon Nanotubes. *Angew. Chem.* **2007**, *119*, 4448–4451, DOI: 10.1002/ange.200700333.

(378) Chemical Computing Group ULC. *Molecular Operating Environment (MOE)*; 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2018.

(379) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004**, *25*, 1605–1612, DOI: 10.1002/jcc.20084.

(380) Gerber, P. R.; Müller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *Journal of computer-aided molecular design* **1995**, *9*, 251–268.

(381) Case, D. A.; Darden, T. A.; Cheatham; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W. *et al. AMBER 10*; University of California, San Francisco, 2008.