

**Genomic single molecule analysis of
chromatin structure around DNA
replication origins in *Saccharomyces
cerevisiae***



Heather Mary Jeffery

Sir William Dunn School of Pathology

University of Oxford

This dissertation is submitted for the degree of
Doctor of Philosophy (DPhil) in Interdisciplinary Bioscience
(Doctoral Training Programme)

University College Oxford

August 2021

Acknowledgements

Firstly, I would like to thank my supervisor Professor Conrad Nieduszynski for his continued support, patience and down-to-earth approach to problem solving. The enthusiastic and insightful discussions we have had throughout my PhD motivated and inspired me from the start. I would also like to thank Dr David Sims for the computational training, discussions and feedback which has reinforced my interest and abilities in bioinformatics. I would also like to thank Professor Nicholas Proudfoot for his experimental insights and feedback in the later stages of my PhD which provided new perspectives.

I will always be thankful to the Replication Group who have been there through thick and thin, always there with technical advice and encouragement. In particular I would like to thank Dr Carolin Müller and Dr Dzmitry Batrakou for their day-to-day experimental support, in addition to Stephanie Barker, Dr Michael Boemo, Dr Jamie Carrington, Dr Emma Heron, Dr Phoebe Oldach, Dr Ana Wallis, and Dr Rosemary Wilson for never seeing a question as too small to be answered and for making the lab group a community.

I am incredibly grateful to collaborations with Dr Afaf El-Sagheer and Paolo Spingardi whose mass spectrometry expertise provided important ways to validate my research. I also appreciate the shared resources received from Dr Georgi Marinov and Professor David MacAlpine.

I am grateful to both the Dunn School and University College Oxford for creating supportive communities through which I have had many exciting opportunities both academically and personally, which have helped me to learn about myself and develop networks. I would

particularly like to thank my tutor, Professor Catherine Pears and my programme director, Professor Esther Becker, for their mentorship, support and thoughtfulness throughout my PhD. I would also like to thank both University College Oxford and the BBSRC for their financial support.

I am grateful to Singer Instruments, with whom I did my internship, for the opportunities they gave me and their continued confidence in me. In particular, I am incredibly thankful to my internship supervisor, Dr Oliver Severn, for going above and beyond his role and continuing to provide mentorship and to be an inspirational role model once I returned to my PhD studies.

On a personal note, I am incredibly indebted to my partner, Samuel Kirby, who has been there for me at every turn and never doubted me. He has always had my back and been the voice of reason when I have needed it. I am also incredibly thankful to my parents, my brother, my grandparents and Sam's family for their continued support and encouragement from near and far through the ups and downs of this adventure. My friends have also been amazing, particularly Natalie Bevan, Renee Tonkin, Gemma Daszki, Beatriz Moreno Garcia, Mirjam Hazenbosch, Claire Hill, Luciné Karanfilian, Emma Walmsley and Niall Bootland among others, who have always provided listening ears, fresh perspectives and a smile.

Table of contents

List of figures	xi
List of tables	xv
Abbreviations	xvii
1 Introduction	2
1.1 DNA replication	2
1.1.1 Overview	2
1.1.2 DNA replication initiation in <i>Saccharomyces cerevisiae</i>	3
1.1.3 DNA replication initiation among other yeast	7
1.1.4 DNA replication initiation in bacteria and archaea	8
1.1.5 DNA replication initiation among other eukaryotes	9
1.1.6 Clinical outcome of mutations in replication initiation proteins . . .	11
1.1.7 Effect of chromatin structure on DNA replication	12
1.2 <i>S. cerevisiae</i> as a model system	15
1.3 Population-based chromatin accessibility assays	15
1.3.1 Population-based investigation of DNA replication	17
1.3.2 Population vs single cell vs single molecule methods	18
1.4 Nanopore sequencing	19
1.5 Aims	22

2	Materials & Methods	25
2.1	General	25
2.2	DNA quantification	26
2.3	Primers	26
2.4	Enzymes	29
2.5	sacCer3 reference genome analysis	29
2.6	Psoralen crosslinking	30
2.6.1	Annealing	30
2.6.2	Crosslinking with TMP	30
2.6.3	Filtration	31
2.6.4	Crosslink reversal	31
2.6.5	Polyacrylamide denaturing gels	32
2.6.6	Mass spectrometry	33
2.6.7	Reannealing of oligonucleotides	33
2.6.8	Phosphorylation	34
2.6.9	Ligation	34
2.6.10	Agarose gels	35
2.6.11	Agarose gel extraction	35
2.6.12	Sanger sequencing	36
2.6.13	Nanopore sequencing	36
2.6.14	DNA modification with angelicin	36
2.7	<i>S. cerevisiae</i> and methylation	39
2.7.1	Yeast strains	39
2.7.2	Yeast growth media	42
2.7.3	DNA extraction	43
2.7.4	Recombinant nucleosomes	44

2.7.5	Droplet Digital PCR	45
2.7.6	Chromatin extraction for ddPCR	47
2.7.7	MNase digestion of chromatin	50
2.7.8	Methylation of chromatin with M.CviPI	51
2.7.9	SMAC-seq variant	53
2.7.10	Yeast strain generation	55
2.7.11	Growth curves	57
2.7.12	α -factor arrest of <i>S. cerevisiae</i> cells	58
2.7.13	Flow cytometry	58
2.7.14	Protein extraction	59
2.7.15	Western blots	59
2.7.16	Mass spectrometry	62
2.7.17	Nanopore sequencing	64
2.7.18	Nanopore data analysis	64
2.8	Other bioinformatics	66
2.9	Code availability	66
3	Use of psoralens to probe <i>S. cerevisiae</i> chromatin structure <i>in vitro</i>	67
3.1	Introduction	67
3.1.1	Chemical crosslinking	67
3.1.2	Furanocoumarins	68
3.1.3	Furanocoumarins as markers of chromatin accessibility	70
3.1.4	Nanopore sequencing of modified DNA	75
3.1.5	Aim of this chapter	77
3.2	Distribution of modification sites across the <i>Saccharomyces cerevisiae</i> genome	77
3.2.1	Potential modification sites within nucleosome-bound DNA	78
3.2.2	Potential modification sites within DNA linkers	80

3.3	4,5',8-trimethylpsoralen (TMP) modification of DNA	84
3.3.1	Overview of approach for TMP modification of DNA	84
3.3.2	Polyacrylamide denaturing gels detect TMP crosslinking	87
3.3.3	Mass spectrometry detects TMP modification	93
3.3.4	Generation of psoralen-treated sample for nanopore sequencing . . .	98
3.3.5	Nanopore sequencing does not detect TMP modification on DNA . .	111
3.4	Angelicin modification of DNA	118
3.4.1	Overview of approach for angelicin modification of DNA	118
3.4.2	Sanger sequencing does not detect angelicin modification on DNA .	121
3.4.3	Mass spectrometry does not detect angelicin modification on DNA .	123
3.5	Discussion	125
4	Use of adenine methylation to probe chromatin structure both <i>in vitro</i> and <i>in vivo</i> in <i>S. cerevisiae</i>	132
4.1	Introduction	132
4.1.1	Natural methylation in a variety of organisms	133
4.1.2	M.EcoGII, a non-specific adenine methyltransferase	136
4.1.3	Methylation as a marker of chromatin accessibility	137
4.1.4	Methylation detection by nanopore sequencing	139
4.1.5	Methylation detection by PacBio SMRT-sequencing	142
4.1.6	Aim of this chapter	143
4.2	Qualitative analysis of methylation with enzyme digests	143
4.2.1	Methylation inferred from restriction enzyme protection assay . . .	144
4.3	Quantitative analysis of methylation using Droplet Digital PCR (ddPCR) . .	147
4.3.1	Quantitative ddPCR on plasmid DNA	149
4.3.2	Quantitative ddPCR on genomic DNA	155
4.3.3	Quantitative ddPCR on recombinant nucleosomes	161

4.3.4	Quantitative ddPCR on chromatin	169
4.4	<i>In vitro</i> assessment of <i>S. cerevisiae</i> chromatin structure	176
4.4.1	Validation of <i>S. cerevisiae</i> chromatin preps by MNase treatment . .	176
4.4.2	Validation of <i>S. cerevisiae</i> chromatin preps by protein detection . .	177
4.4.3	Mass spectrometry shows very low cytosine methylation of spheroplasts	179
4.4.4	Variant of SMAC-seq approach to methylate chromatin <i>in vitro</i> does not completely replicate SMAC-seq results	181
4.5	Verification of inducible Dam methyltransferase <i>S. cerevisiae</i> strain and assessment of motif-specific adenine methylation detection by nanopore sequencing	185
4.5.1	Inducible Dam methyltransferase identified by western blot	185
4.5.2	Effect of inducible Dam methyltransferase on <i>S. cerevisiae</i> growth .	187
4.5.3	DNA methylation by inducible Dam methyltransferase identified in <i>S. cerevisiae</i> by enzyme digests and nanopore sequencing	188
4.6	<i>In vitro</i> M.EcoGII treatment of <i>S. cerevisiae</i> genomic DNA	200
4.6.1	Mass spectrometry of <i>in vitro</i> M.EcoGII methylated genomic DNA	200
4.6.2	Nanopore sequencing of <i>in vitro</i> M.EcoGII methylated genomic DNA	201
4.7	Creation and validation of inducible M.EcoGII <i>S. cerevisiae</i> strain and assess- ment of non-specific adenine methylation detection by nanopore sequencing	208
4.7.1	Inducible M.EcoGII <i>S. cerevisiae</i> strain generation and verification	208
4.7.2	Inducible M.EcoGII identified by western blotting	211
4.7.3	Backcrossing of inducible M.EcoGII <i>S. cerevisiae</i> strain	212
4.7.4	Effect of inducible M.EcoGII on <i>S. cerevisiae</i> growth	215
4.7.5	DNA methylation by inducible M.EcoGII methyltransferase identi- fied in <i>S. cerevisiae</i> by mass spectrometry and nanopore sequencing	216
4.8	Inducible M.EcoGII <i>S. cerevisiae</i> strains with replication protein mutations	223

4.8.1	Strain generation and verification	223
4.8.2	Effect of replication protein mutations combined with inducible M.EcoGII on <i>S. cerevisiae</i> growth	226
4.8.3	Cell cycle progression of strains with replication protein mutations and inducible M.EcoGII	229
4.8.4	Mass spectrometry of strains with replication protein mutations and inducible M.EcoGII	231
4.9	Inducible M.EcoGII <i>S. cerevisiae</i> strain arrested in G1 phase	233
4.9.1	Experimental procedure and cell cycle analysis	233
4.9.2	Nanopore sequencing of inducible M.EcoGII <i>S. cerevisiae</i> strain arrested in G1 phase	235
4.10	Detecting <i>S. cerevisiae</i> chromatin structure with adenine methylation per- formed <i>in vivo</i> by an inducible M.EcoGII	237
4.11	Discussion	251
5	Discussion and Conclusion	257
5.1	Discussion	257
5.2	Conclusion	265
	References	266
	Appendix A Code for producing figures	287
A.1	Chapter 3	287
A.1.1	Tables	287
A.1.2	Figures	288
A.2	Chapter 4	294
A.2.1	Tables	294
A.2.2	Figures	295

Table of contents	x
-------------------	---

Appendix B Licence for reproducing figure 4.18a	313
--	------------

List of figures

1.1	Initial steps of DNA replication in <i>S. cerevisiae</i>	5
1.2	Nucleosome positioning around <i>S. cerevisiae</i> replication origins	13
1.3	Nanopore sequencing concept	22
1.4	Single molecule chromatin accessibility assay	23
1.5	Single molecule chromatin accessibility assay applied to DNA replication origin licensing	24
3.1	Furanocoumarin chemical structures	69
3.2	TMP as a marker of chromatin accessibility	72
3.3	Structures of base modifications	76
3.4	Chemical modification target sites in the <i>Saccharomyces cerevisiae</i> genome	79
3.5	Characterisation of linker regions in the <i>Saccharomyces cerevisiae</i> genome	81
3.6	Potential modification sites in linker regions with replication origins	83
3.7	Experimental workflow of TMP modification of DNA	86
3.8	TMP crosslink on DNA can be detected by polyacrylamide denaturing gels	89
3.9	Potential concatemerisation of oligonucleotides	90
3.10	Column filtering retains crosslinked DNA	92
3.11	Mass spectrometry detects TMP modification on DNA	97
3.12	TMP interstrand crosslink can be reversed with NaOH or 254 nm UV treatment	101
3.13	Ligation schematic producing desired product	104

3.14 Self-ligation schematic	105
3.15 Ligation of TMP modified DNA to create longer substrate	106
3.16 Linearisation of TMP modified DNA	107
3.17 Effectiveness of restriction enzymes to prevent large fragment self-ligation .	109
3.18 Sanger sequencing does not detect TMP modification on DNA	111
3.19 TMP modification does not affect nanopore sequencing coverage or indels .	113
3.20 TMP modification does not cause signal perturbation in nanopore sequencing	117
3.21 Experimental workflow for angelicin modification of DNA	120
3.22 Sanger sequencing does not detect angelicin modification on DNA	122
3.23 Mass spectrometry does not detect angelicin modification on DNA	124
4.1 Restriction enzyme protection assay identifies adenine methylation sensitive restriction enzymes	146
4.2 Ability of PCR primers to bind to methylated DNA	149
4.3 Optimisation of plasmid DNA concentration for ddPCR	151
4.4 Detection of enzyme digestion by ddPCR	153
4.5 Detection of 6mA in plasmid DNA by ddPCR	155
4.6 Testing ddPCR accuracy on methylated DNA	156
4.7 ddPCR quantification of 6mA at one site in genomic DNA	157
4.8 ddPCR quantification of 6mA at multiple sites in genomic DNA	160
4.9 Recombinant nucleosome plasmid	161
4.10 ddPCR assessment of the methylated adenine sensitivity of AluI	163
4.11 ddPCR quantification of methylation on recombinant nucleosome plasmid .	168
4.12 ddPCR quantification of methylation on <i>S. cerevisiae</i> chromatin	170
4.13 ddPCR quantification of methylation on <i>S. cerevisiae</i> spheroplasts	173
4.14 ddPCR quantification of methylation on <i>S. cerevisiae</i> chromatin with pUC19 spike-in	175

4.15	Detection of nucleosome ladder by MNase digest of DNA extracted from <i>S. cerevisiae</i> chromatin	177
4.16	Western blot shows correct separation of proteins in chromatin extraction from <i>S. cerevisiae</i>	179
4.17	Mass spectrometry of chromatin treated with a GC methyltransferase (M.CviPI)	181
4.18	Nanopore sequencing of SMAC-seq variant experiment	184
4.19	Western blotting detects inducible Dam methyltransferase in timecourse . .	186
4.20	Growth analysis of inducible Dam methyltransferase <i>S. cerevisiae</i> strain . .	188
4.21	Methylation investigated by restriction enzyme digests of inducible Dam methyltransferase throughout timecourse	190
4.22	Nanopore sequencing detection of methylation in Dam methyltransferase timecourse	191
4.23	Nanopore sequencing sequence context analysis of Dam methyltransferase methylation	193
4.24	Nanopore sequencing model specificity for adenine methylation	195
4.25	Nanopore sequencing detection of 6mA in GATC sites around nucleosomes	197
4.26	Nanopore sequencing detection of 6mA in GATC sites around <i>ARS1513</i> . .	199
4.27	Mass spectrometry of <i>in vitro</i> genomic DNA treated with M.EcoGII dilution series	201
4.28	Nanopore sequencing threshold comparisons to mass spectrometry data . .	204
4.29	Nanopore sequencing 4-mer analysis of <i>in vitro</i> M.EcoGII-treated DNA . .	205
4.30	Nanopore sequencing 6-mer analysis across 6mA classification thresholds .	207
4.31	Plasmid containing M.EcoGII for transformation	209
4.32	Generation and verification of an inducible M.EcoGII <i>S. cerevisiae</i> strain .	210
4.33	Detection of inducible M.EcoGII by western blotting	212
4.34	Backcrossing of an inducible M.EcoGII <i>S. cerevisiae</i>	214

4.35	Growth of an <i>S. cerevisiae</i> strain with an inducible M.EcoGII	216
4.36	Mass spectrometry of <i>in vivo</i> inducible M.EcoGII timecourse	219
4.37	Nanopore sequencing 6mA detection of <i>in vivo</i> inducible M.EcoGII timecourses	222
4.38	Generation and verification of <i>orc1-161</i> strain with an inducible M.EcoGII .	224
4.39	Generation and verification of <i>cdc6-1</i> strain with an inducible M.EcoGII . .	225
4.40	Spot assay assesses temperature sensitive replication protein mutant strains	226
4.41	Doubling times of replication protein mutant and inducible M.EcoGII strains	228
4.42	Flow cytometry of replication protein mutant and inducible M.EcoGII strains	230
4.43	Mass spectrometry of replication protein mutant and inducible M.EcoGII strains	232
4.44	Light microscopy images of <i>S. cerevisiae</i> G1 α -factor arrest	234
4.45	Overall methylation detected in nanopore sequencing of G1 arrested cells with an inducible M.EcoGII	236
4.46	<i>In vivo</i> M.EcoGII adenine methylation around nucleosome dyads	238
4.47	<i>In vivo</i> M.EcoGII adenine methylation around transcription start sites . . .	240
4.48	<i>In vivo</i> M.EcoGII adenine methylation around Eaton replication origins . .	242
4.49	<i>In vivo</i> M.EcoGII adenine methylation around Berbenetz replication origins	243
4.50	Cumulative <i>in vivo</i> adenine methylation around three replication origins . .	245
4.51	Cumulative <i>in vivo</i> adenine methylation around three replication origins . .	246
4.52	Cumulative <i>in vivo</i> adenine methylation around <i>ARS1513</i>	247
4.53	<i>In vivo</i> adenine methylation around clustered replication origins	248
4.54	M.EcoGII adenine methylation around <i>ARS1513</i>	250
4.55	M.EcoGII adenine methylation around <i>ARS413</i> , <i>ARS1</i> and <i>ARS437</i>	251

List of tables

1	Abbreviations	xx
1.1	Population-based chromatin accessibility methods	17
2.1	Primer list	28
2.2	Enzyme list	29
2.3	Psoralen crosslinking annealing buffer	30
2.4	Phosphorylation of small fragment DNA	34
2.5	Ligation reaction for producing a long fragment with a TMP modification .	35
2.6	PCR reaction composition for amplification of DNA for angelicin treatment	37
2.7	pGEM-T ligation reaction	37
2.8	PCR reaction composition for production of DNA for angelicin treatment .	38
2.9	PCR program for production of DNA for angelicin treatment	39
2.10	<i>S. cerevisiae</i> strains	41
2.11	Yeast growth media	43
2.12	Droplet Digital PCR (ddPCR) program	47
2.13	M.CviPI methylation of genomic DNA	52
2.14	Primary antibodies used in western blots	61
2.15	Secondary antibodies used in western blots	61
2.16	Western blot gel	62

2.17	Mass spectrometry adducts	64
2.18	Bioinformatic software websites	66
3.1	Detection methods for furanocoumarin interstrand crosslinks or monoadducts	75
3.2	Chemical modification target sites in the <i>Saccharomyces cerevisiae</i> genome	82
3.3	Molecular weight of potential substrates analysed by mass spectrometry . .	94
3.4	Nanopore sequencing read counts	114
3.5	Nanopore sequencing k-mer analysis overview	116
4.1	Chromatin accessibility assays involving methylation	139
4.2	Nanopore software for detecting methylation	141
4.3	Genomic DNA regions for ddPCR quantification of methylation	159
4.4	M.EcoGII <i>in vitro</i> nanopore sequencing overview	202
4.5	M.EcoGII <i>in vitro</i> 6-mer methylation across 6mA classification thresholds .	208
4.6	M.EcoGII <i>in vivo</i> nanopore sequencing overview	220
4.7	EcoGII <i>in vivo</i> nanopore sequencing overview of G1 arrest experiment . . .	235

Abbreviations

The abbreviations used in this thesis are summarised in table 1.

5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
6mA	6-methyladenine
8-MOP	8-methoxypsoralen
ACS	ARS consensus sequence
AID-tag	Auxin-inducible degradation (AID) tag
APS	Ammonium persulphate
ARS	Autonomously replicating sequence
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	Adenosine triphosphate
BAH	Bromo-adjacent homology domain of ORC
BLAST	Basic Local Alignment Search Tool
bp	base pairs
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>
Cdc6	Cell division cycle protein 6
Cdc7	Cell division cycle protein 7
Cdc45	Cell division control protein 45
CDK	Cyclin-dependent kinase
Cdt1	Chromatin licensing and DNA replication factor 1 protein
CG or CpG	Cytosine Guanine dinucleotide
-CH ₃	Methylation (1 carbon, 3 hydrogen atoms)
ChIP	Chromatin Immunoprecipitation

CMG	Cdc45, Mcm2-7, GINS complex
<i>D. melanogaster</i>	<i>Drosophila melanogaster</i>
DAMT-1	DNA adenine methyltransferase in <i>C. elegans</i>
Dbf4	DDK kinase regulatory subunit DBF4 / Dumbbell forming protein 4
DDK	Cdc7-Dbf4 kinase
ddPCR	Droplet Digital PCR
DMAD	DNA 6mA demethylase in <i>D. melanogaster</i>
dMRM	Dynamic Multiple Reaction Monitoring
DMSO	Dimethyl sulfoxide
D-NAscent	Detecting Nucleotide Analog Signal Currents on Extremely long Nanopore Traces
DNaseI	Deoxyribonuclease I
DNA	Deoxyribonucleic Acid
Dpb11	DNA replication initiation protein
dsDNA/RNA	double-stranded DNA/RNA
<i>E. coli</i>	<i>Escherichia coli</i>
ECL	Enhanced Chemiluminescent detection
EM-seq	Enzymatic Methyl-seq
ESI	Electrospray ionisation
ESI-AJS	Electrospray ionisation with Agilent Jet Stream
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements using sequencing
FORK-seq	Nanopore sequencing method to map replication of single DNA molecules
GATC	Guanine, Adenine, Thymine, Cytosine sequence context
gDNA	genomic DNA
GINS	Sld5, Psf1, Psf2, Psf3 complex
-HF	High Fidelity enzyme
HCl	Hydrochloric Acid
HPLC	High Performance Liquid Chromatography
HPLC-QQQ	HPLC combined with a triple quadrupole mass spectrometer
IHF	Integration Host Factor
IPD	Inter-pulse duration
KAc	Potassium Acetate plate for sporulation of <i>S. cerevisiae</i>
kb	kilobases
M	Molar
MBL	Company that sells antibodies

MCM	Mini-Chromosome Maintenance helicase complex
MMR	Mismatch repair system
MNase	Micrococcal nuclease
N6AMT1	N-6 adenine-specific DNA methyltransferase 1 in <i>H. sapiens</i>
NEB	New England Biolabs
NMAD-1	Adenine DNA demethylase in <i>C. elegans</i>
OD ₆₀₀	Optical density at 600 nm
ORC	Origin Recognition Complex
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PAGE	Polyacrylamide Gel Electrophoresis
PBST	Phosphate Buffered Saline with Tween
PCR	Polymerase Chain Reaction
PEG	Polyethylene glycol
PMSF	Phenylmethylsulfonyl Fluoride (serine protease inhibitor)
PNK	Polynucleotide kinase
Pol	Polymerase
pre-RC	pre-Replication Complex
pre-IC	pre-Initiation Complex
Psf1/2/3	Partner of Sld5 1/2/3
PUVA	Psoralen ultraviolet A
PVDF	Polyvinylidene fluoride
rDNA	Ribosomal DNA
RM system	Restriction modification system
RNA	Ribonucleic Acid
RNAi	RNA interference
<i>S. cerevisiae</i>	<i>Saccharomyces cerevisiae</i>
<i>S. pombe</i>	<i>Schizosaccharomyces pombe</i>
SAM	S-adenosylmethionine
SDS	Sodium Dodecyl Sulfate
Sld2/3	DNA replication initiation proteins
Sld5	Component of GINS complex
-seq	-sequencing (often added to another technique)
SMAC-seq	Single Molecule Accessible Chromatin sequencing

SMRT-sequencing	Single Molecule Real Time sequencing (performed by PacBio)
ssDNA/RNA	single-stranded DNA/RNA
TAPS	TET-assisted pyridine borane sequencing
TBE	Tris/Borate/EDTA buffer
TE	Tris-EDTA buffer pH 8.0
TEMED	Tetramethylethylenediamine
TET	Ten-Eleven Translocation cytosine demethylase
TMP	4,5',8-trimethylpsoralen
TSS	Transcription Start Site
UHPLC	A version of HPLC using a smaller column and smaller particles
UV	Ultraviolet light
<i>X. laevis</i>	<i>Xenopus laevis</i>
YPA	Yeast-Peptone-Adenine yeast media
YPAD	Yeast-Peptone-Adenine-Dextrose yeast media

Table 1 List of abbreviations

Abstract

Origin licensing initiates DNA replication by loading proteins on to DNA. In eukaryotes origin licensing varies among inactive replication origins and can be analysed by looking at chromatin structure. This has predominantly been studied using population-based approaches including MNase-seq. These mask the heterogeneity in interactions between DNA-binding events that occur on the same DNA molecule.

In this thesis the development of a single molecule genomic footprinting approach will be discussed, with a focus on identifying an appropriate marker of accessible DNA. Firstly, two chemicals, 4,5',8-trimethylpsoralen (TMP) and angelicin, were assessed using polyacrylamide denaturing gels and mass spectrometry, which showed that TMP could modify DNA while angelicin modification of DNA was not observed. Subsequently, nanopore sequencing did not detect the TMP modification. Secondly, DNA methylation was investigated which involved validating the sequence specificity of M.EcoGII, a non-specific adenine methyltransferase, and developing a Droplet Digital PCR assay to quantitatively assess methylation levels at individual loci. Finally, methylation was used in combination with nanopore sequencing to analyse chromatin structure in *Saccharomyces cerevisiae* through both *in vitro* chromatin preparations and novel *in vivo* inducible Dam and M.EcoGII systems. These showed that *S. cerevisiae* tolerates expression of an adenine methyltransferase, methylation in a variety of sequence contexts is detectable by nanopore sequencing and some chromatin structure can be observed however further optimisation is required.

Chapter 1

Introduction

1.1 DNA replication

1.1.1 Overview

DNA replication is an essential process in the cell cycle that allows a second copy of the DNA to be produced and subsequently passed to a daughter cell. Defects in DNA replication can lead to hallmarks of cancer through replication stress [1, 2] and genomic instability [3]. Replication starts in specific regions of the genome termed origins. There are two main stages to DNA replication: origin licensing and origin firing. These correspond to preparing the cell for replication, and performing replication respectively. Origin licensing occurs in G1 when there is low CDK (cyclin-dependent kinase) activity, while origin firing occurs during S phase when there is high CDK activity.

Whilst replication proteins are highly conserved among eukaryotes, including the budding yeast *S. cerevisiae*, the way in which origins are specified varies. Bacteria with comparatively small, circular genomes require one origin. Conversely, larger eukaryote genomes have

multiple linear chromosomes therefore require multiple origins for efficient DNA replication. These origins are unevenly spaced due to factors including chromatin structure. Additionally, origin efficiency is varied: many are not active in every cell, and do not fire simultaneously during S phase.

1.1.2 DNA replication initiation in *Saccharomyces cerevisiae*

Origin selection

In *Saccharomyces cerevisiae* origins are marked by ORC (Origin Recognition Complex) binding at specific 11-17 bp sequence motifs ((T/A)TTTAT(A/G)TT-T(T/A)) known as ACS (ARS (autonomously replicating sequence) consensus sequence) sites [4, 5]. There are over 10,000 occurrences of the ACS motif in the *S. cerevisiae* genome [6]. However, only 228 sites have been experimentally shown to act as DNA replication origins [5]. A position-weight matrix from 253 ORC binding sites in *S. cerevisiae* highlights the AT-rich composition of the ACS, with many highly conserved thymine bases, and one highly conserved adenine at position 7 [7]. It is likely that factors - including the surrounding chromatin structure - affect the ability of a predicted ACS site to be a functional origin. This is supported by context-based computational analysis which compared characteristics of ACS sites that bind ORC (ORC-ACS) with those that do not. ORC-ACS sites contained an A-rich motif within a nucleosome-free region which facilitated replication protein binding/assembly and were more likely to be found far from gene start sites [8]. DNA helical twist also differed at ORC-ACS sites [8]. Computational model accuracy can be improved by using larger sequences. Oriscan used 268 bp of DNA for each origin (including the T-rich ACS and a large 3' A-rich region), with 94 of the top 100 predictions matching known ACS sites with origin activity, or replication protein binding sites [6]. Therefore, in *S. cerevisiae* DNA replication origins can be reasonably reliably identified by sequence alone.

Origin licensing

Origin licensing consists of loading the pre-replication complex (pre-RC) on to the DNA (figure 1.1). The eukaryotic pre-RC consists of the Origin Recognition Complex (ORC), Cell division cycle 6 (Cdc6), Mini-Chromosome Maintenance (MCM) helicase and Chromatin licensing and DNA replication factor 1 (Cdt1). In *S. cerevisiae*, ORC is the first protein complex to bind and determines the location of the replication origin [9]. While the ACS is the main sequence requirement for ORC binding, three B elements also play an important role in the recruitment of pre-RC proteins [10].

ORC is a hexameric complex which encircles the DNA, with five of the proteins involved in binding DNA in the order Orc1, Orc4, Orc5, Orc2, Orc3, while Orc6 binds to Orc2 and is not part of the ring [11]. Orc2 also acts as a transcriptional silencer [12] but this thesis will discuss ORC and its role in DNA replication.

Despite the binding of ORC to DNA being the first step of DNA replication, in *S. cerevisiae* ORC is bound to DNA throughout most of the cell cycle, not just in G1 [13]. It was observed that the pre-replicative chromatin state of an origin, which is found from the end of mitosis and persists throughout G1 but disappears when the cell enters S phase, was characterised by an increase in the region of protected DNA bound by ORC [14]. This was caused by Cdc6 binding to Orc1, altering the orientation of Orc1 within the Orc1-6 complex [11]. Therefore, the binding of Cdc6 to Orc1 is a marker for the initiation of origin licensing. This is supported by single-molecule analysis which shows that Cdc6 drives pre-RC formation [15].

Instead of binding DNA, Orc6, in a new orientation after Cdc6 binding, is responsible for binding Cdt1 which allows Mcm2-7 helicase loading on to the DNA [16]. Cdt1 is required for Mcm2-7 recruitment as it overrides an autoinhibitory domain in Mcm6 [17]. Activation

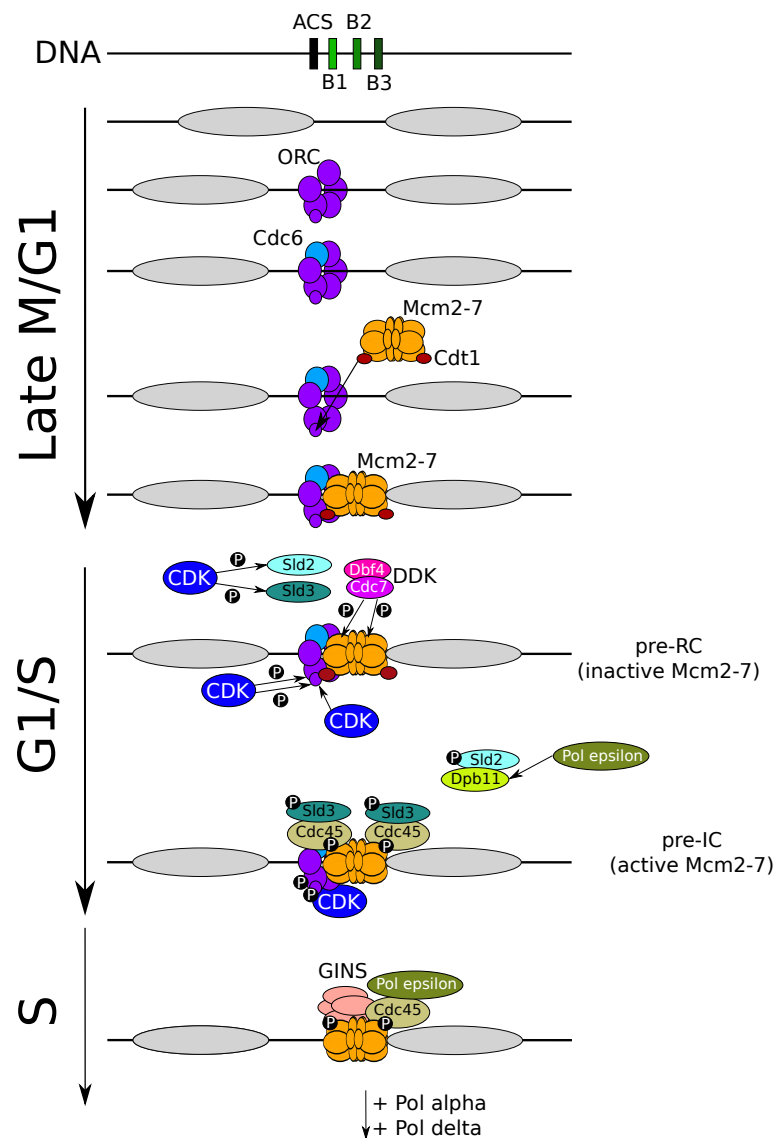


Fig. 1.1 Initial steps of DNA replication in *S. cerevisiae*

The DNA replication origin consists of four genetic elements (ACS, B1, B2 and B3). Across eukaryotes the initial steps of DNA replication are reasonably conserved, thus *S. cerevisiae* is used as an example organism. However, the genetic elements are specific to *S. cerevisiae* and some protein names may differ between organisms. In late M/G1 phase, the replication initiation proteins load on to DNA sequentially to form the pre-replication complex (pre-RC) which contains an inactive Mcm2-7 helicase. In the G1/S phase transition CDK levels increase, leading to phosphorylation (shown by the black circles with a white 'P') of various proteins performed by both CDK and DDK (a heterodimer of Dbf4 and Cdc7). This forms the pre-initiation complex (pre-IC) in which the Mcm2-7 helicase is now active. In S phase, the initiation factors leave and DNA polymerases are recruited. Protein names are shown for *S. cerevisiae* and protein sizes are not to scale. Black lines = double-stranded DNA; grey = nucleosomes. Other proteins are labelled with their names, Pol = Polymerase; GINS = a complex of Sld5, Psf1, Psf2 and Psf3. Cell cycle stage labelled on the left hand side of the arrows. Replication fork progression, DNA unwinding and production of replicated strands not shown.

of Cdc6, ORC or MCM ATP hydrolysis aids Mcm2-7 loading and ring closure after which Cdt1 dissociates [17–19].

Two MCM hexamers are required per origin as each associates with one replication fork moving away from the origin. The Orc6-Cdt1-Mcm2-7 complex is only sustained for the loading of one MCM, therefore the Orc6-Cdt1 interaction must be dynamic for multiple Mcm2-7 complexes to be loaded [16]. Head-to-head Mcm2-7 double hexamers may be formed prior to loading [20] or rapid binding of different Cdc6 and Cdt1 molecules may recruit the hexamers sequentially with the first loaded MCM aiding the binding of the second MCM [21]. Therefore, it is not clear exactly when the Mcm2-7 double hexamer forms or how it is recruited, but ORC, Cdc6 and Cdt1 are all involved in coordinating this process. The pre-RC is complete upon loading of ORC, Cdc6 and MCM on the DNA (figure 1.1), which is inactive during G1 of the cell cycle when CDK levels are low.

Origin firing

The next stage is formation of the pre-initiation complex (pre-IC), which consists of the CMG complex (Cdc45, Mcm2-7, GINS (Sld5, Psf1, Psf2, Psf3)) and travels with the replication fork [22]. When the cell cycle progresses to S phase, CDK levels rise and the pre-RC is activated by two protein kinases, CDK and DDK (Dbf4-dependent protein kinase) (see figure 1.1). CDK phosphorylates Sld2 and Sld3 [23, 24], leading to the binding of phosphorylated Sld2 to Dpb11 [25, 26], and the association of phosphorylated Sld3 with Cdc45 [27], generating a Dpb11-binding site [23, 24]. Dpb11 is required for the subsequent association of DNA polymerase α and ϵ with the replication machinery at origins [28]. DNA polymerase α is responsible for initiating DNA replication by acting as a primase [29, 30] while polymerase δ and ϵ can replicate longer regions of DNA. Polymerase δ requires interactions with proliferating cell nuclear antigen (PCNA) to continue DNA replication over long regions [31] while polymerase ϵ does not [32]. DNA polymerase ϵ is predominantly

associated with leading strand synthesis [33] while DNA polymerase δ is involved with both leading and lagging strand synthesis [34]. The Sld3-Cdc45 complex is particularly important for maintaining the interaction between Cdc45 and Mcm2-7 [27]. Activation of the Mcm2-7 helicase exposes single-stranded DNA which disrupts the interaction between Sld3 and Mcm2-7 while promoting interactions between GINS and Mcm2-7, promoting the formation of the CMG complex [22]. At the start of S phase DDK phosphorylates Mcm4, which stabilises the Cdc45-Mcm2-7 complex [35]. CDK inhibits Orc6-Cdt1 interactions by phosphorylation, preventing additional Mcm2-7 loading during S phase, which in turn prevents over-replication errors [17, 36]. However, it has been suggested that the loading of multiple MCM's on to origins may be associated with earlier firing origins [37]. Furthermore, *in vitro* studies show that RNA polymerase and other DNA translocases can push inactive MCM double hexamers along the DNA, allowing replication to initiate away from origins [38]. Once the replication forks have formed, DNA replication occurs bidirectionally, producing a continuous DNA fragment in the 5' to 3' direction and short Okazaki fragments [39] in the 3' to 5' direction which are subsequently ligated [40].

1.1.3 DNA replication initiation among other yeast

Some yeast, for example *Schizosaccharomyces pombe*, have origins in AT-rich nucleosome-free regions that lack a defined ACS sequence motif [41]. In *S. pombe* 50% of intergenic regions can act as origins, provided they have a high AT-content with runs of A's or T's and sufficient total length to allow the replication proteins to bind [42]. *Kluyveromyces lactis*, a yeast species predominantly known for its ability to convert lactose into lactic acid, has a similar genome size to *S. cerevisiae* but half the number of DNA replication origins and an ARS of 50 bp compared to the *S. cerevisiae* 200 bp ARS [43]. Both species contain ARS's predominantly in non-syntenic intergenic regions, which are less conserved during evolution [43]. Therefore, even if the exact conserved sequence is not present, there are numerous

similarities between the characteristics of origins; they are in nucleosome-depleted AT-rich intergenic regions.

1.1.4 DNA replication initiation in bacteria and archaea

In bacteria replication of the circular genome starts at a single origin (OriC) [44] with replication forks moving in a bi-directional manner [45]. Three tandem repeats of an AT-rich 13 bp sequence are required for DnaA binding, which subsequently unwinds the DNA [46]. DnaC recruits DnaB helicases to stabilise the resulting open DNA structure [46].

Archaea have a circular genome yet DNA replication appears to have more in common with eukaryotes than bacteria, including the replication proteins involved. Archaea have one gene encoding a protein with homology to eukaryotic Orc1 and Cdc6, and most archaea have one to three Orc1/Cdc6 orthologs [47]. In *Pyrococcus abyssi* Cdc6 and MCM homologs bind to the OriC region [48]. A potential homolog of eukaryotic Cdt1 has also been identified in archaea [49], meaning that all four of the pre-replication complex components are at least partly represented in archaea.

Unlike bacteria, archaea can have multiple replication origins, with many species having between two and four origins [49–51]. Replication forks in two of these species were identified as moving at similar rates to each other, an order of magnitude slower than bacterial fork speeds, and closer to eukaryotic replication fork speeds [50]. Not all archaea have multiple origins, for example *P. abyssi* only has a single origin [52]. The diversity in archaea origin number suggests that the existence of multiple origins may have first arisen in archaea as eukaryotes are descended from archaea [53].

1.1.5 DNA replication initiation among other eukaryotes

Xenopus laevis has homologues of both ORC and MCM protein complexes, with particular similarities of the ORC to yeast ORC [54]. Both ORC and MCM are essential and non-redundant and they bind in a sequential manner (ORC followed by MCM) [55]. In *S. cerevisiae* ORC is bound to DNA throughout the entire cell cycle while in *X. laevis* ORC is displaced from the chromatin and located in the cytoplasm during mitosis [55]. In *X. laevis* the minimum length of DNA required to recruit both ORC and MCM was 80 bp, which produced a ratio of 1 ORC to 1 MCM on the DNA [56]. An increase in the length of DNA substantially increased the number of MCM complexes on the DNA, while the quantity of ORC remained the same [56]. Therefore, additional MCM loading occurred away from the ORC binding site. Only a few of the MCM's were subsequently bound by Cdc45, leading to replication, however all the MCM complexes were phosphorylated and capable of being induced to accept Cdc45 loading [56].

In *Drosophila melanogaster* DNA replication origins are marked by DNA shape and chromatin configurations rather than specific DNA sequences [57, 58]. *D. melanogaster* ORC has a 30-fold higher affinity for negatively supercoiled DNA than relaxed or linear DNA, while there was a maximal 6-fold difference in affinity between ORC binding to origin DNA and random DNA fragments [57]. Transcription can also affect origin activity, as cell-type variations in origin activity have been linked to cell-type-specific transcriptional programs [58]. Additionally, origin-proximal G-quadruplexes can transiently stall replication forks *in vivo* [58]. Regarding the proteins involved, isolation of the *D. melanogaster* CMG complex confirmed ATP-dependent DNA helicase activity [59] as found in *S. cerevisiae*. RNAi of Cdc45 and GINS components in *D. melanogaster* confirmed these proteins are required for the transition from G1 into S phase [59] showing conservation from *S. cerevisiae*.

In human cells, DNA replication is not initiated at a specific sequence. Instead, replication initiates from broader AT-rich, non-transcribed regions termed initiation zones and can be up to 150 kb [60, 61]. These initiation zones contain multiple redundant sites for protein binding [60]. Interestingly, there are two origins in *S. cerevisiae*, *ARS101* and *ARS310* which have multiple ORC binding sites, similar to metazoan origins [60]. Metazoan initiation zones are normally bordered by transcribed genes and enriched in open chromatin marks, enhancer marks and ORC binding sites [61]. Initiation zones which replicate at a similar time to each other may also co-localise in 3D space during S phase [62]. The same key proteins (ORC, Cdc6 and MCM) are involved in origin licensing in human cells [63], with the structures conserved from *S. cerevisiae*. Approximately 52,000 Orc2 binding sites have been identified in the human genome and these were characteristically broader than equivalent profiles for sequence-specific transcription factors [64]. Similar to *Drosophila* ORC, human ORC does not differentiate between origin-containing and random DNA fragments, although it does have a preference for AT-rich DNA [65]. This generality is exemplified by the observation that purified human ORC can fully replicate ORC-depleted *X. laevis* DNA [65]. Human ORC can bind nonspecifically to DNaseI hypersensitive open chromatin sites with chromatin marks including H3 acetylation and H3K4 methylation [64]. In HeLa cells origins were significantly associated with transcription as initiation zones often overlapped transcribed genes [66]. Origins that fire early in S phase were found to have more ORC binding sites suggesting that replication timing may be caused by ORC density combined with stochastic origin firing which was supported by a computational model [64]. In addition, regions which lacked ORC binding sites were associated with common fragile sites and recurrent deletions in cancer [64]. It has been hypothesised that ORC is recruited to binding sites within the initiation zones by activator proteins at enhancers and promoters that remodel the chromatin environment to allow ORC binding [64]. Furthermore, origins colocalising with the

transcription start site (TSS) of coding RNA with moderate/high transcription levels were found to replicate early in S phase [67].

Western blotting of chromatin and soluble protein fractions identified that human cells behave similarly to *S. cerevisiae* cells with regards to pre-RC formation; ORC stays bound throughout the human cell cycle while Cdc6 accumulates on chromatin in late mitosis and MCM is added to chromatin in G1 [55, 63]. Permeable HeLa G1- and G2-phase nuclei can replicate in ORC-depleted *Xenopus* egg extract, indicating that ORC is bound to DNA during origin licensing and post-replication in HeLa cells [55]. An excess of MCM proteins has been identified in HeLa cells with cell proliferation being possible several days after RNAi reduction of MCM level by 90%. Over time the MCM reduction caused a hypersensitivity to DNA replication stress, accumulated DNA lesions and activation of a checkpoint response, eventually preventing cells from entering further rounds of mitosis [68]. DNA combing single molecule analysis showed that replication forks progressed as normal, however dormant origins could not be activated, therefore the backup system was removed [68]. Another implication of removing excess MCM proteins was a reduction in homologous recombination and non-homologous end joining, observed when Mcm2 or Mcm3 were targeted by shRNA [69], therefore leading to increased DNA damage. Mcm2 also has a role in cilia formation in both humans and zebrafish [70] therefore an excess of MCM may not be solely linked to its role in DNA replication.

1.1.6 Clinical outcome of mutations in replication initiation proteins

Specific recessive mutations in Orc1, Orc4, Orc6, Cdc6 and Cdt1 have been implicated in the developmental disorder Meier-Gorlin syndrome [71]. Patient-specific Orc4 mutations cause tissue-specific replication defects in *Drosophila melanogaster* [72] and mutations in Orc1 and Cdc6 cause growth defects in zebrafish [73, 74]. Cdc6 mutants can lead to

embryonic lethality, with one mutation causing defective reproduction in males [74]. It is likely that these perturbations to pre-RC formation reduce DNA replication activation at origins, slowing the entire cell cycle as the forks will have to travel further to replicate the entire genome. This would elongate the time DNA replication takes, leading to a reduction in overall organism growth.

1.1.7 Effect of chromatin structure on DNA replication

Chromatin structure can impact licensing, initiation and elongation. *S. cerevisiae* origins are characterised by an ACS in a nucleosome-free region [75], surrounded by well positioned nucleosomes [7, 76]. The chromatin structure could be regulated by DNA sequence [76] or ORC binding which requires an asymmetric nucleosome-depleted region around the ACS (figure 1.2a) [7, 76]. Once bound to the DNA, ORC acts as a physical block, positioning the surrounding nucleosomes, with the help of nucleosome modifiers and remodelers, to create a phasing effect (figure 1.2a). This aids replication by aiding pre-replication complex formation [7, 76, 77]. Berbenetz *et al* performed k-means clustering on 228 *S. cerevisiae* origin DNA sequences and identified four main groups with distinct chromatin structures (figure 1.2b), which were linked to the presence of transcription start and stop sites [76]. For example, in the red and yellow clusters in figure 1.2b there is a second nucleosome-depleted region, shown in dark blue, upstream of the ACS nucleosome-depleted region. These two additional nucleosome-depleted regions correspond with the presence of transcription start sites directly upstream [76] which may perturb the chromatin structure due to the requirement for accessible DNA for transcription factor binding. Meanwhile, in *Drosophila* it has been identified that ORC binding sites are associated with a subset of RNA polymerase II binding sites, indicating that transcription may act locally to influence origin selection. Transcription was also shown to act globally to regulate origin activation as the state of transcription can affect replication timing over regions as large as 100 kb [78].

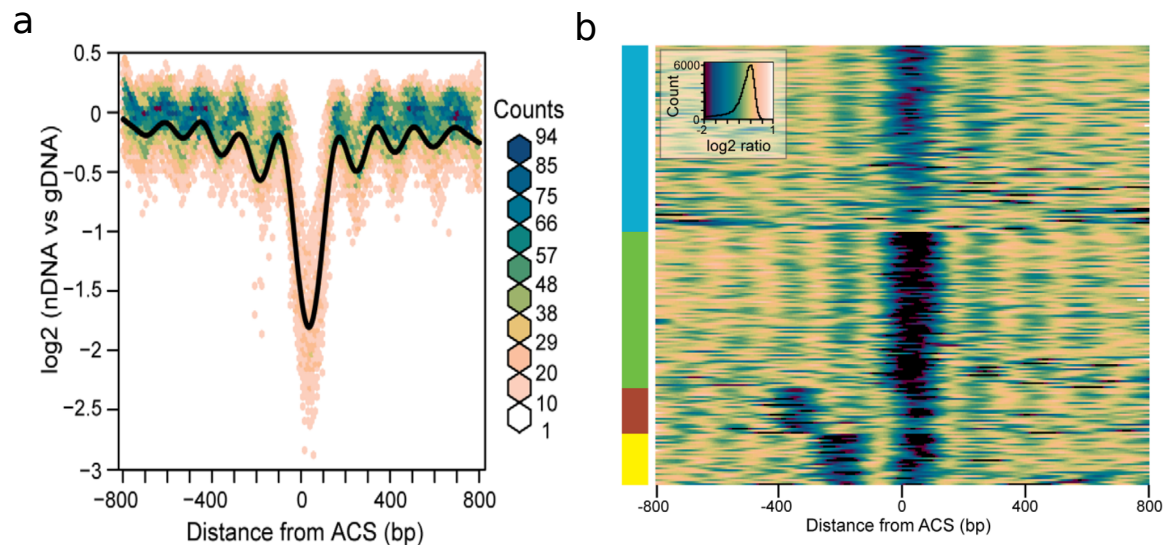


Fig. 1.2 Nucleosome positioning around *S. cerevisiae* replication origins

a) Nucleosome positions from 228 replication origins in *S. cerevisiae* mapped to the sacCer3 (S288C) *S. cerevisiae* reference sequence, aligned by the ACS and oriented by the T-rich strand. The colours are a bivariate histogram, with the colour of each point showing the density of data present. Black line = average; nDNA = nucleosomal DNA; gDNA = genomic (non-nucleosomal) DNA. Figure 1a from Berbenetz *et al* (2010) [76]. **b)** Heatmap of k-means clustering of the origins used in **a** into four groups based on their chromatin structure. Each line represents a strand of DNA encompassing an origin, aligned by the ACS and oriented by the T-rich strand. Colours represent log2 values for data at a specific position, with yellow representing nucleosome bound and green/blue representing accessible DNA. Cluster classification is represented by the colour bar on the y axis. Figure 3a from Berbenetz *et al* (2010) [76]. Both figures come from the following open-access article which is distributed under the terms of the Creative Commons Attribution Licence: Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure (2010) Berbenetz *et al*, PLoS Genet 6(9): e1001092.

The conservation of chromatin structure around replication origins shows that chromatin structure is important for DNA replication. For example, a broad easily unwound region of DNA is essential to the 3' of the ACS to facilitate DNA unwinding in DNA replication [4, 75]. The DNA around replication origins is particularly rigid compared to the surrounding genomic regions, however rigid DNA is also observed at promoters which indicates that this rigidity might aid protein binding [79]. It has also been proposed that chromatin proteins can target ORC to DNA in metazoan cells. For example, HMGA1a (a high mobility group protein) may target ORC to origin sites as they are both found in AT-rich heterochromatin regions and have

been shown to interact [80]. HMGA1a has also been reported to recruit ORC via an RNA-dependent mechanism and it has been shown that RNaseA treatment of chromatin releases a subset of the bound ORC, thus suggesting that RNA may stabilise ORC binding to DNA [81]. ORC may also be recruited to DNA by chromatin and DNA modifications since in HeLa cells it can recognise nucleosomes more effectively in the presence of CpG methylation and repressive histone marks including H3K9me3 and H3K27me3 which mark heterochromatin [82]. ORC has also been found co-localising and sharing cell-cycle dynamics with LRWD1 (Leucine-rich repeat and WD repeat-containing protein 1) in HeLa cells, therefore LRWD1 could play a role in recruiting ORC to chromatin [82, 83]. Another co-localisation of ORC within heterochromatin is with HP1 in *Drosophila* and complexes of ORC and HP1 have been isolated suggesting that ORC may play a larger role in chromatin organisation than just remodelling around replication origins [84]. Furthermore, there was a high overlap between sites of ORC binding and cohesin loading in *Drosophila* and it was suggested that ORC may aid cohesin loading [85]. In addition, the conserved N-terminal bromo-adjacent homology (BAH) domain of Orc1 is required for the stable interaction between ORC and DNA and is particularly important for recruitment of *S. cerevisiae* ORC to origins with a distinct local nucleosome organisation [86]. This indicates that origins may vary in the relative importance of different factors for recruiting replication proteins [86]. The BAH domain of human Orc1 is also important for facilitating the interaction between human ORC and DNA as a mutated version of the BAH domain reduced the ability of Orc1 to promote binding of Orc2 to chromatin at the end of mitosis [87]. A mutated BAH domain in Orc1 did not affect nuclear localisation, association of Orc1 with other ORC subunits or degradation of Orc1 in S-phase [87]. Further refinement identified that it is a particular histone mark (H4K20me2), which is enriched at replication origins, that is recognised by the BAH domain [88]. Furthermore, mutations in the BAH domain associated with a loss of recognition of

H4K20me2 are associated with Meier-Gorlin syndrome [88], highlighting the importance of chromatin structure in recruiting ORC and facilitating DNA replication.

1.2 *S. cerevisiae* as a model system

S. cerevisiae was chosen for the study of DNA replication because it is easily genetically manipulated, can be arrested in different stages of the cell cycle, has experimentally confirmed specific DNA replication origin locations and is well-studied in terms of DNA replication and chromatin structure. There are a variety of publicly available datasets available for *S. cerevisiae*, for example nucleosome positions [7, 89] and ChIP-seq datasets for ORC and MCM [90].

1.3 Population-based chromatin accessibility assays

ChIP [91] and ChIP-seq [92] both provide information on binding sites of a particular protein, either across a selection of regions or more genome-wide. ChIP-seq involves formaldehyde cross-linking proteins to DNA, nuclear isolation and lysis, chromatin isolation, DNA fragmentation, immunoprecipitation and DNA sequencing [93]. The reliance on antibodies renders this method protein-specific and may be subject to variable off-target effects. It does not assess the total chromatin environment and is unlikely to perform well for every protein of interest. Another popular pair of methods for identifying protein binding sites are DamID [94] and DamID-seq [93]. These require the protein of interest to be fused to a Dam adenine methyltransferase which methylates adenines in a GATC sequence context [94, 93]. When the protein binds to its native binding sites on the DNA, the methyltransferase will be brought into proximity of this DNA, thus able to methylate it [93, 94]. The binding sites can then be detected by microarray hybridisation of the methylated DNA [94], or identification of regions of methylation in sequencing data [93]. Again this is a

protein-specific method and this requires quite a lot of prior manipulation of the protein of interest.

A variety of population-based approaches to investigate chromatin accessibility in a less protein-specific manner have been developed over the last couple of decades (table 1.1). Since their initial development, these population-based approaches have been combined with next generation sequencing, to increase throughput and allow visualisation of protein binding events across the entire genome rather than a specific region. However, these methods provide a summary of a population of DNA molecules therefore they cannot be used to identify what events occur simultaneously on the same molecule. Common methods, like MNase-seq and DNaseI-seq, involve digestion of accessible DNA, leaving protein-bound DNA fragments for analysis. Other techniques have evolved to either retain DNA molecules from accessible DNA, as done in ATAC-seq, or allow either subset of DNA molecules to be retained, as in FAIRE-seq. FAIRE-seq, involving formaldehyde crosslinking to maintain the chromatin structure, followed by sonication and DNA extraction provides an advantage over the digestion methods as the effectiveness of the 1% formaldehyde treatment is consistent over a range of incubation times [95]. In addition, FAIRE-seq needs few reagents and is relatively quick [95]. Conversely, for enzyme digestions, the concentration and incubation time may need to be optimised due to commercial variation in products and in nuclear preparations [95]. A major advantage of ATAC-seq is that it can be performed on a relatively small number of cells: 500-50,000 [96] rather than tens of millions [97, 98].

Method	Year developed	Description	Output
MNase-seq [98]	2012	Digest accessible DNA with MNase	DNA fragments mapped to regions of protected (protein-bound) DNA
DNaseI-seq [97]	2010	Digest accessible DNA with DNaseI	DNA fragments mapped to regions of protected (protein-bound) DNA
FAIRE-seq [95]	2007	Formaldehyde crosslink, sonication, phenol-chloroform extraction density separation (protein bound DNA vs naked DNA from open chromatin)	DNA fragments mapped to regions of protected (protein-bound) and accessible (non-protein-bound) DNA (in the two separate fractions)
ATAC-seq [96]	2013	Tn5 transposase to put sequencing adaptors into open chromatin, PCR amplify and sequence	DNA fragments mapped to regions of accessible (non-protein-bound) DNA

Table 1.1 Population-based chromatin accessibility methods

All of these population-based approaches can be combined with next-generation sequencing to provide information on the genomic location of the chromatin structure. The dates and papers cited relate to the ‘-seq’ version of each method. Previously, the methods used either microarrays or southern blots to produce the final readout of chromatin structure. FAIRE = Formaldehyde-Assisted Isolation of Regulatory Elements; ATAC = Assay for Transposase-Accessible Chromatin; -seq = using sequencing.

1.3.1 Population-based investigation of DNA replication

Previous studies have investigated the chromatin structure surrounding DNA replication origins in yeast using population-based methods. ATAC-seq was used to identify that origin licensing is limited by chromatin compaction in cells exiting mitosis, in order to maintain genomic integrity at the mitosis to G1 transition [99]. MNase digestion was used to show that ORC can coordinate the surrounding nucleosomes to produce a chromatin structure that favours pre-RC formation [77]. Belsky *et al* also used MNase to demonstrate that ORC binding can protect DNA from digestion by MNase [90]. They show that ORC binding at

the ACS produced small DNA fragments of less than 100 bp which disappeared when the experiment was performed in an *orc1-161* mutant yeast strain where ORC was unable to bind the DNA. They also showed that there was a change in chromatin structure around the ACS between the wild-type and *orc1-161 S. cerevisiae* strains [90]. Pairing MNase digestion with Illumina sequencing allowed them to observe where ORC bound in a population of cells and to identify that there is an interaction between ORC binding and the surrounding chromatin structure. However, they were unable to identify how ORC was interacting on each DNA molecule as the DNA fragments for each nucleosome cannot be recombined to look at the whole region from one molecule.

1.3.2 Population vs single cell vs single molecule methods

Population-based approaches have provided an overview of the DNA replication process but they mask the variation that occurs within a population of DNA molecules, which can be identified with single cell or single molecule approaches. Single cell analysis can give information on cellular localisation and spatial interactions while single molecule can give insights into DNA and chromatin structure and interactions. Methods should be selected based on the biological question of interest.

Single-cell ATAC-seq (scATAC-seq), developed from ATAC-seq [100], involves using the Tn5 transposase to insert sequencing adapters into accessible regions of the genome [100]. Single cells are then captured and assayed using microfluidics, during which the transposase is released from the DNA and the Tn5 adapters are extended and PCR amplified [100]. This produces short sequencing reads for which it is known that the ends of the reads were in accessible DNA. Single-cell chromatin flow cytometry has also been used to study DNA replication. By measuring DNA content with DAPI, DNA synthesis with EdU incorporation and loaded MCM with anti-Mcm2 antibodies they identified that Mcm2-7 is loaded on to

DNA faster in pluripotent stem cells with a shorter G1 phase than in differentiated cells. As the Mcm2-7 loading slows down, the length of G1 phase extends and differentiation proceeds, suggesting that rapid replication inhibits differentiation [101].

Microscopy-based single molecule methods have been used to investigate DNA replication, including DNA combing which is designed for looking at replication fork movement along single molecules [102, 103] and has been optimised to visualise DNA molecules up to 12 Mb [104]. Another microscopy-based single molecule method is DNA curtains which provides a high throughput visualisation of DNA-protein interactions in aligned single molecules [105]. DNA curtains have been applied to DNA replication to identify the recruitment of ORC to origin DNA and subsequent formation of the pre-RC, with the identification that Mcm2-7 double hexamers form preferentially at native origin sequences [15]. Single-molecule Förster resonance energy transfer (FRET) has been used to determine the arrival and departure orders of proteins during pre-RC formation in addition to structural changes of the proteins during this process [19, 21].

Single molecule methods can also be sequencing based, for example D-NAscnt which uses BrdU incorporation at replication origins to measure replication fork movement [106].

1.4 Nanopore sequencing

Nanopore sequencing is a technology produced by Oxford Nanopore Technologies. The main advantage of nanopore sequencing is its ability to produce long-read single molecule data without amplification, which facilitates the detection of modified bases. PacBio SMRT-sequencing is the only other technology that can also be used to produce information on modified bases in long-read single molecules. Nanopore sequencing was selected for this

thesis due to timely examples of detection of the modified base of interest [107, 108] which will be discussed in Chapter 4.

Nanopore sequencing primarily involves a single-strand of DNA passing through a pore which sits in a membrane. As the DNA passes through the pore it causes a change in the electrical signal based on the DNA sequence. During library preparation an adaptor is ligated on to a double-stranded DNA molecule (dsDNA). This adaptor includes a motor protein (a helicase) that brings the dsDNA molecule to the pore and separates the two DNA strands, feeding one into the pore, while the other is released. Figure 1.3 shows an example of one pore, though in a flow cell there are 512 channels, each containing four pores. One pore can be active in each channel at any point in time. Therefore, lots of DNA molecules will be sequenced at the same time on a single nanopore flow cell. A voltage is applied to the membrane, creating a potential difference across the membrane. The negatively charged DNA molecules are attracted to the positively charged side, therefore, with the help of a motor protein, pass through a pore to get to the other side. While in the pore the DNA disturbs the flow of electrons going in the same direction. Each DNA base has a unique structure and size relative to the other DNA bases, therefore these each produce different blockages for the electrons and a different electrical signal is produced. However, the DNA is a continuous molecule, therefore there are multiple bases within the narrowest part of the pore at the same time. This means that the electrical signal is disturbed by a group of bases (a k-mer) rather than a single base. The process of converting the electrical signal to a DNA sequence is called basecalling. Oxford Nanopore Technologies selected a 6-mer for their basecalling model as this captures the majority of the information without adding excess computations. As the DNA passes through the pore, each base will be included in six 6-mers, as each position in the DNA can be the start of a new 6-mer. Therefore, there are multiple chances for gaining information on each base which increases the accuracy of each basecall.

DNA base modifications affect the structure of the base therefore they also affect the electrical signal, facilitating the detection of modified bases in addition to the canonical bases. Many DNA modifications have been detected by nanopore sequencing including BrdU [106, 109, 110], methylation [107, 108] and a range of other base analogues including covalently attached oligonucleotides [110]. Nanopore sequencing of modifications relevant to this thesis will be discussed in the results chapters. Nanopore sequencing can also be used to sequence RNA, either as cDNA or native RNA, although different library preparation protocols are needed for the latter. When sequencing DNA using nanopore technology there is a lot of information encoded in the electrical signal and a major challenge in the field has been developing ways to decode this signal into the canonical and modified bases.

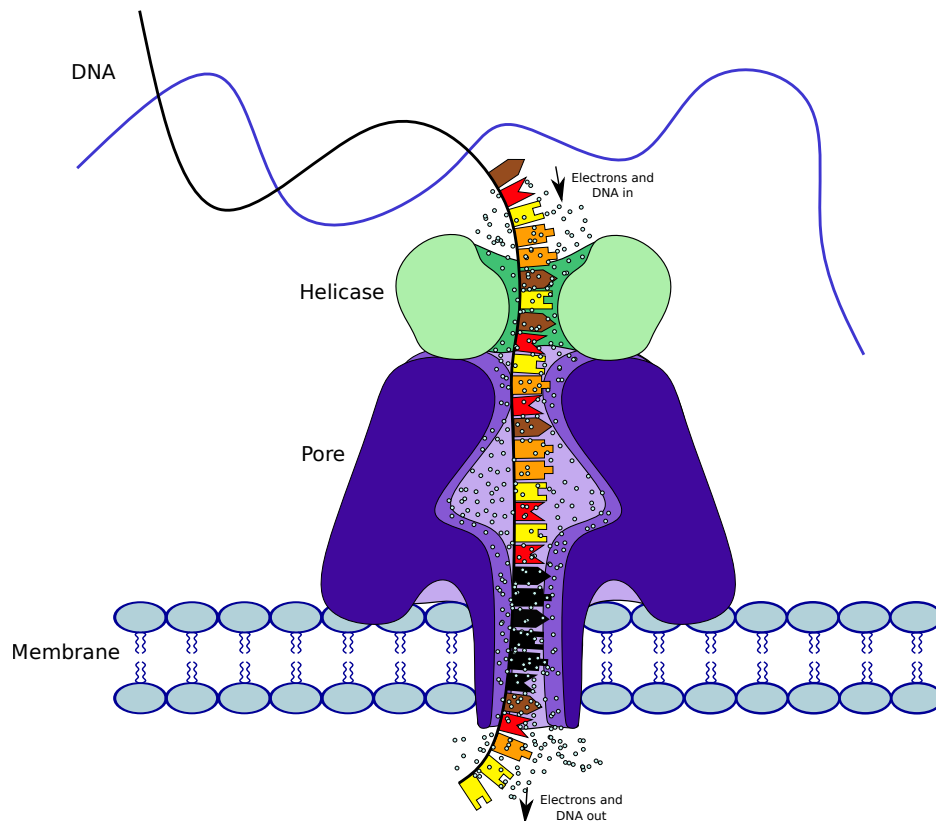


Fig. 1.3 Nanopore sequencing concept

Nanopore sequencing involves a double-stranded molecule of DNA being recruited to a pore in a membrane, the strands separated by a helicase and one strand passed through the pore where it alters the electrical signal by distorting the passage of electrons through the pore. Dark blue and black lines = DNA strands; brown, red, yellow and orange shapes = DNA bases; black shapes = six DNA bases in the narrowest part of the pore which will be responsible for the resulting electrical signal at this point; green (all shades) = helicase to separate the two DNA strands; purple (all shades) = pore to pass the DNA and electrons through the membrane; light blue dots = electrons; blue ovals and wavy lines = membrane. Bases further up the DNA strands, away from the pore, are not shown.

1.5 Aims

The aim of this thesis is to develop a single molecule chromatin accessibility assay that can be applied to DNA replication. This will allow the process of origin licensing to be probed in single molecules, in order to investigate the heterogeneity within a population of *S. cerevisiae* DNA molecules. Single molecule rather than single cell was selected as origins can be compared across molecules without needing to know which DNA molecules

came from the same cell, as each origin is only present once in each cell. The concept of this single molecule chromatin accessibility assay is shown in figure 1.4 and works on the principle that proteins binding to DNA can protect the DNA from modification. This creates footprints where the proteins were bound, allowing protein binding sites in a single molecule to be identified in the DNA after the proteins have been removed. The first time genomic *in vivo* footprinting was applied to DNA replication in *S. cerevisiae* was in 1992 by Diffley and Cocker who identified the presence of ARS-binding factor 1 (Abf1) and ORC at the *ARS1* origin by the absence of DNaseI digests at these binding positions [13]. This thesis will attempt to develop a method using the same principles of protection, but to apply it across the genome, at a higher resolution and in single molecules.

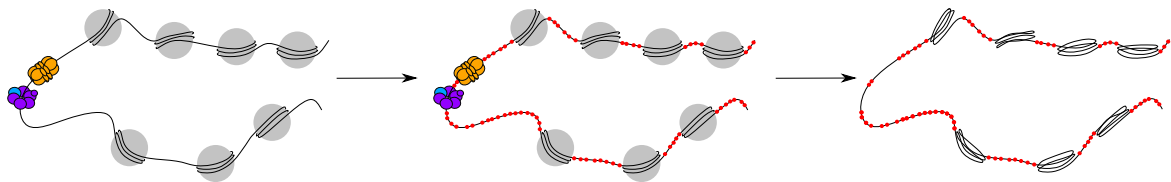


Fig. 1.4 Single molecule chromatin accessibility assay

Schematic of the concept of a single molecule chromatin accessibility assay that will be discussed in this thesis. It consists of three steps; obtain chromatin, modify accessible DNA, extract DNA. Protein binding sites can subsequently be identified by the footprints (regions of no modification) which are created by protein-binding events protecting the DNA from modification. Grey = nucleosomes; black = DNA; red = modification of accessible DNA; orange, purple and blue = DNA replication proteins.

To view the chromatin structure on both sides of a DNA replication origin, nanopore sequencing has been selected as the final readout due to the ability to generate long reads while detecting base modifications on native DNA. The main challenge for developing this assay is selection of an appropriate modification. The ideal modification would not be present in *S. cerevisiae* DNA prior to the assay, have a high frequency of target sites in the DNA and be detectable by nanopore sequencing. Furthermore, the detection of the modification needs to have a high true positive rate and a low false positive rate, as some protein footprints may be relatively small. During pre-RC formation, ORC binds around 48 bp of DNA, which is enlarged to 76-82 bp when Cdc6 binds [111]. This is larger than would be expected based

on the protein structures, thus suggesting that the ORC-Cdc6 complex is bending DNA to encompass a larger region [111]. Mcm2-7 has been shown to produce a footprint of 68 bp [20] however this binding event is likely to be close to a nucleosome therefore there may not be a linker present between Mcm2-7 and the nucleosome, or if there is it may be too small to be accessible for modification. The changes in these footprints during pre-RC assembly can be seen in figure 1.5. The effect of the size of the protein footprint on the requirement for good detection also needs to be balanced with the size of the surrounding accessible regions and the frequency of the modification target site as the less frequent the target site, the more accurate the detection would need to be. This thesis explores options for modifying DNA from *S. cerevisiae* with detection by nanopore sequencing, with the intention of using this modification as a marker of accessible DNA in single molecules.

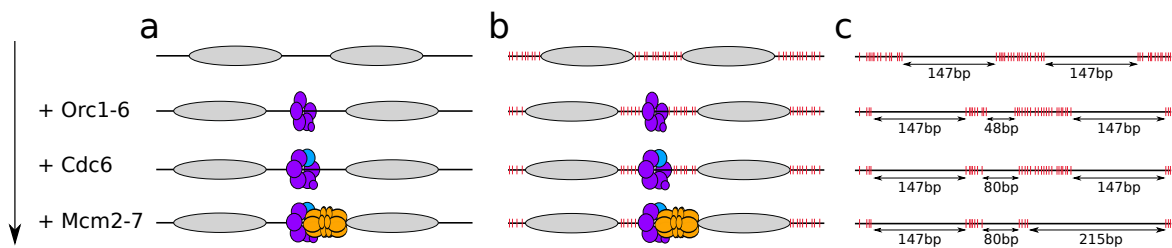


Fig. 1.5 Single molecule chromatin accessibility assay applied to DNA replication origin licensing
The initial steps of origin licensing involve the binding of Orc1-6, Cdc6 and Mcm2-7. Using the same steps as in 1.4 a unique pattern is produced after the binding of each protein, allowing detection of which protein loading step a DNA molecule has reached in the process of origin licensing. **a)** Replication proteins binding to DNA in the normal process of origin licensing. **b)** Chromatin with accessible DNA modified at specific bases, depending on the modifying agent. **c)** DNA extracted from chromatin with accessible DNA modified (with red bars) and previously protein-bound DNA unmodified. Footprints of unmodified DNA represent protein-binding events, of which the sizes are labelled, for example footprints of 147 bp represent nucleosome binding sites. Previously bound proteins can be identified using the location and size of each footprint. This requires a high true positive and low false positive rate for detection of the modified sites. Footprint size of ORC (48-49 bp) and ORC-Cdc6 (76-82 bp) are from Speck *et al* (2005) [111] and the footprint size of Mcm2-7 (68 bp) is from Remus *et al* (2009) [20]. Black = DNA; grey = nucleosomes; purple = ORC; blue = Cdc6; orange = MCM; red = modification of accessible DNA.

Chapter 2

Materials & Methods

The methods have predominantly been divided into psoralen and methylation sections (linking to chapters 3 and 4 respectively) to facilitate easier reproduction of experiments. Abbreviations are located at the start of this thesis.

2.1 General

Centrifugation steps were performed at either room temperature or 4 °C, with a Eppendorf 5424R centrifuge with a FA-45-24-11 rotor for eppendorf tubes or a ThermoScientific Heraeus Multifuge X3R centrifuge with TX-750 rotor for falcon tubes. All restriction enzymes, methyltransferases, cofactors and associated buffers have been purchased from New England Biolabs (NEB) except for BioTaq polymerase (see table 2.2). Unless otherwise stated, reactions were performed in line with NEB recommendations. Standard lab chemicals were purchased from Sigma, Thermo Fisher Scientific or VWR. Throughout these methods, unless specified otherwise, water refers to double distilled water.

2.2 DNA quantification

Double-stranded DNA quantification was performed with a DeNovix DS-11 FX spectrophotometer and fluorometer using the double-stranded DNA High-Sensitivity kit (Invitrogen #Q32854). A master mix was made for the two standards (0 and 10 ng/ μ l) and the samples. This consisted of 199 μ l HS buffer and 1 μ l Qubit dye per sample or standard. For the standards, 190 μ l master mix was used with 10 μ l DNA. For the samples, 199 μ l master mix was used with 1 μ l DNA. These were all stored in darkness for five minutes prior to measurements. Each sample was measured four times, with rotation of the tube, and an average calculated.

2.3 Primers

The primers in table 2.1 were used in this study. Primers with 'HJ' in the name were designed in this study. All primers were diluted to 100 μ M in water.

Amplified region	5' primer	3' primer
Annealed primers for psoralen modification	HJ001 TCGAGGAGAGTAGGACCG	HJ002 GATCCGGTCCTACTCTCC
gDNA psoralen large fragment	HJ003 ACTAGTTGTCTCGGCATCTTCC	HJ004 AATAACTCTGTGCCCCTCGGTC
Sanger sequencing psoralen ligation product	HJ011 CTGAATCTCAGTATTCTTTACT	HJ012 AGCACAGTAAATAATGCGTAGT
PCR and Sanger sequencing for angelicin treatment	HJ014 TTTGAAGGGGGTGTCTGCC	HJ017 AACAGCCGCTTCCCTGTT
ddPCR pUC19 PstI-HF, SphI-HF, EcoRI-HF, BamHI, AluI and SmaI sites	HJ019 GCGGATAACAATTTACACAGG	HJ020 AGGCGATTAAAGTTGGGTAACG
ddPCR pUC19 PstI-HF, SphI-HF, EcoRI-HF, BamHI, AluI and SmaI sites	HJ021 TCACACAGGAAACAGCTATGA	HJ022 AAACGACGGCCAGTGAAT
gDNA PCR on methylated DNA	CA1029 CAGGAAACAGCTATGAC	CA1028 GTAAAACGACGGCCAG
ddPCR pUC19 no RE sites	DB79 GGATGGAGGCGGATAAAGTTG	DB78 AGATAACTACGATACGGGAGGG
ddPCR pUC19 and pG5E4 ScaI-HF and HpaII site	DB31 ATCCTTGAGAGTTTTTCGCCC	DB30 GATGCTTTTCTGTGACTGGTG
ddPCR pUC19 and pG5E4 AluI site	DB49 GAGCGCAGCGAGTCAGT	DB515 ATTAATTGCGTTGCGCTCAC
ddPCR pG5E4 AluI site in 5S rDNA	HJ075 AGGGATTATAAGCCGATGACGT	HJ076 TTCGGTATTTCCAGGCGG
ddPCR pG5E4 BamHI and PstI-HF sites	HJ077 CCCGAATCTTTAACTCGAGTGC	HJ078 AGGGCCAAGTGCAGAGC
ddPCR pG5E4 AluI site	HJ079 GCTCTGCACTTGGCCCT	HJ080 ACAGCGGCAGCCATAACA
ddPCR pG5E4 AluI site	HJ081 TGTTATGGCTGCCGCTGT	HJ082 GAATTCCTCGGATCCAGCG
pG5E4 no RE sites	CN2309 GTGGCACTTTTCGGGGAAATG	CN2301 AAAGGGAATAAGGCGACAC
ddPCR gDNA ScaI site	DB85 TTTAGCTGGGTTTATGGGAGG	DB86 TGAATTATACTGCCGCACG
ddPCR gDNA no RE sites	DB9 TGTGAAATGCAGTCAAAAGGTAC	DB10 GCTTTAAGGCTTTTATTTGGGATG

Amplified region	5' primer	3' primer
ddPCR gDNA PstI site in well-positioned nucleosome region	HJ037 GTAGAATGCCCAATGTCCACTA	HJ038 CTGCCAGAATTGCAGGATTAAA
ddPCR gDNA PstI site in well-positioned nucleosome region	HJ039 GGTTCTATTGTTGGCTCCACT	HJ040 CTGTCAAGCTGGGACCTTTC
ddPCR gDNA PstI site in well-positioned nucleosome region	HJ041 CTTTGCAGCAGCCTGAATATC	HJ042 CTGCTACAAGATCGATAACAAGATT
ddPCR gDNA PstI site in nucleosome-free region	HJ053 CTTTGACAGCTTAAGAGCAAGC	HJ054 TTTCACGCTGTATACCATGATTC
ddPCR gDNA PstI site in nucleosome-free region	HJ055 AATGGCGCACAACAACCTATTAC	HJ056 TCGGATGGAGCGAGTATGA
ddPCR gDNA PstI site in nucleosome-free region	HJ057 TCCCGTAAATTTGGATGTTGAAAG	HJ058 CACTGTTTGTAAGTGAAGTAACTGCT
gDNA Ura3 locus M.EcoGII insertion 5' region	CA033 (in gDNA) GCACATCAATGCATCAATCC	DB508 (in <i>M.EcoGII</i>) TACGATTAAGTCTTGAAGTTCGC
gDNA Ura3 locus M.EcoGII insertion 5' region	CA033 (in gDNA) GCACATCAATGCATCAATCC	DB559 (in <i>M.EcoGII</i>) CGTCACCTTTCCACTGGTTATC
gDNA Ura3 locus M.EcoGII insertion 3' region	DB526 (in <i>M.EcoGII</i>) GCTCTGGAAGCACGGTTAAA	CA038 (in gDNA) GAACCCGATACCGATTACTGC
Sanger sequencing of gDNA Ura3 locus M.EcoGII insertion	DB428 ACGGATTAGAAGCCGCCGAGC	DB558 CCGGCTGTCTTGCTTAGTT

Table 2.1 Primer list

The name of the primer is written in bold with the sequence below. All primer sequences are written as 5' to 3'. gDNA = genomic DNA (*S. cerevisiae*); pG5E4 = recombinant nucleosome plasmid; PCR for angelicin is for both steps, from MRC5 DNA and after miniprep from *E. coli*; RE = restriction enzyme; no RE means that there are no recognition sites for the restriction enzymes assessed in the experiment.

2.4 Enzymes

Enzyme	Function	Concentration	Company	Catalogue Number
BamHI	Restriction enzyme	20 units/μl	NEB	R0136S
BglII	Restriction enzyme	10 units/μl	NEB	R0144S
BioTaq	Polymerase	5 units/μl	BioLine reagents Ltd	BIO-21060
EcoRI-HF	Restriction enzyme	20 units/μl	NEB	R3101S
HpaII	Restriction enzyme	10 units/μl	NEB	R0171S
M.EcoGII	Methyltransferase	5 units/μl	NEB	M0603S
M.CviPI	Methyltransferase	4 units/μl	NEB	M0227S
Micrococcal Nuclease (MNase)	Endonuclease	2,000 gel units/μl	NEB	M0247S
NotI	Restriction enzyme	10 units/μl	NEB	R0189S
PstI-HF	Restriction enzyme	20 units/μl	NEB	R3140S
PvuII	Restriction enzyme	10 units/μl	NEB	R0151S
SalI	Restriction enzyme	20 units/μl	NEB	R0138S
SalI-HF	Restriction enzyme	20 units/μl	NEB	R3138S
Scal-HF	Restriction enzyme	20 units/μl	NEB	R3122S
SmaI	Restriction enzyme	20 units/μl	NEB	R0141S
SphI-HF	Restriction enzyme	20 units/μl	NEB	R3182S
T4 polynucleotide kinase	Kinase	10 units/μl	NEB	M0201L
T4 DNA ligase	Ligase	400 units/μl	NEB	M0202S
XhoI	Restriction enzyme	20 units/μl	NEB	R0146S

Table 2.2 Enzyme list

All enzymes were used as prepared by the company and according to their recommendations unless stated otherwise in other sections. NEB = New England Biolabs.

2.5 sacCer3 reference genome analysis

For *S. cerevisiae* genome analysis, the sacCer3 reference sequence from UCSC (<http://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/bigZips/>) was used, with the exclusion of mitochondrial DNA. For nucleosome positions, 147 base nucleosome binding sites were generated based on consensus nucleosome dyad positions from Jiang *et al* (2009) [89]. For ACS positions which represent DNA replication origins, wild-type cells in G2 stage at 23 °C

(GSM424494) from Eaton *et al* (2010) were used [7]. See appendix 1 for the commands and scripts used.

2.6 Psoralen crosslinking

2.6.1 Annealing

Two 18 base oligonucleotides (HJ001 and HJ002) were annealed by using 25 µl of each 100 µM oligonucleotide (14.2 µg of HJ001 and 13.6 µg of HJ002) and 2.5 µl 20x annealing buffer (see table 2.3). The annealing reactions underwent a thermocycler program of 92 °C for 1 minute followed by a drop in temperature of 0.1 °C every 10 seconds until it reached 4 °C, in a BioRad DNA Engine Dyad Peltier Thermal Cycler machine.

Reagent	20x stock concentration	1x final concentration
Magnesium acetate tetrahydrate	40 mM	2 mM
HEPES-KOH pH 7.5	600 mM	30 mM
Potassium acetate	2 M	100 mM

Table 2.3 Psoralen crosslinking annealing buffer

Composition of 20x stock and 1x final annealing buffer for psoralen crosslinking reaction.

2.6.2 Crosslinking with TMP

The crosslinking step is based on Zhang *et al* [112]. 4,5',8-trimethylpsoralen (TMP, Sigma-Aldrich #T6137) was dissolved in dimethyl sulfoxide (DMSO) to a stock concentration of 2 mg/ml and stored at -20 °C in black eppendorfs as furanocoumarins are light-sensitive. Next, 26 µl annealed primers (6.8 µg), 10 µl 2 mg/ml TMP and 1 ml crosslinking buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 50 mM NaCl) were mixed and left on ice for 10 minutes in black eppendorfs. Then the samples were added to 6-well plates (Costar 6 well clear TC-treated multiple well plates individually wrapped sterile, Scientific Laboratory Supplies,

#3516) on ice with one condition per well. The crosslinking ability of TMP was activated by 365 nm ultraviolet light (UV) using a 6 watt UVP UVGL-58 handheld UV lamp at either 2 or 10 cm above the plate. UV treatment was performed in 15 minute cycles, with 2 minute rest intervals in between, during which the plate was gently rocked to mix the samples and 10 µl of 2 mg/ml TMP was added to samples due to have further UV treatment. Samples could also be removed at this point, to collect samples with a range of UV exposure times. UV treatment was performed in the dark by using a cardboard box over the samples and UV lamp. Samples were exposed to normal light during mixing and TMP additions.

2.6.3 Filtration

Filtration was performed with the use of 3,000 Nominal Molecular Weight Limit Amicon Ultra 0.5 mL Centrifugal Filters (Merck, #UFC500324) following the kit protocol (spin for 30 minutes at 14,000 g, then invert column in a new collection tube and spin for 2 minutes at 1,000 g).

2.6.4 Crosslink reversal

Two methods of crosslink reversal were investigated: heat and alkali treatment or 254 nm UV. 50 µl of DNA (14.5 µg) which had been exposed to six 365 nm UV cycles (1.5 hours) was used for each condition. For heat and alkali, 70 and 90 °C combined with final concentrations of 10 and 100 mM NaOH, followed by neutralisation with 1 M HCl (prior to running on a denaturing gel) were investigated. The alternative method involved 254 nm UV using the same set up (samples in a 6-well plate on ice) and same handheld UV lamp as for the 365 nm UV crosslinking reaction (section 2.6.2), with cycle length reduced to 5 minutes rather than 15 minutes. No additional treatment was performed between cycles except for removing samples and gentle mixing.

2.6.5 Polyacrylamide denaturing gels

Running polyacrylamide denaturing gels

4 µl of each sample was mixed with 4 µl Novex TBE (Tris/Borate/EDTA)-Urea Sample Buffer (2x) (Invitrogen #LC6876) in eppendorf tubes. Roughly 10 ng of sample DNA was loaded per expected band. Two ladders were used (10/60 and 20/100 Oligo Length Standards from IDT, #51-05-15-01 and #51-05-15-02 respectively) with 4 µl of 1:100 diluted ladder added to 4 µl sample buffer. After mixing with sample buffer, all eppendorf tubes containing sample or ladder DNA were boiled for 5 minutes at 99 °C and snap-cooled on ice prior to loading on the denaturing gel.

The 15% polyacrylamide 8M TBE-Urea denaturing gel (Invitrogen #EC68855BOX) was pre-warmed at 65 °C for a couple of minutes. Wells were pre-rinsed with 1x TBE running buffer prior to sample loading. The tank contained 1x TBE buffer during the run. 8 µl of the boiled sample or ladder DNA with sample buffer was loaded into each well and run at 100 V for 30 minutes, then 150 V for 50 minutes. Gels were stained with 1 µl SYBR Gold Nucleic Acid Gel Stain (ThermoFisher Scientific #S11494) in 10 ml 1x TBE running buffer, wrapped in clingfilm and protected from light on a rocker for 15 minutes to allow visualisation of DNA bands with a BioDoc-It 210 Imaging system (UVP).

Gel extraction from polyacrylamide denaturing gels

The crosslinked band (28 bases) was excised from the SYBR-Gold stained polyacrylamide denaturing gel using a scalpel and placing the gel above a blue light (475 nm) plate. Electroelution was used to extract the DNA from the gel. This involved placing the gel slice in a small bag of 3,500 molecular weight cut-off SnakeSkin Dialysis Tubing (Thermo Scientific #68035) sealed with SnakeSkin Dialysis Clips (Thermo Scientific #68011). A small amount of 0.1x TBE was added and all air bubbles removed prior to sealing the other end of the

snakeskin bag with another clip. The clips sealing the snakeskin bag containing the gel slice were taped to the bottom of an empty gel tank with the DNA towards the negative (black) electrode. The gel tank was then filled with 0.1x TBE and run at 2 V/cm (using the distance between the electrodes for the calculation) for 3 hours. Upon completion, the buffer (now containing the DNA) within the bag was extracted by pipetting and the bag and gel slice were rinsed with 1 ml 0.1x TBE which was added to the previous extracted buffer to increase DNA recovery. DNA was subsequently filtered through an Amicon 3k spin column (Merck, #UFC500324) to concentrate the DNA (section 2.6.3). The same column was used for multiple 30 minute spins (with extra addition of eluted material) prior to inversion.

2.6.6 Mass spectrometry

Sample requirements were clean DNA eluted in water at a concentration of at least 20 μ M. DNA was filtered through a NAP-25 column filter (Merck, #GE17-0852-02) following kit instructions, with the sample volume made up to 1.5 ml with water. NAP-25 column filtered DNA was concentrated using a GeneVac at 65 °C. Mass spectrometry was performed on an electrospray ionisation MS-TOF (mass spectrometry - time of flight) Xevo oligonucleotide mass spectrometer. Data was analysed with MassLynx software with a resolution of 1 Da/channel. The resulting graphs show Electrospray Maximum Entropy calculated by a Uniform Gaussian Damage model and iterated until convergence. The samples were run on the mass spectrometer by Dr Afaf El-Sagheer.

2.6.7 Reannealing of oligonucleotides

Reannealing was performed with 45 μ l of DNA (4.2 μ g) and 2.25 μ l 20x annealing buffer (table 2.3). The same thermocycler program was used as for the initial annealing step (section 2.6.1).

2.6.8 Phosphorylation

The 5' ends of the 18 base annealed oligonucleotides were phosphorylated by T4 polynucleotide kinase (PNK). The reaction (table 2.4) was incubated at 37 °C for one hour. Then 1 ml Tris pH 7.0 was added and the sample was filtered through an Amicon 3k column as in section 2.6.3.

Reagent	Stock concentration	Final concentration	Volume (µl)
Water	-	-	4.8
DNA	90 ng/µl	70.5 ng/µl	47
PNK	10 units/µl	0.2 units/µl	1.2
PNK buffer	10x	1x	6
rATP	10 mM	0.166 mM	1

Table 2.4 Phosphorylation of small fragment DNA

The small fragment DNA used in this reaction is the two 18 base annealed oligonucleotides which have been crosslinked, gel extracted, filtered, reversed and re-annealed. PNK = T4 polynucleotide kinase (see table 2.2 for details).

2.6.9 Ligation

The large fragment was a PCR product from *S. cerevisiae* gDNA (chr10:245,842-246,715) amplified with HJ003 and HJ004 with an annealing temperature of 58 °C. This was digested with BglII and SalI to produce complementary ends. The small fragment was the 18 base oligonucleotides which had been annealed, crosslinked with TMP, run on a denaturing gel, gel extracted, crosslink reversed with 100 mM NaOH at 90 °C for ten minutes and re-annealed.

Reagent	Stock concentration	Volume (µl)
Water	-	48
Large fragment	35 ng/µl	20
Small fragment	70 ng/µl	10
T4 DNA ligase	3 Weiss units/µl	2
T4 DNA ligase buffer	10x	10
BglII	10 units/µl	2
SalI-HF	20 units/µl	2
BamHI	20 units/µl	2
XhoI	20 units/µl	2
rATP	10 mM	2

Table 2.5 Ligation reaction for producing a long fragment with a TMP modification

The small fragment is annealed oligonucleotides containing a TMP crosslink which have been gel extracted, filtered and phosphorylated at the 5' ends. Details of company and product code for commercial reagents can be found in table 2.2.

The ligation reaction in table 2.5 was run at 23 °C for 48 hours. XhoI and BamHI were included to prevent self-ligation of the small fragment. The ligase was heat-inactivated by incubation at 65 °C for ten minutes. Linearisation was performed with 100 µl of the ligation reaction and 5 µl PvuII in 1x Buffer 3.1 (NEB #B7203S) at 37 °C overnight. The DNA was run on a 3% agarose gel and the linearised ligation product of 448 base pairs (bp) was gel extracted as in section 2.6.11.

2.6.10 Agarose gels

Agarose gels were run at 90 V in 1x TBE and DNA was visualised using ethidium bromide staining and a wash with water post-electrophoresis. Visualisation was performed with a BioDoc-It 210 Imaging system (UVP).

2.6.11 Agarose gel extraction

Agarose gel extraction was performed as recommended by Qiagen, with the QIAquick Gel Extraction Kit (#28704). All centrifugation steps were performed at 17,900 g.

2.6.12 Sanger sequencing

Sanger Sequencing of PCR products was outsourced to Source Bioscience. Submitted samples were 10 ng/ μ l. Submitted primers were 3.2 pmol.

2.6.13 Nanopore sequencing

Library preparation was performed using a ligation 1D sequencing kit (SQK-LSK108). 1D nanopore sequencing was performed on a FLO-MIN107 MinION flowcell and the data was basecalled with Albacore v2.0.2, using the r95_450bps_linear.cfg configuration file. The psoralen crosslinked material was run alongside yeast genomic DNA thus was extracted by mapping to a reference sequence with demultiplex.py. Within this script 15, 30 and 80% 'query_length' values were tested. Indexing of reference sequences (forward and reverse) and alignment of reads to these references were performed with BWA (v0.7.12). Reads were plotted in IGV (v2.4.1) and event data was extracted using nanopolish (v0.13.2). ONT model data was obtained from [https://github.com/nanoporetech/kmer_models/blob/master/r9.4_180mv_450bps_6mer/](https://github.com/nanoporetech/kmer_models/blob/master/r9.4_180mv_450bps_6mer/template_median68pA.model) template_median68pA.model on 31/10/2017.

2.6.14 DNA modification with angelicin

Verification of angelicin (Sigma #A0956) modification of DNA involved two approaches (figure 3.21).

For approach 1, a 654 bp region of MRC5 genomic DNA (chr14:75,277,283-75,277,937) was selected based on published work on angelicin modification of DNA [113] and TA/AT frequency. This region of MRC5 gDNA was PCR amplified with HJ014 and HJ017 (see table 2.6) using the PCR program in table 2.9 with annealing temperature of 62 °C rather than 59.3

°C. The amplified DNA was purified using a GenElute PCR clean-up kit (Sigma #NA1020). The DNA band at 654 bp was extracted from a 1.5% TBE-agarose gel as in section 2.6.11.

Reagent	Company	Stock concentration	Volume (µl)
Water	-	-	27.4
NH ₄ reaction buffer	NEB	10x	5
MgCl ₂	NEB	50 mM	2
dNTPs	NEB	10 mM	2.5
Forward primer	IDT (HJ017)	100 µM	1.2 of 1:10
Reverse primer	IDT (HJ014)	100 µM	1.2 of 1:10
MRC5 gDNA	-	8.13 ng/µl	8.2
BioTaq polymerase	BioLine reagents Ltd (#BIO-21060)	5 units/µl	0.5

Table 2.6 PCR reaction composition for amplification of DNA for angelicin treatment

The NEB reagents are all part of a #E5000S PCR kit. gDNA = genomic DNA. See table 2.1 for primer details.

Ligation of the amplified 654 bp DNA region into a pGEM-T vector (table 2.7) was incubated at 16 °C overnight. Subsequent transformation into DH5α *E. coli* cells was performed using the pGEM-T Promega transformation protocol with the addition of X-Gal plates in step 10.

Reagent	Company and kit	Stock concentration	Standard reaction volume (µl)	Positive control volume (µl)	No insert control volume (µl)
Rapid ligation buffer	Promega #A3600	2x	2.5	2.5	2.5
pGEM-T vector	Promega #A3600	50 ng/µl	0.5	0.5	0.5
PCR product	-	11 ng/µl	1.5	-	-
Control insert DNA	Promega #A3600	4 ng/µl	-	1	-
T4 DNA ligase	Promega #A3600	3 Weiss units/µl	0.5	0.5	0.5
Water	-	-	0	0.5	1.5

Table 2.7 pGEM-T ligation reaction

Promega #A3600 is a kit and the protocol from this was used to perform the ligation and transformation steps. Vector:Insert ratio was 1:3.

Plasmid DNA was extracted from a set of *E. coli* colonies from the transformation plate using a NucleoSpin Plasmid DNA purification kit (Macherey-Nagel #740588). The pres-

ence and orientation of the 654 bp insert was assessed by a PvuII digest following NEB recommendations and visualisation on a 1.2% TBE-agarose gel run at 90 V. The insert was PCR amplified from the mini-prepped pGEM-T plasmid using the reagents in table 2.8 and PCR program in table 2.9. The amplified DNA was purified using a GenElute PCR clean-up kit (Sigma #NA1020). Angelicin stocks were made up at 1 mM in 100% ethanol in black eppendorfs. Similar to section 2.6.2, 83 µl crosslinking buffer and 17 µl amplified DNA (58.8 ng/µl) was added to each well of the plate on ice for UV treatment or an eppendorf on ice for the no UV conditions, all plus/minus 2 µl 1 mM angelicin in 100% ethanol. UV treatment of 30 or 60 minutes was performed on the plate. All samples were then filtered with Amicon 3k columns (see section 2.6.3). Sanger sequencing was performed with HJ014 (see section 2.6.12).

Reagent	Company	Stock concentration	Volume (µl)
Water	-	-	35.3
NH ₄ reaction buffer	NEB	10x	5
MgCl ₂	NEB	50 mM	2
dNTPs	NEB	10 mM	77.5
Forward primer	IDT	100 µM	1.2 of 1:10
Reverse primer	IDT	100 µM	1.2 of 1:10
DMSO	Sigma (#D4540)	≥99.5%	2
Plasmid DNA	-	0.08 ng/µl	0.3
BioTaq polymerase	BioLine reagents Ltd (#BIO-21060)	5 units/µl	0.5

Table 2.8 PCR reaction composition for production of DNA for angelicin treatment

The NEB reagents are all part of a #E5000S PCR kit. The primers used were HJ014 and HJ017.

Temperature (°C)	Time	Cycles
95	5 minutes	1
95	30 seconds	35x
59.3	30 seconds	
72	30 seconds	
72	5 minutes	1
15	Forever	1

Table 2.9 PCR program for production of DNA for angelicin treatment

Forever represents until the samples are collected from the PCR machine.

For approach 2, the oligonucleotides were annealed as in section 2.6.1. Angelicin stocks of 2 mg/ml in DMSO were made and stored in black eppendorfs at 4 °C. Eppendorfs with 26 µl annealed primers (6.8 µg), 1 ml crosslinking buffer and with or without the starting addition of 2 mg/ml angelicin were incubated on ice for ten minutes. In a similar design to section 2.6.2, a 24-well plate was used, with each condition split across two wells of 500 µl to reduce bias from plate location. Next, 365 nm UV treatment was performed with the lamp touching the plate for one hour. During the UV treatment, the UV was paused every five minutes to gently mix the samples and add extra angelicin. A total of none, 1 nmol, 10 nmol, 100 nmol or 1 µmol 2 mg/ml angelicin was added in total. For the condition of angelicin with no UV, the angelicin was all added at the start. For reference, 88 nmol of TMP was used when crosslinking was observed. Mass spectrometry was performed as in section 2.6.6.

2.7 *S. cerevisiae* and methylation

2.7.1 Yeast strains

The *S. cerevisiae* strains used are detailed in table 2.10. All strains have a W303 background [114] and were stored frozen at -80 °C. To freeze a strain the culture was grown on a YPAD plate at either 23 °C or 30 °C. A sample was picked with a pipette tip, added to two autoclaved

tubes of 1 ml 25% glycerol and vortexed. When obtaining a strain from a frozen stock it was streaked on to a YPAD plate and grown at either 23 °C or 30 °C for a few days prior to any experiments.

Strain name	Genotype	Mating type	Temperature sensitive	Source
HJ001	<i>RAD5, BUD4, orc1-161</i>	a	Yes	<i>orc1-161</i> strain from D. MacAlpine lab [90]
HJ002	<i>RAD5, BUD4, ura3::GAL1-ECOGII</i>	a	No	This study
HJ004	<i>RAD5, BUD4, ura3::GAL1-ECOGII</i>	a	No	This study
HJ005	<i>RAD5, BUD4, ura3::GAL1-ECOGII</i>	a	No	This study
HJ006	<i>RAD5, BUD4, ura3::GAL1-ECOGII</i>	alpha	No	This study
HJ007	<i>RAD5, BUD4, leu2, ura3, trp1, ade2, his3, orc1-161, ura3::GAL1-ECOGII</i>	alpha	Yes	This study
HJ008	<i>RAD5, BUD4, leu2, ura3, trp1, ade2, his3, orc1-161, ura3::GAL1-ECOGII</i>	a	Yes	This study
HJ009	<i>RAD5, BUD4, leu2, ura3, trp1, ade2, his3, cdc6-1, ura3::GAL1-ECOGII</i>	a	Yes	This study
SBY024	<i>RAD5, BUD4, cdc6-1</i>	a	Yes	Stephanie Barker*
CAY336	<i>RAD5, BUD4, leu2, trp1, ade2, his3, URA3-ADH1-promotor-OsTIR1</i>	a	No	Made by Carolin Müller
CAY434	<i>RAD5, BUD4, leu2, trp1, ade2, his3, ura3, lys2, ura3::GALS-DAM</i>	a	No	Made by Carolin Müller
T7107	<i>RAD5, BUD4, leu2, ura3, trp1, ade2, his3</i>	a	No	T. Tanaka lab
AUY116	<i>RAD5, BUD4, leu2, ura3, trp1, ade2, his3</i>	alpha	No	The lab (Amy Upton)

Table 2.10 *S. cerevisiae* strains

All strains are *S. cerevisiae* with a W303 background. HJ001 was sent from the MacAlpine lab, but given a ‘HJ’ name when entered in our lab system. Temperature sensitive strains were grown at 23 °C for experiments unless stated otherwise, while non-temperature sensitive strains were grown at 30 °C. * The mutation in the SBY024 strain was originally described by Hartwell (1976) but was made *BUD4* and *RAD5* wild-type by multiple backcrosses performed by Stephanie Barker.

2.7.2 Yeast growth media

Media detailed in table 2.11 was made up according to manufacturer's instructions, autoclaved at 117 °C for fifteen minutes and stored at room temperature.

For 2% raffinose and 2% galactose, YPA media was used as a base. Stocks of 20% raffinose and 20% galactose were freshly prepared in water for each experiment. These were filter sterilised with a Terumo syringe without needle (#SS+50ES1) with a Minisart single use filter unit CA 0.2 (#16534-K) from Sartorius Stedim Biotech. In experiments, 2% galactose was used to induce expression of a gene of interest, therefore the 20% galactose stock was added to the 2% raffinose culture to create a culture with a final concentration of 2% galactose.

Acronym	Full name	Company	Catalogue number	Components	Format
YPA	Yeast-Peptone-Adenine	Thermo Fisher Scientific	BD212750 BD211820	10 g/l yeast extract 20 g/l peptone 0.5% adenine solution	Broth
YPAD	Yeast-Peptone-Adenine-Dextrose	Formedium	CCM1010	10 g/l yeast extract 20 g/l peptone 20 g/l glucose 0.04 g/l adenine sulphate	Broth
YPAD	Yeast-Peptone-Adenine-Dextrose	Formedium	CCM1010	10 g/l yeast extract 20 g/l peptone 20 g/l glucose 0.04 g/l adenine sulphate 20 g/l agar	Agar
KAc	Sporulation medium	Formedium	CCM0802	10 g/l yeast extract 0.2 g/l glucose 10 g/l potassium acetate 20 g/l agar	Agar
-	Minimal medium	Formedium	CSM0105	Contains a yeast nitrogen base with no glucose or amino acids	Agar
LB	Luria-Bertani Minimal medium	Melford	L24040-1000.0	5 g/l yeast extract 10 g/l casein digest peptone 10 g/l sodium chloride	Broth
LB	Luria-Bertani Minimal medium	Sigma	L3147-250G	5 g/l yeast extract 10 g/l tryptone 10 g/l sodium chloride 15 g/l agar	Agar

Table 2.11 Yeast growth media

Minimal and sporulation media were used for generation of new yeast strains. YPA media was made to pH 6 - 6.5 with 1 M HCl and the 0.5% adenine solution used adenine (Sigma A9126) in water with 5% 1M HCl and was filter sterilised prior to addition to YPA and subsequent autoclaving. YPAD was used for normal yeast growth and tetrad dissections. LB broth and agar media used to grow *E. coli*.

2.7.3 DNA extraction

All yeast cultures grown for DNA extraction were treated with 1:1,000 ampicillin from 100 mg/ml aliquots. Yeast Smash and Grab protocol (based on [115]) involved combining the

cells with one volume of lysis buffer (10 mM Tris pH 8.0, 100 mM NaCl, 1 mM EDTA, 1% w/v SDS, 2% v/v Triton X-100). If using an ethanol precipitation at the end, volumes of 0.2 ml were used, or 0.3 ml if using an isopropanol precipitation. One volume of phenol-chloroform-isoamyl alcohol (25:24:1, pH 8.0) was added in addition to 0.2 ml zirconia/silica beads. The cells were lysed in a FastPrep-24 at maximum power for 20 seconds, repeated three times with a ten second gap, put on ice for one minute and then the lysing step was repeated twice more. One volume of TE buffer was added to the cells and mixed by shaking. Cells were then spun at 13,523 g for five minutes, the upper aqueous layer was collected and moved to a new tube. Next two volumes of chloroform were added and mixed vigorously by shaking. The tubes were spun at 13,523 g for five minutes, and the upper aqueous layer transferred to a new tube. Next, 10 µl 10 mg/ml RNaseA was added and incubated at 37 °C for 30 minutes. Then 10 µl of 20 mg/ml proteinase K was added and incubated at 50 °C for one hour. Two volumes of chloroform were added, shaken and spun at 13,523 g for five minutes with transferral of the upper aqueous layer to a new tube. DNA was then precipitated. For ethanol precipitation, two volumes of ethanol were added, mixed by inversion, spun at 20,238 g for 15 minutes and the supernatant removed. For isopropanol precipitation, 0.1 volume of 3 M KoAc pH 5.5 and one sample volume of isopropanol were added to the DNA, mixed by inversion, incubated at room temperature for ten minutes, spun at 20,238 g for 15 minutes at 4 °C and supernatant removed. Finally, for both precipitation methods the pellet was washed with 1 ml 70% ethanol, air dried and resuspended in 0.1 % TE.

2.7.4 Recombinant nucleosomes

For investigating methylation of accessible DNA compared to nucleosome-bound DNA, a plasmid with human recombinant nucleosomes (Recombinant Polynucleosomes (H3.1) #31466 from Active Motif) was used. This pG5E4 plasmid consisted of 5 kb of DNA

from multiple sources and was reported to form eleven nucleosomes. These recombinant nucleosomes were stored in 5 µl aliquots at -80 °C.

To methylate the recombinant nucleosomes, 0.5 µl (330 ng) was used with 1x Cutsmart buffer, 160 µM S-adenosylmethionine (SAM, NEB #B9003S) and 1 µl (5 units) M.EcoGII or water in a total volume of 20 µl. This was incubated at 37 °C for one hour, followed by overnight incubation at 65 °C with 76 µl TE buffer and 4 µl 20% SDS. The next day, the samples were cooled to room temperature, 40 µg of proteinase K was added and incubated at 50 °C for one hour, followed by incubation at 65 °C for one hour. Phenol-chloroform DNA extraction and ethanol precipitation was performed as in section 2.7.3.

2.7.5 Droplet Digital PCR

Droplet Digital PCR (ddPCR) was used to quantitatively measure DNA concentrations to infer the level of methylation present at specific loci.

Sample preparation

For plasmid DNA, linearisation was performed prior to M.EcoGII methylation. After methylation reactions M.EcoGII was heat-inactivated (10 minutes at 65 °C) and PCR cleaned up with GenElute PCR clean-up kit (Sigma #NA1020). The DNA was then digested with adenine-sensitive and control restriction enzymes, PCR cleaned up and fragmented (if genomic DNA) with a restriction enzyme (ScaI-HF, BamHI or HpaII). All reactions were performed according to New England Biolabs (NEB) recommendations. DNA was subsequently diluted in water for ddPCR analysis.

Plasmid DNA contained the site of interest at a substantially higher frequency than in genomic DNA, therefore large dilutions were required. These dilutions were performed step-wise with a series of 1:10 dilutions (10 µl DNA and 90 µl water) to reach the correct order of

magnitude. The final concentration was optimised using ddPCR and varied depending on whether the starting material was pure plasmid DNA or a spike-in of plasmid DNA.

For genomic *S. cerevisiae* DNA, concentrations of 0.5 ng/μl and between 1.5 and 2 ng/μl were used for each sample (in two experiments respectively). The latter led to ddPCR readings in a better concentration range. For chromatin and spheroplast DNA approximately 1 ng/μl DNA was used per reaction.

Running ddPCR

The ddPCR equipment used was the Bio-Rad QX200 Droplet Digital PCR System. The first step for ddPCR was droplet generation. For this a solution of 11 μl EvaGreen ddPCR mixture, 9 μl of DNA template and 2 μl of 1:40 dilution of 100 μM IDT primers was prepared for each sample. Subsequently, 20 μl of this mixture was added to each well in the top row of the cartridge and 70 μl droplet generation oil (Bio-Rad #1863005) was added to each well in the bottom row. The cartridge was topped with a rubber cover and placed in the droplet generator. Once complete, 40 μl of the newly formed droplets were taken out for each well using a wide-mouth tip and moved to the next available column of the plate. The plate was covered with foil, sealed in the plate sealer and then put in the thermocycler where a PCR reaction was performed (see table 2.12). Next the plate was moved into the droplet reader where the fluorescence was measured and the data was analysed with Bio-Rad QuantaSoft software.

Temperature (°C)	Time	Cycles
95	5 minutes	1
95	30 seconds	40
58	1 minute	
4	5 minutes	1
90	5 minutes	1
10	Forever	1

Table 2.12 Droplet Digital PCR (ddPCR) program

Steps involved in the PCR used in ddPCR analysis of the methylation state of plasmid DNA, genomic DNA and chromatin.

2.7.6 Chromatin extraction for ddPCR

For the ddPCR experiments testing methylation on chromatin there were three methods investigated, two involving the formation of spheroplasts with zymolyase treatment, and one involving formaldehyde crosslinking to maintain the protein-DNA interactions during methylation followed by mechanical cell lysis.

Formaldehyde crosslinking chromatin extraction

For the formaldehyde crosslinking method, 25 ml of wild-type *S. cerevisiae* cells were grown in YPAD at 30 °C to OD₆₀₀ 0.7 - 1.0, transferred into a 50 ml falcon tube to which 700 µl 37% formaldehyde was added, to a final concentration of 1%. The cells were incubated on a rocker at room temperature for 30 minutes. Formaldehyde was subsequently quenched with 1.35 ml 2.5 M glycine (in water) to a final concentration of 125 mM and incubated at room temperature for five minutes. Next the cells were spun at 4 °C at 1,962 g for three minutes and the supernatant was discarded. Then the cells were resuspended in 10 ml ice-cold TBS and spun again. The resuspension and subsequent spin was repeated twice more and cells were then resuspended in 0.4 ml cold lysis buffer (50 mM HEPES-KOH pH 7.5, 140 mM

NaCl, 1 mM EDTA, 1% Triton-X100, 0.1% Na-deoxycholate) and 2 μ l 200 mM PMSF and 8 μ l 50x anti-proteolytic cocktail were added, to final concentrations of 1 mM and 1x respectively. Then this was transferred to a 2 ml tube with a screw cap, glass beads were added and cells were broken open using the FastPrep-24 at maximum speed (6.5 m/s) for eight cycles of 30 seconds with 30 seconds on ice in between. The tube was wiped and a 0.45 mm needle was used to make a hole in the bottom of the tube. The tube was inserted into a 15 ml falcon and spun at 4 °C at 1,962 g for one minute. Then the 2 ml tube was removed and any liquid in the falcon tube was transferred to an eppendorf. The falcon tube steps were repeated twice more. The sample was then spun at 4 °C at 20,238 g for one minute. For the methylation step, 1 μ l 32 mM SAM and 0, 1 or 4 μ l 5 units/ μ l M.EcoGII were added to 100 μ l cell extract and incubated at 37 °C for one hour. In this methylation step a spike-in of approximately 50 ng ScaI-HF linearised pUC19 DNA can be added. Then 380 μ l TE plus 20 μ l 20% SDS was added to each sample and they were incubated at 65 °C overnight. The next day samples were removed from the heatblock and allowed to cool to room temperature. To digest the RNA, 10 μ g 10 mg/ml RNase A was added and incubated at 37 °C for two hours. To degrade proteins 80 μ g 20 mg/ml proteinase K was added and incubated at 50 °C for one hour. The temperature was increased to 65 °C for one hour and phenol-chloroform DNA extraction and ethanol precipitations were performed as described in section 2.7.3.

Spheroplast chromatin extraction

For the ddPCR experiments testing methylation at the spheroplast stage the protocol was also used for the SMAC-seq variant (section 2.7.9), with the addition of a post-methylation step after the DNA extraction and ethanol precipitation. This involved treatment of mock-methylated DNA (only previously treated with water) with 0.5 μ l 5 units/ μ l M.EcoGII in a 20 μ l reaction with 1x Cutsmart buffer and 160 μ M SAM. This was incubated at 37 °C for one hour. Next, 10 ng of DNA from this additional methylation step, and from the previous water

and M.EcoGII treated DNA samples, were subsequently treated with either 0.5 µl PstI-HF or water in 1x Cutsmart buffer and incubated at 37 °C for one hour. Then samples were treated with 0.5 µl ScaI-HF at 37 °C for one hour and diluted to 0.1 ng/µl for ddPCR.

Spheroplast chromatin extraction using a sucrose cushion

For the chromatin analysed by western blotting chromatin needed to be extracted from the cells. One of the ddPCR experiments also tested chromatin extraction using this method. *S. cerevisiae* cells were grown overnight in 5 ml YPAD, diluted to OD₆₀₀ 0.1 in 25 ml of YPAD and grown to OD₆₀₀ 0.4. The 25 ml culture was transferred into a 50 ml falcon tube with 20 ml ice, 450 µl 10% sodium-azide (final concentration of 0.1%), 225 µl of 200 mM PMSF (final concentration of 1 mM). Cells were then pelleted at 1,962 g at room temperature for five minutes and the supernatant was subsequently removed. Cells were resuspended in 0.2 ml of cold Wash buffer A (0.1% sodium-azide, 1 mM sodium-vanadate, 1mM PMSF, total volume made up with water) and transferred to eppendorf tubes. Cells were then pelleted (2,000 g for three minutes at 4 °C) and supernatant removed. The cell pellet was then weighed to calculate the amount of zymolyase required. The cell pellet was loosened by tapping the bottom of the tubes, resuspended in 0.2 ml spheroplasting buffer (50 mM potassium phosphate buffer pH 7.5, 0.6 M sorbitol, 10 mM DTT, 1x Complete ULTRA Mini EDTA-free protease inhibitor, 0.4 mg zymolyase 20-T per 100 mg wet cell pellet), and incubated at 30 °C for ten minutes or until spheroplasting was complete. Spheroplasting was monitored under a light microscope (5 µl cells with 1 µl 10% Triton X-100). When spheroplasting was sufficient, the tubes of spheroplasts were put on ice and then pelleted (2,000 g at 4 °C for three minutes), and supernatant removed. Spheroplasts were resuspended in 0.2 ml of 0.4 M Sorbitol/EB (0.4 M Sorbitol, 1x EB, 1 mM PMSF, where 1x EB is 50 mM HEPES-KOH pH7.5, 100 mM KCl, 2.5 mM MgCl₂, 1x Complete ULTRA Mini EDTA-free protease inhibitor). Cells were then pelleted (2,000 g at 4 °C for three minutes), supernatant

removed and cell pellet weighed again. Assuming 100 mg = 100 μ l, the volume of the pellet was made up to 200 μ l by resuspending in buffer EB (1x EB, 1 mM PMSF). Next 5 μ l of 10% Triton X-100 was added to a final concentration of 0.25%. The tube was mixed gently by inversion three to four times then incubated on ice for five minutes. Next 150 μ l of the lyase was put on top of a 500 μ l sucrose cushion (30% sucrose, 1x EB, 1 mM PMSF), and lids were weighted down to ensure they would not open when subsequently boiled. Cells were then pelleted at 12,000 g at 4 °C for ten minutes. Meanwhile, 45 μ l of the remaining lysate (residual spheroplasts) from on top of the sucrose cushion were mixed with 15 μ l 4x Laemmli loading buffer to create a whole cell extract sample. This was boiled for five minutes. From the main sample, 75 μ l was removed from the top layer of the supernatant, mixed with 25 μ l of 4x Laemmli loading buffer and boiled for five minutes, to create the soluble sample. The rest of the supernatant was removed from the main sample, and the pellet washed with 500 μ l EBX (1x EB, 1 mM PMSF, 0.25% Triton X-100). The cells were pelleted at 3,000 g at 4 °C for three minutes and resuspended in 45 μ l EBX. Subsequently, 1 μ l benzonase was added and the sample was incubated on ice for thirty minutes. Then 15 μ l of 4x Laemmli loading buffer was added and the sample boiled for five minutes, spun at 3,000 g at room temperature for three minutes, and the supernatant collected which is the chromatin sample.

2.7.7 MNase digestion of chromatin

For the MNase digests, the chromatin did not need to be removed from the cells as only the DNA was needed, as long as MNase digest occurred while the chromatin structure remained intact. A 25 ml culture of wild-type (T7107) *S. cerevisiae* culture was grown in YPAD to OD₆₀₀ 0.7. This was crosslinked by the addition of 2.5 ml 11% formaldehyde (750 μ l 37% formaldehyde (Sigma #F8775), 1.75 ml diluent (71.43 mM HEPES pH 7.5, 1.43 mM EDTA, 143 mM NaCl)) to a final concentration of 1%. This was incubated for one hour at 25 °C. Cells were then pelleted at 1,962 g for three minutes at room temperature prior to

resuspending in 1 ml spheroplasting buffer (1 M Sorbitol, 50 mM HEPES pH 7.5). Cells were then pelleted at 2,348 g for three minutes at room temperature followed by resuspension in 0.49 ml spheroplasting buffer supplemented with 5 μ l 1 M DTT to a final concentration of 10 mM, and 10 μ l 10 mg/ml zymolyase 100T. The cells were incubated at 30 °C until the zymolyase reaction had sufficiently spheroplasted the cells. OD₆₀₀ was regularly measured using 800 μ l 1% SDS and 5 μ l cells. When OD₆₀₀ reached 0.05 cells were spun at 2,348 g for five minutes at room temperature and the supernatant was saved (S1). The spheroplasts were resuspended in 0.5 ml MNase buffer (50 mM Tris-HCl pH 7.9, 5 mM CaCl₂) and 5 μ l 10% IGEPAL CA-630 was added. The tubes were gently mixed by tapping and incubated on ice for five minutes. The spheroplasts were then spun at 2,348 g for five minutes at room temperature and the supernatant was saved (S2). The spheroplasts were resuspended in 0.5 ml MNase buffer and split into five tubes with 99 μ l in each. Next, 1 μ l of either 2,000 gel units/ μ l MNase (NEB #M0247S), or a dilution (1:10, 1:100, 1:1000) of MNase made with 50% glycerol, or 50% glycerol alone was added to each tube. They were incubated at 37 °C for fifteen minutes prior to addition of 100 μ l Stop buffer (20 mM Tris pH 8.5, 0.6 M NaCl, 1% SDS, 10 mM EDTA) and 300 μ l TE buffer. RNA was degraded by addition of 10 μ l 10 mg/ml RNaseA which was incubated for fifteen minutes at 55 °C. Next 10 μ l 20 mg/ml Proteinase K was added and incubated at 55 °C overnight. The next day the DNA was incubated at 65 °C for thirty minutes. Phenol-chloroform extraction and isopropanol DNA precipitation was performed as in section 2.7.3.

2.7.8 Methylation of chromatin with M.CviPI

For *in vitro* methylation of genomic DNA with M.CviPI see table 2.13. This was incubated at 37 °C for one hour, with M.CviPI subsequently removed with the GenElute PCR Clean Up Kit.

Reagent	Stock concentration	Final concentration	Volume (µl)
Water	-	-	19.7
GC reaction buffer	10x	1x	6
SAM	32 mM	160 µM	0.3
DNA	95 ng/µl	50 ng/µl	32
M.CviPI	4 units/µl	0.133 units/µl	2

Table 2.13 M.CviPI methylation of genomic DNA

Details of company and product code for M.CviPI can be found in table 2.2. SAM (S-adenosylmethionine) and GC reaction buffer come with M.CviPI from NEB, although SAM can also be purchased separately (NEB #B9003S).

For methylation in spheroplasts with M.CviPI, an 80 ml culture of wild-type *S. cerevisiae* (T7107) was grown in YPAD to provide 25 ml for methylation, 25 ml for a negative control, 10 ml for an MNase digest to validate the chromatin and some extra to allow for OD₆₀₀ measurements to be taken. To preserve the chromatin structure, a crosslinking step was added when the culture got to OD₆₀₀ 0.8. This involved adding 8 ml of 11% formaldehyde solution (50 mM HEPES-KOH pH 7.5, 1 mM EDTA, 100 mM NaCl, 11% formaldehyde) for one hour at 30 °C. This was divided into two 50 ml falcon tubes, spun down at 1,962 g for three minutes at room temperature, resuspended in 0.8 ml spheroplasting buffer (50 mM HEPES-KOH pH 7.5, 1 M Sorbitol) and recombined into one tube which was repelleted at 4,816 g for three minutes and resuspended in 0.98 ml spheroplasting buffer supplemented with 10 µl 1 M DTT (10 mM final concentration) and 20 µl 10 mg/ml Zymolyase 100T. Cells were incubated at 30 °C and OD₆₀₀ was regularly measured by using 0.8 ml 1% SDS and 5 µl of the cells. When OD₆₀₀ reached 0.05 the spheroplasts were split into three tubes (200 µl for MNase digest, 550 µl for M.CviPI methylation, 550 µl for no methylation control). All spheroplasts were pelleted at 4,816 g for five minutes at room temperature. For the MNase digest the spheroplasts were resuspended in 0.5 ml of MNase buffer (50 mM Tris-HCl pH 7.9, 5 mM CaCl₂) and 5 µl of 10% IGEPAL CA-630 was added. For the M.CviPI GC methylation and no methylation control, spheroplasts were resuspended in 0.5 ml of GC methyltransferase

buffer (160 μ M SAM, 1x GC methyltransferase buffer from NEB) and 5 μ l of 10% IGEPAL CA-630 was added. All three tubes were gently mixed by tapping and incubated on ice for five minutes, then pelleted at 2,348 g for five minutes at room temperature. For the MNase digest the spheroplasts were resuspended in 0.4 ml MNase buffer and split into four tubes (99 μ l each) to which 1 μ l of stock MNase or its dilutions in 50% glycerol (1:10, 1:100), or 50% glycerol (negative control), were added. These tubes were incubated at 37 °C for fifteen minutes. Next 0.1 ml Stop buffer (20 mM Tris-HCl pH 8.8, 0.6 M NaCl, 1% SDS, 10 mM EDTA) and 0.3 ml TE buffer was added to each tube. For the M.CviPI GC methylation, spheroplasts were resuspended in 0.49 ml GC methyltransferase buffer and 10 μ l M.CviPI was added. For the no methylation control spheroplasts were resuspended in 0.49 ml GC methyltransferase buffer and 10 μ l 50% glycerol was added. For both the M.CviPI and no methylation control the spheroplasts were incubated at 37 °C for one hour, incubated at 65 °C for twenty minutes (for both tubes, although the purpose is to inactivate the M.CviPI). Then 500 μ l TE buffer was added. For all three tubes (MNase digest, M.CviPI and no methylation conditions) 50 μ l 10 mg/ml RNaseA was added and spheroplasts were incubated at 55 °C for fifteen minutes. Next 50 μ l 20 mg/ml Proteinase K (in TE) was added and tubes were incubated at 55 °C overnight. On the next day all tubes were incubated at 65 °C for thirty minutes. The M.CviPI and no methylation conditions were split into two tubes each for phenol-chloroform extraction and isopropanol precipitation (section 2.7.3).

2.7.9 SMAC-seq variant

For the variation on SMAC-seq, 60 ml wild-type (T7107) *S. cerevisiae* cells were grown in YPAD media to OD₆₀₀ 0.7 (mid-log phase), from which 10 ml (approximately 1×10^8 cells) were pelleted for each condition, therefore 40 ml in total (one negative control to be treated with water, one with M.CviPI, one with M.EcoGII and one to be digested with MNase as a control for the quality of the chromatin). The cells were spun at 1,962 g for

5 minutes, media was removed and cells kept on ice. For each 10 ml culture, cells were resuspended by pipetting in 1 ml of spheroplasting solution (1 M sorbitol, 5 mg/ml zymolyase for every 100 mg of cell pellet, 10 mM DTT) and transferred to an Eppendorf where the cells were incubated for one hour at room temperature with constant gentle inversion of the tube. Lysis was regularly monitored by light microscopy. When sufficiently digested, the spheroplasts were spun at 13,500 g for thirty seconds and the spheroplasting solution was removed. The spheroplast pellet was briefly washed twice with 500 µl 1 M sorbitol which was subsequently removed after another spin at 13,500 g for five seconds. Spheroplasts were quickly resuspended in 190 µl methylation buffer (1 M sorbitol to stabilise the spheroplasts, 1x Custmart buffer, 160 mM SAM for the methyltransferases, 10 mM DTT for M.CviPI, 0.5 mM spermidine to stabilise the chromatin, 0.075% IGEPAL to permeabilise the membrane to allow the methyltransferases to access the chromatin). The spheroplasts were left on ice for 15 minutes to allow IGEPAL to work, then 2.5 µl RNaseA was added (without salt) to a final concentration of 0.1 mg/ml and incubated at 37 °C for five minutes. Next 30 units of each methyltransferase (M.EcoGII or M.CviPI) or water were added and incubated at 37 °C for one and a half hours. The reaction was terminated by lysing the cells with 20 µl 250 mM EDTA/5% SDS and rapid mixing. The lysed cells were left overnight in 20% SDS in TE at 65 °C. The next day 120 µg RNaseA and 30 µl 1 M NaCl was added to each of the cooled tubes to digest ssRNA and dsRNA and incubated at 37 °C for two hours. Next, 240 µg Proteinase K was added to each tube and incubated at 50 °C for one hour. The temperature was increased to 65 °C and incubated for one hour. Subsequently a phenol-chloroform extraction followed by ethanol precipitation was performed (section 2.7.3). The pellet was finally resuspended in 20 µl 1x TE and frozen until nanopore sequencing library preparation.

2.7.10 Yeast strain generation

In order to generate an *S. cerevisiae* strain with an inducible M.EcoGII (non-specific adenine methyltransferase) a plasmid (pDB81) comprising a *GALI* promoter, the M.EcoGII gene, 3x FLAG tags, 6x HIS tags, *TDHI* terminator, *URA3* homology regions and a kanamycin resistance gene with promoter and terminator was used. This plasmid was generated using MoClo components by Dr Dzmitry Batrakou. Linearisation of the plasmid was performed with NotI (where it cuts in the two *URA3* homology regions) and involved 1 µg pDB81, 5 units of NotI (NEB, R0189S, 10,000 units/ml) and 1x NEBuffer 3.1 in 50 µl total reaction volume for one hour at 37 °C. This created two fragments, of which only the larger one (containing the M.EcoGII gene) had homology with the *URA3* locus thus allowing it to insert into the *S. cerevisiae* genome.

Yeast transformation was then performed with the linearised plasmid (pDB81) into wild-type *S. cerevisiae* (T7107 strain, W303 background) at the *URA3* locus. Wild-type yeast (5 ml in YPAD) was grown from a single colony, diluted to 50 ml (OD₆₀₀ 0.1), salmon sperm DNA was denatured (boiled at 100 °C for 5 minutes then placed on ice), cells pelleted when around OD₆₀₀ 0.7 (1,962 g for 5 minutes at 4 °C), supernatant removed, pellet resuspend in 25 ml water, cells pelleted (same as above), supernatant removed, cells resuspended in 1 ml water, transferred to eppendorf, pelleted (15,871 g for 30 seconds), supernatant removed, cells resuspended in 1 ml water and 100 µl cells transferred to a new tube for each transformation. Cells were then pelleted at 15,871 g for 30 seconds, supernatant removed and a master mix made for transformation (for 1 transformation: 240 µl PEG 3350 (50% (w/v)), 36 µl LiAc 1.0M, 50 µl single strand carrier DNA (sperm DNA)). Next 326 µl transformation master mix and 34 µl DNA (taken from a mix of 50 µl linearised pDB81, 5 µl pEDH3 (gRNA against Ura3) and 5 µl Cas9-Trp) or 34 µl water (negative control) was added to each tube of cells, vortexed to resuspend cells, placed at 42 °C for 40 minutes, pelleted at 15,871 g for 30

seconds, supernatant removed and prototrophic gene selection was performed on Ura-Trp-plates at 30 °C (10 and 90% plates set up).

S. cerevisiae strains with mutations in replication origin licensing proteins were mated with the strain with an inducible M.EcoGII, the creation of which is described above.

Mating, sporulation and tetrad dissection

For mating, *S. cerevisiae* strains were streaked onto a YPAD plate, overlaying each other and left to grow at 30 °C. For all growth steps if one of the strains was temperature sensitive 23 °C was used and times were extended.

For sporulation some of the mating patch on YPAD was streaked onto a potassium-acetate (KAc) sporulation plate and left at room temperature for three to seven days. These were checked for tetrads by adding 5 µl water to a glass microscope slide and scraping a small amount of cells from the sporulation plate with a pipette tip and adding them to the water. These were assessed using a light microscope at 40x objective magnification (400x total magnification).

Tetrad dissection involved adding a small amount of cells from the KAc plate to a 20 µl solution with 19 µl dissection buffer (0.8 M sorbitol, 10 mM EDTA, 10 mM NaPO₄ pH 7.2) and 1 µl 5 mg/ml T20 Zymolyase. This was incubated at 37 °C for 10-15 minutes. Meanwhile a flat YPAD plate was warmed up to room temperature (from storage at 4 °C), and left without a lid in the laminar flow hood to dry. Two grooves were cut along the diameter of the plate, not going completely through the YPAD media, creating a central line. The zymolyase/yeast mixture was diluted with 100 µl water and a couple of drops were added to the central section of the YPAD plate. These were spread with a pipette tip and the YPAD plate was left to dry without a lid on for 15 minutes. Separation of the tetrads was

performed with a Singer Instruments Spore Play dissection microscope with a glass needle. The tetrad plate was incubated for approximately four days at 30 °C.

Strain verification

For testing mating type, two strains were streaked on top of each other on a YPAD plate and left to grow at a temperature (either for a couple of days at 23 °C or overnight at 30 °C) that is compatible with the growth of both strains. The YPAD plate was then replica-plated on to minimal media and incubated at the same temperature as the previous step. Growth will only be observed if the two strains were of opposite mating types (therefore one of type alpha and one of type a) and both haploid. For testing temperature sensitivity, the plate containing the spores from tetrad dissection were replica-plated on to YPAD plates and incubated at 23, 30 and 37 °C. Temperature sensitive strains should be able to grow at 23 °C but reduced or no growth should be observed at higher temperatures. Integration of genes into the genome were tested by PCR over the integration sites with one primer in the *S. cerevisiae* genome and one primer in the inserted region of DNA.

2.7.11 Growth curves

For the inducible Dam methyltransferase growth curve, CAY434 *S. cerevisiae* cells were grown in 2% raffinose overnight, then transferred to 10 ml YPAD, 2% galactose or fresh 2% raffinose. Addition of IAA (Indole-3-acetic acid) was used to degrade the Dam methyltransferase. IAA powder (Sigma-Aldrich (1375-0)) was dissolved in 100% ethanol to 0.5 M final stock concentration which was frozen in 200 µl aliquots at -20 °C in black tubes as IAA is light sensitive. Four raffinose samples were measured; one control, one with an initial addition of IAA, one with hourly additions of IAA and one with additions of IAA every two hours. Every hour 25 µl of either 0.5M IAA (to degrade the Dam methyltransferase) or 100% ethanol (as a control as IAA is dissolved in 100% ethanol) was added to each sample.

For inducible M.EcoGII growth curves cultures were initially grown in 5 ml YPAD, then scaled up to larger volumes in either YPAD or 2% raffinose overnight. Growth curves were not performed in 2% galactose when the M.EcoGII was induced. Cultures had doubled more than twice in the final media prior to starting the OD₆₀₀ measurements which were taken from around 0.1 to 0.7, with at least five measurements for each sample.

2.7.12 α -factor arrest of *S. cerevisiae* cells

For this experiment *S. cerevisiae* cells with an inducible M.EcoGII (strain HJ005) were grown in YPAD to mid-log phase (to OD₆₀₀ 0.5), pelleted, washed with water, resuspended in 150 ml 2% raffinose at OD₆₀₀ 0.05 and grown to OD₆₀₀ 0.2. Then 0.15 μ l of α -factor for every ml of culture was added at 50 and 100 minutes, then every 30 minutes until 520 minutes. Small samples (10 μ l) of cells were regularly sonicated and visually inspected under a light microscope to assess the state of cell cycle arrest. Galactose was added to a final concentration of 2% at 430 minutes when the cells appeared to be predominantly arrested. The cultures were maintained at 30 °C throughout. DNA extraction performed as in section 2.7.3.

2.7.13 Flow cytometry

1.5 ml cell samples were collected, pelleted at 845 g for 5 minutes, supernatant removed by pipetting, pellet washed with 500 μ l water, respun, supernatant removed by pipetting, samples resuspended in 1 ml 100% ethanol, vortexed and stored in the fridge. On the day before flow cytometry analysis, the cells were pelleted, washed twice in 1 ml 50 mM NaCitrate (filter-sterilised), resuspended in 975 μ l 50 mM NaCitrate and 225 μ l 10 mg/ml RNase A (final concentration of 0.25 mg/ml) and incubated at 50 °C for one hour. Then 20 μ l of 20 mg/ml Proteinase K was added and the samples were incubated at 50 °C for one hour. Subsequently, the cells were pelleted and resuspended in 1 ml of 50 mM NaCitrate with 1x

Sytox Green (1 μ l per 5 ml 50 mM NaCitrate), vortexed and stored in the fridge overnight. On the day of the flow cytometry analysis, the samples were sonicated with a Biorupter UCD-200 (diagenode) on low setting for 3-5 minutes, with sonication at 0.5 minute intervals. Flow cytometry was performed with a BDX20 machine, measuring 100,000 events for each sample, with between 3,000 and 6,000 events per second. Data was analysed with FlowJo (version 10.7.1).

2.7.14 Protein extraction

Approximately 2.5×10^7 cells at OD₆₀₀ 0.7 were pelleted at 1,962 g for five minutes at room temperature. The cell measurement is based on OD₆₀₀ 0.7 being approximately 1×10^7 *S. cerevisiae* cells/ml based on conversions from <http://www.pangloss.com/seidel/Protocols/ODvsCells.html>. The pelleted cells were rinsed in water. If cells were collected as part of a timecourse, the cells were pelleted as above and frozen at -20 °C until all samples had been collected. Cell pellets were then resuspended in 100 μ l water and 0.2 M sodium hydroxide was added to a final concentration of 0.1 M. Cells were incubated at room temperature for five minutes, pelleted, resuspended in 50 μ l SDS sample loading buffer (0.06 M Tris-HCl, pH 6.8, 5% glycerol, 2% SDS, 4% β -mercaptoethanol, 0.0025% bromophenol blue) and heated at 100 °C for 5 minutes. Cells were then pelleted and the supernatant transferred to a fresh tube and stored at -20 °C for further analysis.

2.7.15 Western blots

Three western blots were produced, one for Dam methyltransferase induction, one for M.EcoGII induction and one for chromatin extraction. A 12.5% Laemmli gel was used for the Dam and M.EcoGII experiments (see table 2.16), while a 12% Laemmli gel was used for the chromatin extractions. For all, isopropanol was used as a buffer on top of the resolving gel while it dried, and was washed off with water prior to addition of the stacking gel. All

used 5 µl Page Ruler (Thermo Fisher Scientific # 26617) and 5 µl of each sample (see section 2.7.14) for Dam, M.EcoGII or the ‘total’ sample in the chromatin extraction, with 10 µl loaded for each other chromatin extraction sample. The gels were run at 120 V for 30 minutes for Dam and M.EcoGII followed by 150 V for one hour 25 minutes for Dam and one hour for M.EcoGII. The chromatin extraction gel was run at 120 V for one hour 25 minutes. The gel running buffer for all was 0.192 M glycine, 0.025 M Tris, 0.1% SDS. The polyvinylidene fluoride (PVDF) membrane (Amersham Hybond-P 0.45 µm pore size VWR #10600023) to be used for transfer was soaked in methanol for 30 seconds, rinsed in water and soaked in transfer buffer (48 mM Tris, 39 mM glycine, 0.036% SDS, 20% methanol) for a few minutes. Transfer with a BioRad Mini Trans-Blot Electrophoretic transfer cell involved 30 V on ice for two hours for Dam, 30 V overnight in a cold room for M.EcoGII or 100 V at room temperature for one hour for chromatin extractions. To visualise overall protein loading the membrane was washed in water and stained with Ponceau Red (0.1% Ponceau Red powder, 5% acetic acid). Washes were one quick and three five-minute washes each time a wash is mentioned. First washes were with phosphate-buffered saline with 0.1% Tween (PBST). At this point the M.EcoGII membrane was cut horizontally at 25 kDa to facilitate probing with different antibodies. Multiple antibodies were probed on the chromatin membrane by successive incubation and detection steps (tubulin followed by histone H3). Blocking was performed at room temperature for one hour with 10% milk for Dam or 5% milk for the others. All milk was made up with PBST. See tables 2.14 and 2.15 for details on antibodies used. Primary antibodies were incubated in 2 ml milk (5% for Dam and M.EcoGII, 1% for chromatin extractions) in a 50 ml falcon on rollers for either overnight at 4 °C (Dam and chromatin) or two hours at room temperature (M.EcoGII). Washes were performed with PBST for Dam, 5% milk for M.EcoGII and 1% milk for chromatin extractions. Secondary antibody treatment involved incubation in 2 ml milk (5% for Dam and M.EcoGII, 1% for chromatin extractions) at room temperature for one hour. The membranes were washed

once with PBST prior to protein detection with the GE Healthcare Amersham ECL Prime Western Blotting Reagents chemiluminescence kit (RPN22325K) and LAS4000 machine. The chromatin extraction membrane was subsequently washed with PBST and stored at 4 °C. Later it was washed with fresh PBST, blocked with 5% milk for one hour at room temperature, incubated with the secondary antibody for histone H3 for one hour at room temperature, washed with PBST and detected in the same way as above.

Protein of interest	Antibody	Dilution	Company (mg/ml)	Catalogue number
Dam methyltransferase	anti-mini-AID-tag	1:1,000	MBL	#M214-3
Histone H3	Rabbit anti-histone H3	1:10,000	abcam	#ab1791
M.EcoGII	anti-FLAG	1:2,000	Sigma	#F1804
Tubulin	anti-tubulin	1:5,000	abcam	#ab6161

Table 2.14 Primary antibodies used in western blots

Protein of interest	Antibody	Dilution (mg/ml)	Company	Catalogue number
Dam methyltransferase	Goat anti-mouse HRP	1:10,000	Jackson Laboratories	#115-035-062
Histone H3	Goat anti-rabbit HRP	1:20,000	Jackson Laboratories	#111-035-144
M.EcoGII	Goat anti-mouse HRP	1:10,000	Jackson Laboratories	#115-035-062
Tubulin	Goat anti-rat HRP	1:20,000	Jackson Laboratories	#112-035-167

Table 2.15 Secondary antibodies used in western blots

Gel type	Reagent	Stock concentration	Volume
Resolving gel	Water	-	3.76 ml
	Tris pH 8.8	1.5 M	3 ml
	Bis-acrylamide	37.5:1 30%	5 ml
	SDS	10%	120 µl
	APS	10%	12 µl
	TEMED	-	4.8 µl
Stacking gel	Water	-	1.69 ml
	Tris pH 6.8	0.5 M	0.75 ml
	Bis-acrylamide	37.5:1 30%	0.5 ml
	SDS	10%	30 µl
	APS	10%	30 µl
	TEMED	-	3 µl

Table 2.16 Western blot gel

Reagent volumes used to make a 12.5% Laemmli (Tris-glycine) gel. TEMED from ThermoFisher Scientific #17919. 37.5:1 30% bis-acrylamide from Sigma #A3699.

2.7.16 Mass spectrometry

Sample preparation

DNA was extracted as in section 2.7.3. Then 1 µg of DNA was added to DNA hydrolysis buffer (100 mM NaCl, 20 mM MgCl₂ and 20 mM Tris-HCl, pH 7.9) to a total of 50 µl. Then 50 µl digest mix (0.1 µl Benzonase, 5 µl Phosphodiesterase, 0.05 µl Alkaline Phosphatase, 0.13 µl 7.2 mg/ml EHNA, 0.16 µl 2.2 mM DFOM, 45 µl DNA hydrolysis buffer) was added and incubated at 37 °C for four hours to digest the DNA to nucleosides. Next 400 µl HPLC water was added to an Amicon ultra-0.5 (3 kDa) column and centrifuged for 15 minutes at 20,238 g, then the collection tube was replaced, sample made up to 400 µl with HPLC water and added to column. This was centrifuged for 15 minutes at 20,238 g to separate the enzymes (retained in the column) from the sample (flow-through). The sample was then lyophilized in a GeneVac miVac DNA concentrator at 65 °C for one hour.

Mass spectrometry analysis

Mass spectrometry was performed by Paolo Spingardi and method information provided by him. The lyophilisate was resuspended in 50 µl of buffer A (10 mM ammonium acetate, pH 6), spiked with 12.5 nM mdA +3 Da. This was transferred into an LC-MS vial for analysis.

For the analysis by HPLC–QQQ mass spectrometry, a 1290 Infinity UHPLC was fitted with a Zorbax Eclipse plus C18 column, (1.8 µm, 2.1 mm 150 mm; Agilent) and coupled to a 6495a Triple Quadrupole mass spectrometer (Agilent Technologies) equipped with a Jetstream ESI-AJS source. The data were acquired in dMRM mode using positive electrospray ionisation (ESI1). mdA was quantified by mass spectrometry and HPLC-UV, whereas dC, dG, dT and dA were quantified solely by HPLC-UV.

Next, 5 µl of the sample was injected. The gradient used to elute the nucleosides started with a 5 minute isocratic gradient composed of 100% buffer A (as above) and 0% buffer B (100% methanol) with a flow rate of 0.400 ml min⁻¹ and was followed by the subsequent steps: 5–8 minutes with 94.4% buffer A; 8–9 minutes with 94.4% buffer A; 9–16 minutes with 86.3% buffer A; 16–17 minutes with 0% buffer A; 17– 21 minutes with 0% buffer A; 21–24.3 minutes with 100% buffer A; 24.3–25 minutes with 100% buffer A. The gradient was followed by a 5 minute wait to re-equilibrate the column. The AJS ESI settings were as follows: drying gas temperature 230 °C, the drying gas flow 14 lmin⁻¹, nebulizer 20 psi, sheath gas temperature 400 °C, sheath gas flow 11 lmin⁻¹, Vcap 2,000 V and nozzle voltage 0 V. The iFunnel parameters were as follows: high pressure RF 110 V, low pressure RF 80 V. The fragmentor of the QQQ mass spectrometer was set to 380 V and the delta EMV set to +200.

The raw mass spectrometry data was analysed using the MassHunter Quant Software package (Agilent Technologies, version B.08.01). The transitions and retention times used

for the characterization of nucleosides and their adducts are summarized in table 2.17. For the identification of compounds, raw mass spectrometry data was processed using the dMRM extraction function in the MassHunter software.

Compound	Adduct	Precursor ion (m/z)	MS1 resolution	Product ion (m/z)	MS2 resolution	Retention time (minutes)	Dwell time (ms)	CE (V)	CA (V)	Polarity
mdA	mdA-H ⁺	266	Unit	150	Unit	19.1	500	12	4	Positive
	mdA-Na ⁺	288	Unit	172	Unit	19.1	500	12	4	Positive
mdA + 3 Da	mdA + 3 Da-H ⁺	269	Unit	153	Unit	19.1	500	10	4	Positive
	mdA + 3 Da-Na ⁺	291	Unit	175	Unit	19.1	500	10	4	Positive

Table 2.17 Mass spectrometry adducts

Table from Paolo Spingardi of the m6A adducts investigated in the HPLC-QQQ mass spectrometry.

2.7.17 Nanopore sequencing

Nanopore sequencing library preparation for all experiments was performed using the 1D Native barcoding genomic DNA (with EXP-NBD104, EXP-NBD114, and SQK-LSK109) protocol with Ligation Sequencing Kit (SQK-LSK109) and either EXP-NBD104 or NBD114 barcoding kits in addition to a MinION R9.4 flow cell. The only experiment that was not barcoded was the G1 α -factor arrest of M.EcoGII inducible *S. cerevisiae*. The library for this experiment was prepared according to the Genomic DNA by Ligation (SQK-LSK109) protocol from Oxford Nanopore Technologies. 1 μ g of input DNA for each sample was used as input to the library preparation.

2.7.18 Nanopore data analysis

All data was basecalled with Guppy, versions detailed below. Modified base data was outputted into new fast5 files by Guppy basecalling. The read id's and modified base probability tables were subsequently extracted from these fast5's and stored in protocol buffers (protobufs) to limit the space required and to prevent duplication of information storage. Both of these tasks were performed with a 'Guppy_extract_from_fast5_new.py' script written by Heather Jeffery and stored in the Github repository (section 2.9). For

barcoded sequencing runs, the protobufs were reorganised into barcode-specific protobufs using the ‘barcode_protobuffers.py’ script. Each protobuf was given a maximum number of reads (1,000) to reduce memory usage when accessing data from these files, thus creating multiple protobufs per barcode. There are six columns in the modified base table (A, 6mA, C, 5mC, G, T). There is a row for every position within each read. In the modified base tables each position in every read is given a probability of being either methylated adenine or methylated cytosine. Each position in the table has a value between 0 and 255, and for each position the probabilities of being a canonical base (i.e. A) or modified (i.e. 6mA) will add up to 255. For example, A plus 6mA will equal 255, C plus 5mC will equal 255, G will be 255 as there is no modified option, same for T.

Another output from Guppy basecalling were fastq files which were concatenated into one file per barcode using this command:

```
cat *.fastq.gz > sample.fastq.gz.
```

The individual fastq’s were subsequently deleted with:

```
rm fastq*.fastq.gz
```

For basecalling and modified base detection with Guppy, version 3.3.0 and version 3.6.0 were used for analysis of the Dam methyltransferase and SMAC-sequencing data [116] (as detailed in chapter 4 figures). Guppy version 3.6.0 was used for *M.EcoGII in vitro* and *in vivo* experiments with the exception of the α -factor arrest experiment for which version 4.4.0 was used. For version 3.3.0 analysis the dna_r9.4.1_450bps_modbases_dam-dcm-cpg_hac.cfg configuration file provided with Guppy was used. For version 3.6.0 and above the res_dna_r941_min_modbases-all-context_v001.cfg configuration file from Oxford Nanopore Technologies Rerio research models was used. Mapping of the fastq reads to the sacCer3 reference (from UCSC genome browser) was performed with Minimap2 version 2.17-r941 [117] with the following command:

```

minimap2 -ax map-ont -t 50 sacCer3.fa sample.fastq |
samtools view -Sb - | samtools sort - sample_alignments.sorted

```

2.8 Other bioinformatics

Benchling was used to design primers, select restriction enzymes and align Sanger sequencing results to a reference sequence. Samtools (v1.6) [118] was used to count reads and index bam files. Bedtools (v2.26.0) [119] was used to calculate distances between potential sites of modification and biological regions of interest including replication origins and nucleosomes. BLAST (Basic Local Alignment Search Tool) [120] was used to identify species of nanopore sequencing reads that did not map to the sacCer3 reference. See table 2.18 for links.

Tool	Website
Bedtools	https://bedtools.readthedocs.io/en/latest/
Benchling	www.benchling.com
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Minimap2	https://github.com/lh3/minimap2
Python	https://www.python.org/
R	https://www.R-project.org/
Rerio basecalling research models	https://github.com/nanoporetech/rerio
Samtools	http://www.htslib.org/doc/samtools.html
UCSC genome browser	https://genome.ucsc.edu

Table 2.18 Bioinformatic software websites

2.9 Code availability

See appendix 1 for commands used to analyse and plot data for each figure.

See https://github.com/HJeffery/Thesis_supplementary for custom scripts.

Scripts were written in python version 3.8.6 or R version 4.0.2.

Chapter 3

Use of psoralens to probe *S. cerevisiae* chromatin structure *in vitro*

3.1 Introduction

The first part of this thesis describes a chemical approach to modifying accessible DNA. In this chapter, DNA modification with two chemicals, 4,5',8-trimethylpsoralen (TMP) and angelicin, will be described.

3.1.1 Chemical crosslinking

Chemical crosslinking is the formation of a covalent bond between two or more molecules. It is commonly used to maintain interactions between proteins and DNA during extraction from cells as membranes are permeable to the crosslinking agent, formaldehyde. This allows biological interactions to be captured in an *in vivo* setting and later probed in an *in vitro* setting. For example, ChIP-seq uses formaldehyde crosslinking to identify where proteins are binding to DNA in populations of cells.

3.1.2 Furanocoumarins

The furanocoumarin class of compounds, to which TMP and angelicin belong, can modify DNA at particular sites, leading to cell death. They originate from plants [121–123] where they are a defence mechanism against fungi, insects and predators. These compounds are structurally characterised by a furan ring (C_4O_1) and a coumarin compound ($C_9H_6O_2$), resulting in a final structure of three aromatic rings. Two important families in the furanocoumarin compound class are the linear furanocoumarins and the angular furanocoumarins, for which psoralen and angelicin are the parent compounds respectively (figure 3.1a). This difference in structure, being either linear or angular, leads to different DNA modifications (figure 3.1b and c). Both psoralen and angelicin chemically modify DNA by forming covalent bonds with thymine bases when exposed to UV light. Psoralen can form chemical crosslinks between DNA strands while angelicin only modifies one DNA strand.

Psoralen can react with DNA in the presence of long wavelength UV light (UV-A) between 320 and 400 nm, with a wavelength between 360 nm and 366 nm most commonly used [124–127]. The psoralen reaction with DNA consists of three steps. Firstly, the psoralen molecule intercalates between the DNA strands in the double helix. Secondly, upon UV-A activation, the psoralen molecule covalently binds to the 5,6-double bond of a thymine [128] in one strand of DNA (figure 3.1b). UV light itself can also cause neighbouring pyrimidines (cytosine and thymine) to dimerize [129]. Finally, if it is a linear furanocoumarin and a furan-side monoadduct was formed, it may covalently bind to the other strand, forming a crosslink between the two DNA strands (figure 3.1c). This also requires another thymine to be present close by on the unmodified strand. Each molecule of psoralen or angelicin can intercalate in either orientation, thus if the resulting modification is a monoadduct, either strand may be modified. Therefore, each binding event is independent. An inter-strand

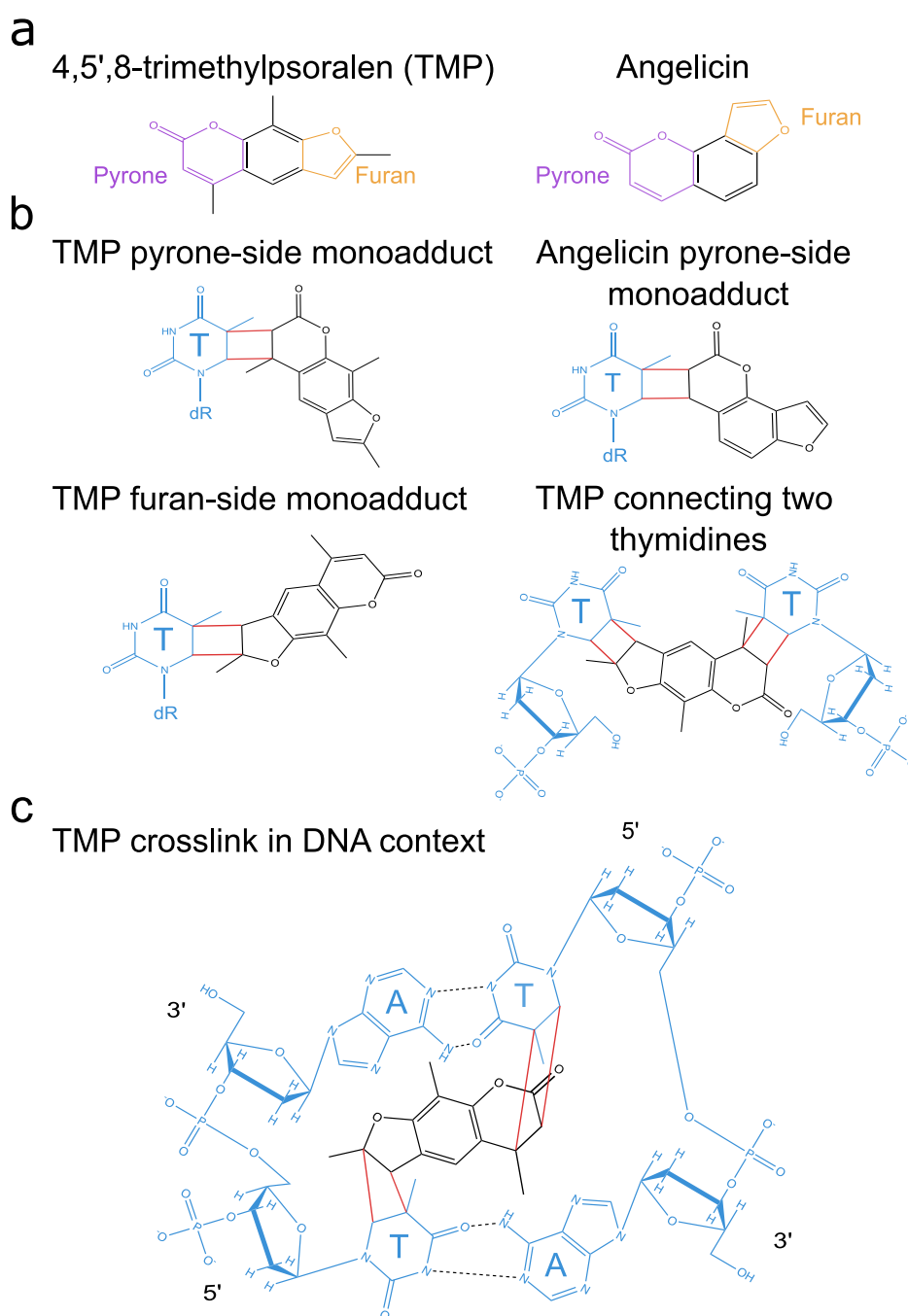


Fig. 3.1 Furanocoumarin chemical structures

a) Structures of a linear furanocoumarin (4,5',8-trimethylpsoralen (TMP)) and an angular furanocoumarin (angelicin). **b)** Structures of the furanocoumarin chemicals bound to a thymine to make a monoadduct with either the pyrone or the furan ring of the furanocoumarin. The TMP compound can bind in two positions while the angelicin compound can only bind in one position. **c)** Representation of a TMP crosslink at a TA site, indicating how the TMP covalently binds to the thymine residues, positioning itself between base pairs and distorting the DNA shown by the long bond between the carbon and oxygen connecting the DNA bases on the right. Furanocoumarin structure = black; pyrone ring = purple; furan ring = orange; DNA = blue; hydrogen bond = black dashed line; covalent bond = red line.

crosslink can also be reversed, with either shorter wavelength UV (~250 nm) or alkali and high temperatures, to reform a monoadduct [113, 124, 128, 130, 131].

For both chemicals, TMP and angelicin, the initial covalent binding reaction preferentially occurs at thymine bases with strongest preference for a thymine in a TA site, with modifications occurring at AT sites at a lower frequency, and TG and GT sites at even lower frequencies [113, 125, 132]. The rate of crosslinking at these dinucleotide sites can also be affected by flanking bases up to 14 bp from the dinucleotide site [125].

Psoralens are used in the medical world in the form of psoralen ultraviolet A (PUVA) therapy. This is used for the treatment of a range of skin conditions including psoriasis [133], parakeratosis variegata [134], light sensitive disorders [135] and vitiligo [135]. Psoriasis involves rapidly growing skin cells while vitiligo is a loss of melanin production resulting in pigmentation loss. PUVA therapy using 8-methoxypsoralen (8-MOP), a linear psoralen, was first used to treat vitiligo in 1953 [135] and psoriasis in 1979 [134]. PUVA increases cell death by inducing p53-independent apoptosis in cutaneous T-cell lymphoma cell lines [136]. Furthermore, interstrand crosslinks block transcription, leading to Ser-15 phosphorylation of the tumour suppressor p53, inducing apoptosis preferentially in S-phase. At high doses the psoralen interstrand crosslinks may also block DNA replication, therefore further induce p53 and apoptosis. Monoadducts were less effective at producing an apoptotic response [137]. These effects on the cell would counteract the excess growth of cells observed in psoriasis. A comparison of TMP and 8-MOP showed TMP to react with DNA 10x faster than 8-MOP, and indicated that 8-MOP is less specific to TA sites within the DNA [124].

3.1.3 Furanocoumarins as markers of chromatin accessibility

Psoralen modification of DNA can lead to bends of up to 70° [138, 139], increased DNA rigidity [140] and partial unwinding of the DNA [140]. Additionally, the presence of a

covalently bound psoralen intercalated in the DNA is comparable to the presence of an additional base pair [132, 140]. These factors may affect the regions of DNA suitable for psoralen modification, for example linkers would be more suitable than protein-bound DNA, and psoralen modification may in turn regulate protein binding if performed *in vivo*. The level of protection from psoralen modification may vary depending on how the protein binds the DNA and the resulting DNA structure at the potential furanocoumarin binding site. Furanocoumarins have the potential to modify accessible DNA while protein-bound DNA may be protected from modification. Therefore they could be used to detect chromatin structure.

Psoralen has previously been used as a marker for chromatin accessibility, for which the final readout has been electron microscopy [127, 141] or primer extension assays [113, 127] (figure 3.2a). Primer extension assays explore the level of modification in a population of cells while electron microscopy provides a good proof of concept that psoralen modification can be used to detect protein binding sites on single molecules. Two independent electron microscopy studies by Hanson *et al* and Wellinger *et al* used 4,5',8-trimethylpsoralen (TMP) to treat either *Drosophila* DNA or the yeast minichromosome YRpTRURAP respectively, with Hanson *et al* also reporting observations of the same patterns in HeLa cells and *in vitro* SV40 chromosomes [141, 127]. They both showed that the psoralen crosslinks could hold the two DNA strands together in denaturing conditions. Hanson *et al* looked generally at protein-bound DNA while Wellinger *et al* specifically showed that nucleosomes could protect DNA from psoralen modification. Hanson *et al* found that even when saturating the chromatin with psoralen and UV-A, at least 60% of the DNA remained non-crosslinked [141], most likely due to the presence of nucleosomes. The amount of protein-bound DNA and the distribution of fragment lengths of the protected regions were in agreement with those obtained from MNase digestions [141]. Therefore, these studies validate the use of TMP, a linear furanocoumarin, as a marker of chromatin accessibility in single molecules.

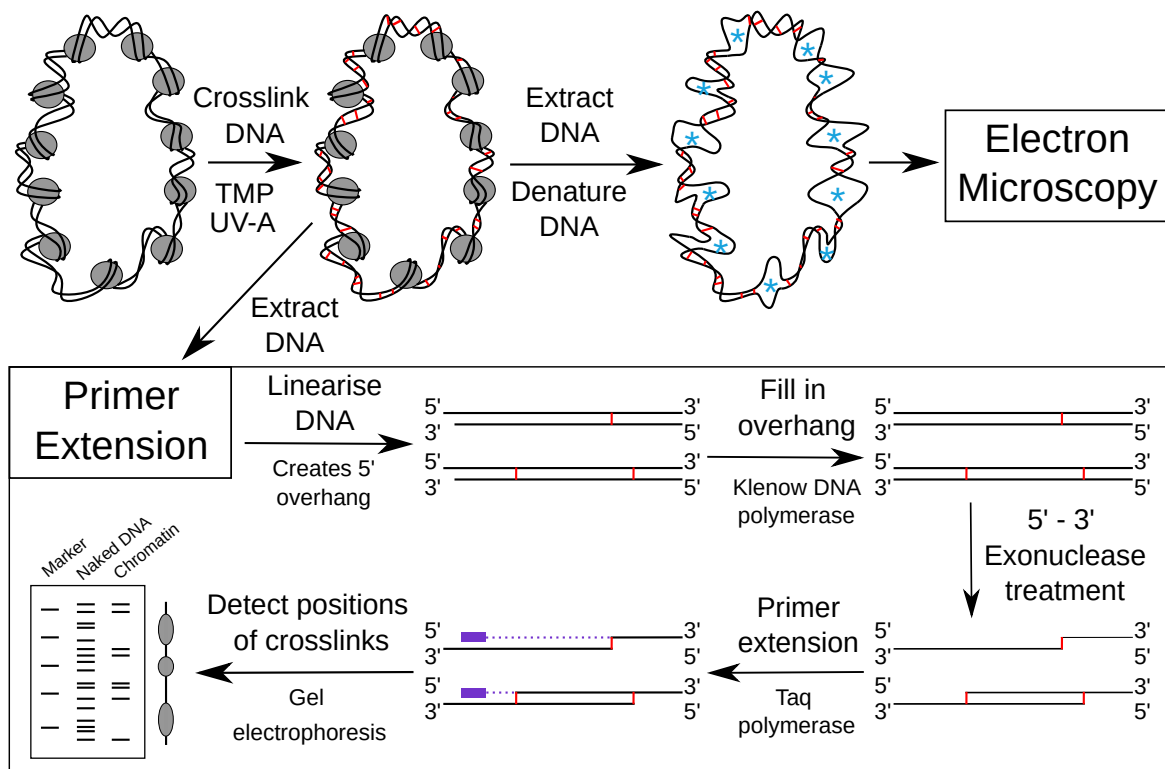


Fig. 3.2 TMP as a marker of chromatin accessibility

Schematic of detecting crosslinks in single molecules by electron microscopy, and in populations of molecules by a primer extension assay. Firstly, plasmid DNA is treated with TMP which is activated by UV light. This forms crosslinks between the two DNA strands in accessible DNA where proteins were not bound. Then the DNA is extracted and either visualised by electron microscopy under denaturing conditions or undergoes primer extension steps (shown with two example DNA molecules with either 1 or 2 crosslinks). In electron microscopy, the two DNA strands separate when denatured if not crosslinked, from which it can be inferred that bound proteins had protected the DNA at these locations. Black lines = DNA; grey circles = nucleosomes; red lines = TMP crosslinks; blue asterisk = nucleosome positions; purple blocks = primers; purple dotted lines = new DNA. Schematic based on [127]).

Angelicin and TMP have both been used as a marker of chromatin accessibility in primer extension assays where nucleosome binding sites have been detected in a population of molecules by reduced signal at the bound TA sites [113, 126, 127]. A primer extension assay to detect modified locations involves DNA modification, extraction, linearisation (if starting with a plasmid), overhang removal, exonuclease treatment, primer extension and visualisation of fragment sizes by gel electrophoresis, potentially as part of a southern blot (figure 3.2). This detects the modified locations because TMP crosslinks and TMP or angelicin monoadducts act as physical barriers to the exonuclease and polymerase [126]. Therefore, the resulting primer extension product starts at the selected primer and terminates at the crosslink site. As the second strand synthesis ends when a modification is hit, only one modification per molecule is detected. Therefore, a population of molecules is required to infer protein-binding sites and the overall modification level needs to be low, preferably one or two modified sites per molecule, to allow a range of modification sites to be detected across the population of molecules. This would result in primer extension products of a variety of lengths. This contrasts to electron microscopy and the new single molecule chromatin accessibility method discussed in this thesis, where a high number of modifications per molecule are vital. In a population of cells, primer extension assays can give an insight into the genomic location of the modifications and can therefore inform which regions of DNA are accessible and which are protein-bound. However, this method is limited to specific regions, due to the need for specific primers, which prevents its application to repetitive regions of the genome such as the rDNA, centromeres and telomeres.

An advantage of using TMP or angelicin as a chromatin accessibility marker is that they are not endogenous to most eukaryotes. Therefore, the technique could be applied to multiple model organisms or cell cultures, with the exception of plants where furanocoumarins are found naturally [121–123]. One important consideration of the application of this technique to eukaryotes is the requirement to pass through both cellular and nuclear membranes. These

have been shown to be permeable to psoralens [142], therefore *in vivo* biological insights could potentially be gained. The *in vivo* approach is also reasonably rapid, as incubation of *S. cerevisiae* cells with TMP for 5 minutes followed by a maximum of 12 minutes UV-A led to successful DNA crosslinking [127]. Short treatment durations will be important for the new nanopore sequencing single molecule chromatin accessibility method presented in this thesis. They facilitate the capture of biologically accurate snapshots of *in vivo* chromatin states, as the longer the treatment takes, the more DNA binding events may change and the more challenging they will be to detect.

It has been shown that electron microscopy and primer extension assays can be used to detect furanocoumarin modification of DNA however, these methods have different strengths and weaknesses (summarised in table 3.1). Both of these methods have shown that furanocoumarins can be used effectively as a marker of accessible chromatin. However, electron microscopy studies don't provide information regarding the genomic position while primer extension assays can only provide insightful information when used on a population of molecules. By combining furanocoumarin modifications with nanopore sequencing, it should be possible to retain the location information for each molecule and to look at a large number of modifications on each molecule. In addition, nanopore sequencing provides high coverage across the whole genome, whereas primer extension assays are limited to single loci. Application of nanopore sequencing would therefore facilitate insights into specific protein binding events across genomic locations of interest and in individual cells within a population.

Status	Detection method	Molecules analysed	Modification type	Genomic position	Scale
Published [113, 126, 127]	Primer extension assay	Population-based	Monoadducts and interstrand crosslinks	Yes	Single loci
Published [127, 141]	Electron microscopy	Single-molecule	Interstrand crosslinks	No	10s of molecules
Proposed	Nanopore sequencing	Single-molecule	Monoadducts	Yes	>10 ⁶ molecules across the whole genome

Table 3.1 Detection methods for furanocoumarin interstrand crosslinks or monoadducts

Summary of the methods available to detect furanocoumarin modification of DNA in terms of the requirements and information that can be obtained from each.

3.1.4 Nanopore sequencing of modified DNA

Nanopore sequencing involves DNA being pulled through a helicase and a pore in a membrane. Therefore, the modified DNA needs to fit within both the helicase and the pore. B-form DNA has a width of 2.37 nm. Only one strand will be going through the pore at any one time. Therefore, the addition of a furanocoumarin modification, which could add an additional 1.5 nm [139], should not prevent the DNA from passing through the 3.5 nm inner pore.

Previously, furanocoumarin modified DNA has not been nanopore sequenced, however other modifications have including small modifications like cytosine or adenine methylation [107, 108] and BrdU [106], in addition to a range of thymine analogs of increasing size [110] and PEG-labeled nucleotides [143]. If the modification has a large effect on the nanopore signal, there may be a negative impact on the ability to map modified reads to the reference genome, thus there may be a bias against large modifications. In terms of molecular weight,

TMP and angelicin monoadducts fit in the middle range of modifications that are detectable by nanopore sequencing (figure 3.3).

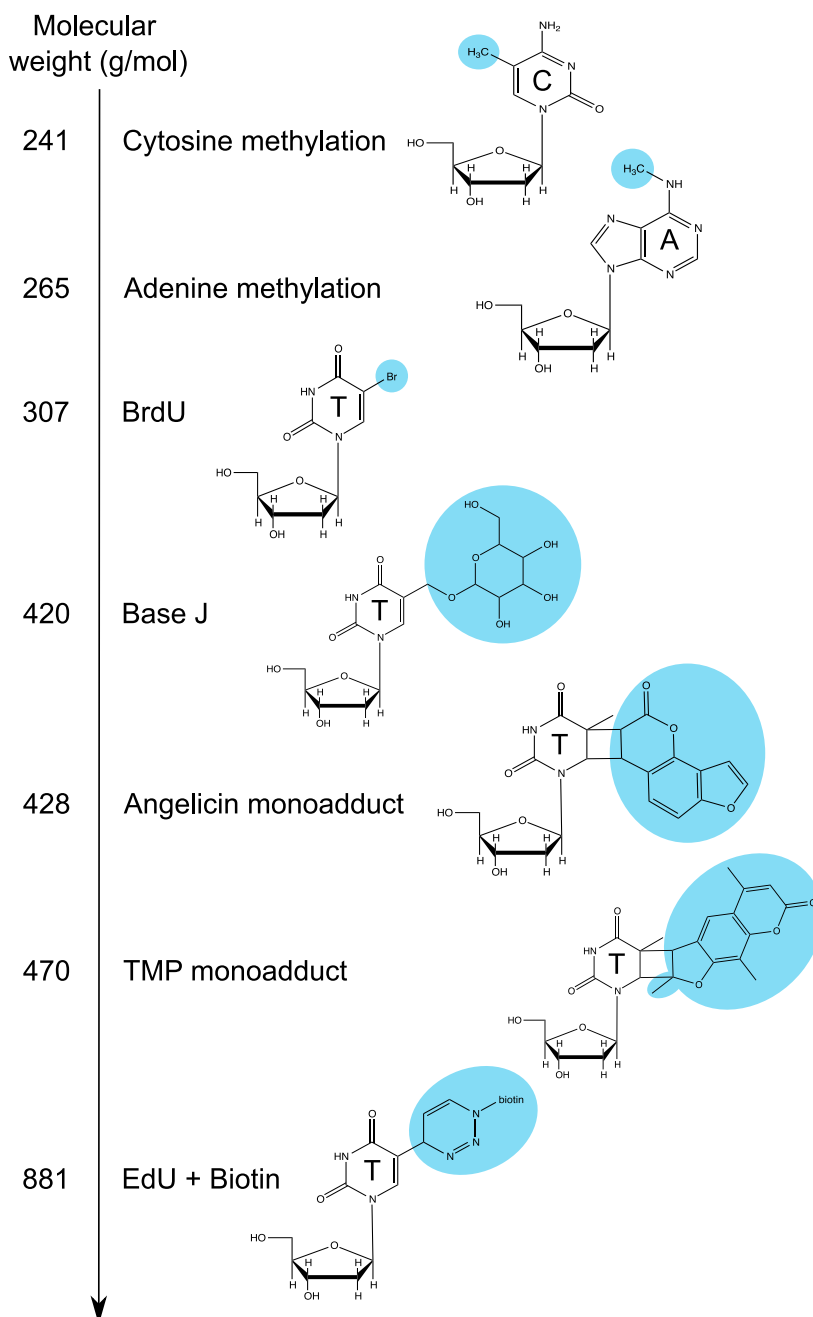


Fig. 3.3 Structures of base modifications

Examples of common base modifications ordered by molecular weight. Some have been successfully detected by nanopore sequencing (cytosine methylation [107], adenine methylation [108], BrdU [106] and biotin [110]). Detection of Base J (a DNA modification found in kinetoplasts), angelicin monoadducts and TMP monoadducts by nanopore sequencing have not been reported.

3.1.5 Aim of this chapter

This chapter aims to determine whether or not furanocoumarins (in particular TMP and angelicin) are good markers for chromatin accessibility in single molecules, with detection performed by nanopore sequencing. This requires investigation of three main aspects: how suitable modification at TA sites would be for the proteins of interest, what the appropriate conditions for chemical modification of DNA would be, and characterisation of these chemical modifications in nanopore sequencing data.

3.2 Distribution of modification sites across the *Saccharomyces cerevisiae* genome

The preferred DNA sequence context for furanocoumarin modification is an adjacent thymine and adenine (in the 5' to 3' direction), also termed a TA site. Across all 3,718,146 thymines in the *S. cerevisiae* genome (using the sacCer3 reference sequence (forward strand only) and excluding the mitochondrial DNA) 23.8% are in a TA site, while 29.0% are in AT sites, which is the second preference dinucleotide for modification. Thymines in an ATA site account for 7.9% of thymines and were counted in both categories, therefore 44.9% of thymines are in either a TA or AT site. The *S. cerevisiae* genome is AT-rich with 61.7% of the DNA being either adenine or thymine, of which 30.8% are thymines. Therefore, there is a 13.8% chance of a randomly selected base being a thymine in a potential modification site (TA or AT). Therefore the number of bases between consecutive modification sites is low with a mean distance of 11 bases between TA sites.

A marker of chromatin accessibility requires sufficient sites across the genome and within regions of interest. Furanocoumarins have been shown to bind at TA, AT, TG and GT dinucleotides, ordered in preference from most to least preferred [113, 125, 132]. Therefore,

the genomic context of these sites was investigated. Using a consensus nucleosome dyad dataset [89], 147 base nucleosome-bound DNA regions were made *in silico*, centered on each dyad. This produced 61,103 nucleosomes and 60,847 linker regions across the *S. cerevisiae* genome, representing 74.4% and 25.6% of the genome respectively. Fewer linker regions were identified than nucleosomes as some consensus nucleosomes directly touch each other. Overall, approximately 75% of all 4 potential modification sites (TA, AT, TG and GT) are present within nucleosome bound sites, leaving approximately 25% of sites in linker regions which may also be other protein-binding sites (figure 3.4a).

3.2.1 Potential modification sites within nucleosome-bound DNA

In the chromatin accessibility assay being developed here, nucleosome-bound DNA would be protected from furanocoumarin modification thus could be identified by a collection of unmodified dinucleotide sites within a range of 147 bases. Most TA/AT/TG/GT dinucleotides are within nucleosome-bound DNA (figure 3.4a) therefore are less than 73 bases from the nearest nucleosome dyad (figure 3.4c). Detection of each individual nucleosome would be dependent on the number of dinucleotide sites it contains. The majority of nucleosomes have between 7 and 13 sites of each dinucleotide, with a mean of 16.4 sites if including both TA and AT or 21.2 sites if including all four dinucleotides of interest (table 3.2 and figure 3.4b). The distance of each dinucleotide to the nearest nucleosome dyad [89] either upstream or downstream was calculated based on the thymine within each dinucleotide. For each dinucleotide the distributions displayed reasonably straight vertical sides from 0 (x axis) to 73 bases (black line) away from the closest nucleosome dyad, indicating that these dinucleotide sites are reasonably evenly distributed within nucleosomes (figure 3.4c). However, there is an increase in the width of the distribution around the nucleosome boundaries (the horizontal line at 73 bases) for both TA and AT sites. This suggests TA and AT occur more frequently at nucleosome boundaries which was not previously reported

in a study analysing TA/AT/AA/TT periodicity in nucleosome-bound DNA [144]. Using a furanocoumarin as a chromatin accessibility marker would involve modification of accessible sites while protein-bound sites remain unmodified. Therefore modification would be targeting less than 25% of the genome, meaning a high true positive rate would be needed for detection of the modifications.

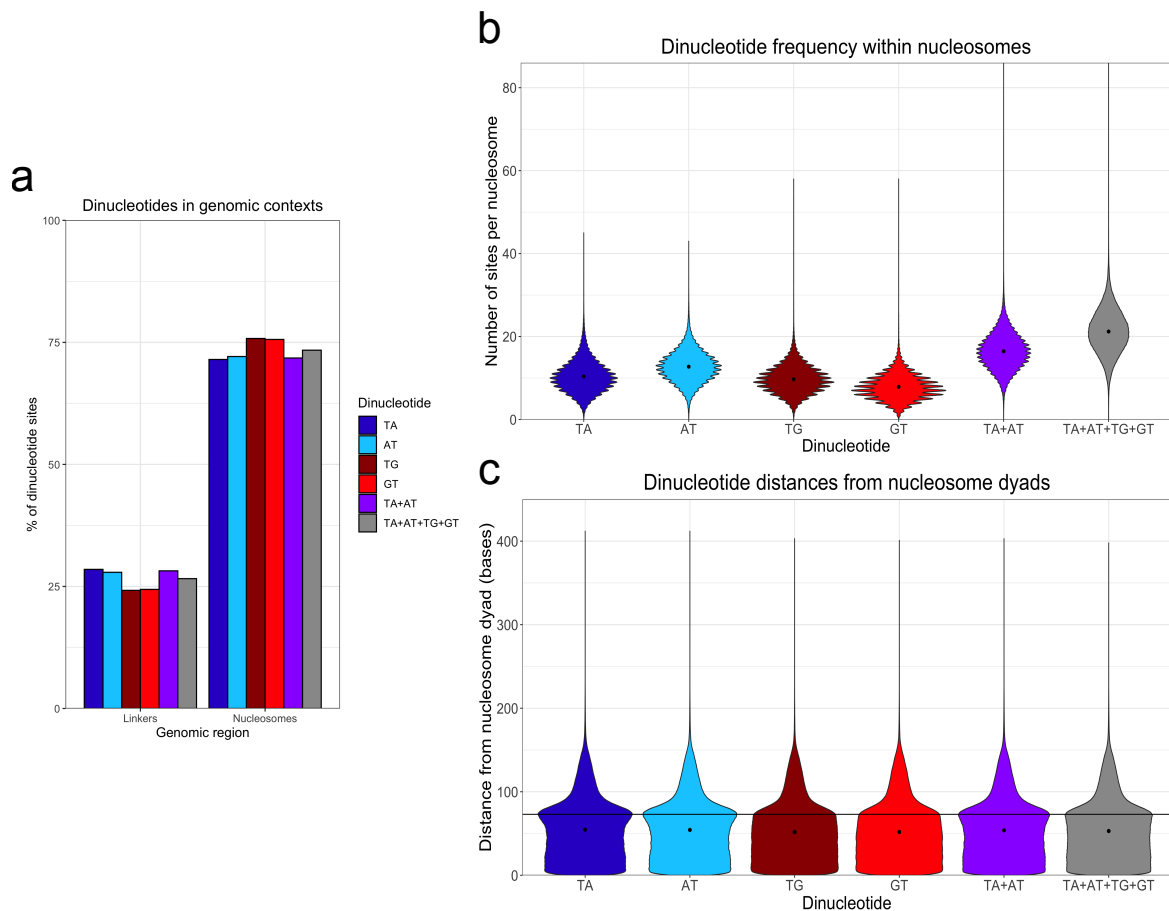


Fig. 3.4 Chemical modification target sites in the *Saccharomyces cerevisiae* genome

All analysis involves the forward strand of the sacCer3 reference genome sequence, excluding the mitochondrial DNA. Nucleosome positions were set as 147 bases centered on consensus nucleosome dyad positions from Jiang *et al* (2009) [89]. Sequence contexts looked at are a thymine in four dinucleotides preferentially bound by furanocoumarins (TMP and angelicin) in the preference order TA > AT > TG > GT [113, 125, 132] and combinations of either the top two preferred dinucleotides (TA and AT) or all four dinucleotides. **a**) Percentage of sites in nucleosome-bound or linker regions for the different sequence contexts. **b**) For each 147 base region of nucleosome-bound DNA, the number of sites of each dinucleotide or combination of dinucleotides was counted and this distribution is shown as a violin plot. Black dot = mean. **c**) Distribution plot of the distance to closest nucleosome dyad (either upstream or downstream) from each dinucleotide site. The horizontal black line is at 73 bases, marking the edge of a 147 base region of nucleosome-bound DNA. Black dot = mean.

3.2.2 Potential modification sites within DNA linkers

Linker regions could be potentially non-protein bound, and thus accessible to DNA modification. Therefore, it is these regions that will need to be detected by nanopore sequencing in the proposed method. Electron microscopy has shown that the majority of linkers can be modified by psoralen in plasmid DNA [127] and *Drosophila* DNA [141]. Therefore, it may be possible to modify the majority of linkers on a genome-wide scale in *S. cerevisiae*, given optimisation of the furanocoumarin concentration and length of UV treatment. The length of linker regions is variable in the *S. cerevisiae* genome, with most linkers being less than 50 bases (figure 3.5a). As most linker regions are considerably smaller than nucleosome-bound DNA regions it is more important to evaluate the number of potential modification sites within each linker. Figure 3.5b shows there is a range of TA and AT sites within linkers, with 4,717 linkers (7.8% of total linkers) containing no TA or AT sites (table 3.2). This correlates with linkers with a mean length of less than 100 bases (figure 3.5b). Furthermore, 5,539 linkers (9.1% of total linkers) contain just one TA or AT site (table 3.2), leaving 50,591 (83.1% of total linkers) with two or more TA/AT sites. If the modification rate is reasonably efficient there should be sufficient sites to identify linkers in most regions, except for regions of particularly dense heterochromatin.

To determine the ability of this method to be used to understand the function of DNA replication origins, these regions were investigated. Replication origins are predominantly in regions of accessible DNA to facilitate the loading of replication proteins on to DNA. Of 252 ACS sites investigated, 181 were in linker regions. This may be lower than expected due to the nucleosome dataset being a consensus generated from asynchronous cultures [89] whereas origin licensing only occurs in the subset of cells that are in G1 phase. These 181 ACS sites tend to be in linkers of above average length, with a mean linker length of 115.2 bases (figure 3.5c) compared to a mean of 50.8 bases for all linkers, supporting the idea that

replication origins are in regions of accessible DNA. In the linkers containing an ACS, there are a mean of 12.3 TA sites and 13.0 AT sites per linker, which is much higher than the mean numbers for all linkers (table 3.2). ACS sites recruit ORC and MCM, but other proteins may also bind in these nucleosome-free regions, thus protecting some of the surrounding modification sites. The modification sites are relatively evenly distributed across the linkers, with only a few linkers having large regions without any potential modification sites (figure 3.6). Therefore, there should be adequate modification sites for this method to provide some biological insight with regards to origin licensing.

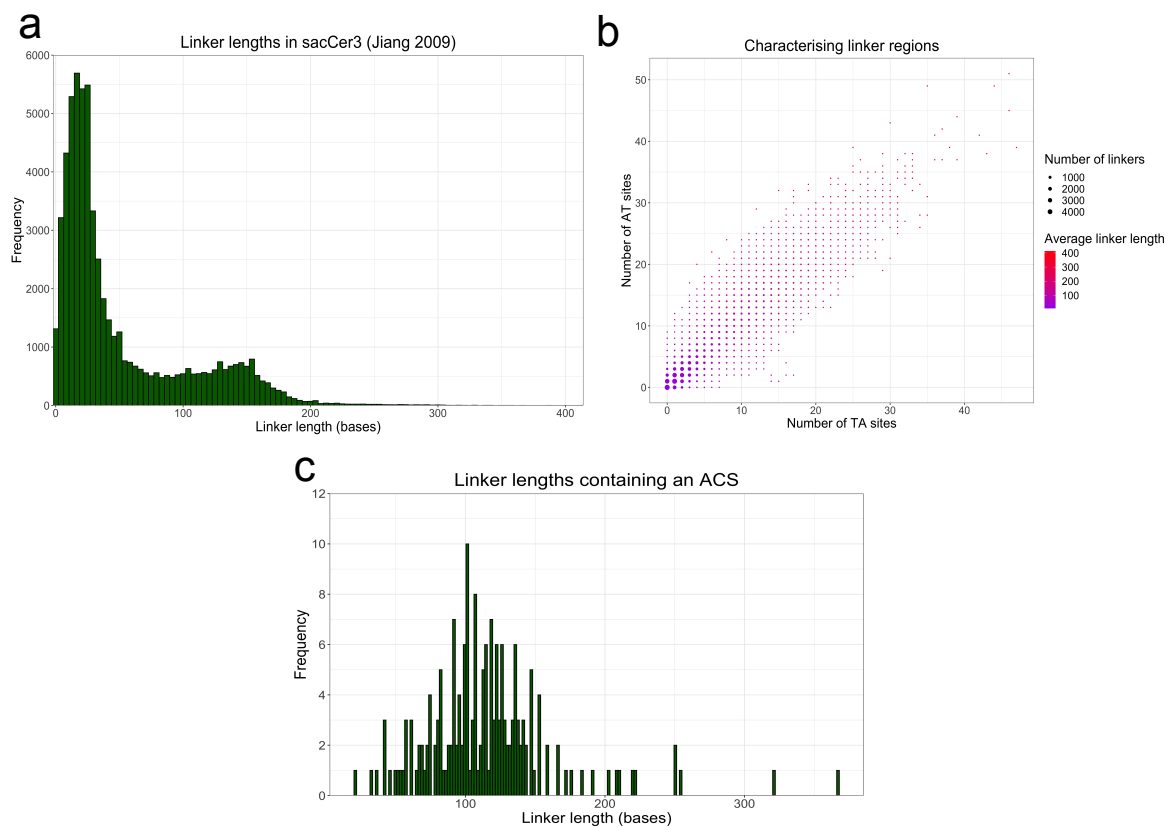


Fig. 3.5 Characterisation of linker regions in the *Saccharomyces cerevisiae* genome

All analysis involves the forward strand of the sacCer3 reference sequence excluding the mitochondrial DNA. 147 base nucleosome positions were generated using the consensus nucleosome dyad positions from Jiang *et al* (2009) [89], ACS sites were from wild-type *S. cerevisiae* in G2 phase, grown at 23 °C (GSM424494), published by Eaton *et al* (2010) [7]. **a)** Length of linker regions in sacCer3 between consensus nucleosome positions. **b)** Characterisation of linker regions, with the number of TA (most preferred dinucleotide for furanocoumarin modification) or AT (second preferred dinucleotide for furanocoumarin modification) sites and the correlation with the number of linkers and mean linker length. **c)** Length of the 181 linker regions that contain an ACS site.

Feature	Bases	Statistic
Number of sites in the genome	T	3,718,146
	TA	884,876 (23.8% of all T's)
	AT	1,079,128 (29.0% of all T's)
	TG	780,450 (21.0% of all T's)
	GT	634,941 (17.1% of all T's)
	TA+AT	1,377,762 (37.1% of all T's)
	TA+AT+TG+GT	1,750,235 (47.1% of all T's)
Number of nucleosomes	-	61,103
Number of linkers	-	60,847
Mean length of linkers (all/with ACS)	-	50.8 bases / 115.2 bases
Mean number of sites per nucleosome	TA	10.4
	AT	12.7
	TG	9.7
	GT	7.9
	TA+AT	16.4
	TA+AT+TG+GT	21.2
Mean number of sites per linker	TA	4.1
	AT	5.0
	TG	3.1
	GT	2.5
	TA+AT	6.1
	TA+AT+TG+GT	7.5
Number of linkers	0 TA + 0 AT	4,717 (7.8% of all linkers)
	0 TA + 1 AT	3,503 (5.8% of all linkers)
	1 TA + 0 AT	2,036 (3.3% of all linkers)
	1 TA + 1 AT	4,370 (7.2% of all linkers)

Table 3.2 Chemical modification target sites in the *Saccharomyces cerevisiae* genome

Statistics were calculated using the forward strand of the sacCer3 reference genome, excluding mitochondrial DNA and the consensus nucleosome dataset from Jiang *et al* (2009) [89] with nucleosome dyads made in to 147 base nucleosome regions. Dinucleotide frequencies were calculated in Python (see appendix 1). Bedtools intersect was used to identify the sites within specific genomic regions and the counts were performed in R. Means and percentages are rounded to 1 decimal point.



Fig. 3.6 Potential modification sites in linker regions containing a replication origin

Each black line represents a linker centered around the ACS within it. The linkers are the same subset of 181 linkers analysed in figure 3.5c and are grouped by chromosome. For each linker the chromosome and start position were concatenated to form the linker ID. The grey box represents a 35 base ORC footprint (size from Eaton *et al* (2010) [7]). The two most preferred sites for furanocoumarin modification (TA and AT) are shown in red and blue respectively. ACS positions are from wild-type *S. cerevisiae* in G2 phase, grown at 23 °C (GSM424494), published by Eaton *et al* (2010) [7] and are oriented by the T-rich strand.

3.3 4,5',8-trimethylpsoralen (TMP) modification of DNA

3.3.1 Overview of approach for TMP modification of DNA

In line with previous electron microscopy single molecule chromatin accessibility studies with furanocoumarins [127, 141], TMP was selected as the DNA-modifying chemical. The ability of a TMP crosslink to persist in denaturing conditions [127, 141] indicated that TMP crosslinked DNA should be distinguishable from non-crosslinked DNA on a polyacrylamide denaturing gel [124]. In this case, crosslinked DNA would run at a higher molecular size than the same DNA without the crosslinks, where there is nothing to prevent the two strands denaturing and running separately through the gel. The same principle can be applied to mass spectrometry, where crosslinks would hold the two DNA strands together, thus generating a product of higher molecular weight than the individual DNA strands. This technique, which preserves oligonucleotides in their entirety, will be used in this chapter.

To optimise crosslinking conditions, a substrate with a single TA site was used. This substrate was made from two 18 base oligonucleotides annealed together with a 4 base overhang on each end and has been used in previous TMP studies [124]. By limiting modifications to one site, there are a discrete number of possible options to distinguish between on a gel and by mass spectrometry. The overhangs ('End B' and 'End S' in figure 3.7) present in the annealed oligonucleotides are both palindromic and are identical to the overhangs produced by DNA digestion with BglII and SalI-HF restriction enzymes, thereby allowing a longer section of DNA digested with these enzymes to be ligated to the annealed oligonucleotides. This is important as nanopore sequencing, one of the detection methods used in this study, requires long fragments of DNA.

Polyacrylamide denaturing gels and mass spectrometry can identify the presence of a furanocoumarin modification on DNA but they cannot identify the position of the modification. This may be possible with nanopore sequencing. However, it is currently unknown whether TMP monoadducts can be detected by nanopore sequencing. The crosslinks, as detected by polyacrylamide denaturing gels and mass spectrometry, will need to be further processed in order to fulfil the requirements for nanopore sequencing. These include: being able to separate the two DNA strands in order to pass one through the pore, thus the crosslink must be reversed to form a monoadduct; being double-stranded for adapter ligation, thus the DNA must be reannealed post-crosslink reversal; and the DNA must be of sufficient length (>200 bp) thus a ligation step is required. An overview of the workflow can be seen in figure 3.7.

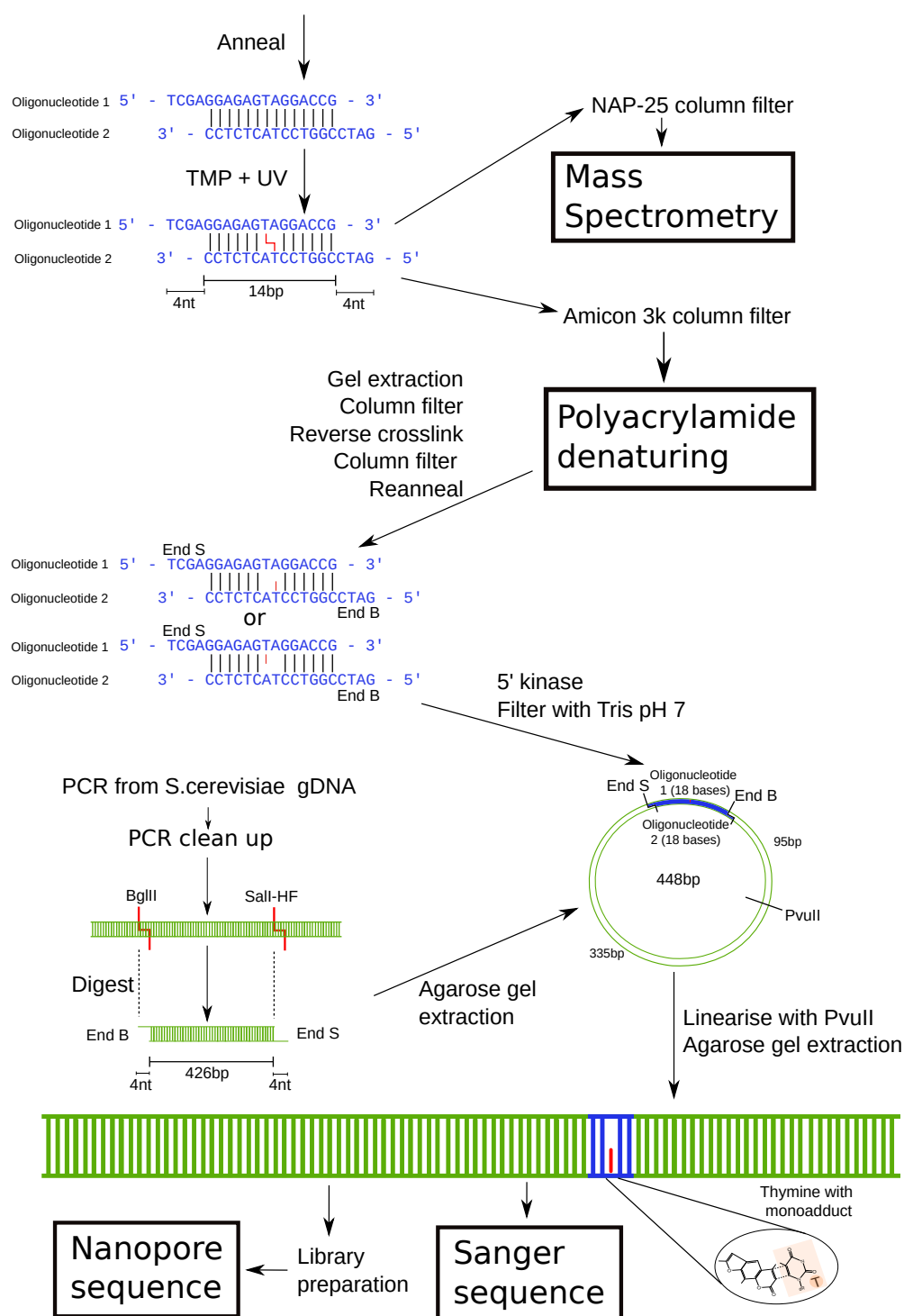


Fig. 3.7 Experimental workflow of TMP modification of DNA

Overview of the experimental approach taken to identify if TMP can modify DNA (test by denaturing gels and mass spectrometry) and whether or not this modification can be detected by nanopore sequencing. Green = large fragment of DNA; blue = insert (18 base oligonucleotides containing one target site for modification by TMP); red = TMP modification in the insert or recognition sites for BglII and SalI-HF in the large fragment; gDNA = genomic DNA; black lines between DNA bases are hydrogen bonds.

3.3.2 Polyacrylamide denaturing gels detect TMP crosslinking

To test the level of TMP modification achieved by the experimental setup used in this study, DNA crosslinks were assessed by polyacrylamide denaturing gels. The substrate is two 18 base oligonucleotides which contain one target (TA) site (figure 3.8a). These oligonucleotides were annealed and mixed with TMP and a crosslinking buffer to allow the TMP to intercalate into the DNA. The samples were then transferred to a 6-well plate on ice where 365 nm UV treatment was performed. The amount of UV was tested using cycles of 15 minutes on, 2 minutes off, with the UV lamp positioned 10 cm above the plate containing the samples (figure 3.8b). The 'off' periods provided an opportunity to safely remove a sample so that a range of UV treatment lengths could be analysed. Between cycles additional TMP, of the same amount as had been originally added, was given to the remaining samples (section 2.6.2). These additional TMP treatments did not have a separate incubation time to intercalate but each UV treatment cycle was longer than the original incubation period so there should have been time for intercalation and DNA binding within each cycle.

Samples that had been subjected to various number of cycles of UV treatment were heat-denatured, snap-cooled and separated on a 15% denaturing polyacrylamide gel. DNA was visualised using SYBR Gold, a sensitive nucleic acid stain. The oligonucleotides contained one TA site and no other sites of lower frequency modification therefore allowing easy determination of crosslink presence or absence. The expected band for non-crosslinked single-stranded oligonucleotides (or single-stranded oligonucleotides with a TMP monoadduct) was observed around 18 bases (figure 3.8c and d). The continued presence of this throughout TMP and 365 nm UV treatment indicates that the crosslinking reaction was not able to modify all the DNA molecules. The band at 36 bases may represent the annealed oligonucleotides or single-strand DNA dimers (figure 3.8c and d) and was observed at a reasonably constant level across all conditions including the annealed sample that was not exposed to any TMP

or UV. This indicated that the sample denaturing process was not 100% efficient. As the number of UV cycles (and thus UV exposure time) increased, a band at 28 bases appeared and increased in intensity (figure 3.8c and d). This band was not seen in the annealed or 0 minute UV sample. Therefore, no crosslinking was observed in the absence of UV treatment, where TMP was inactive, while the level of crosslinking did increase with an increase in length of UV treatment. The different molecular weight observed between the crosslinked material and annealed material may be because the annealed are held together at multiple sites, while the crosslinked material may be fully denatured except for the crosslink, causing it to run through the gel in a more X-like shape.

The increase in UV treatment was also associated with the minor increase of higher weight molecular bands (>50 bases). These may be due to additional crosslinks holding multiple molecules together. This was not investigated further experimentally, but conceptually the following concatemerisation could have occurred due to the presence of an AT site in the 4 base overhang at the 5' end of oligonucleotide 2. An extra molecule of oligonucleotide 2 could bind to the oligonucleotide 1-oligonucleotide 2 duplex due to the palindromic nature of the overhang sequence. TMP could then intercalate and form a crosslink at the AT site. Subsequently, another molecule of oligonucleotide 1 could anneal to the new oligonucleotide 2 molecule, creating a molecule with 2 copies of each oligonucleotide (figure 3.9). These concatemerisation reactions could have happened in any order, however the annealing reaction of oligonucleotide 1 and 2 will be more stable than annealing of two identical oligonucleotides due to a longer sequence of complementary bases. Once there are two copies of each oligonucleotide, with three potential TMP crosslinks (the bottom molecule in figure 3.9), additional molecules could anneal to the resulting 4 base overhangs but they would not remain attached in denaturing conditions as the overhangs (TCGA) do not contain any potential TMP modification sites. This concatemerisation process could explain the high molecular weight bands in figure 3.8c and d. A higher number of cycles

does not show much increase in crosslinked material and the single oligonucleotide band is still present after 180 minutes (12 cycles) of 365 nm UV treatment. Therefore, to optimise for a sufficient amount of TMP-crosslinked DNA, but to minimise the presence of larger concatemers, a range of 4 to 8 UV cycles was used to optimise downstream steps.

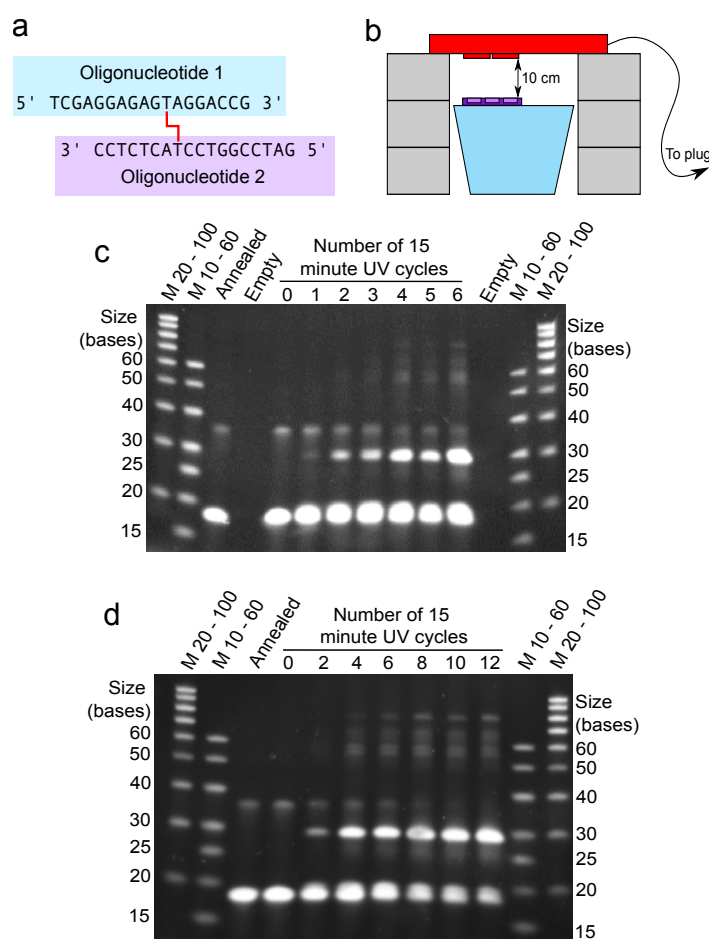


Fig. 3.8 TMP crosslink on DNA can be detected by polyacrylamide denaturing gels

a) Schematic of the two 18 base oligonucleotides crosslinked at the TA site. Red = TMP crosslink. **b)** Schematic of the setup used for UV treatment of the DNA to activate the TMP. Red = UV lamp; blue = ice bucket; purple = 6-well plate containing samples; grey = tip boxes used to position the UV lamp. **c)** Polyacrylamide denaturing gel showing optimisation of the amount of 365 nm UV in 15 minute cycles with 2 minutes of no UV between each cycle. 10 µl of 2 mg/ml TMP was added to each remaining sample between each cycle of UV treatment. The band around 36 bases represents annealed oligonucleotides, the band around 28 bases represents crosslinked oligonucleotides and the band around 18 bases represents individual single-stranded oligonucleotides. **d)** Further optimisation of the number of 15 minute 365 nm UV cycles shown on a polyacrylamide denaturing gel. Same treatment as in c.

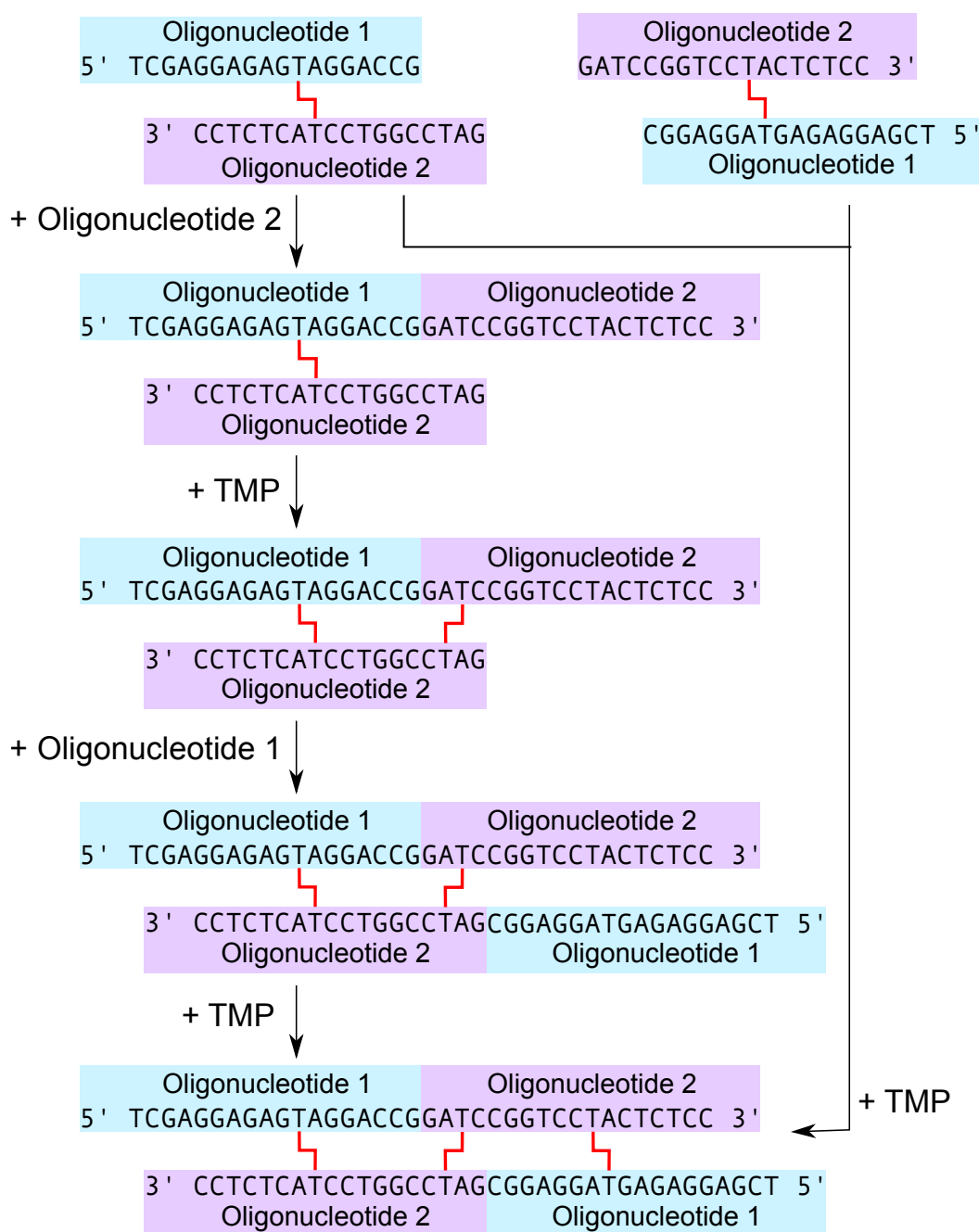


Fig. 3.9 Potential concatemerisation of oligonucleotides

Schematic of the potential concatemerisation sequence of events of the two 18 base oligonucleotides to form higher molecular weight molecules that could persist in denaturing conditions. Red = TMP crosslink.

Furanocoumarins are toxic chemicals which could damage the mass spectrometry equipment therefore TMP was removed prior to further sample processing. The small size of TMP

(228.24 g/mol) compared to the oligonucleotides (approximately 5,500 g/mol each) allowed size separation by filtration with size exclusion columns (figure 3.10c). For this, two types of size exclusion columns (Amicon 3k and NAP-25) were tested, for different spin durations. The NAP-25 columns required four steps (column preparation, column equilibration, sample application and elution) while the Amicon spin columns only required sample application and elution. The polyacrylamide denaturing gel in figure 3.10a shows that the DNA is maintained in the Amicon column and increases in concentration with longer spin times, with 30 minute spin duration being almost qualitatively equivalent to the non-filtered sample. TMP-treated DNA filtered by either column appeared comparable and increased for both columns with increased spin times (figure 3.10b), therefore 30 minutes was selected. As TMP is not directly visualised on a polyacrylamide denaturing gel, it was not possible to identify the loss of this chemical in the filtered samples. However, the retained DNA was observed and it was assumed that the TMP would be lost as Pathak *et al* previously showed that Sephadex G-100 columns eluted DNA and unbound TMP in different fractions [142].

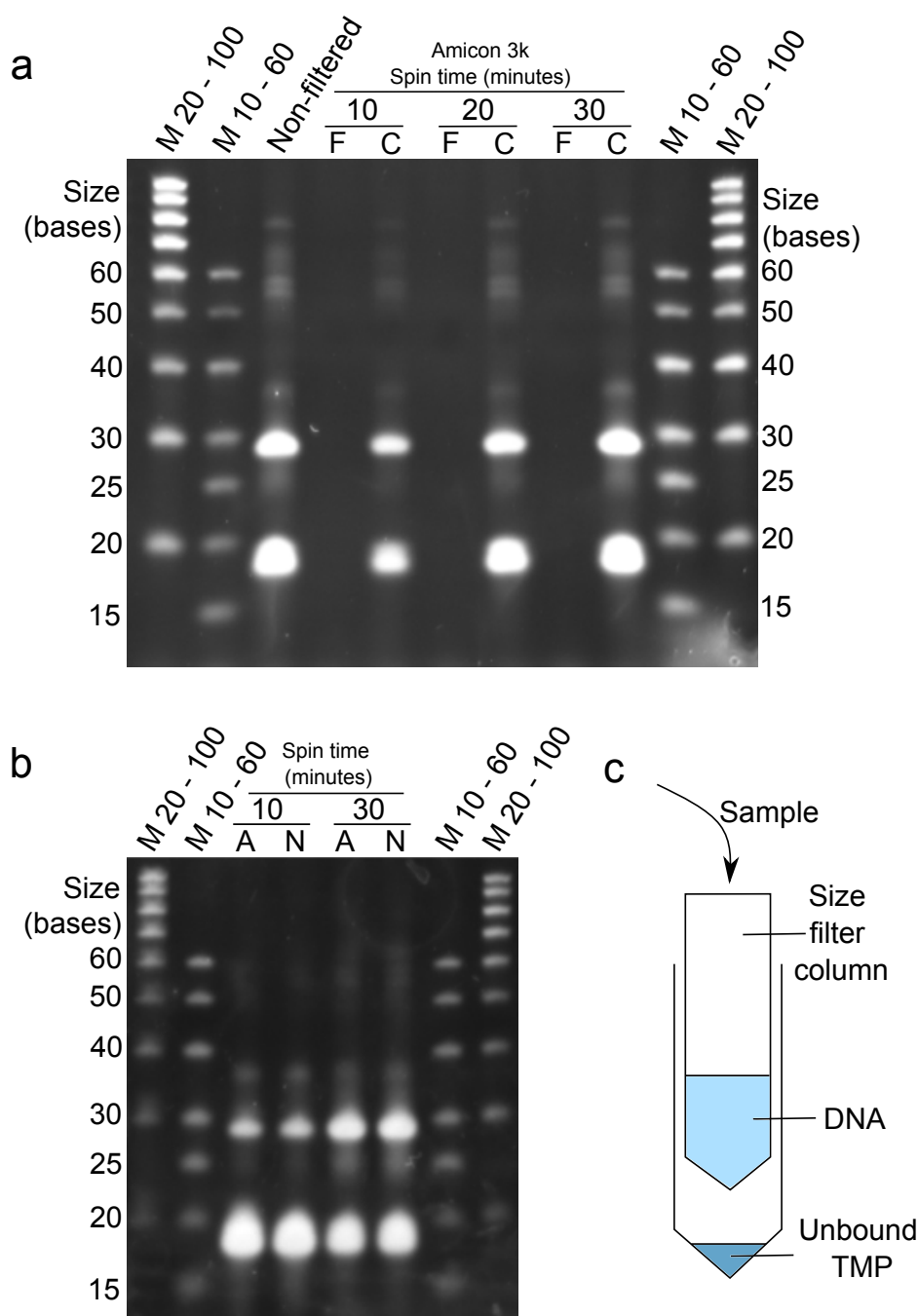


Fig. 3.10 Column filtering retains crosslinked DNA

Samples were column filtered to remove the unbound TMP and to concentrate the DNA. **a)** Optimisation of spin time for filtering the DNA after TMP and UV treatment. The DNA used were annealed oligonucleotides that had undergone 6 x 15 minute 365 nm UV cycles on ice. Spins were performed at 14,000 g and a final spin of 2 minutes at 1,000 g with the column inverted in a new tube was performed for elution of the DNA. F = Filtrate (passed through column); C = Concentrate (retained in column); M = Marker (10/60 and 20/100 Oligo Length Standards from IDT). **b)** Comparison of Amicon 3k columns (A) and NAP-25 columns (N) showed equivalent retention of the DNA in both columns at the two spin durations by similar appearance of bands on a denaturing gel. M = Marker (10/60 and 20/100 Oligo Length Standards from IDT). **c)** Schematic of column filtering to remove unbound TMP and retain DNA.

3.3.3 Mass spectrometry detects TMP modification

The presence of a TMP crosslink has been inferred on polyacrylamide denaturing gels by the presence of a band that appeared only after TMP and 365 nm UV treatment. This was further verified by mass spectrometry which can distinguish molecules by their molecular weights to a higher precision than a polyacrylamide denaturing gel.

Oligonucleotides were annealed, then incubated with or without TMP, and subjected to either 0, 10 or 30 minutes of 365 nm UV treatment. Controls without TMP and UV (in combination and separately) were performed to ascertain the level of DNA damage introduced by either treatment. Compared to previous experiments (section 3.3.2) 5 minute 365 nm UV cycles were used to identify if shorter UV treatment could be used to optimise the quantity of crosslinked material produced versus the risk of DNA damage. The band at 28 bases in figure 3.11a shows that crosslinked material was produced by shorter UV treatments, with a reduction in the number of high molecular weight bands. Samples were then filtered with NAP-25 columns to remove unbound TMP and to concentrate the DNA. Any remaining TMP could be detected by mass spectrometry but analysis of low molecular weight molecules was not performed. To allow all potential products and remaining substrates to be observed in the mass spectrometry data no gel extraction step was performed.

Mass spectrometry, performed by Dr Afaf El-Sagheer, was used to confirm the presence of modified oligonucleotides. Electrospray ionisation was used to ionise the oligonucleotides. This technique has a very low frequency of fragmentation, therefore covalent bonds should be maintained while hydrogen bonds will be broken. Consequently, annealed oligonucleotides will denature but not fragment into individual nucleosides, while oligonucleotides with a TMP crosslink covalently bound between them will remain linked. The electrospray ionisation produces protonated molecules. These are attracted towards a negatively charged plate which

accelerates them through a hole towards the detector. Here they hit a negatively charged plate and displace the positive ions by gaining electrons. The size of the resulting electric current represents the number of ions hitting the plate and thus the number of molecules in the sample with that mass to charge ratio. As electrospray ionisation mostly produces ions with a +1 charge, the mass to charge ratio is comparable to the mass of each ion. The mass of each molecule determines the velocity, with lighter molecules moving faster, thus arriving first at the detector. Therefore, the unmodified oligonucleotides would arrive first at the detector, followed by the oligonucleotides with a TMP monoadduct and then the crosslinked oligonucleotides. The concentration of the molecules cannot be assessed quantitatively as the instrument may have had altered sensitivity over time and no standards were run. Furthermore, the modified and unmodified oligonucleotides were not proven to have the same sensitivity and to ionise in the same way. It is not known how the modification would affect the flight characteristics of the oligonucleotide, therefore this experiment is predominantly useful for identifying the presence or absence of a modification, rather than the exact abundance.

Furanocoumarin	DNA substrate	Molecular weight (g/mol)
None	Oligonucleotide 1	5,612.6 *
	Oligonucleotide 2	5,385.6
TMP	Oligonucleotide 1 with monoadduct	5,840.84
	Oligonucleotide 2 with monoadduct	5,613.84 *
	Oligonucleotide 1 and 2 with crosslink	11,226.44
Angelicin	Oligonucleotide 1 with monoadduct	5,798.76
	Oligonucleotide 2 with monoadduct	5,571.76

Table 3.3 Molecular weight of potential substrates analysed by mass spectrometry

Molecular weights of the potential substrates analysed by mass spectrometry. * = These two molecular weights are very similar, meaning they appear as one peak in the mass spectrometry analysis.

There are five main possible detectable substrates in this mass spectrometry experiment: the two unmodified original oligonucleotides; each oligonucleotide with a single TMP monoadduct; or the two oligonucleotides crosslinked with TMP (table 3.3). Mass spectrometry identified only the unmodified oligonucleotides in the control conditions, indicating that both TMP and 365 nm UV individually do not alter the oligonucleotides to a detectable level (figure 3.11b-d). With both TMP and 365 nm UV treatment the presence of additional peaks representing DNA with monoadducts or crosslinks were observed (figure 3.11e-f). The relative abundance of crosslinked material (shown by the red arrow) increased with a longer UV treatment while the relative abundance of unmodified DNA (blue arrows) or DNA with a monoadduct (orange arrows) reduced with an increase in UV treatment. This is likely to be real but due to the inability of this method to be used for quantitative assessment, it should be treated with caution. This indicates that 10 minutes of UV did not reach saturation and a higher level of crosslinking can be achieved by 30 minutes UV treatment, supporting a similar increase on a polyacrylamide denaturing gel (figure 3.11a). The presence of unmodified DNA in all conditions indicates that the crosslinking reaction was not 100% efficient. Monoadducts were also detected, although the molecular weight of oligonucleotide 2 with a TMP monoadduct was identical to that of unmodified oligonucleotide 1 therefore they appear as one peak and the presence of the two cannot be distinguished from each other. This data shows that 10 and 30 minutes of 365 nm UV treatment with 88 nmol TMP dissolved in DMSO produces TMP-crosslinked DNA. However, there is a considerable amount of unmodified oligonucleotides or oligonucleotides with a monoadduct bound present in both conditions, consistent with results from polyacrylamide denaturing gels.

The 365 nm UV (UV-A) used in this experiment is less likely to damage DNA than UV-B as UV-A is only weakly absorbed by DNA, and instead acts on DNA indirectly through external chromophores [145]. However, UV-A can still cause a variety of types of DNA damage (single-strand breaks, oxidized pyrimidines, oxidized purines (e.g. 8-oxo-

7,8-dihydroguanine) and cyclobutane pyrimidine dimers) which have been detected in a 1:1:3:10 ratio [145]. The most common site for pyrimidine dimers caused by UV-A are TT dinucleotides [145] which are absent from both oligonucleotides. Therefore, the sequences used reduce the risk of UV-A damage. However, signs of DNA damage may be observed in the mass spectrometry results. Firstly, single strand breaks would be detectable as peaks at smaller fragment sizes, for example the peak at 3,742 g/mol is present in the TMP and UV treated conditions and not the controls (figure 3.11), despite one control involving 30 minutes UV-A treatment. Therefore, TMP may increase the susceptibility of DNA to single-strand breaks under UV-A light. Other lower abundance peaks may be results of occasional fragmentation events. Secondly, oxidized bases would produce very small modifications to the overall mass of the fragments thus would not be detected on this resolution scale. Thirdly, pyrimidine dimers form between adjacent bases on the same strand thus would not mimic the double-stranded effect observed by TMP crosslinks. They may cause a small change to the molecular weight, but it would be difficult to detect on the resolution used. There were no detectable signs of DNA damage in the control with 30 minutes UV treatment and no TMP, therefore it may be possible to safely extend the duration of UV treatment, allowing an increase in the level of crosslinked material without damaging the DNA. The presence of the 3,472 g/mol peak in the TMP and UV-A conditions indicates that the combination of treatments may lead to minimal DNA fragmentation which did not increase with UV duration. Therefore, in order to maximise the amount of crosslinked material in future experiments, longer UV treatment and a gel extraction step could be performed.

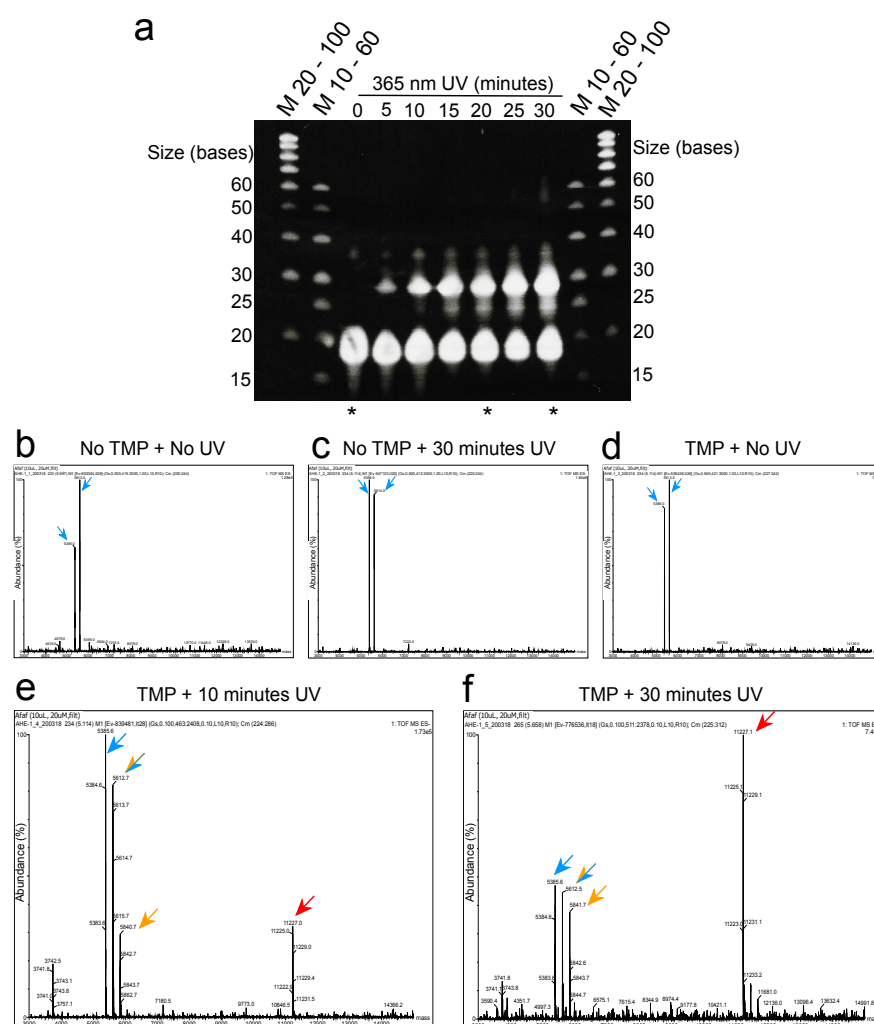


Fig. 3.11 Mass spectrometry detects TMP modification on DNA

A variety of conditions were tested by mass spectrometry, where DNA strands are separated but are not hydrolysed into individual bases. The DNA sample used was two 18 base oligonucleotides with one potential site of TMP modification (a TA dinucleotide). **a)** Annealed oligonucleotides incubated with TMP and crosslinked with 365 nm UV were run on a 15% polyacrylamide denaturing gel. * = samples taken forward into mass spectrometry analysis (samples not treated with TMP are not shown on this gel). **b-d)** Controls for mass spectrometry include DNA with no TMP or UV treatment, DNA with TMP but no UV (therefore the TMP is not activated) and DNA with no TMP but with 365 nm UV treatment (to assess UV damage). Only unmodified oligonucleotides are detected in all controls. **e-f)** TMP with 10 or 30 minutes 365 nm UV produces peaks for unmodified oligonucleotides plus oligonucleotides with monoadducts and crosslinked oligonucleotides. The monoadduct causes a shift in molecular weight relative to the size of the TMP molecule. The TMP crosslinked material appears as a peak at a much higher molecular weight due to the crosslink holding the two DNA oligonucleotides together. The relative height of the peaks changes between the 10 and 30 minute UV conditions, with the crosslinked peak increasing in height with longer exposure to UV. This indicates that the increased 365 nm UV exposure has increased the number of TMP crosslinks formed. TMP treatments are all 88 nmol. UV treatment was always 365 nm with the UV lamp touching the 6-well plate containing the sample. See table 3.3 for molecular weights of expected products. Arrows: blue = unmodified oligonucleotides; blue/orange = unmodified oligonucleotide 1 and oligonucleotide 2 with a monoadduct have very similar molecular weights therefore they appear as a combined peak; orange = oligonucleotides with monoadduct; red = crosslinked oligonucleotides. Mass spectrometry was performed by Dr Afaf El-Sagheer (Department of Chemistry, University of Oxford).

3.3.4 Generation of psoralen-treated sample for nanopore sequencing

Introduction of a TMP crosslink into a pair of annealed oligonucleotides has been optimised and validated by polyacrylamide denaturing gels and mass spectrometry. However, these methods cannot be applied to genomic DNA. Both methods are population-based and rely on the overall mass of the molecule. If multiple modification sites were present mass spectrometry may facilitate distinction of products with different numbers of sites modified however no information would be gained on the identity of the sites in each molecule. Additionally, the more sites, the more challenging the resulting data is to interpret. Both methods also have maximum size restrictions that prevent application to regions of DNA larger than have been assessed here. To use DNA modifications to mark accessible DNA and provide information about multiple protein binding sites on single molecules, long molecules and a large number of modifications per molecule will be required. Nanopore sequencing may be able to fulfil these requirements, as information on the DNA base (plus any modifications) is gained at each position within each DNA molecule, and long DNA molecules (>10 kb) can be analysed. However, it is not currently known if TMP monoadducts can be detected by nanopore sequencing. Therefore, now that TMP crosslinking of a single site in DNA has been optimised and validated, the detection ability of nanopore sequencing can be tested. As the signal perturbation for nanopore sequencing is unknown, only one modified site was introduced which also allows previous optimised conditions to be used.

Two 18 base DNA oligonucleotides were annealed, incubated with TMP to allow intercalation, crosslinked with 365 nm UV, gel extracted by electroelution to select for crosslinked DNA and filtered to concentrate the DNA. To perform nanopore sequencing, additional steps were required to meet three main criteria: the DNA strands must be able to separate; the DNA needs to be double stranded; and the DNA must be >200 bp in length. Firstly, the ability of the DNA strands to separate requires the TMP modification to be reversed from an

interstrand crosslink to a monoadduct on one of the strands. Secondly, the sample must be double stranded in order to allow the adapters, which transport the DNA to the pore, to be ligated therefore the oligonucleotides must be reannealed after crosslink reversal. Thirdly, to prevent the 18 bp annealed oligonucleotides being lost in the library preparation steps, a ligation step was performed to exceed the minimum DNA molecule length.

Reversal of a TMP interstrand crosslink

One of the main requirements for nanopore sequencing was the reversal of the crosslink to leave a monoadduct on one DNA strand. This would allow the DNA strands to separate so that one could pass through the pore during sequencing. There are two methods for reversing TMP crosslinks: alkali and high temperature treatments, or 254 nm UV light [113, 124, 128, 130, 131]. There is variation in the literature over exact conditions for both methods. For example, alkali reversal can involve 100 mM NaOH at 90 °C for 30 minutes [131] or 6 mM NaOH at 70 °C for 30 minutes [113]. Therefore, a combination of tests were performed for alkali reversal, with 10 mM and 100 mM NaOH for 30 minutes at both 70 °C and 90 °C. Similarly, there is variation in reported 254 nm UV exposure time for crosslink reversal, with a range of 20 minutes to two hours [131]. Therefore, a range of time points encompassing both of these were investigated for the application of 254 nm UV to reverse the crosslinks.

Prior to reversal, the oligonucleotides were annealed, incubated with TMP to allow intercalation of TMP into the DNA, treated with 365 nm UV to form the crosslink, filtered to remove unbound TMP, gel extracted to select for crosslinked DNA and filtered to concentrate the DNA. For 254 nm UV reversal of the crosslink, the setup was the same as for crosslinking (figure 3.8b), where the DNA was in a 6-well plate on ice with the UV lamp positioned above. Polyacrylamide denaturing gels were used to assess the extent of reversal for both alkali and 254 nm UV methods, in addition to evaluating the effect of acid neutralisation of

the alkali reversal samples (figure 3.12). For initial optimisations of reversal treatments non-gel extracted material was used (figure 3.12a and c), however for subsequent optimisation (figure 3.12b, d and e) and proceeding to the next step in substrate preparation for nanopore sequencing, gel extraction of the crosslinked band was performed. The alkali treatment with 100 mM NaOH at 90°C for 30 minutes was the most effective for crosslink reversal (figure 3.12a) although two hour treatments with 254 nm UV light produced a similar result (figure 3.12b). Figure 3.12c shows that 10 minutes treatment with 100 mM NaOH at 90 °C was sufficient for effective crosslink reversal. Additionally, the alkali crosslink reversal was not affected by neutralisation with 100 mM HCl as the band at 18 bases appears in the alkali reversed sample and remains upon HCl neutralisation (figure 3.12d). The alkali treatment with 100 mM NaOH for 10 minutes at 90 °C was used on the sample for future nanopore sequencing because the duration of this was much shorter, and prevented possible DNA damage from further UV treatment. Furthermore, the alkali reversal method was selected because if this is scaled up to genomic DNA, alkali treatment is less affected by variation in the total volume than the 254 nm UV treatment. In addition, future scaling up requires a highly efficient reversal step as a single non-reversed crosslink could prevent the entire molecule or at least part of it from being read by the pore during nanopore sequencing. When nanopore sequencing encounters an intact crosslink the DNA molecule is likely to be ejected from the pore. It is not known if the signal up until that point is stored or discarded and this may be length-dependent. Therefore, the optimal solution would be to have highly efficient reversal to ensure there are no intact crosslinks in the sample. The process of steps can be seen in figure 3.12e where the DNA is crosslinked, gel-extracted (GE), reversed (R) and filtered (R + F), producing single oligonucleotides, of which half are expected to contain monoadducts as each crosslink is left on a single strand, causing the other strand to become unmodified.

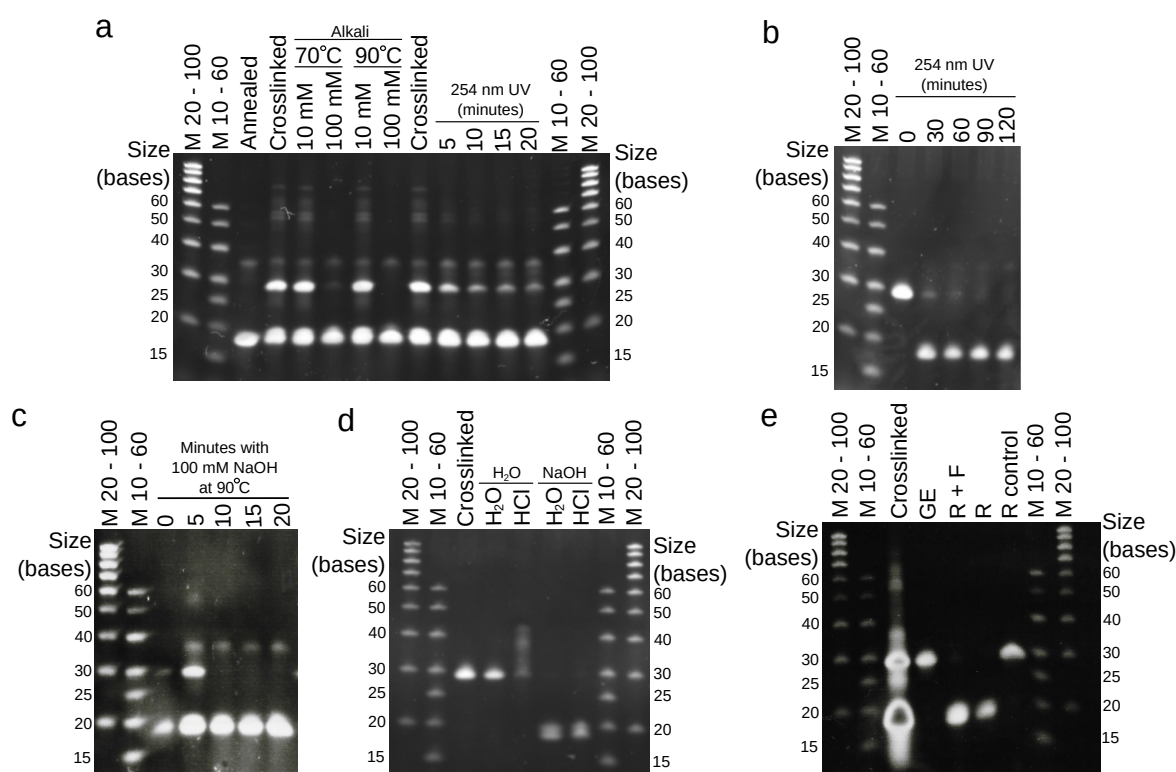


Fig. 3.12 TMP interstrand crosslink can be reversed with NaOH or 254 nm UV treatment

a) Optimisation of two methods of crosslink reversal using alkali (NaOH) and high temperatures or 254 nm UV light. Both treatments showed a decrease in the crosslinked material (band around 28 bp) but the alkali treatment showed more efficient reversal (shown by the absence of the 28 bp band in the 100 mM 90°C condition). Alkali treatment was performed for 30 minutes and neutralised with either 10 or 100 mM HCl (concentration equal to the concentration of NaOH used) prior to running on the denaturing gel. **b)** Longer durations of 254 nm UV produced crosslink reversal equivalent to that of 100 mM alkali at 90°C. The UV lamp was positioned 10 cm above the sample plate which was on ice (same setup as for crosslinking in figure 3.8b). **c)** Further optimisation of the alkali reversal, optimising the time for the best temperature and concentration from panel a. This showed that reversal was obtained after 10 minutes of treatment with 100 mM NaOH and 90 °C therefore 30 minutes treatment was unnecessary. After alkali treatment the samples were neutralised with 100 mM HCl. A loading issue with the 0 minute sample led to less DNA being loaded into this well. **d)** Neutralisation with HCl after reversal doesn't affect the state of the modification. Reversal reactions were performed with double distilled water (control) or 100 mM NaOH at 90°C for 10 minutes, then neutralised with either double distilled water (control) or 100 mM HCl prior to running on the denaturing gel. **e)** Overview of samples through the work flow, from the mixture of products after 365 nm UV crosslinking (crosslinked), purified crosslinked DNA after gel extraction by electroelution (GE), crosslink reversal with 100 mM NaOH for 10 minutes at 90 °C followed by neutralisation with 1 M Tris pH 7 (R) and filtration with an Amicon 3k column (R + F). R control = Control reversal step with water for 10 minutes at 90 °C. The initial DNA for a and c were two 18 base annealed oligonucleotides which had undergone 6 x 15 minute 365 nm UV cycles. The initial DNA for b and d were two 18 base annealed oligonucleotides which had undergone 8 x 15 minute 365 nm UV cycles, filtered through an Amicon 3k column to remove unbound TMP, run on a 15% polyacrylamide denaturing gel, gel extracted by electroelution and 3k filtered to concentrate the DNA. The initial DNA for e were two 18 base annealed oligonucleotides which had undergone 4 x 15 minute 365 nm UV cycles, filtered through an Amicon 3k column to remove unbound TMP, and run on a 15% polyacrylamide denaturing gel. M = Marker (10/60 and 20/100 Oligo Length Standards from IDT).

Reannealing of TMP modified DNA

Efficient reversal with alkali treatment and high temperatures produces a TMP monoadduct on one strand, however the high temperatures also denature the two DNA strands. At the time of performing this sequencing, the DNA needed to be double-stranded in order to allow the double-stranded adapter to be bound during library preparation. Therefore, a reannealing step identical to the initial annealing reaction (section 2.6.1) was performed to return the oligonucleotides to a double-stranded state. This step is possible with a relatively homogenous solution of two 18 base oligonucleotides, however may not be suitable for scaling up to genomic DNA.

Ligation and linearisation reactions to create a long linear molecule with a TMP modification

Now that the DNA has the TMP as a monoadduct on one strand and is double-stranded once again, the remaining problem was its length. Therefore, a ligation step was performed using a 873 bp PCR product amplified from chromosome 10 of *S. cerevisiae* genomic DNA. Prior to ligation this PCR product was digested with BglII and SalI-HF to produce ends complementary to the single-strand overhangs in the annealed oligonucleotides. This region of *S. cerevisiae* genomic DNA was selected because it contained a BglII and a SalI-HF digest site approximately 430 bp apart, with no other BglII or SalI-HF digest sites within this region. The 430 bp digested material facilitated visual detection on agarose gels, and was above the minimum length criteria for nanopore sequencing. The BglII digested end is complementary to the overhang of oligonucleotide 2 (which is a BamHI recognition site), while the SalI digested end is complementary to the overhang of oligonucleotide 1 (which is an XhoI recognition site). This, and the phosphorylation of the 5' ends of the insert, facilitated ligation of the insert (annealed oligonucleotides) to the large fragment (digested PCR product) as in figure 3.13. Self-ligation reactions could occur for both the insert and

large fragment as they both have palindromic 4-base overhangs (figure 3.14) and these were observed by the higher weight molecular bands in figure 3.15a and b which predominantly indicate the self-ligation of large fragment molecules. This was validated by the observation of the higher molecular weight bands in the absence of the insert (LF ligation condition in figure 3.15b). When two identical ends ligate together in self-annealing, the original restriction enzyme recognition site is reformed, facilitating subsequent digestion with BamHI and XhoI for the insert, or BglII and SalI for the large fragment (figure 3.13). However, if the insert and large fragment ligate together, the recognition sites are modified, therefore cannot be digested by BamHI, XhoI, BglII or SalI. Consequently, these four enzymes were added to the ligation reaction to prevent self-annealing and to bias towards ligation of the large fragment and insert molecules although these may not be 100% efficient. The intended product of 448 bp contains one insert and one large fragment. Once one end of the large fragment and insert has ligated, the other reaction should be relatively fast as intra-molecular reactions occur on a faster timescale than inter-molecular reactions. The desired product can be obtained by circularisation once the insert and large fragment have ligated, or by extracting the desired product from a linear molecule with additional ligated molecules (figure 3.13).

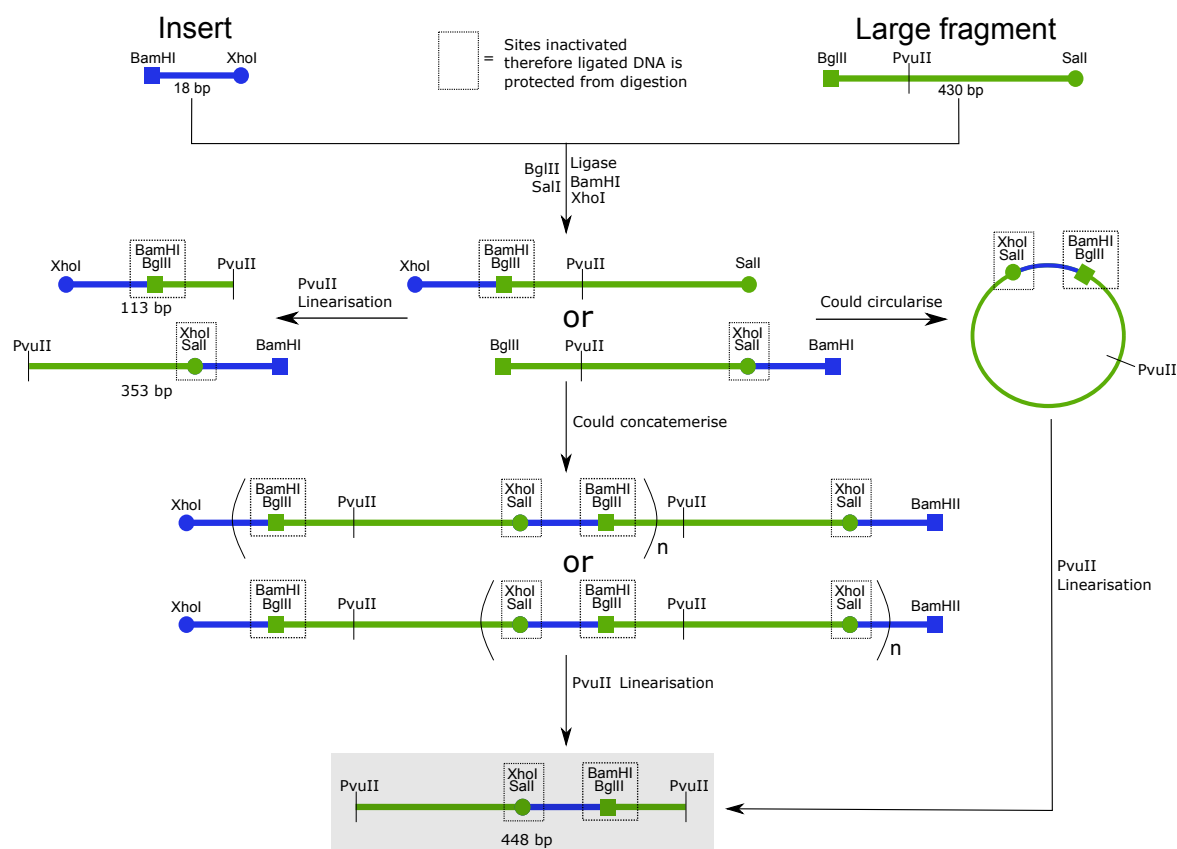


Fig. 3.13 Ligation reaction producing desired TMP-modified product

An overview of the ligation reaction leading to the desired product, including restriction enzymes and the inactivation of their sites through desired ligation reactions. Insert (blue) is the 18 base oligonucleotides, annealed to each other. Large fragment (green) is the digested PCR product. The desired product is shown in the grey box. Each blue or green line represents double-stranded DNA. Brackets with 'n' represent potential repeats of the region inside the brackets. Complementary overhanging ends are represented by the same shape (e.g. overhangs produced by digestion with BamHI and BglII are complementary, therefore these are represented by a square). Dotted box represents inactivated restriction enzyme sites.

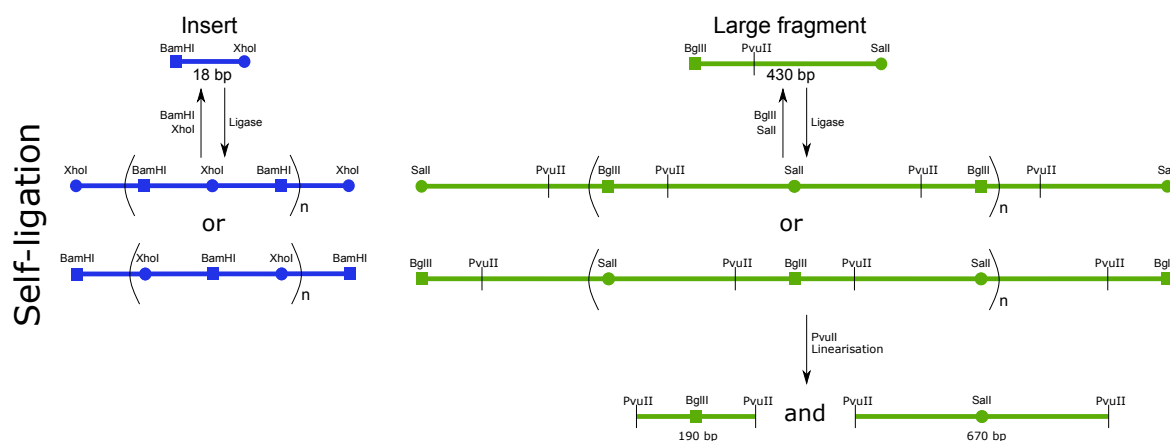


Fig. 3.14 Potential self-ligation products

An overview of the potential self-ligation products that could be generated in the same ligation reaction as shown in figure 3.13. Insert (blue) is the 18 base oligonucleotides, annealed to each other. Large fragment (green) is the digested PCR product. Each blue or green line represents double-stranded DNA. Brackets with 'n' represent potential repeats of the region inside the brackets. Complementary overhanging ends are represented by the same shape (e.g. overhangs produced by digestion with BamHI and BglII are complementary, therefore these are represented by a square).

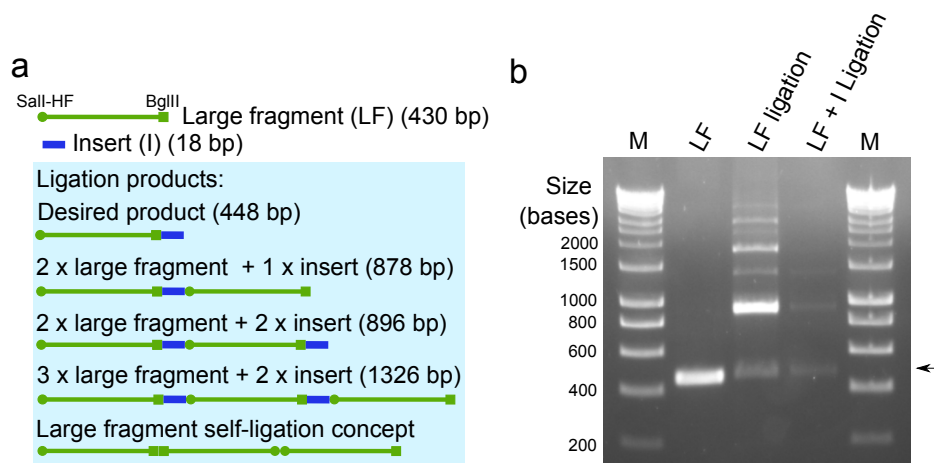


Fig. 3.15 Ligation of TMP-modified DNA to create longer substrate

a) Schematic of potential products from the ligation reaction between large fragment in green (PCR product from *S. cerevisiae* digested with BglII and SalI-HF to create ends complementary to the insert; circle and square indicate digestion with different enzymes) and insert in blue (18 base oligonucleotides annealed, treated with TMP and 365 nm UV, Amicon 3k filtered, gel extracted, Amicon 3k filtered, reversed and re-annealed). DNA was not subsequently linearised. **b)** Ethidium bromide stained 3% agarose gel showing the large fragment (LF) as a control, large fragment in the ligation reaction without the insert (LF ligation), and the ligation reaction containing both the large fragment and insert (LF + I ligation). Ligation was performed at 23 °C for 2 days. Bands are very similar between the two ligation reactions as the insert is small (18 bp). M = Bioline Hyperladder 1kb; LF = large fragment. The arrow indicates the linear form of the desired ligated product of one large fragment and one insert at 448 bp. The circular form would be expected to appear as a band at a larger size as circular DNA migrates more slowly than linear DNA. However, no band representing the circular form is observed.

The ligation products could be either linear or circular depending on whether or not the ends are complementary. To ensure the desired product with 1 large fragment and 1 insert molecule was linear and to ensure the insert was away from the ends of the DNA molecule, a linearisation step was performed (figure 3.16a and b). If the circularisation of the molecule was highly efficient with 1 large fragment and 1 insert molecule there would not be the plethora of higher molecular weight bands observed (figure 3.15b, 3.16b), therefore it can be assumed that this circularisation was not particularly efficient. There was a single PvuII digest site within the large fragment, therefore the non-ligated large fragment would be cut into two fragments during the PvuII linearisation step as it was unable to circularise due to non-complementary ends. The two fragments of the PvuII-digested large fragment would be

95 bp and 335 bp, of which the former is too small to be seen in figure 3.16b. In this agarose gel, the ligation linearised condition also contained the band at 335 bp, suggesting that within the ligated sample some of the large fragment had not ligated to the insert. At around 448 bp, the expected size of the desired product, there was a band in the non-linearised ligated sample and a fainter band in the linearised ligated sample which would represent one insert and one large fragment. This band was subsequently gel-extracted and was confirmed to contain both the large fragment and insert by Sanger sequencing (figure 3.18). Therefore, this band was used as the nanopore sequencing substrate.

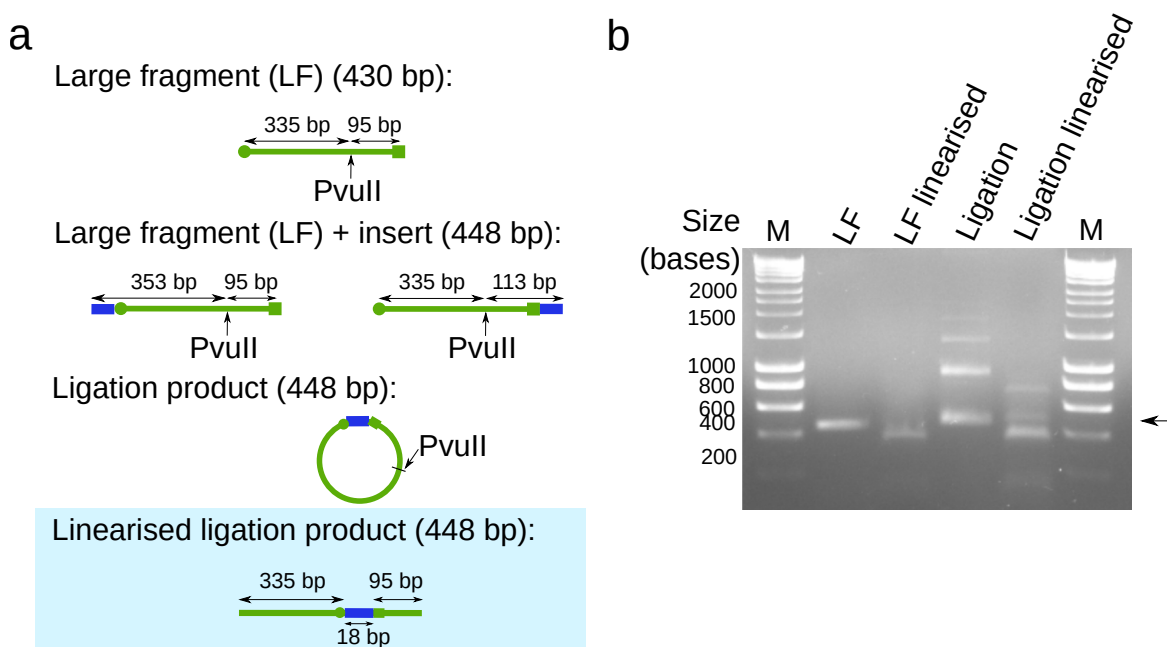


Fig. 3.16 Linearisation of TMP-modified DNA

a) Schematic of the linearisation reaction using PvuII (same colour coding as in figure 3.15). This linearisation step relies on the ligation product circularising. The end result is with the insert towards the middle of the product. **b)** Ethidium bromide stained 3% agarose gel showing the linearisation reaction with a faint band at the desired product size of 448 bp. Ligation was performed at 23 °C for 2 days (same ligation sample as in figure 3.15b). Linearisation was performed at 37 °C for 4 hours. M = Bioline Hyperladder 1kb; LF = large fragment. The arrow indicates the desired ligated product of one large fragment and one insert at 448 bp.

The implemented strategy involved the addition of restriction enzymes to prevent self-ligation. The ability of BglII and SalI to prevent self-ligation of the PvuII digested large

fragment was assessed by performing the ligation reaction in the presence and absence of these enzymes. For simplification, PvuII linearised DNA was used to reduce the number of possible products and the insert was not included for lanes 2-6 of figure 3.17b. In figure 3.17, lanes 3 and 4 (without and with BglII respectively) show almost identical bands. BglII should be able to digest the band at 190 bp to form one at 95 bp, however in lane 4 the 190 bp band persists and the 95 bp band is very faint therefore BglII is only performing minimal digestion of the self-ligated large fragment. In comparison, lanes 3 and 5 (without and with SalI-HF respectively) show quite different band patterns. In lane 5, the 670 bp band becomes faint upon SalI-HF digestion and a band at 335 bp increases in intensity, therefore SalI-HF mostly, but not completely, prevents self-ligation at the large fragment ends that were produced by digestion with this enzyme. These findings highlight the potential for other ligation products to be formed in the ligation reaction with large fragment and insert, due to self-ligation of large fragment molecules. Finally, despite these self-ligation events, the desired ligation product (448 bp with one insert and one large fragment) can be observed at the position of the arrow in lane 1 of figure 3.17b. This band is slightly higher up the gel than the fainter band in lane 2 which corresponds to the 430 bp product representing the linearised but undigested large fragment.

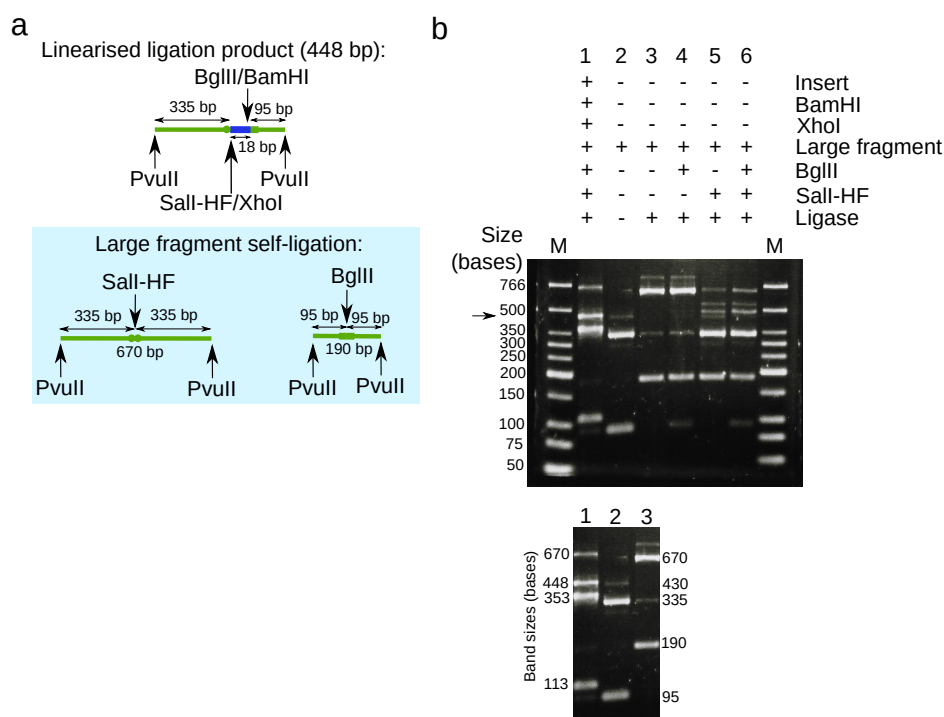


Fig. 3.17 Effectiveness of restriction enzymes to prevent large fragment self-ligation

BglII and Sall-HF are used to generate the large fragment and can also be used to digest self-annealed large fragment molecules. **a)** Schematic of the large fragment self-ligation that can occur. BamHI and XhoI are complementary to BglII and Sall-HF respectively, and can be used to prevent the insert self-ligating. PvuII cuts both strands at the same position, producing blunt ends, therefore there should not be any ligation at PvuII sites. **b)** Visualisation of potential undesired products from ligation reactions on a 3% agarose gel stained with ethidium bromide. Samples were ligated for 4 days at 23 °C and linearised with an overnight PvuII digest at 37 °C. M = Quickload low molecular weight marker (NEB); LF = large fragment. The arrow indicates the desired ligated product in lane 1 of one large fragment and one insert at 448 bp.

Sanger sequencing verification of ligation product

Sanger sequencing was used to verify the sequence of the new substrate prior to nanopore sequencing. Annealed oligonucleotides were crosslinked with four 15 minute 365 nm UV cycles and run on a polyacrylamide denaturing gel from which crosslinked material was gel extracted by electroelution and filtered to concentrate the DNA. The crosslink reversal was performed with 0.1 M NaOH at 90 °C for 10 minutes, filtered with 1 M Tris pH 7, ligated, run on an agarose gel, gel extracted and linearised with PvuII. The control sample underwent the same conditions but the crosslink step was performed in the absence of TMP. The shift

from 430 bp to 448 bp on an agarose gel is minimal (figure 3.15b, d and f) therefore Sanger sequencing was used to confirm that the correct product had been gel extracted. In theory, Sanger sequencing should also provide information on the modification status of the thymine in the insert. In Sanger sequencing the DNA is extended with a polymerase and fluorescently labelled terminator dideoxynucleotides (ddNTPs) are introduced at different points in different molecules, allowing the base at each position to be identified by the terminator colour. The use of a polymerase indicates a similarity to primer extension assays, where TMP can block the polymerase [127, 113]. Therefore, in Sanger sequencing, the peak height may reduce at the point of the modification as the polymerase would only be able to proceed past this point on unmodified molecules. In figure 3.18a there is sign of a reduction around the position of the modification in both the forward and reverse strands. This indicated that the sample may be modified thus the DNA was taken forward into nanopore sequencing library preparation. However, when an unmodified DNA sample was later analysed by Sanger sequencing (figure 3.18b), a reduction in peak height equivalent to that in the TMP-modified sample was also observed. This indicated that the reduction may not be indicative of the presence of a TMP monoadduct at this position. It may instead be due to being close to the end of the sequencing region as the potentially modified TA site is >300 bases away from the forward primer and <50 bp away from the reverse primer. However, if this was the case this would be expected to be a more gradual reduction and a greater mixture of base calls would be expected at each position. Another explanation could also be a mixture of products, with some modified (causing a reduction) and others unmodified (not causing a reduction). Alternatively, it may have been a side effect of the ligations that had occurred at each end of the insert, thus were only seven bases away from the TA site.

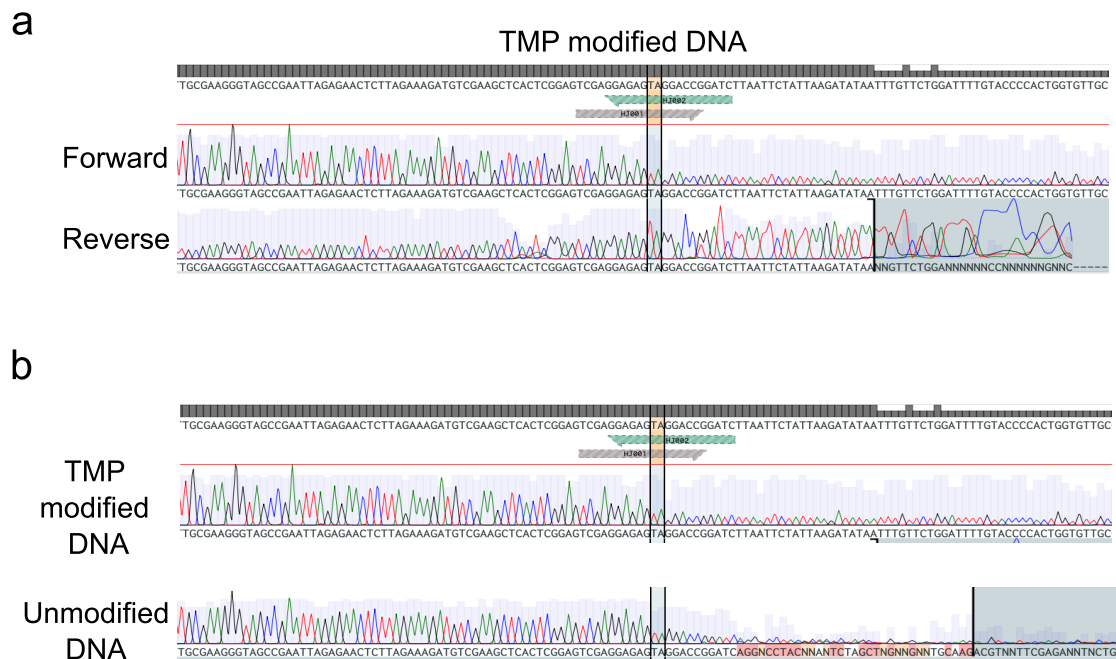


Fig. 3.18 Sanger sequencing does not detect TMP modification on DNA

Sanger sequencing was used to assess the presence of a TMP modification on DNA as the TMP modification may physically block progression of the polymerase. **a)** Sanger sequencing is shown for the ligation product, made from a PCR fragment digested with BglII and SalI, ligated to an insert consisting of two annealed 18 bp oligonucleotides which have been crosslinked by TMP, gel extracted, crosslinks reversed to form monoadducts and then the oligonucleotides were re-annealed. The ligation product of the desired size for 1 PCR fragment and 1 insert (448 bp) was gel extracted. This DNA was Sanger sequenced on both strands (represented by the terms forward and reverse) which confirms the presence of the 18 bp insert in the ligated product. They also indicate there might be a reduction in peak height around the TA site within the insert (the selected region) which indicates the potential presence of a TMP monoadduct based on the principle of the monoadduct acting as a block to the polymerase in primer extension assays [127, 113]. **b)** The TMP modified DNA sequence (upper) is the same as in a. However, this time it is compared to a forward sequence of the control DNA which showed a similar reduction in peak height at the selected region.

3.3.5 Nanopore sequencing does not detect TMP modification on DNA

Nanopore sequencing library preparation was performed on DNA that had been annealed, incubated with TMP, UV crosslinked, gel extracted, concentrated by filtration, reverse crosslinked with heat and alkali treatment, reannealed, ligated to a double-end digested PCR product, gel extracted and linearised. As a furanocoumarin DNA modification has not previously been detected by nanopore sequencing, there is no software designed to

detect this modification, and the effect on the signal was not known. The data was first basecalled to convert the detected electrical signals to DNA sequence, and then mapped to the compiled reference sequence of the linearised ligation product (448 bp). Possible effects the monoadduct might have on the data are: (i) a reduction in coverage if the monoadduct blocked the pore and caused the DNA molecule to be ejected (assuming that the data recorded up until that point on the molecule is stored); (ii) an increase in mismatches or indels due to the monoadduct being an uncharacterised event; (iii) an inability to map reads over the modified position; (iv) a change in the mean and/or standard deviation recorded around the modified position in the raw data.

Firstly, the mpileup function within samtools was used to quantify the coverage and identify if each read had a match, mismatch or indel at each position. Figure 3.19 shows that there was not a substantial reduction in coverage (red line) at the potential modification site (grey dashed line). Therefore, either the pore is not being blocked by the monoadduct, or the data up until the blockage event was not recorded. Secondly, these results also show that the incidence of indels and mismatches was not noticeably different at or close to the site of potential modification when compared to the rest of the substrate. Therefore, the monoadduct does not seem to be disrupting the ability of the basecaller to call the bases surrounding the modified site.

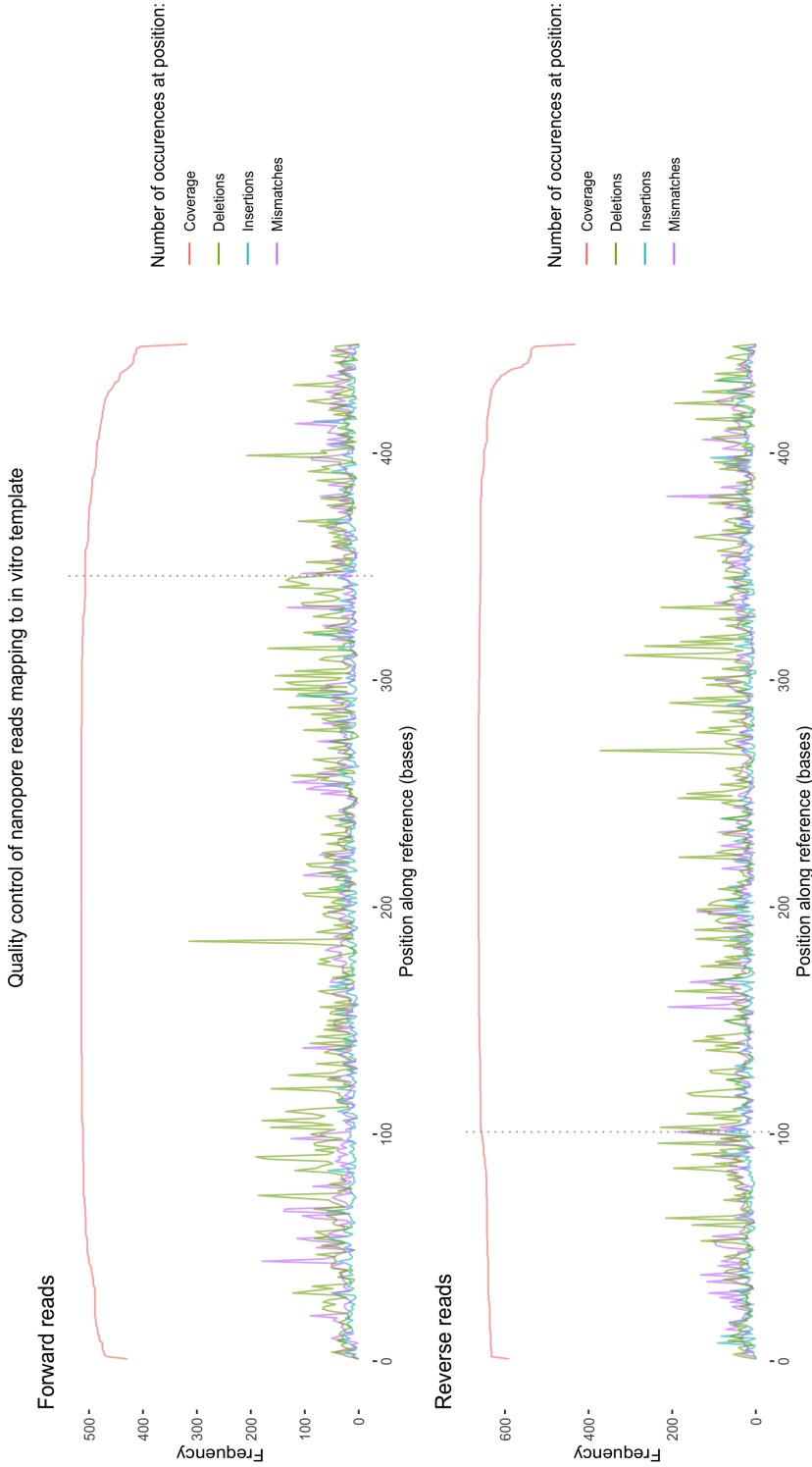


Fig. 3.19 TMP modification does not affect nanopore sequencing coverage or indels
Nanopore sequencing of potentially TMP-modified DNA. Reads mapped to a minimum of 30% of the reference sequence. Coverage and indel/mismatch data were collated using samtools mpileup. Red = coverage; green = deletions; blue = insertions; purple = mismatches; grey dotted line = site of potential furanocoumarin modification.

Thirdly, to address the possibility that the nanopore reads weren't mapping correctly, three thresholds for the minimum coverage of the reference sequence were tested. As shown in table 3.4 there was no difference between using a threshold of 15% or 30%, and only 16 or 20 reads were gained by increasing the threshold to 80% for the forward and reverse complement templates respectively. The potential modification site was at 346 or 102 bases into the substrate of total length 448 bp, for the forward and reverse complement reference sequences respectively. Therefore, the shortest fragment to one side of the modification site is 101 bp. This represents 22.5% of the total substrate length. Therefore, if there were reads that were able to only map to this part of the reference but are then blocked by the modification and thus sequencing is disrupted, the 15% threshold should have detected at least some of them. The fact that there wasn't any difference between the 15% and 30% thresholds indicates that this scenario either did not occur, or the regions mapped to less than 15% of the reference. Lowering the threshold increases chances of reads incorrectly mapping to the reference sequences as the nanopore run was pooled with yeast genomic DNA samples involved in another experiment.

Minimum coverage of reference sequence (%)	Reads aligned to forward strand	Reads aligned to reverse complement
15	514	663
30	514	663
80	498	643

Table 3.4 Nanopore sequencing read counts

Number of reads mapped to either of the two references (forward strand or reverse complement) under variable requirements for minimum coverage of the reference sequence.

Fourth, to look at the data in a less processed form, a fastq file of the reads mapping to each strand was produced with Albacore and the reads in the fastq were then mapped to the reference sequences with BWA-MEM to produce a bam file. This was then passed to

nanopolish eventalign which was used to align the signal events to a reference sequence (either the forward strand or reverse complement, see table 3.5 for read numbers at each stage). This goes through the reference sequence one 6-base k-mer at a time (sliding along by one base each time), and for each read, there might be multiple events per k-mer, or there may be some k-mers with no events. K-mers are used rather than single bases, as the nanopore signal is affected by a group of bases in the pore, not a solitary base. Each event has a mean and standard deviation associated with it and a position in the reference it has been mapped to, in addition to an Oxford Nanopore Technologies (ONT) model k-mer and the details of that k-mer. Some events were given an unknown model k-mer of 'NNNNNN'. There was a potential increase in frequency of this occurring upstream (to the 5' end of the target thymine) in the forward strand (figure 3.20c), but no increase in frequency was observed for the reads aligned to the reverse complement (figure 3.20f). Using the aligned events, the region surrounding the target thymine was analysed. Individual k-mer analysis involved the potentially modified thymine plus 10 bases upstream and 11 bases downstream (encapsulating all k-mers derived from the two initial annealed oligonucleotides including their ligation sites to the rest of the substrate). The signal detected for each k-mer across the reads aligning to the forward and reverse references was compared to the Oxford Nanopore Technologies (ONT) standard 6-mer model. This showed no differences between the observed and standard ONT results for each k-mer on reads mapped to either the forward or reverse strands when looking at the events given a non-'NNNNNN' ONT model k-mer (figure 3.20a and d). For the events that were given 'NNNNNN' as an ONT model k-mer the range of mean signal values was much larger and the mean of these often deviated substantially from the ONT mean (figure 3.20b and e). There does not appear to be a larger disagreement between the experimental values and ONT model data for the k-mers containing the target thymine (in the grey box), than neighbouring k-mers not containing the TMP modification (figure 3.20b and e), therefore it is likely a feature of the 'NNNNNN' classification rather than the

TMP modification. This is also supported by the absence of a substantial increase in the frequency of 'NNNNNN' ONT model k-mers over the TMP target site. Previous detection of DNA modifications by nanopore sequencing has identified a bimodal distribution in the nanopore signal (representing modified and unmodified DNA) [106–108]. However, no bimodal distributions were observed here, for the nanopore signals for the region around the thymine targeted for TMP modification, therefore there was no indication of two populations, one unmodified and one modified. In summary, there was no indication that the nanopore signal was perturbed at the potential site of TMP modification.

Condition for k-mer analysis	Reads aligned to forward strand	Reads aligned to reverse strand
Number of reads in fastq files (pass and fail)	498	643
Number of reads in fastq files (pass)	490	639
Number of alignments in bam files	595	768
Number of mapped primary alignments in bam files	536	714
Number of nanopolish reads that failed qc	5	4
Number of nanopolish reads that it could not calibrate	1	1
Number of nanopolish reads with no alignment	3	19
Nanopolish number of reads outputted	432	571
Nanopolish number of events	340,214	481,699
Nanopolish events with NNNNNN model k-mer	11,305 (3.3%)	16,634 (3.5%)

Table 3.5 Nanopore sequencing k-mer analysis overview

Summary figures describing the data used for k-mer analysis when looking at the raw data to identify if there is any signal perturbation by TMP modification on DNA. Reads were basecalled with Albacore v2.0.2, indexed with nanopolish v0.13.2, mapped with BWA-MEM and the events were aligned to the reference sequence with nanopolish eventalign with the `–scale-events` argument.

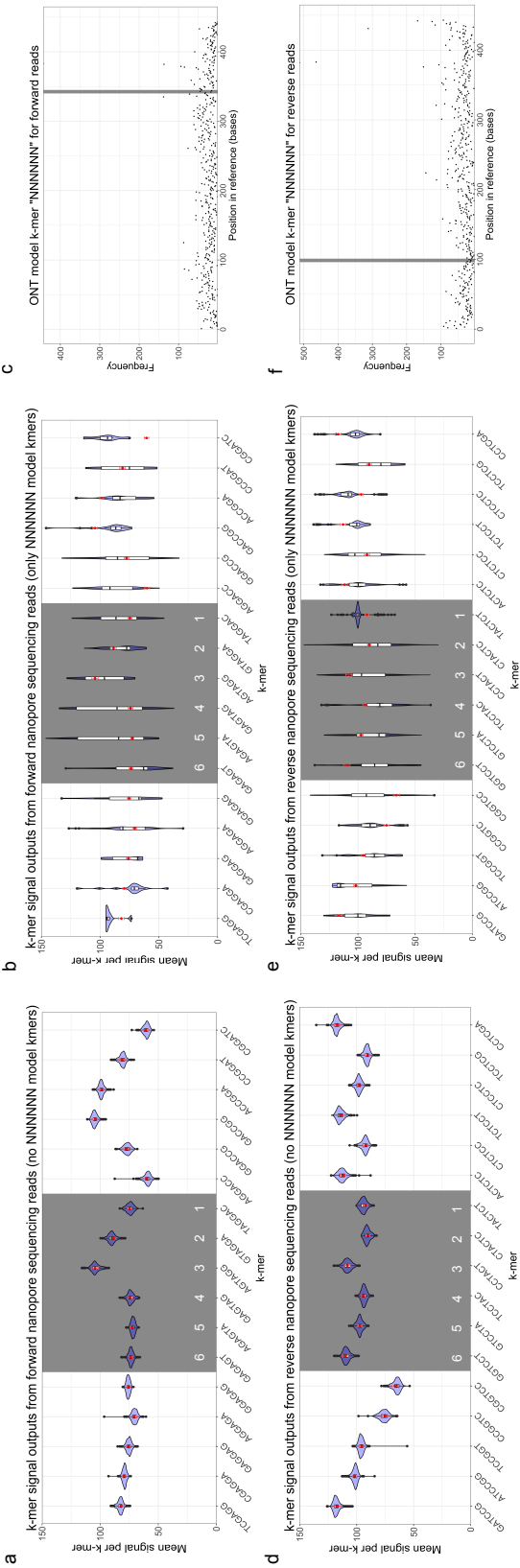


Fig. 3.20 TMP modification does not cause signal perturbation in nanopore sequencing
Nanopore sequencing reads aligned to the reverse strand ($n = 571$) are analysed in d-f, with the same analysis run for both. **a & d**) Violin plots focusing on the region of the DNA containing the target thymine. The mean values of the TMP-treated DNA events (blue) and the mean and standard deviation from the Oxford Nanopore Technologies (ONT) model (red) are shown for the events that were not given an ONT k-mer model of 'NNNNNN'. The grey box highlights the k-mers that contain the potentially modified thymine and the white numbers indicate the position of the potentially modified thymine within the k-mer. Reads were mapped with a minimum of 30% coverage of the reference sequence. ONT model was 'template_median68pA.model' using k-mers of 6 bases, r9.4 pore, 180 millivolts and 450 bases per second. All k-mers are written in 5' - 3' direction. **b & e**) Same analysis as in a & d except now only showing events that were given an ONT k-mer model of 'NNNNNN'. For each k-mer, the ONT model data is taken from other reads that have a non-'NNNNNN' ONT k-mer model at that position. **c & f**) Frequency of 'NNNNNN' ONT model k-mer throughout the entire substrate for reads that mapped to the forward or reverse strands. The grey box highlights the k-mers that include the target thymine.

Previously the Sanger sequencing result had indicated that the TA site might be modified (figure 3.18a). However, when a control was run, after the nanopore sequencing was performed, a similar reduction in peak height was observed at the position of potential modification (figure 3.18b). This indicated that the reduced peak height was not due to the presence of a TMP monoadduct on the DNA. Therefore, suggesting that the substrate for nanopore sequencing may not have been modified by TMP. However, the Sanger sequencing may not have shown a difference at this position in the molecule, when peak height is reducing anyway therefore it is not clear whether the substrate was not modified or nanopore sequencing was unable to detect the modification.

3.4 Angelicin modification of DNA

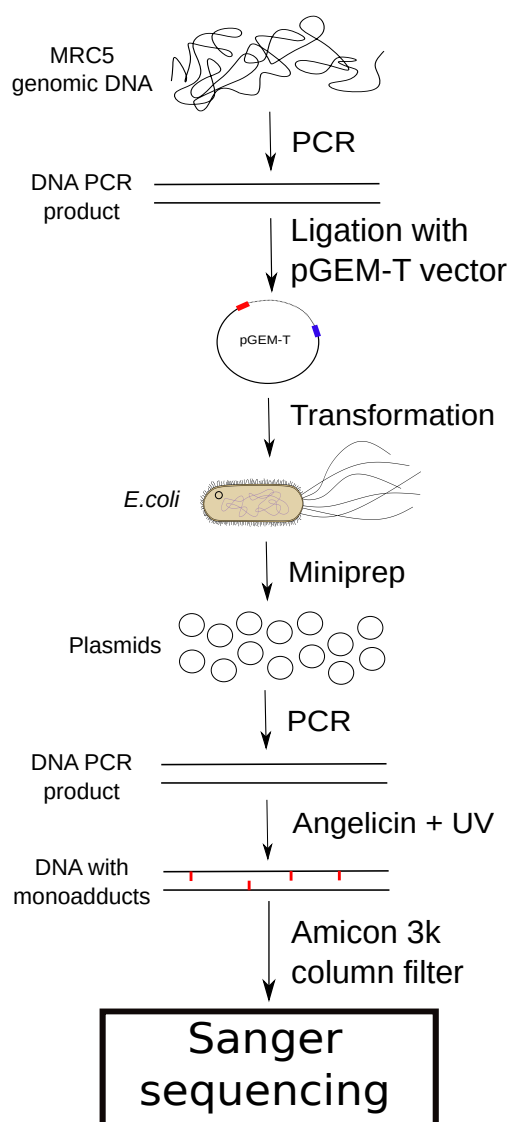
3.4.1 Overview of approach for angelicin modification of DNA

In the previous section, TMP was shown capable of modifying DNA at a single site which was validated by polyacrylamide denaturing gels and mass spectrometry, however the modification was not detected in the nanopore sequencing signal. This may have been due to the high number of processing steps required to produce a sample that was suitable for nanopore sequencing. To simplify the system, an angular furanocoumarin called angelicin (figure 3.1a) was used. Angelicin binds to thymine in the same DNA sequence contexts as TMP, although angelicin binds to DNA more weakly than TMP and is unable to form interstrand crosslinks [113]. Therefore, no reversal step would be required as modifications would only be in the monoadduct form. Conversely, visualisation and purification of modified DNA could not be performed with polyacrylamide denaturing gels as monoadducts were indistinguishable from unmodified DNA for TMP modification (figure 3.8c and d).

The other method used for validating the presence of a TMP modification on DNA was mass spectrometry. In addition to the crosslinked product, one of the oligonucleotides with a monoadduct formed a new peak when treated with both TMP and 365 nm UV. The second oligonucleotide could not be assessed in this way as the presence of a monoadduct caused its mass to be indistinguishable from the first unmodified oligonucleotide. As the binding of each angelicin molecule is independent and there should not be a strong bias in strand preference, mass spectrometry can be used to detect monoadduct formation achieved by angelicin and 365 nm UV treatment.

Sanger sequencing was also performed to see if there were any perturbations at TA sites. To get material of sufficient length for Sanger sequencing and to avoid the presence of ligation sites close to the TA site of interest, a 654 bp PCR product was treated with angelicin and 365 nm UV. The larger size of the substrate increased the number of potential modification sites while avoiding potential sequencing distortions caused by ligation and gel extraction steps. The overview of the angelicin workflow can be seen in figure 3.21.

Approach 1:



Approach 2:

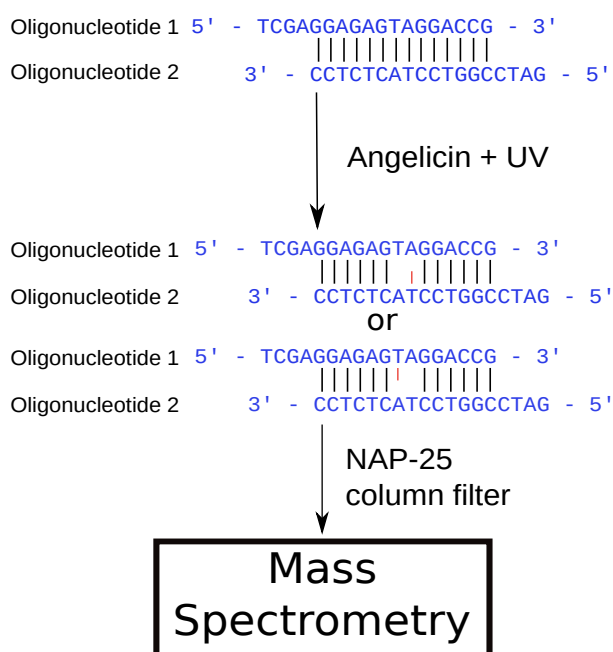


Fig. 3.21 Experimental workflow for angelicin modification of DNA

Overview of the experimental approaches taken to identify if angelicin can modify DNA, with testing by Sanger sequencing and mass spectrometry. This is different to the TMP workflow due to the inability of angelicin to form crosslinks. UV = 365 nm UV. In approach 1: black lines = DNA; blue and red blocks on pGEM-T plasmid mark the ligation sites. In approach 2: blue = DNA; black lines between DNA bases = hydrogen bonds. In both approaches: red lines = monoadducts.

3.4.2 Sanger sequencing does not detect angelicin modification on DNA

A region of MRC5 (diploid human fibroblast cell line) genomic DNA was PCR amplified, ligated with a pGEM-T vector and transformed into *E. coli* for amplification. The plasmid DNA was then extracted using a miniprep and a 654 bp region was PCR amplified for testing the modification ability of angelicin. The 654 bp PCR product was then incubated with angelicin on ice for 5 minutes for intercalation prior to 365 nm UV treatment using the setup in figure 3.8b, and subsequently filtered to remove unbound angelicin. The PCR product contained 10 TA sites (of which 4 are in ATA sites and 2 are in TAT sites) and 12 AT sites, all of which are potential target sites for angelicin. Visualisation of the samples on an agarose gel did not show any detectable size changes due to the angelicin or UV treatment (figure 3.22a) as expected for modifications of this magnitude. Similarly to TMP, angelicin blocks polymerases in primer extension assays [113]. Therefore, the polymerase in Sanger sequencing may be similarly blocked by the presence of angelicin. For the four TA sites away from the edges of the sequencing region two were in ATA sites, thus more likely to be modified. However, there were no noticeable differences between the controls (without angelicin) and samples treated with angelicin at any of these sites (figure 3.22b). Additionally, no UV damage was observed when comparing the 0, 30 and 60 minute UV treatments in either the absence or presence of angelicin (figure 3.22b). This indicates that the DNA has not been modified by angelicin or damaged by UV.

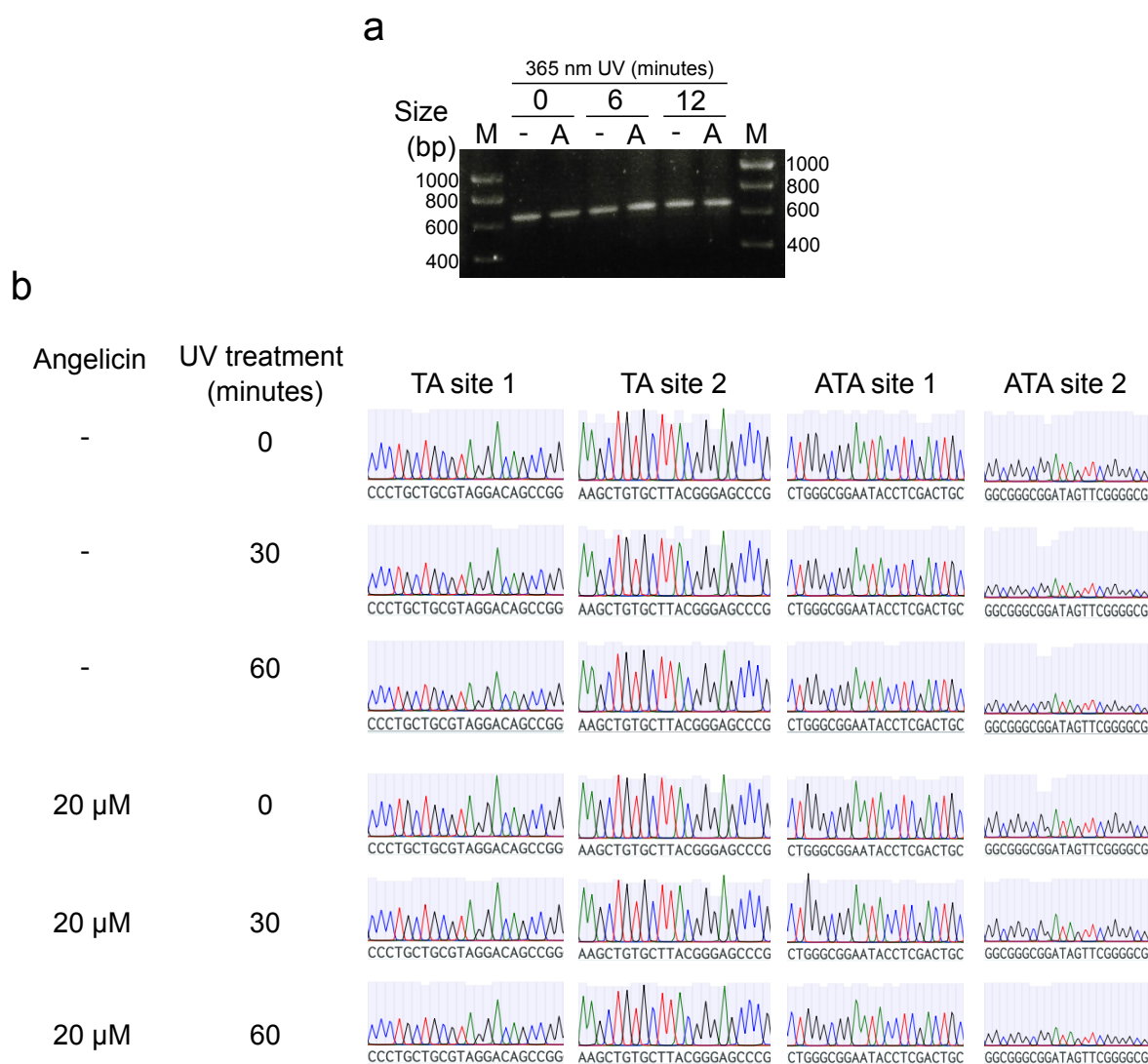


Fig. 3.22 Sanger sequencing does not detect angelicin modification on DNA

a) Sanger sequencing samples run on a 1.2% agarose gel run at 90 V. M = Bioline Hyperladder I; - = No angelicin added; A = 0.2 μ l of 1 mM angelicin added. **b)** Sanger sequencing was performed on a 654 bp PCR product amplified from pGEM-T using HJ014 and HJ017 and four TA potential modification sites, two of which are ATA sites, are shown with 10 flanking bases on each side. For each condition 1 μ g DNA was treated with 87.6 nmol angelicin from an initial 10 mM stock. Each sample was in 1 ml crosslinking buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0, 50 mM NaCl) in a 6-well plate on ice where it was treated with 365 nm UV with the lamp touching the 6-well plate. 87.6 nmol of angelicin was used to be comparable to the 10 μ l additions of 2 mg/ml TMP previously used. Sequencing trace height was unchanged between taking images of the different sites and conditions on Benchling.

3.4.3 Mass spectrometry does not detect angelicin modification on DNA

Mass spectrometry analysis was performed on the same oligonucleotides used in the TMP experiments (figure 3.8a). Samples were incubated with a 10-fold dilution series (1 nmol - 1 μ mol) of angelicin concentrations and treated with 365 nm UV for one hour to have a good chance of identifying angelicin monoadducts. Unlike with TMP, mass spectrometry of the angelicin and UV treated oligonucleotides only detected two peaks in each condition (figure 3.23). The mass of each peak corresponded to the two unmodified oligonucleotides (table 3.3). The one hour UV treatment without angelicin did not show any sign of DNA damage (figure 3.23a) in agreement with the same for 30 minutes UV (figure 3.11c). Likewise, the presence of angelicin without 365 nm UV treatment did not produce any other peaks (figure 3.23b), therefore the controls behaved as expected. Despite a range of angelicin concentrations, and a longer UV treatment than used for TMP, no modifications were detected (figure 3.23c-f). However, the presence of angelicin in the solution was not investigated. For TMP, 88 nmol was used for successful modification, therefore 1 μ mol should have been in large excess if angelicin was present in the intended quantities. The lack of observed DNA modification by angelicin suggests there may have been an issue with the angelicin stock. Therefore, this line of enquiry was not pursued any further.

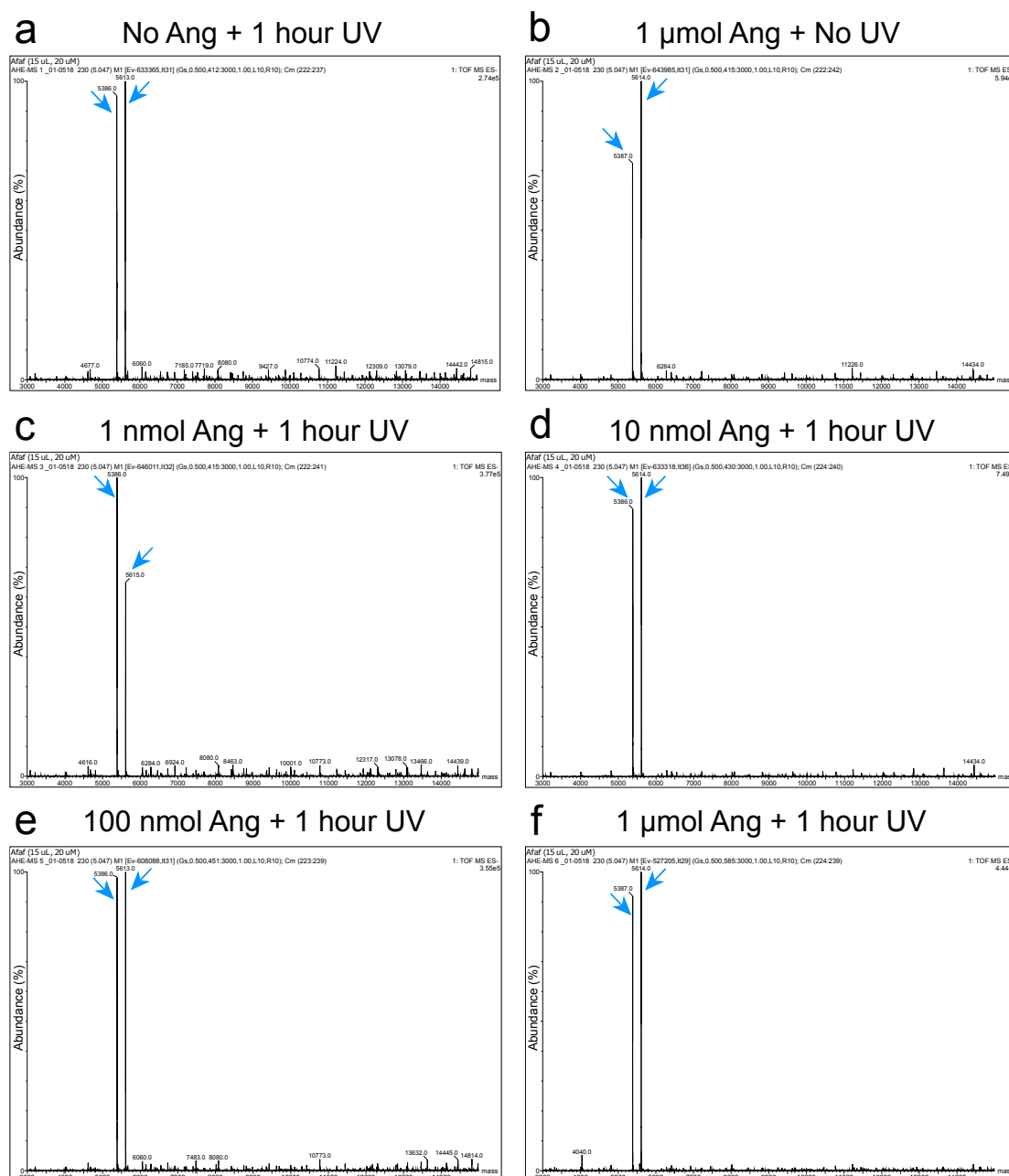


Fig. 3.23 Mass spectrometry does not detect angelicin modification on DNA

A variety of conditions were tested by mass spectrometry, where DNA strands are separated but are not hydrolysed into individual bases. The DNA sample used was two 18 bp oligonucleotides with one potential site of angelicin modification. **a - b**) Controls include DNA with no angelicin but UV treatment (to assess UV damage) and DNA with angelicin but no UV (therefore the angelicin is not activated). All controls only produce unmodified oligonucleotides. **c - f**) Angelicin 10-fold dilution series (from 1 nmol to 1 μ mol), all with 1 hour of 365 nm UV with the lamp touching the 6-well plate containing the samples. All treatments only show unmodified oligonucleotides, indicating that angelicin is not modifying the DNA in any of these conditions. Blue arrows = unmodified oligonucleotides; Ang = Angelicin. See table 3.3 for molecular weights of expected products. Mass spectrometry was performed by Dr Afaf El-Sagheer (Department of Chemistry).

3.5 Discussion

In this chapter, two furanocoumarin chemicals were tested as potential markers for accessible DNA: 4,5',8-trimethylpsoralen (TMP) and angelicin. Both chemicals had previously been shown to be detectable in populations of molecules by primer extension assays [113, 127]. TMP was also shown to be able to mark accessible DNA in single molecules, leading to the identification of protein binding sites in electron microscopy studies [127, 141]. This chapter concludes that these two chemicals, TMP and angelicin, are not suitable chromatin accessibility markers when combined with nanopore sequencing.

The first chemical to be trialled, TMP, is a linear furanocoumarin that is able to form crosslinks between two DNA strands. In polyacrylamide denaturing gels, the presence of TMP crosslinks holding the two DNA strands together was detected as a band of 28 bases which increased in intensity with increased cycles of 365 nm UV treatment (figure 3.8). A TMP monoadduct should cause the oligonucleotide to run through the gel as though it had one extra base [124] however, in the current study the bands representing unmodified oligonucleotides and oligonucleotides with a monoadduct were indistinguishable due to their high intensity and similar size. Therefore, from polyacrylamide denaturing gels TMP crosslinks were detected but TMP monoadducts were not.

Another technique used to confirm the ability of TMP to modify DNA was mass spectrometry. These experiments confirmed the presence of both TMP crosslinks and monoadducts (figure 3.11) as mass spectrometry provides more accurate molecular weight information than polyacrylamide gels, thus allowing the oligonucleotides with monoadducts to be distinguished from unmodified oligonucleotides. A complicating factor in this mass spectrometry analysis was the overlap in molecular weights of oligonucleotide 2 with a monoadduct and the unmodified oligonucleotide 1 which was also observed in the initial use of these

oligonucleotides [124]. However, oligonucleotide 1 with a monoadduct produced a clear, separate peak thus indicating that monoadducts were present in the mixture (figure 3.11e-f). Monoadducts are precursors of interstrand crosslinks, therefore the detection of them indicates ongoing modification. Conversely, they could also indicate inefficient crosslinking as an increase in the length of UV treatment led to a reduction in monoadducts and an increase in crosslinks (figure 3.11e-f). Additionally, there should not be any strand bias among the monoadducts as TMP can orient itself in either direction in relation to the two DNA strands. The conversion of a monoadduct to interstrand crosslink depends on the initial TMP-DNA binding event. If this produced a furan-side monoadduct the coumarin part of the chemical would have been able to react with a photon from UV light to form the crosslink [113]. Conversely, if the initial monoadduct was a pyrone-side monoadduct, the coumarin part of the chemical was disrupted, thus unable to form a crosslink [113]. Approximately 98% of TMP monoadducts are bound on the furan-side, therefore allowing a high proportion of thymine modifications to progress from monoadducts to interstrand crosslinks [124, 128]. Therefore, the majority of TMP modifications are expected to form interstrand crosslinks, but there will always be some residual pyrone-side monoadducts.

To make a sample for nanopore sequencing there were many processing steps (figure 3.7). This includes reversing the crosslink to make a monoadduct on one DNA strand. If the reversal is achieved through treatment with alkali and high temperatures, as was used in this study for the nanopore sequenced substrate, the products are predominantly pyrone-side monoadducts [130]. Conversely, 254 nm UV reversal has been linked to the production of either a more even distribution of pyrone- and furan- monoadducts [124] or predominantly furan-side monoadducts [130]. Here, the reversed material subsequently underwent re-annealing, ligation, linearisation, gel extraction and nanopore sequencing library preparation steps. The resulting nanopore sequencing data did not show any sign of a perturbation at the thymine in the 18 bp substrate (which was at this point part of a larger 448 bp substrate

containing no extra modifications) (figures 3.19 and 3.20). Potential reasons for this include (i) the monoadduct being present but not perturbing the nanopore signal, (ii) the monoadduct being present and preventing the molecule from being sequenced due to an inability to pass through the pore, (iii) the monoadduct preventing the molecule from being mapped to the reference sequence, or (iv) the monoadduct being absent in the sequenced material due to a bias towards unmodified oligonucleotides in the experimental pre-processing steps.

Firstly, it is highly unlikely that the TMP monoadduct would not affect the nanopore signal as substantially smaller modifications like methylation [107, 108] and BrdU [106] can perturb the nanopore signal. A psoralen monoadduct is similar to the presence of an extra base [124] and the bulkier the modification, the larger the perturbation to the nanopore signal, and the more neighbouring bases are affected [110]. By occupying substantially more space in the pore than an unmodified thymine and subsequently reducing the flow of ions, an effect on the resulting signal would be expected.

Secondly, theoretically the monoadduct could block the pore, causing the DNA molecule to be ejected once it reached the monoadduct, and the data collected on that molecule erased as it was incomplete. However, this is unlikely as larger modifications (large thymine analogs and PEG-labelled substrates) have been detected by nanopore sequencing [143, 110] therefore they must have successfully passed through the pore.

Thirdly, the read number was not substantially affected by changes in coverage requirements, suggesting that the monoadduct does not prevent the modified reads from being mapped to the reference genome. The sequenced DNA only contained one site that may have been modified, in a substrate of 448 bases. Reads with 15% and 30% coverage of the reference sequence were analysed, which should have facilitated mapping at a location away from the site of modification, therefore should have maintained TMP-modified reads.

The ability to map the reads is of high importance for identifying where in the genome the modifications are. If modified reads were unable to be mapped in this context, there would be no ability to map them if the rate of modification was dramatically increased.

Fourthly, the crosslinked material may have been lost in the experimental pre-processing steps. This is the most likely reason, as the reannealing reaction after crosslink reversal may have favoured the unmodified oligonucleotides. The monoadduct is relatively bulky and there are only 14 complementary bases on each oligonucleotide to be annealed. The nanopore sequenced material contained the insert and large fragment, as verified by Sanger and nanopore sequencing, but most likely did not contain the TMP monoadduct. This would explain why no signal perturbation or bimodal distribution was observed in the nanopore sequencing data (figure 3.20) and why the Sanger sequencing was identical to the control (figure 3.18b).

Future experiments could include performing mass spectrometry analysis on both the reversed and the re-annealed substrate (selecting for this over the separate oligonucleotides by a non-denaturing gel extraction step) to confirm, as far as possible, that the DNA still contained the TMP modification. However, the reannealed oligonucleotides with a monoadduct may still be unstable and biased against in the ligation step. The final 448 bp product could not be analysed by the same mass spectrometry procedure as used in this chapter as this was restricted to short oligonucleotides. However, mass spectrometry involving digestion to single nucleosides could be performed if a standard of the modified base could be provided.

The long term goal of these experiments was to identify whether TMP could be used to mark accessible DNA in a genomic context. If the TMP approach was scaled up to work on genomic DNA, to assess the chromatin state of a cell, there would be many more modifications on each DNA molecule. Each processing step adds an opportunity to bias

against the modified DNA. Crosslink reversal would need to be 100% efficient to allow the DNA strands to separate during nanopore sequencing as both polyacrylamide denaturing gels and mass spectrometry showed that a single crosslink can hold two DNA strands together in denaturing conditions, which are harsher conditions than experienced during nanopore sequencing. The reannealing step after reversal is also likely to bias against the presence of monoadducts in the final substrate. If the reannealing in this chapter biased against the modification with only one monoadduct it would not be possible to apply this approach to molecules with multiple modification sites. Therefore, TMP is not a suitable candidate for marking accessible chromatin if the final readout is going to be performed by nanopore sequencing.

Nanopore sequencing was selected because it can directly detect modified bases on long DNA molecules without an amplification step, allowing for single molecule chromatin accessibility data to be generated. TMP has been shown it can form monoadducts but there is no way to prevent the formation of crosslinks, therefore, to reduce the number of processing steps involved, a chemical that cannot form interstrand crosslinks, angelicin, was investigated. This required working with a new chemical, requiring re-optimisation and validation of DNA modification, building on the optimisation of TMP modification of DNA.

The second chemical to be trialled, angelicin, is an angular furanocoumarin that cannot form crosslinks between two DNA strands, but can form a monoadduct on one strand. This gave it the advantage of already being only on one strand, thus the two strands could be separated during nanopore sequencing without any concerns regarding inefficiently reversed crosslinks. However, when tested experimentally by Sanger sequencing and mass spectrometry, angelicin modification of DNA was not observed despite testing a range of conditions (figure 3.22 and 3.23). In previous studies angelicin has been used at a concentration ten times higher than the TMP concentration with suggestions that this is required due to angelicin

being more hydrophobic and having a lower affinity for DNA [113]. In this chapter, the maximum concentration of angelicin (1 μmol) tested by mass spectrometry was more than ten times the TMP concentration (88 nmol) that produced modified DNA. However, suitable conditions for producing an angelicin monoadduct at a single site on a pair of annealed oligonucleotides were not found. Therefore, angelicin was not taken forward as a candidate for genomic DNA modification.

If the furanocoumarin modification of DNA and subsequent detection had been successful *in vitro*, translating this to *in vivo* in *S. cerevisiae* would have introduced further challenges. Firstly, yeast cells have a cell wall which the molecules would be required to pass through. However, the addition of a furanocoumarin and subsequent UV treatment of *S. cerevisiae* facilitates furanocoumarin modification of DNA within the yeast cell [127, 146, 147], therefore the yeast cell wall does not prevent furanocoumarin entry to the cell. Once in the cell there is a high level of RNA, which TMP is able to modify although with a lower efficiency than for DNA modification [142]. In an *in vivo* system the furanocoumarin modifications could be repaired by the nucleotide excision repair pathway [148]. The repair of crosslinks in *S. cerevisiae* can also lead to occasional mutations of thymine to cytosine at the original site of modification [148]. For *S. cerevisiae*, the *pso2-1* and *pso3-1* mutants with mutations in rad genes could be used to prevent this as these are specifically sensitive to modification by psoralen derivatives and have an inability to repair the introduced modifications [149]. Interstrand crosslinks (and monoadducts to a lower extent) have also been shown to be detrimental to the cell by an observed increase in apoptosis [137]. Therefore it is most likely that the cells would be killed upon treatment with furanocoumarins and UV. An *in vivo* approach would have a very limited time window in which it would be possible to capture a representative biological snapshot of chromatin in unperturbed cells.

In summary, the furanocoumarin treatment of DNA had the potential to work as a chromatin accessibility marker. However, the two selected compounds both had their own technical issues, either not allowing them to fit in the desired pipeline of the overall assay in the case of TMP, or failing to provide initial DNA modification in the case of angelicin. Therefore neither TMP or angelicin were suitable as a marker of accessible DNA. It was identified that nanopore detection of the modification would be one of the biggest challenges in the overall assay and to reduce potential complexities when combining with nanopore sequencing a smaller modification would be optimal. Therefore, in 2017, when publications from Simpson *et al* [107] and Rand *et al* [108] showed the ability of nanopore to detect methylated cytosine and adenine respectively, it was decided to investigate DNA methylation as an alternative accessible DNA marker.

Chapter 4

Use of adenine methylation to probe chromatin structure both *in vitro* and *in vivo* in *S. cerevisiae*

4.1 Introduction

In the previous chapter two chemicals were evaluated for their potential use as chromatin accessibility markers in combination with detection by nanopore sequencing. Neither was successful, therefore new approaches were investigated. In 2017 a paper featuring a new adenine methyltransferase (M.EcoGII) that was not sequence specific [150], was published. In the same year two papers were published showing the ability of nanopore sequencing to detect methylated cytosine [107] and methylated adenine [108]. This inspired a methylation (-CH₃) approach, combining M.EcoGII adenine methylation with nanopore sequencing. This chapter will assess the suitability of M.EcoGII as a marker of accessible chromatin

and explore *in vitro* and *in vivo* methods for using methylation to investigate the chromatin structure of *S. cerevisiae*.

4.1.1 Natural methylation in a variety of organisms

DNA methylation predominantly occurs on two bases: cytosine and adenine. Cytosine methylation is most commonly found in eukaryotic DNA, whereas adenine methylation is most commonly found in prokaryotic DNA.

Methylation is added to DNA by methyltransferases. These vary in their base specificity, and their ability to methylate unmethylated or hemi-methylated DNA (thus to perform *de novo* methylation or to maintain methylation marks during DNA replication). In human cells there are three cytosine methyltransferases. Dnmt3a and Dnmt3b target unmethylated DNA therefore predominantly perform *de novo* methylation [151, 152]. Dnmt1, the most abundant methyltransferase and a component of the human DNA replication complex, targets hemi-methylated DNA therefore maintains methylation marks across DNA replication [153, 152].

Methylation patterns can also play a role in human disease. Absence of DNA methylation is a marker of cancer development [154] however, tumour-suppressor genes in cancer cells are often methylated at their CpG islands [155], leading to transcriptional silencing.

Cytosine methylation (5mC) has been well studied in eukaryotes and plants, predominantly in the CG sequence context. This dinucleotide often forms a repeated palindromic unit, termed a CpG island, where the p represents the phosphate connecting the two bases. CpG islands act as promoters for 60% of human genes [156] and CpG methylation can inactivate the entire X chromosome [157]. Active genes often have unmethylated promoters and methylated gene bodies as methylation at promoter regions can prevent transcription factors from binding although some development-associated transcription factors bind methylated

CpG sites more efficiently [158]. Therefore, cytosine methylation plays an important role in the regulation of gene expression in eukaryotes.

In *E.coli* the two main methyltransferases are Dam and Dcm. They methylate an adenine in a GATC sequence context [159, 160] and a cytosine in a CC(A/T)GG sequence context (where the second cytosine is the target) [159] respectively. Dam methyltransferase can also methylate adenines in other sequence contexts, for example the adenine in the second position of GACC, AATC, CATC and TATC. However, this occurs at a lower frequency than modification of adenine in GATC sequence motifs [161]. Additionally, Dam methyltransferase is unable to methylate adenines in a GATC site at the ends of DNA molecules [159]. Both Dam and Dcm showed similar efficiencies when methylating unmethylated or hemi-methylated DNA, suggesting that the two methylation events, one on each strand, are independent [159].

Dam methyltransferase is involved in the repair of DNA mismatches [162, 163] and the regulation of DNA replication [164, 165] and gene expression [166]. Immediately after DNA replication, GATC sites will be hemi-methylated, with methylation present on the parental DNA strand and absent on the nascent DNA strand [167]. This allows the mismatch repair (MMR) system to recognise the nascent strand, and thus identify the erroneous base at a mis-match site [162]. Therefore, by maintaining methylation patterns across DNA replication, Dam methyltransferase increases the rate of mismatch repair. An overexpression of Dam would lead to an excess of methylation, causing a loss of hemi-methylation and an increase in mismatches as the information on which base is erroneous is lost [162, 163]. DNA methylation by Dam can also act as a marker to ensure that the single DNA replication origin in *E. coli* is activated once and only once per cell cycle [164, 165].

In prokaryotes, DNA has evolved to have high levels of methylation as a defence mechanism against viral invasion. This methylation forms part of restriction modification (RM) systems which comprise a restriction endonuclease and a methyltransferase that both act on the same DNA sequence. The restriction endonuclease is methylation-sensitive, therefore cannot digest methylated host DNA, however is able to fragment unmethylated viral DNA. Not all methyltransferases in the host organisms are involved in RM systems, as they don't all have a matched restriction endonuclease. For example, in *E. coli* Dam methyltransferase is not part of an RM system [168].

Adenine methylation has been detected in a variety of eukaryotes. In *Drosophila*, 6mA is regulated by DNA 6mA demethylase (DMAD) [169]. In *C. elegans* both an adenine DNA methyltransferase (DAMT-1) and DNA demethylase (NMAD-1) have been identified and it was proposed that 6mA could play a role in epigenetic inheritance [170]. In *Chlamydomonas* methylation of adenine appears biased towards ApT dinucleotides [171, 172] and there is a bimodal distribution of 6mA around transcription start sites (TSSs) of active genes in the nuclear genome, in addition to a preference of 6mA towards linker DNA [171]. This indicates that 6mA may work in an opposing fashion to 5mC, by activating gene expression rather than repressing it. Other organisms have evolved to have either no methylation or very minimal levels of methylation. In summary, DNA methylation has existed for a long time in terms of evolution and is present across a divergent range of organisms. However, DNA methylation has also been lost multiple times through evolution, for example in *S. cerevisiae* [173, 174]. The lack of methylation in *S. cerevisiae* [173, 174] allows methylation to be used as a marker, for example for chromatin accessibility studies [94].

In relation to chromatin accessibility, DNA methylation (cytosine or adenine) can induce or enhance DNA curvature and this varies depending on the specific sequence context [175]. For example, the GATC motif itself can modulate DNA curvature, and this effect can be

increased by adenine methylation [176]. This can in turn affect the DNA-binding ability of transcription factors, particularly if the protein relies on both sequence and structural information for binding site identification [176].

There has been controversy over whether or not methylated adenine on DNA is found naturally in humans. Some studies suggest that methylated adenine exists at very low levels (<0.1% of total adenines) [177–179] and is reduced even further in cancers [178]. However, these have been reported to be false positives due to methodological inaccuracy, antibody non-specificity, RNA or bacterial contamination or potential integration of bacterial 6mA into the human genome [180, 181]. Both RNA and mammalian mitochondrial DNA are potential contaminants as they contain methylated adenine, with the latter having at least a 1,300 fold enrichment compared to human genomic DNA [182]. These controversial studies concluded that there was either no, or very minimal levels of 6mA in human genomic DNA, but that more accurate techniques are required for confirmation [180, 181].

4.1.2 M.EcoGII, a non-specific adenine methyltransferase

M.EcoGII methylation of DNA was originally identified in *E. coli* by genome-wide single-molecule real-time (SMRT) sequencing [183]. This was the first time SMRT sequencing had been applied to detect methylation on a genome-wide scale. In this study, they identified strand-specific information for 49,311 putative 6mA sites within the genome, and assessed the level of methylation at each of these sites to identify target sequence motifs [183]. In *E. coli* M.EcoGII is normally inactive however exogenous expression in *E. coli* facilitated analysis of the sequence specificity. This suggests that *E. coli* is able to express a functional enzyme even though the enzyme is prevented from performing its methyltransferase function when expressed endogenously.

M.EcoGII belongs to the class beta-methyltransferases which use S-adenosylmethionine (SAM) to methylate the amino groups of adenine or cytosine DNA bases. The methylation ability of M.EcoGII was further analysed by expression on a pRRS:M.EcoGII vector in the *E. coli* ER2796 strain, which lacks Dam, Dcm and M.EcoKI methyltransferase activities [150]. This confirmed that M.EcoGII has a diverse ability to methylate adenines, with no required sequence context. This was verified through the inability of a set of restriction enzymes to digest M.EcoGII methylated DNA. These enzymes are sensitive to adenine methylation, therefore are unable to cut DNA containing a methylated adenine. M.EcoGII can also methylate adenines in single stranded DNA at a similar efficiency to adenines in double-stranded DNA [150]. Furthermore, M.EcoGII can methylate RNA bases, either in single-stranded RNA, or in either strand of a DNA:RNA-hybrid oligonucleotide duplex [150]. It is reported that M.EcoGII could methylate >85% of adenines *in vivo* and almost 99% of adenines *in vitro* (plasmid DNA) based on liquid chromatography and mass spectrometry (LC-MS) analysis [150]. Therefore, this enzyme is extremely versatile and highly efficient.

These studies provide an insight into M.EcoGII. However, since the enzyme was still relatively uncharacterised my initial experiments as described in this thesis have aimed to replicate these previous *in vitro* findings [150], and to quantify them, in order to learn more about M.EcoGII.

4.1.3 Methylation as a marker of chromatin accessibility

In recent years methylation has been used as a marker of chromatin accessibility by methylating accessible DNA and producing footprints from proteins protecting the DNA from methylation. There are a variety of techniques rapidly evolving in this field, from the use of a single cytosine methyltransferase with short read sequencing, to multiple methyltransferases with long read sequencing (table 4.1). In 2012, a technique called NOMe-seq, combining

GC methylation and Illumina bisulfite sequencing was developed and applied to mammalian chromatin [184]. The assay involved methylating DNA in a cell nuclei, extracting DNA and bisulfite sequencing to identify the location of the methylated cytosines. By looking at two sequence contexts, endogenous CG methylation (regulation of gene expression) and exogenous GC methylation (chromatin accessibility) information on these two aspects could be gained in the same cells. Many other techniques have focused on cytosine methylation as it is a well known modification, therefore there are previous datasets and knowledge of *in vivo* regions of enrichment (i.e. CpG islands) that can be used for validation. More recently, these assays have started to use nanopore sequencing due to its ability to provide long-read information and to detect base modifications from the original DNA, without requiring amplification. However, detection of methylation with nanopore sequencing is also a developing field, therefore the accuracy for single molecule assays has been debatable. The most recent of these nanopore sequencing chromatin accessibility assays is SMAC-seq (single-molecule long-read accessible chromatin sequencing) [116]. This combines the use of two cytosine methyltransferases (M.CviPI which methylates at GC sites and M.SssI which methylates at CG sites) in addition to an adenine methyltransferase (M.EcoGII which methylates at any adenine) and was applied to *S. cerevisiae* chromatin. The accuracy of this method was not sufficient to look at single molecules, therefore these data were always presented as an ensemble. Even so the technique does detect similar regions of accessible DNA to DNase-seq and some overall trends in accessibility can be observed. The most recent developments combining methylation detection and sequencing have involved methylated adenine detection with PacBio long-read sequencing [185, 186]. These have focused on non-specific adenine methylation, either with M.EcoGII [186] or a range of enzymes including M.EcoGII, M.Hia5, M.Btr192IV, M.EcoGI and M.Hin1523 [185]. This provides them with a high resolution on long reads. There are limitations to these methods as PacBio sequencing is still relatively expensive, for larger genomes target enrichment would need

to be performed and methylation requires solubilisation of the oligonucleosomal fragments therefore weak or transient DNA-protein interactions are likely to be lost [186].

Technique	Year	Methylation	Sequencing method
DamID [94]	2000	GATC	None
MeDIP [187]	2005	CG	None
MeDIP-seq [188]	2008	CG	Illumina bisulfite sequencing
NOMe-seq [184]	2012	GC	Illumina bisulfite sequencing
MAPit-patch [189]	2014	GC + CG	Roche 454 sequencing
DamID-seq [93]	2016	GATC	Illumina sequencing
dSMF [190]	2017	GC + CG	Illumina bisulfite sequencing
scNMT-seq [191]	2018	GC	Illumina bisulfite sequencing + RNA sequencing
MeSMLR-seq [192]	2019	GC	Nanopore sequencing
SMAC-seq [116]	2019	GC + CG + A	Nanopore sequencing
ODM-seq [193]	2019	GC + CG	Illumina bisulfite or Nanopore sequencing
NanoNOMe [194]	2020	GC	Nanopore sequencing
Fiber-seq [185]	2020	A	PacBio SMRT-sequencing
SAMOSA [186]	2020	A	PacBio SMRT-sequencing

Table 4.1 Chromatin accessibility assays involving methylation

Overview of techniques developed to use methylation to investigate chromatin structure or protein binding positions. The sequencing method of choice has shifted in recent years away from Illumina sequencing, towards nanopore and SMRT-sequencing which can both provide longer read information. NOMe-seq = Nucleosome Occupancy and Methylome sequencing; dSMF = dual-enzyme Single Molecule Footprinting; scNMT-seq = single-cell Nucleosome, Methylation and Transcription sequencing; MeSMLR-seq = Methyltransferase treatment followed by Single Molecule Long Read sequencing; SMAC-seq = Single Molecule Accessible Chromatin sequencing; ODM-seq = Occupancy measurement via DNA Methylation and high-throughput sequencing; SAMOSA = Single Molecule Adenine Methylated Oligonucleosome Sequencing Assay.

4.1.4 Methylation detection by nanopore sequencing

There have been a variety of pieces of software developed to identify methylated bases in nanopore sequencing data (table 4.2). They can detect methylation in different conditions,

as they differ in the base of interest, sequence context and target type (DNA/RNA). They also differ in how they report the methylation status, for example Flappie marks methylated positions with a 'Z' whereas Guppy analysis with a Rerio model gives a modification likelihood between 0 and 255 for each position.

Nanopolish and signalAlign were the first two pieces of software published for methylation detection in nanopore sequencing data [107, 108]. From the accuracies reported in these papers, the detection of methylated cytosine in nanopore sequencing data was more accurate than for methylated adenine (87% for 5mC methylation [107] versus 74% and 80% for the mean and median accuracy for 6mA detection [108]). This was likely due to the reduced difference in ionic current between methylated and unmethylated adenines, compared to methylated and unmethylated cytosines [108]. In nanopore sequencing data methylated bases can perturb the signal differently depending on the sequence context as multiple bases will be in the pore at the same time. Therefore, models for detecting methylated bases predominantly use 6-mers, and the accuracy of methylation detection varies between 6-mers. Tombo, released by Oxford Nanopore Technologies, is another option however raw data was difficult to access. Since then, other pieces of software have reported higher levels of accuracy for detecting methylated adenine. For example, DeepSignal reports a 99.9% accuracy at detecting 6mA in GATC motifs, compared to 90.8% achieved by signalAlign [195]. If the software detects methylated adenine, it is predominantly in a GATC sequence context. At the time of initiating this methylation project, there was not a good option for detecting methylated adenine in any sequence-context.

Software	Year released	Pore	Target	Methylation	Sequence contexts	Model Approach	URL
Nanopolish [107]	2017	R7.3, R9	DNA	5mC	CpG	HMM	https://nanopolish.readthedocs.io/en/latest/index.html
signalAlign [108]	2017	R7.3, R9	DNA	5mC, 5hmC, 6mA	CC(A/T)GG motifs and GATC	Variable order HMM and HDP	https://github.com/ArtRand/signalAlign
Tombo	2017	Not reported	DNA, RNA	5mC, 6mA, de novo	CG, CC(A/T)GG, GATC plus all-context	Not reported	https://github.com/nanoporetech/tombo
Flappie	2018	R9.4.1	DNA	5mC	CpG	Not reported	https://github.com/nanoporetech/flappie
mCaller [196]	2019	R7.3, R9	DNA	6mA	Variety of contexts	Neural network	https://github.com/al-mcintyre/mCaller
DeepSignal [195]	2019	R9	DNA	5mC, 6mA	CC(A/T)GG, CG and GATC	Deep learning	https://github.com/bioinformaticsCSU/deepSignal
DeepMod [197]	2019	R9.4	DNA	5mC, 6mA	CG, GCGC, GATC, GAAATTC, TCGA	Bi-directional RNN with LSTM	https://github.com/WGLab/DeepMod
NanoMod* [198]	2019	Not reported	DNA	5mC, 6mA (Detects modifications <i>de novo</i> therefore cannot predict type of modification)	Tested on CG, in theory is sequence agnostic	Statistical (Kolmogorov-Smirnov test)	https://github.com/WGLab/NanoMod
EpiNano [199]	2019	R9.4.1	RNA	6mA	All-context	SVM	https://github.com/enovoai/EpiNano/issues
ELIGOS [200]	2020	R9.4.1	RNA	6mA, de novo	All-context	Statistical	https://bitbucket.org/pitoom/eligos.git
Rerio**	2020	R9.4.1, R10.3	DNA, RNA	5mC, 5hmC, 6mA, dU	All contexts for 5mC and 6mA, CpG context for 5hmC	RNN	https://github.com/nanoporetech/erio

Table 4.2 Nanopore software for detecting methylation

Based on table 1 in review by Xu & Seki (2019) [201]. Some of the listed pieces of software have been developed by Oxford Nanopore Technologies, therefore do not have a publication associated with them and have reduced information available on how they work. For each sequence context, the base in bold is the position of methylation that can be detected. * NanoMod requires a matched control dataset which is known to be unmodified. ** Rerio is a set of research release basecalling models and configuration files that can be used with Oxford Nanopore Technologies basecallers (Rerio is not a standalone piece of software). HMM = Hidden Markov Model; HDP = Hierarchical Dirichlet Process; RNN = Recurrent Neural Network; LSTM = Long short-term memory; SVM = Support Vector Machines.

4.1.5 Methylation detection by PacBio SMRT-sequencing

An alternative long-read sequencing method to detect modified bases is single-molecule real-time (SMRT) sequencing, developed by Pacific Biosciences (PacBio) [202]. SMRT-sequencing is a sequence by synthesis method which uses a polymerase to incorporate fluorescently labelled nucleotides to build a new strand complementary to the single stranded DNA template on which the polymerase is positioned. The DNA sequence is determined by the fluorescence released upon nucleotide incorporation as each base is labelled with a different fluorescent dye. This allows SMRT-sequencing to differentiate between canonical and methylated bases by the length of time (inter-pulse duration (IPD)) taken to incorporate each nucleotide as methylated nucleotides cause different polymerase kinetics during incorporation. Similarly to nanopore sequencing, a methylated base in SMRT-sequencing also affects neighbouring bases as multiple bases will be in contact with the polymerase at the same time. As the IPD is sequence-dependent and varies between occurrences of the canonical bases depending on the surrounding sequence context [203], SMRT-sequencing requires an unmodified sample as a control. In theory, this allows the detection of methylated bases by SMRT-sequencing to be sequence independent which would allow it to be used for detection of M.EcoGII methylated DNA. It may prove challenging to analyse endogenous base modifications if there is no available unmodified control sequencing data. With the increased use of SMRT-sequencing, consensus unmodified samples may be generated and used. However, even within a well studied organism like *S. cerevisiae*, strain differences may make it difficult to look at the IPD for certain regions of the genome.

SMRT-sequencing can detect methylated adenine in a GATC sequence context [203]. It can also identify methylated bases in a variety of contexts to characterise the sequence specificity of DNA methyltransferases in bacteria [161] and archaea [204], investigate restriction-

modification systems (RM-systems) [205, 206] and identify the role of methylation in virulence [207, 208].

Both nanopore sequencing and SMRT-sequencing can be used to directly detect base modifications in long-read, single-molecule data. For practical and logistical reasons, nanopore sequencing was selected to produce the final readout of methylation positions for this thesis.

4.1.6 Aim of this chapter

This chapter aims to establish DNA methylation as a suitable marker for *in vitro* or *in vivo* chromatin accessibility in single molecules of *Saccharomyces cerevisiae* DNA, with detection performed by nanopore sequencing. This involves qualitative and quantitative validation of M.EcoGII mediated adenine methylation, optimisation of *S. cerevisiae* chromatin preparations and development of a method to detect M.EcoGII methylated adenine in nanopore sequencing data. There is no endogenous DNA methylation in *Saccharomyces cerevisiae* [173, 174]. Consequently, there will not be any unaccounted background methylation.

4.2 Qualitative analysis of methylation with enzyme digests

The sequence specificity of M.EcoGII was previously investigated by Murray *et al* through the use of a set of 6mA-sensitive restriction enzymes which are unable to digest DNA containing a methylated adenine within their restriction site [150]. The study showed that unmethylated linearised and digested DNA produced two bands while M.EcoGII methylated linearised and digested DNA produced one band. This indicated that M.EcoGII was sufficiently methylating the enzyme recognition sites to prevent digestion. The enzymes tested had a range of recognition sites, which supports the view that M.EcoGII methylates DNA in a wide range of sequence contexts.

4.2.1 Methylation inferred from restriction enzyme protection assay

As M.EcoGII was newly published at the time of starting to use it for this thesis [150], the effect of adenine methylation on a variety of restriction enzyme digests was tested to see how well the published results could be replicated (figure 4.1). The ability of an enzyme known to be sensitive to adenine methylation, to digest DNA treated with M.EcoGII was used as a proxy for the level of methylation present in a population of DNA molecules. If a restriction enzyme was sensitive to adenine methylation, it would show a reduced ability to cut the DNA, therefore leaving a larger proportion of the DNA as uncut (and thus full length). Therefore, it was important to identify which restriction enzymes were sensitive to adenine methylation.

To test the adenine sensitivity of a set of restriction enzymes, pUC19 plasmid DNA was linearised with ScaI-HF, methylated with M.EcoGII and subsequently digested with the restriction enzymes to be tested, with one enzyme tested per reaction. These restriction enzymes each had a single recognition site within the plasmid. Therefore, restriction digest patterns were a result of the digestion efficiency at that particular site. Digestion would produce two bands whereas undigested DNA would be a single band. Two controls were included, no restriction enzyme ('No RE' in figure 4.1a) and digestion with SmaI (whose restriction site does not contain adenines therefore complete digestion is expected due to an insensitivity to adenine methylation). Each restriction enzyme has a different digestion efficiency in unmethylated conditions, therefore digestion of methylated DNA was compared to digestion of unmethylated DNA for each restriction enzyme. Figure 4.1a shows that both SphI-HF and PstI-HF show high levels of protection from digestion in the methylated conditions. SalI-HF also showed high levels of protection but slightly less than for SphI-HF and PstI-HF as there were still faint bands at the sizes of the digested products. EcoRI-HF showed partial protection suggesting that EcoRI-HF is partially sensitive to adenine

methylation. BamHI potentially showed a slight increase in protection but the enzyme was unable to fully digest the unmethylated DNA which made it harder to interpret the level of protection. The ability of M.EcoGII treated DNA to block PstI-HF digest was further investigated with varying volumes of M.EcoGII used to methylate ScaI-HF linearised pUC19 DNA. Figure 4.1b shows that at least 3 μ l M.EcoGII per 1 μ g of DNA is required in a 100 μ l reaction volume.

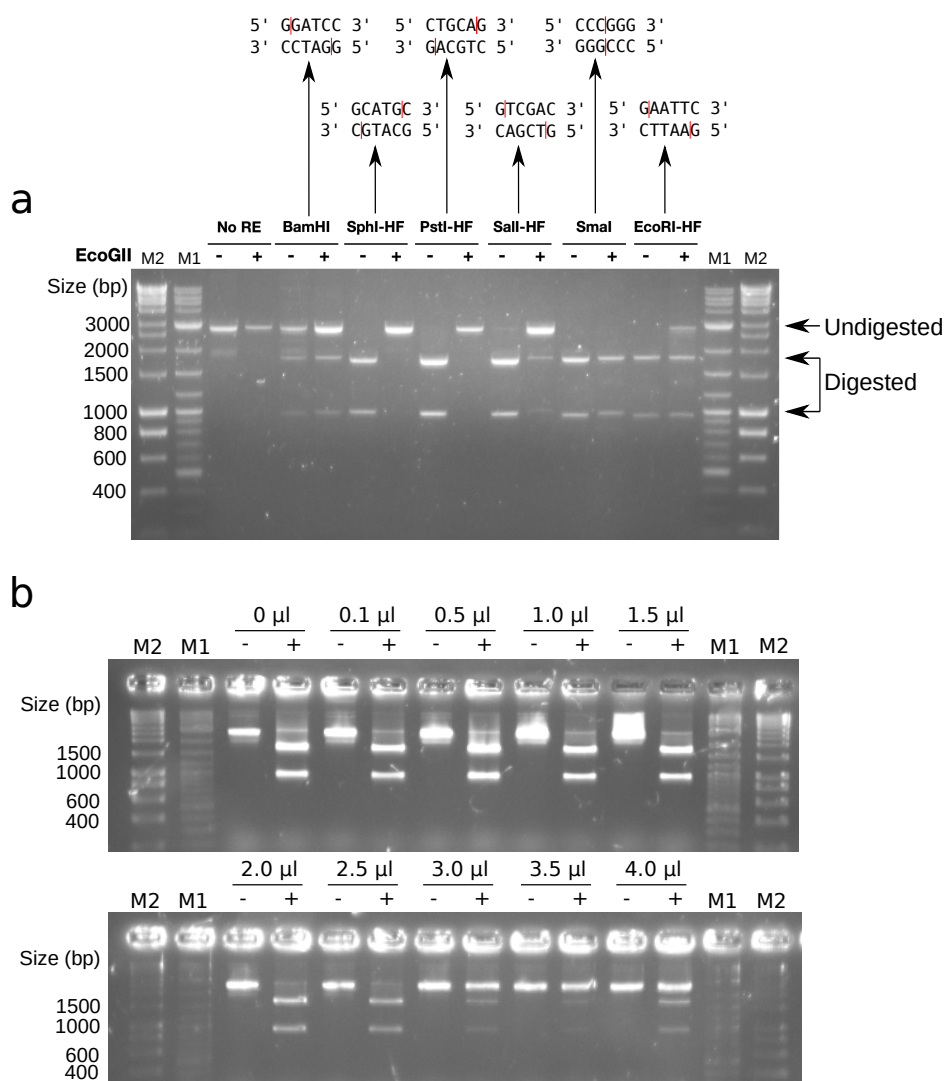


Fig. 4.1 Restriction enzyme protection assay identifies adenine methylation sensitive restriction enzymes

Assay to identify restriction enzymes that are sensitive to adenine methylation for subsequent use as a proxy to measure methylation levels. **a)** 1.5% TBE-agarose gel showing the level to which a variety of restriction enzymes are affected by adenine methylation by M.EcoGII. If the ScaI-HF linearised pUC19 DNA was digested it produced two bands whereas the undigested material produced a single band around 3000 bp. If methylation with M.EcoGII protects the DNA from digestion, the enzyme is sensitive to adenine methylation and can be used to assess the level of methylation at its recognition site. Red lines = restriction enzyme cut sites; - = mock-methylated with water; + = treated with M.EcoGII. **b)** 1.5% TBE-agarose gel showing the effect of different volumes of M.EcoGII per 100 µl reaction volume with 1 µg of pUC19 DNA previously linearised with ScaI-HF. - = mock-digested with water; + = digested with PstI-HF. The volumes relate to the amount of M.EcoGII used in each reaction. For both gels: The gels included Redsafe staining and were run at 90 V. M1 = NEB Quickload 2-log ladder (5 µl); M2 = Bioline Hyperladder 1kb (5 µl).

4.3 Quantitative analysis of methylation using Droplet Digital PCR (ddPCR)

Droplet Digital PCR (ddPCR) is a technique employed to study DNA levels by the absolute quantification of nucleic acids [209] with applications including DNA replication [210]. A comparison with qPCR showed ddPCR to be more accurate, more precise and easier to use [211].

There are three main stages when performing ddPCR experimentally: generate droplets, amplify DNA by end-point PCR and read fluorescence in each droplet. Firstly, the template DNA, primers and dye are split into approximately 20,000 water-oil droplets per sample [209]. Since PCR takes place in droplets, each PCR is an independent reaction. This requires much lower DNA concentrations than standard PCR as each droplet should receive at most one copy of template DNA. If the DNA template is cleaved (i.e. by digestion), or there was no template in the droplet, none or only a low level of DNA will finally be detected ('negative' droplet). In contrast, if the template was complete, then PCR will amplify the DNA, creating a high level of DNA in the droplet ('positive' droplet). Each droplet is passed through a droplet flow cytometer which counts the number of positive and negative droplets. The amount of DNA in each droplet is measured by fluorescent signals as the DNA has been treated with EvaGreen supermix which has a dye for double-stranded DNA. Therefore, as DNA accumulates in the droplets, the green signal increases. In the context of using this as a quantitative method for the restriction enzyme protection assay, the digested material will produce many 'negative' droplets, therefore a low overall DNA concentration, while the undigested material will successfully undergo PCR and so produce many 'positive' droplets, therefore a high DNA concentration. To measure the level of methylation, regions containing one recognition site of each enzyme were selected. To get clear measurements of the level of

methylation PstI-HF was selected for digestion as it has a high level of sensitivity to adenine methylation. Similarly to the restriction enzyme protection assay, the two controls of no restriction enzyme (water) and SmaI digestion were used, for no digestion and complete digestion respectively.

Accurate ddPCR measurements relied upon the ability of PCR to amplify methylated DNA. Therefore, this was assessed by methylating *S. cerevisiae* genomic DNA with M.EcoGII (or mock-methylating with water) at 37 °C overnight to facilitate a high level of methylation, performing a PCR clean up to remove M.EcoGII, followed by PCR and visualisation on an agarose gel. The presence of a band of similar intensity in both the unmethylated and methylated condition in figure 4.2 indicates that the PCR was not affected by methylation. Therefore, primers can bind to methylated DNA and the ddPCR approach can be used to quantitatively measure the level of methylation.

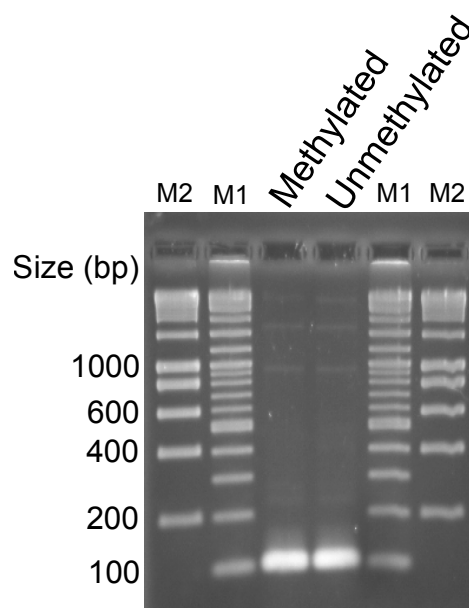


Fig. 4.2 PCR on methylated DNA

2% TBE-agarose gel showing a PCR product obtained from genomic DNA treated with or without M.EcoGII (non-specific adenine methyltransferase). For the PCR, primers CA1029 and CA1028 were used to amplify a 103 bp region, using a 30 second annealing step at 47.5 °C, 30 second elongation step at 72 °C and 35 PCR cycles. DNA was not PCR cleaned prior to running on the gel. The gel was run at 90 V for 1 hour, post-stained with ethidium bromide and washed with water prior to visualisation. M1 = NEB Quickload 2-log ladder (5 µl); M2 = Bioline Hyperladder 1kb (5 µl).

4.3.1 Quantitative ddPCR on plasmid DNA

To test whether or not ddPCR is a suitable method for quantifying the level of methylation by measuring the level of digested template DNA, two criteria needed to be met. Firstly and most importantly, ddPCR needed to be able to distinguish between undigested and digested DNA templates. For this, PstI-HF was selected as this has been shown to be able to efficiently digest DNA and it is blocked by adenine methylation (figure 4.1). Secondly, ddPCR needed to be able to analyse plasmid DNA as this provided a simpler model for testing. Therefore, a region on pUC19 plasmid DNA encompassing one PstI-HF site was selected. ddPCR is often used on genomic DNA, where approximately 0.4 ng of DNA is used per reaction. The

concentration required for plasmid DNA would be substantially less as the ratio of total DNA to desired DNA region for amplification is several orders of magnitude different.

To find an optimal plasmid DNA concentration for ddPCR, a dilution series was performed on undigested and digested DNA samples. When analysing ddPCR data the individual droplets can be viewed prior to the final concentration. An optimal DNA concentration would be characterised by two discrete groups of data points, one for the ‘negative’ droplets (at the bottom), and one for the ‘positive’ droplets (higher up), facilitating robust separation and quantification of the ratio between these two groups. From figure 4.3a and b it can be observed that the dilutions shown provide good separation between the two groups, although there are more droplets in between the two groups for the PstI-HF digested samples than in the controls. A large difference was observed between the concentrations of the water treated and PstI-HF digested samples for all dilutions as seen in figure 4.3c. Therefore, the level of digestion, and hence potentially the level of methylation, can be determined by ddPCR. Furthermore, the optimal concentration will produce a reading between 1,000 and 3,000 copies/ μl which is the range in which ddPCR readings are most accurate. The 1:375,000, 1:500,000 and 1:1,000,000 dilutions, which correspond to 4.1×10^{-5} ng/ μl , 3.1×10^{-5} ng/ μl and 1.6×10^{-5} ng/ μl respectively, were all within this range for the water treated sample. As a result, 2.5×10^{-5} ng/ μl was used in subsequent plasmid ddPCR experiments. The same pattern was observed with both sets of primers, validating the method for use with different primer sets.

Now that a suitable concentration had been determined for ddPCR analysis of plasmid DNA and PstI-HF digestion can be detected, the ability of ddPCR to detect DNA digestion was further investigated with other restriction enzymes. Two controls (water and SmaI (which is insensitive to adenine methylation)) and the two enzymes that showed the greatest sensitivity to adenine methylation (PstI-HF and SphI-HF) in figure 4.1 were used. This

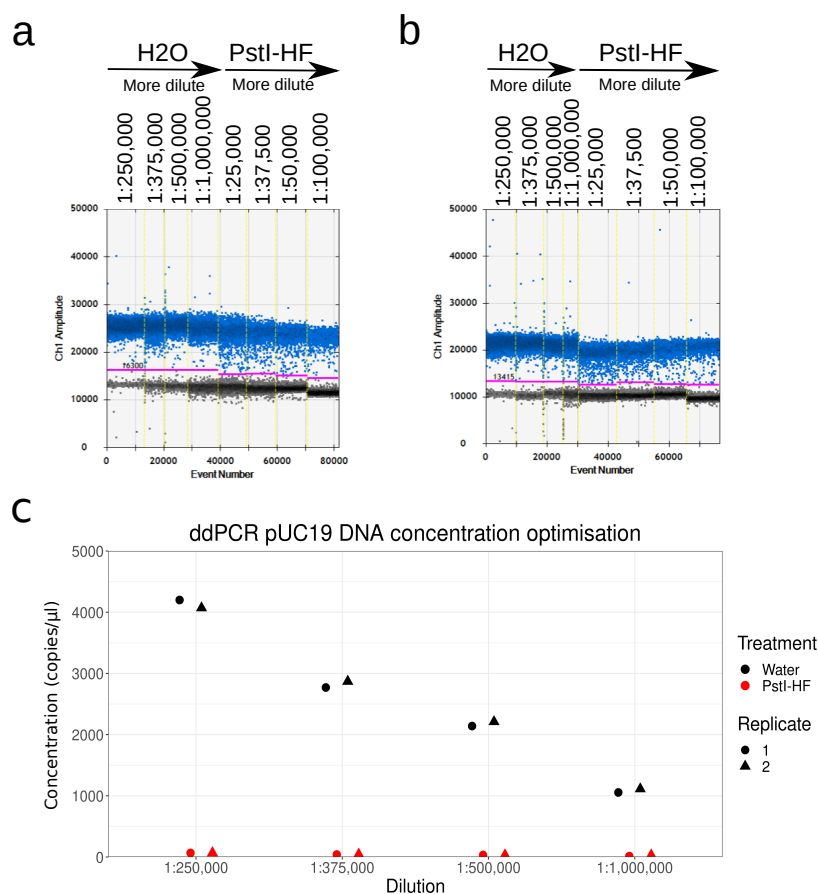


Fig. 4.3 Optimisation of plasmid DNA concentration for ddPCR

ddPCR can be used to quantitatively measure the level of methylation based on digestion by an adenine methylation sensitive restriction enzyme. Digestion of the template prevents amplification in the PCR step of ddPCR, leading to a low concentration readout compared to undigested DNA. Two primer pairs were used to analyse the same region of pUC19 ScaI-linearised DNA (a) HJ019 + HJ020 b) HJ021 + HJ022, forming two replicates. These encompassed a region with a single cut site for PstI-HF. For each primer pair, four samples were treated with water and four were treated with PstI-HF. Each set of four samples was a range of dilutions of the plasmid DNA, however the PstI-HF digested samples were 10x more concentrated than the samples treated with water as the digestion was expected to lead to a lower concentration of amplified DNA. Starting concentrations prior to dilution were 15.5 ng/μl for the water-treated sample, and 13.2 ng/μl for the PstI-HF treated sample, measured using a nanodrop. (a) and (b) show the individual droplet readings. Blue = 'positive' droplets; grey = 'negative' droplets; pink line = cut off point between 'negative' and 'positive' droplets. (c) The ratio of 'positive' to 'negative' droplets is used as the final concentration for each sample. The PstI-HF digested samples were analysed at a 10x higher concentration to allow for more accurate measurements. The data points in (c) have been corrected for the 10x difference in concentration to allow comparison with the water-treated samples. The replicates are shown by the different shapes, and the treatment with either water or PstI-HF is shown by the colour (black = water treatment; red = PstI-HF digest).

experiment involved linearising pUC19 with ScaI-HF prior to treating the DNA with water, SmaI, PstI-HF or SphI-HF. Good droplet separation was observed across all samples (figure 4.4a), with the best separation for primer pair DB78/79. There is a minor reduction in the distinction between ‘negative’ and ‘positive’ droplets for successfully digested samples with low numbers of ‘positive’ droplets. However, for these samples (for example, the samples amplified with DB30/31) the majority of the droplets are clearly seen by the black band among the ‘negative’ droplets. In figure 4.4b the two control primer sets behave as expected, with high concentrations for DB78/79 that encompasses a region with no ScaI-HF, SmaI, PstI-HF or SphI-HF sites, and low concentrations for DB30/31 that encompasses a ScaI-HF site. All samples were linearised with ScaI-HF, therefore this site should always be digested, as is observed in figure 4.4b. In pUC19 there is a SmaI, PstI-HF and SphI-HF recognition site within a region of less than 200 bp, therefore digests at these sites were assessed by ddPCR with the same primers. To assess variation due to different primer binding affinities, two primer sets covering this region (HJ019/20 and HJ021/22) were used. These produced high concentrations when undigested and low concentrations when digested with SmaI, PstI-HF or SphI-HF. Figure 4.4b shows that these two primer sets produce very similar results, therefore the choice of primers does not affect the results. In summary, ddPCR can be used to clearly differentiate between digested and undigested DNA for these enzymes.

Now that ddPCR can differentiate between undigested and digested DNA (figure 4.4) ddPCR was applied to the detection of methylated adenines based on restriction enzyme digests. pUC19 DNA was amplified from *E. coli* expressing Dam methyltransferase therefore will contain methylated GATC sites. MboI, a restriction endonuclease with a recognition site of GATC and a sensitivity to adenine methylation, was used to confirm the presence of GATC methylation. As before, the SmaI, PstI-HF and MboI recognition sites are close enough together on the pUC19 plasmid to allow the same primer sets to be used for all enzyme digests and are the same as was used in figure 4.4. Unlike in figure 4.4, figure 4.5 shows

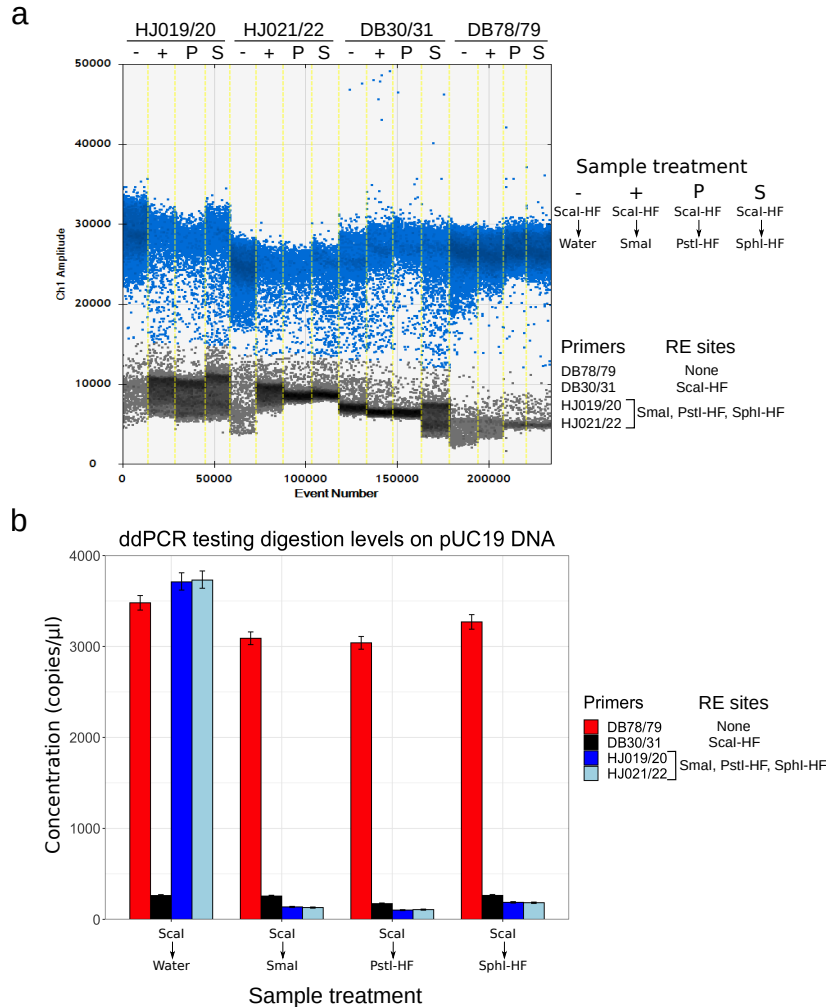


Fig. 4.4 Detection of enzyme digestion by ddPCR

The ability of ddPCR to detect digested versus undigested DNA was assessed on unmethylated pUC19 plasmid DNA. This involved linearising the DNA with ScaI-HF and then treating with water as a negative control, SmaI, PstI-HF or SphI-HF. These enzymes were selected based on figure 4.1 which showed SmaI to be a control that is not affected by adenine methylation, and showed both PstI-HF and SphI-HF to be highly sensitive to adenine methylation. Two primer sets (HJ019/20 and HJ021/22) cover the same region but bind in different places, showing that the results are reproducible across variation in primers. **a)** The DNA concentration measured for each droplet for each sample. ‘Positive’ droplets = blue; ‘Negative’ droplets = grey; - = ScaI-HF linearised and mock-digested with water; + = ScaI-HF linearised and digested with SmaI which is insensitive to adenine methylation; P = ScaI-HF linearised and digested with PstI-HF which is sensitive to adenine methylation; S = ScaI-HF linearised and digested with ScaI-HF which is sensitive to adenine methylation; DB78/79 = The region does not contain a ScaI-HF, SmaI, PstI-HF or SphI-HF recognition site; DB30/31 = The region contains a ScaI-HF recognition site; HJ019/20 and HJ021/22 = Two sets of primers that cover a region containing a single recognition site for each of SmaI, PstI-HF and SphI-HF. **b)** The DNA concentration overall per sample, which is calculated using a ratio of ‘positive’ to ‘negative’ droplets. Error bars represent the Poisson confidence values.

greater variability in the concentration of the region lacking any ScaI-HF, SmaI, PstI-HF or SphI-HF cut sites (red bars) across the different samples. This indicates that digestion levels should be compared within each group of 4 (relative to the water and SmaI digested controls) rather than across conditions. Supporting this, within each group of four samples, the low and high concentration bars correlate to similar heights. From figure 4.5a it can be seen, despite the variation, that MboI is unable to digest both water and M.EcoGII treated DNA thus indicating that the adenine in the GATC site is methylated. The high concentrations in both water and M.EcoGII treated samples for the MboI digests is consistent with the GATC site being almost completely methylated prior to M.EcoGII treatment. However, MboI was not tested on unmethylated DNA, therefore the chance of MboI failing to digest the DNA for reasons other than methylation, for example inefficient enzyme digest, were not eliminated. Importantly, in figure 4.5a and b, PstI-HF has been shown to be able to detect methylated DNA. This is observed by low concentrations detected for both primer sets (light and dark blue bars) in the water (mock-methylation) followed by PstI-HF digestion condition, and high concentrations (relative to the regions with no restriction enzyme sites (red bars)) detected for both primer sets in the M.EcoGII followed by PstI-HF digestion condition. In figure 4.5b, it can be seen that SphI-HF behaves in the same way as PstI-HF. Therefore, this assay demonstrates that M.EcoGII can specifically protect DNA from PstI-HF and SphI-HF digestion, demonstrating the potential for ddPCR to be used in a quantitative methylation-dependent restriction enzyme protection assay.

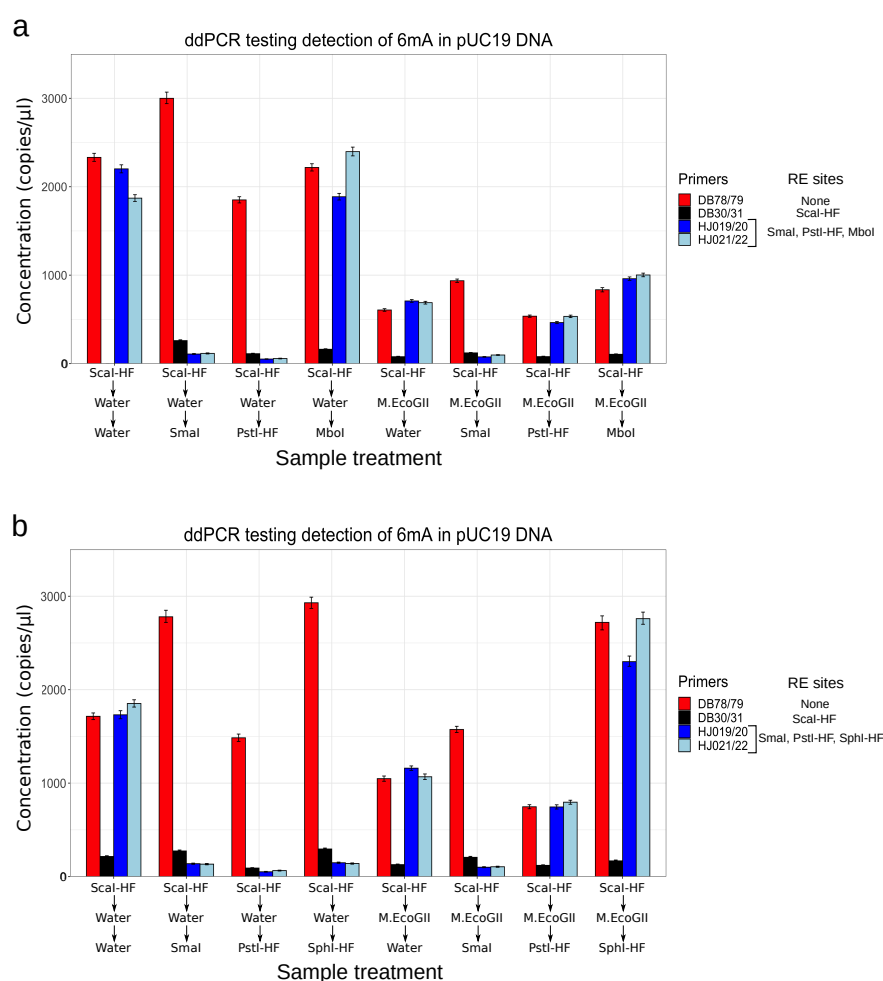


Fig. 4.5 Detection of 6mA in plasmid DNA by ddPCR

pUC19 DNA was ScaI-HF linearised, either treated with M.EcoGII (adenine methyltransferase) or water, and then treated with either water as a negative control, SmaI, PstI-HF, MboI or SphI-HF. The water, SmaI and PstI-HF treatment conditions were included in both of these independent experiments. RE = Restriction enzyme. **a)** DNA copies measured by ddPCR after linearisation with ScaI-HF, methylation by M.EcoGII or mock-methylation with water and subsequent treatment with water, SmaI, PstI-HF or MboI. No data was collected for DB30/31 for the ScaI, water (mock-methylation) and water (mock-digestion) condition due to a problem during the ddPCR process. **b)** Same as for **(a)** but with SphI-HF digest instead of MboI. For both panels the error bars represent the Poisson confidence values.

4.3.2 Quantitative ddPCR on genomic DNA

It was previously shown that PCR can amplify methylated DNA (figure 4.2), however it was important to verify this more quantitatively to detect any subtle changes. Therefore, the

accuracy of ddPCR was tested on methylated and unmethylated DNA, to further validate whether methylation itself would affect the measured DNA concentration. Figure 4.6 shows that two test primer pairs were consistent with each other, there was minimal variation between the three replicates of water and M.EcoGII treated samples and there was no difference between the water or M.EcoGII measured concentrations. The methylation status of the M.EcoGII-treated DNA was not assessed by any other means however the reaction was performed according to New England Biolabs (NEB) instructions therefore some methylation should be present. Therefore, amplification of the DNA in ddPCR is not affected by M.EcoGII methylation.

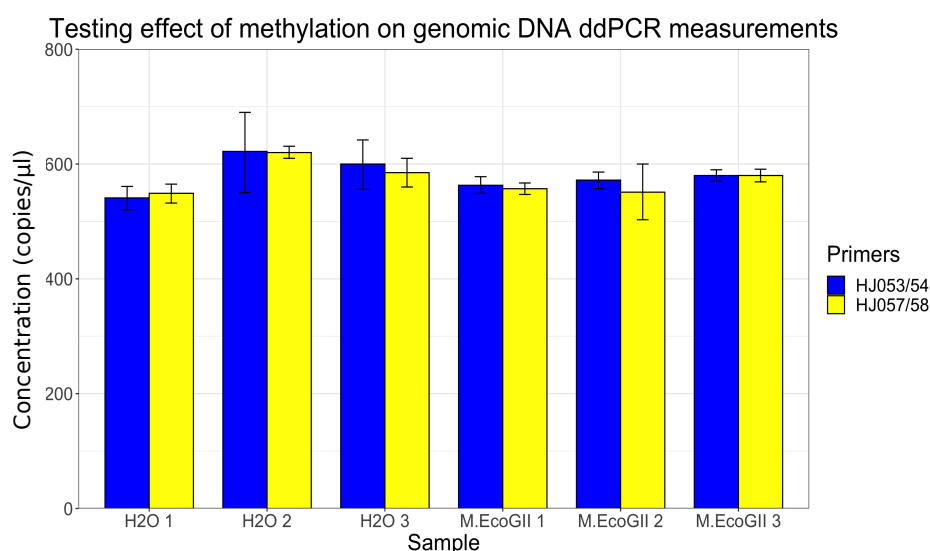


Fig. 4.6 Testing ddPCR accuracy on methylated DNA

Genomic DNA was extracted from T7107 wild-type *S. cerevisiae*, treated with either water or M.EcoGII (in triplicate), and ddPCR employed to amplify the DNA and measure the resulting concentration. Two primer pairs were used to assess reproducibility. Two technical replicates (each sample was run twice on the same ddPCR run) are represented by the error bars.

Another test was performed on *S. cerevisiae* genomic DNA treated with either water or M.EcoGII, followed by water or PstI-HF and ddPCR analysis. Inclusion of no DNA and no primer controls show low DNA concentrations, confirming that these components are required for ddPCR reactions (figure 4.7). This figure also shows that at one site in genomic

DNA amplified by HJ041/42 methylation by M.EcoGII prevents PstI-HF digestion, and the two replicates behave similarly.

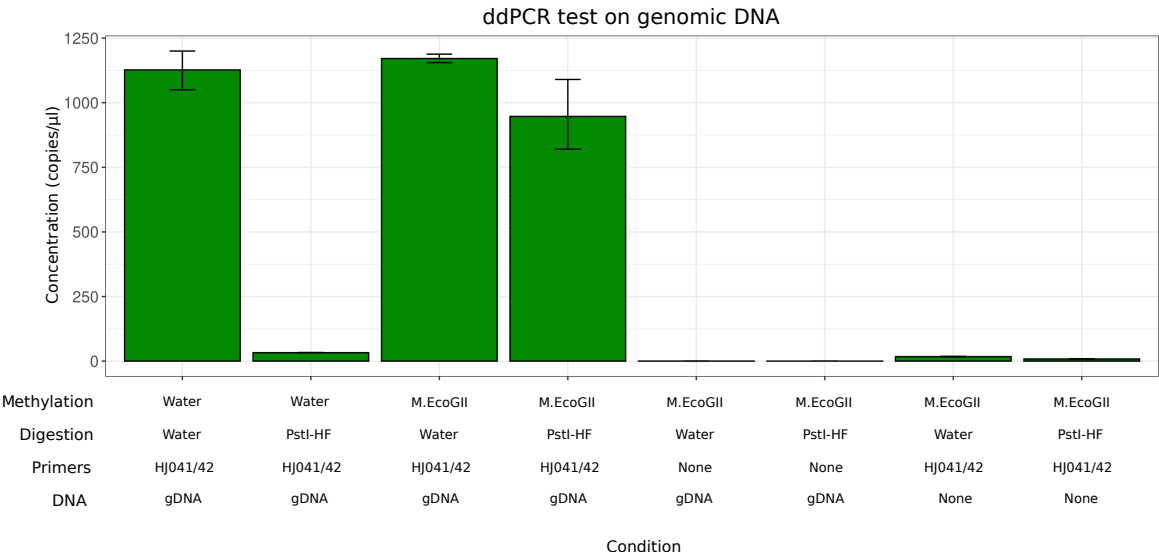


Fig. 4.7 ddPCR quantification of 6mA at one site in genomic DNA

ddPCR was used to assess the level of methylation achieved at a specific loci in genomic DNA treated with M.EcoGII. Genomic DNA was extracted from wild-type T7107 *S. cerevisiae*, fragmented with ScaI-HF, treated with either M.EcoGII or water and then treated with either PstI-HF or water. Controls of no DNA and no primers were included. Error bars represent two duplicates of each sample run on the same ddPCR run.

To test the ability of this ddPCR method in a variety of sequence contexts, and to apply it to a more biologically relevant model the template DNA was switched from plasmid DNA to *S. cerevisiae* wild-type genomic DNA. Similarly to the plasmid experimental design, each region to be analysed contained a single restriction site for PstI-HF. This enzyme was selected as it performed most reliably in the plasmid ddPCR assays. Genomic DNA was extracted from the T7107 wild-type *S. cerevisiae* strain, digested with ScaI-HF to fragment the DNA, incubated at 37 °C for 1 hour with either M.EcoGII or water, incubated at 37 °C for 1 hour with PstI-HF or water and analysed by ddPCR. Starting with the same ScaI-HF fragmented DNA, two different dilutions were used, with either 0.05 ng/μl or 0.17 ng/μl per ddPCR reaction for figure 4.8a and b respectively.

In the ddPCR analysis, eight sets of primers covering eight regions of approximately 200 bp across the genome were selected, including two control regions. The negative control primer pair (DB9/10) covered a region with a ScaI-HF restriction site, as the DNA has been fragmented with ScaI-HF prior to methylation, therefore this template is always digested. The positive control primer pair (DB85/86) covered a region that did not contain either a ScaI-HF or PstI-HF restriction site therefore should not have been digested. In figure 4.8 the two controls behave as expected: the positive control (red bar) is high in all four conditions and the negative control (black bar) is low in all four conditions. The six sites of interest were selected to have different chromatin contexts for future application of the same set of sites to chromatin treated with M.EcoGII. Therefore, the regions assessed with primer pairs HJ037/38, HJ039/40 and HJ041/42 correspond to well-positioned nucleosomes [89] which theoretically protect the DNA from methylation. Conversely, the regions assessed with primer pairs HJ053/54, HJ055/56 and HJ057/58 correspond to nucleosome-free regions [89] which should be accessible to DNA methylation. In the current experiment the DNA was extracted prior to methylation therefore these six positions should all behave similarly as there should not be any remaining nucleosomes. For the condition that was treated with water in both incubation steps (thus mock methylated and mock digested), the six sites of interest produced high concentrations. Conversely, the condition which is unmethylated and digested with ScaI-HF produced low concentrations for all six sites, although the primer pair HJ041/42, and to a lower extent HJ053/54, displayed an indication of incomplete digestion due to a higher concentration, particularly in figure 4.8a. Therefore the methylation status at this site cannot be confidently determined by ddPCR. For the two methylated conditions, both the undigested (water treated) and PstI-HF digested conditions produced high concentrations, indicating that the M.EcoGII treatment methylated the DNA sufficiently at these sites for protection against PstI-HF digestion. For all conditions with high concentrations there was minor variability in absolute values. However, it is easy to distinguish between digested

and undigested DNA conditions, thus allowing methylation status to be inferred. This may suggest that the M.EcoGII methylation has not occurred exactly equally at all genomic sites, which may be due to a minor bias in the sequence specificity of M.EcoGII. It has previously been shown in figure 4.4b that the primer choice does not have an effect on the result, however this was only investigated at one site. The genomic DNA had been treated with proteinase K and RNaseA, thus should not have any proteins bound to block access of the methyltransferase to the DNA. However, the DNA may have formed secondary structures, making some sites more challenging for the primers to access. In addition, the primers may have different DNA binding affinities. This experiment has validated the use of ddPCR to identify if a specific genomic site is generally methylated or not (for most sites), however the variation indicates that it may not be possible to accurately quantify the level of methylation over a population of DNA molecules.

Site	Primers	Genomic location	Region size (bp)	Recognition site
Positive control	DB85/86	ChrVI:198,986-199,176	190	No PstI-HF or ScaI-HF recognition site
Negative control	DB9/10	ChrXIV:249,990-250,139	149	ScaI-HF recognition site
Regions of interest	HJ037/38	ChrII:575,374-575,519	145	PstI-HF recognition site
	HJ039/40	ChrII:617,438-617,578	140	PstI-HF recognition site
	HJ041/42	ChrIII:94,091-94,258	167	PstI-HF recognition site
	HJ053/54	ChrII:595,391-595,548	157	PstI-HF recognition site
	HJ055/56	ChrII:649,394-649,564	170	PstI-HF recognition site
	HJ057/58	ChrII:654,234-654,377	143	PstI-HF recognition site

Table 4.3 Genomic DNA regions for ddPCR quantification of methylation

Six regions of interest containing a PstI-HF recognition site were analysed by Droplet Digital PCR (ddPCR) to identify their adenine methylation status at the PstI-HF recognition site. Region size includes primer binding sites. bp = base pair.

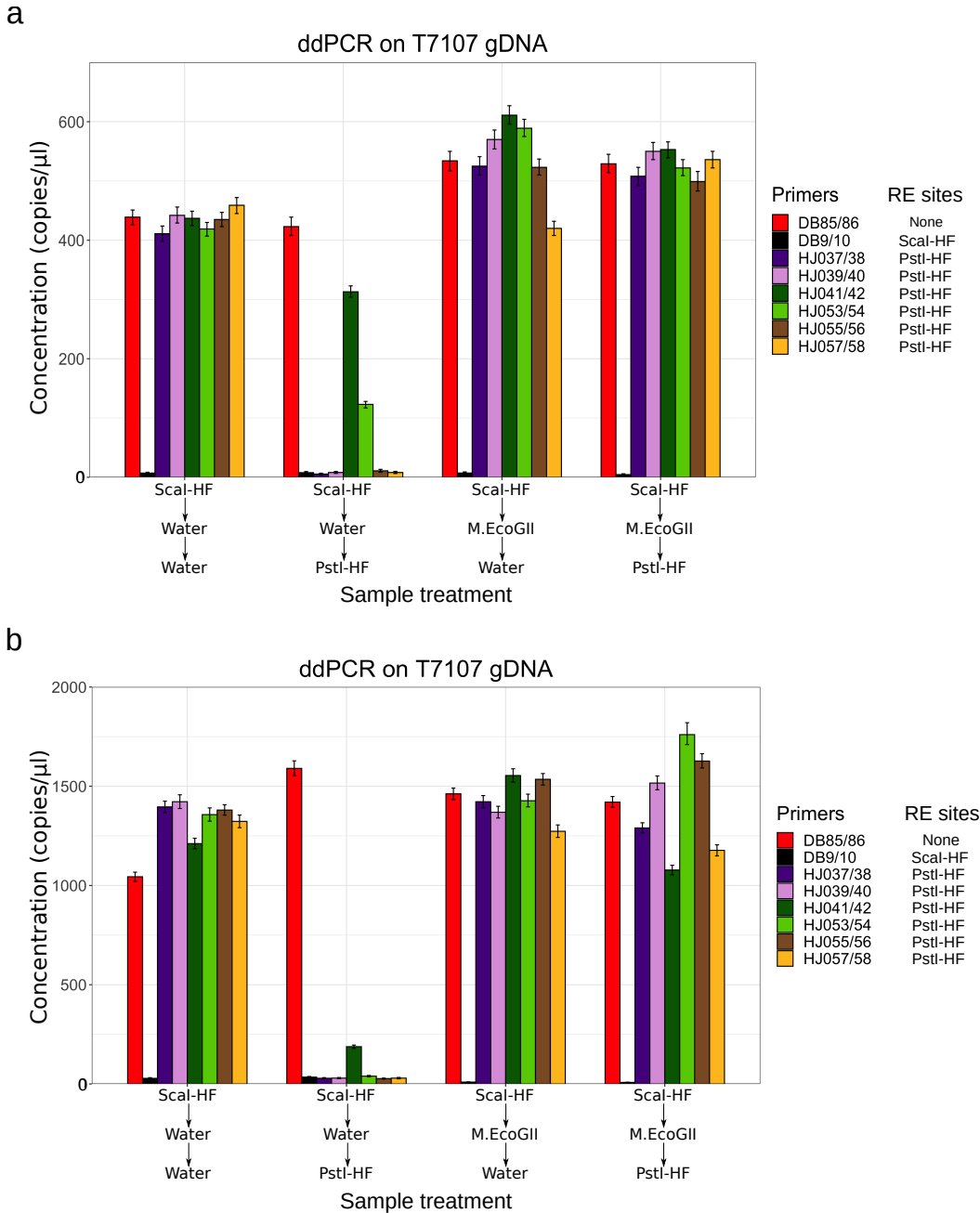


Fig. 4.8 ddPCR quantification of 6mA at multiple sites in genomic DNA

ddPCR was used to assess the level of methylation achieved at specific loci in genomic DNA treated with M.EcoGII. Genomic DNA was extracted from wild-type T7107 *S. cerevisiae* (W303 background), fragmented with Scal-HF, treated with either M.EcoGII or water and then treated with either PstI-HF or water. Two control sites (a negative control containing no PstI-HF or Scal-HF sites and a positive control containing a Scal-HF site) and six sites of interest each containing a single PstI-HF restriction site were analysed to test the reproducibility of the ddPCR quantification of methylation level across multiple sequence contexts (see table 4.3 for genomic locations analysed). RE = Restriction enzyme. Error bars represent the Poisson confidence values. **a)** and **b)** are two repeats of the same experiment. Concentrations were below the optimal levels for most accurate results in **a)** at 0.05 ng/ μ l DNA per sample, therefore a higher concentration of 0.17 ng/ μ l was used for **b)**.

4.3.3 Quantitative ddPCR on recombinant nucleosomes

To assess the ability of nucleosomes to protect DNA from M.EcoGII methylation, a plasmid (pG5E4) with recombinant human H3.1 polynucleosomes (two molecules each of histone H2A, H2B, H3.1 and H4) expressed in *E. coli* was used. This plasmid contained nine 5S rDNA nucleosome positioning sequences, each of which can wrap one histone octamer and five GAL4 binding sites adjacent to an E4 promoter which form a dinucleosome. Active Motif claimed that a total of eleven nucleosomes could be formed (figure 4.9) which they measured by A280 absorbance and SDS-PAGE gels. The nucleosome positions in figure 4.9 may be more stable however the nucleosomes may also be able to move into the other half of the plasmid.

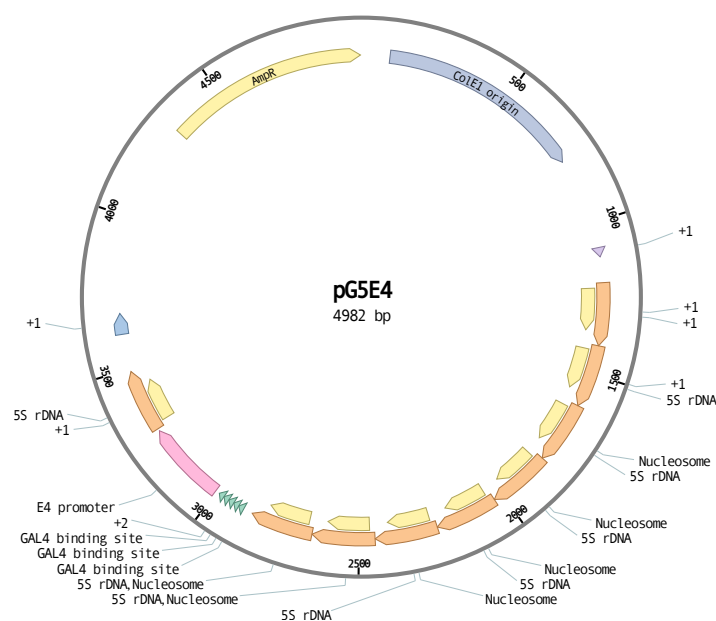


Fig. 4.9 Recombinant nucleosome plasmid

Schematic of the pG5E4 recombinant nucleosome plasmid from Active Motif. The position of nine nucleosomes are shown in yellow in the bottom half of the plasmid. The other two nucleosomes are thought to be positioned over the E4 promoter (pink) and GAL4 binding sites (green). Schematic produced with Benchling.

To test the protection offered by nucleosomes to the DNA, a methylation sensitive restriction enzyme with sites within the nucleosome bound regions of pG5E4 needed to be selected. The pG5E4 plasmid contained a single recognition site for both PstI-HF and SphI-HF and no sites for SallI-HF, all of which had previously been shown by either gel electrophoresis or ddPCR to be sensitive to adenine methylation. There were 26 MboI (GATC) recognition sites however, the plasmid preparation most likely involved amplification in *E. coli*, which would introduce methylation at GATC sites by Dam methyltransferase. This prevents GATC methylation from being used to assess chromatin accessibility. Therefore, a new methylated adenine sensitive restriction enzyme needed to be found. Other restriction enzymes with a large number of recognition sites in pG5E4 included HpaII (21 sites) and AluI (27 sites). Of these, HpaII was unsuitable for detection of methylated adenine due to its recognition site (CCGG) not containing adenines. This made it suitable for linearising and fragmenting the plasmid after methylation, which could not be done prior to methylation as this would have perturbed the chromatin structure. Conversely, AluI has a recognition site of AGCT. Therefore, the methylation sensitivity of AluI was investigated as AluI was not assessed in figure 4.1.

To test the sensitivity of AluI to adenine methylation, pUC19 DNA was linearised with ScaI-HF, methylated with M.EcoGII or mock-methylated with water, treated with either water or AluI and analysed by ddPCR. Similarly to figures 4.3, 4.4 and 4.5, the region in pUC19 encompassed by the primer pair HJ019/20 was assessed. However, this contained three AluI sites, therefore an additional primer pair (DB49/515) which only encompassed one AluI site was also used to simplify the interpretation of the results. In figure 4.10 the ScaI-HF, M.EcoGII, water treated sample had low concentrations for all samples at which point the ddPCR readings are less reliable but the pattern within the four was as expected, with digestion only seen for ScaI-HF linearisation. Figure 4.10 shows that the single AluI site analysed by DB49/515 shows a high level of sensitivity of AluI to adenine methylation due

to the high yellow bar in the ScaI-HF, M.EcoGII, AluI treated sample compared to the low bar in the ScaI-HF, water, AluI treated sample. This suggests that AluI can be used to assess the level of adenine methylation at a specific site. Conversely, the ScaI-HF, M.EcoGII, AluI treated sample shows a low DNA concentration for the region containing three AluI sites, suggesting that at least one of these was often unmethylated, therefore able to be digested. Therefore, ddPCR cannot be used to assess multiple sites within one region. As AluI showed a clear distinction when analysing one recognition site per region, it can be used to probe the ability of M.EcoGII to methylate nucleosomal DNA in the chromatinised pG5E4 plasmid where the AluI sites are sufficiently separated.

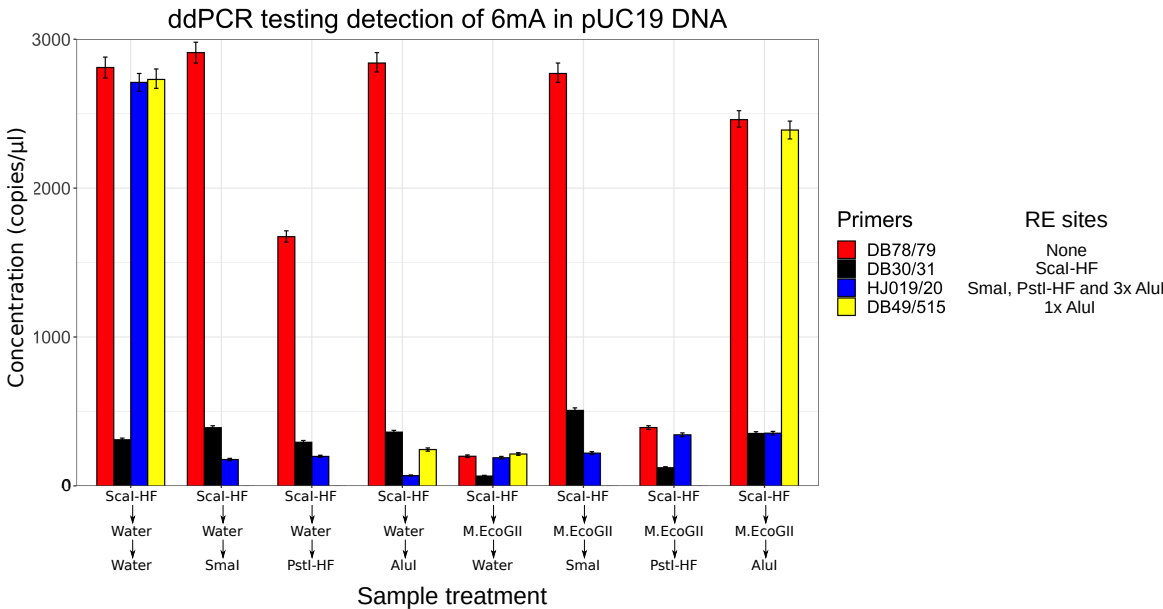


Fig. 4.10 ddPCR assessment of the methylated adenine sensitivity of AluI

pUC19 DNA was linearised with ScaI-HF, treated with water or M.EcoGII, treated with water (mock-digestion control), SmaI, PstI-HF or AluI. The region encompassed by HJ019/20 contained one SmaI, one PstI-HF and three AluI sites, while the region encompassed by DB49/515 only contained one AluI site. RE = Restriction enzyme.

To assess the ability of adenine methylation to probe chromatin structure, the methylation status of AluI and PstI-HF sites in a chromatinised plasmid (pG5E4) were analysed by ddPCR to identify whether these loci were nucleosome bound or not, as nucleosome-bound

DNA should protect DNA from methylation [212]. The pG5E4 plasmid with recombinant nucleosomes was incubated with M.EcoGII or water, heated with SDS to inactivate M.EcoGII and denature the nucleosomes, treated with proteinase K to degrade the nucleosomes, then DNA was extracted and fragmented with either HpaII (figure 4.11a) or BamHI (figure 4.11b), and then treated with either water or a restriction enzyme and analysed by ddPCR to assess the methylation state of each site. A range of sites were chosen throughout the plasmid (figure 4.11c): one positive control with no HpaII, AluI or PstI-HF sites (CN2301/2309), one negative control with a HpaII site (DB30/31), one AluI site in accessible DNA away from the nucleosomes (DB49/515), one AluI site in the nucleosome binding site within the 5S rDNA (HJ075/76), one PstI-HF site in the linker between the 5S rDNA nucleosomes and the dinucleosome over the GAL4 binding sites and E4 promoter (HJ077/78), one AluI site in the dinucleosome just described (HJ079/80) and one AluI site in the linker at the other end of the dinucleosome (HJ081/82). Nucleosome positions within the 5S rDNA are based on DNaseI digests which identified that nucleosomes protect DNA at positions 20-165 (which in pG5E4 is 22-167 as there is an extra GG at the start of each 5S rDNA repeat) [213]. Due to a HpaII recognition site in the region encompassed by the primer pair HJ075/76, the AluI site in the 5S rDNA could not be assessed with HpaII fragmentation, thus a separate experiment with BamHI fragmentation was performed (figure 4.11b).

In general, high DNA concentrations measured by ddPCR for digests of M.EcoGII treated DNA with methylated adenine sensitive enzymes (PstI-HF or AluI) show the site to be methylated, while low concentrations show the site to be unmethylated. In this experiment HpaII was used for fragmentation of the DNA. The single PstI-HF site present in pG5E4 was hypothesised to be accessible linker DNA between the 5S rDNA and GAL4 binding sites. Figure 4.11a shows this site to be mostly accessible to methylation, shown by the height of the pink bar being about two thirds the height of the red positive control bar in the M.EcoGII and PstI-HF condition. The mismatch in the heights of these two bars may

be due to ddPCR variation or dynamic nucleosome positions. The primer pair HJ081/82 (light green bars) amplified a region containing an AluI recognition site at the other side of the GAL4/E4 dinucleosome. They showed similar findings, with a clear increase in the M.EcoGII treated AluI digested sample indicating reasonable accessibility of this site to M.EcoGII methylation. The methylation observed at these two sites indicates that the dinucleosome (around 300 bases) is likely to be located in the 378 bases between these sites. The AluI site within the HJ079/80 amplified region, (dark green bars in figure 4.11a) was within this potential dinucleosome region and showed lower methylation compared to HJ081/82, indicating greater nucleosome protection. However, compared to the unmethylated matched control for HJ079/80 there was an increase in methylation, indicating that either nucleosomes do not completely protect DNA from methylation or there could be variation in the nucleosome position. This site may be close to the boundary of the nucleosome, therefore less tightly bound and more likely to be methylated. Nucleosomal DNA is wrapped around the histone proteins, therefore some of it may be accessible to DNA modification depending on the particular bases exposed and the structural requirement of the modifying enzyme. Another nucleosome bound AluI site was in the 5S rDNA, however a neighbouring HpaII site prevented analysis in this experiment as, due to HpaII fragmentation the DNA would be digested at this site regardless of the methylation status of the AluI site. The site in the region amplified by DB49/515 should have been accessible to DNA methylation as it was 220 bases away from the nearest proposed nucleosome. However, in the M.EcoGII and AluI digested condition, this region (blue bar) is about half the height of the positive control (red bar) suggesting only partial methylation. This indicates that these nucleosomes may be reasonably dynamic, or there may be wider sequence contexts or secondary structures that affect either the ability of AluI to digest DNA, or M.EcoGII to methylate the DNA.

To test the methylation status of the site within the 5S rDNA, BamHI was used to fragment the DNA. For this experiment, the HJ077/78 region became the negative control as it

contained a BamHI site. The experimental procedure was the same as for figure 4.11a except for the addition of two samples where the initial methylation step was replaced with a water treatment, and an extra methylation step with M.EcoGII was performed after DNA extraction, when all sites should be accessible. The results in figure 4.11b are difficult to interpret as the 5S rDNA site (purple bars) often produced much higher concentrations than the positive control (red bars). This may be due to the nine repeats of the 5S rDNA, which introduces multiple primer binding sites. Therefore, many of the primers may not produce a product of less than 200 bases, as is required for ddPCR. The variation in product size may have led to an inaccurate DNA concentration as a smaller number of longer DNA molecules may be detected as similar to a larger number of shorter DNA molecules. A filtering step for the correct size products was not possible as this would have required screening of the individual droplets. Despite this variation, the water and AluI condition (second condition in figure 4.11b) produce a lower concentration (purple bar) than the positive control (red bar), indicating that the DNA is digested. For the M.EcoGII and AluI condition the purple bar is much closer to the positive control (red bar) than the negative control (black bar), indicating that some methylation has occurred, therefore the DNA was at least partially accessible. This is unexpected as the AluI site within the HJ075/76 region should be nucleosome-bound and thus protected from methylation. When the DNA is methylated with M.EcoGII after DNA extraction (two samples to the right of figure 4.11b), when all sites would be equally accessible, the AluI site (purple bar) in the AluI digest is much higher than the positive control (red bar), indicating that the M.EcoGII methylation of the chromatinised DNA (the M.EcoGII condition) was a relatively low level of methylation compared to what can be achieved on unchromatinised DNA. Therefore, the nucleosome seems to partially protect the 5S rDNA, however the presence of the repeats add a level of complexity to the analysis and interpretation. In summary, the recombinant nucleosomes have shown that it may be possible to detect whether a site is nucleosome-bound or not, but the answer is unlikely to be clear

cut, either because nucleosome binding is a dynamic process that can involve some mobility or because nucleosomes do not completely protect DNA from M.EcoGII methylation.

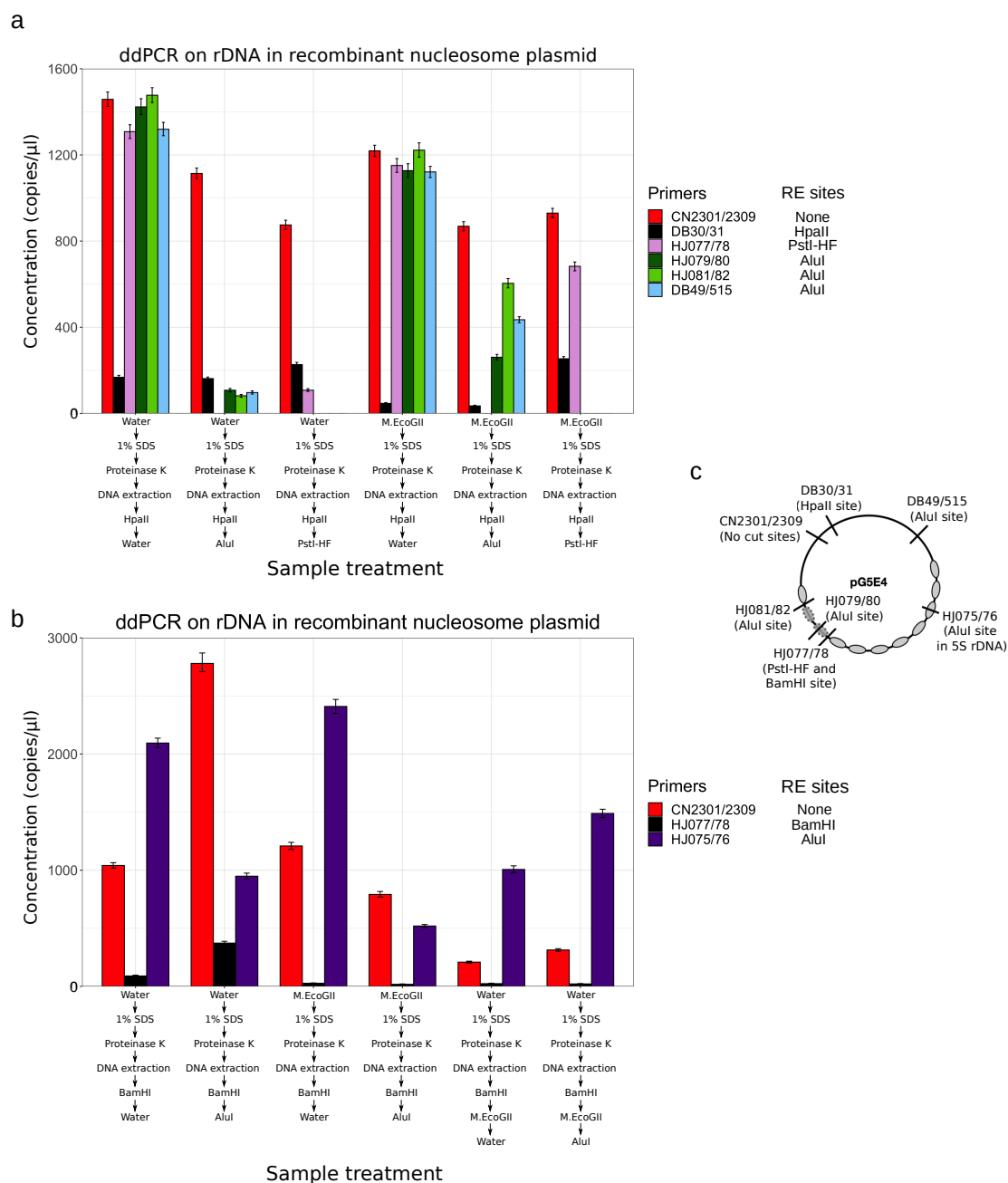


Fig. 4.11 ddPCR quantification of methylation on recombinant nucleosome plasmid

a) pG5E4 recombinant nucleosome plasmid from Active Motif was treated with water or M.EcoGII, nucleosomes were denatured with 1% SDS and degraded with proteinase K prior to DNA extraction and treatment with water, PstI-HF or AluI. All samples were digested with HpaII to fragment the plasmid. HJ077/78 encompasses a PstI-HF site not an AluI site, therefore these primers were tested in the PstI-HF conditions and not the AluI treated ones. Conversely, primers encompassing AluI sites (HJ079/80, HJ081/82, DB49/515) were not tested in the PstI-HF treated conditions. **b)** Same treatment as in **a)** but BamHI used to fragment the plasmid and final digestion to test methylation sensitivity was only performed with water or AluI. The two samples furthest to the right were incubated with M.EcoGII after DNA extraction and BamHI fragmentation, when the DNA should no longer have any bound proteins, prior to AluI or water treatment. **c)** Schematic of the pG5E4 plasmid with DNA in black and the recombinant nucleosomes shown in grey. The dotted line represents the position of the proposed dinucleosome. Location of the digest sites of interest are labelled with the primer pair used to amplify that region and the cut site within the region. Orientation is the same as in figure 4.9.

4.3.4 Quantitative ddPCR on chromatin

Based on the potential detection of nucleosome binding sites in the recombinant nucleosome plasmid (figure 4.11), and the clear evidence in figure 4.8 that genomic DNA can be methylated with M.EcoGII and show high concentrations in subsequent ddPCR analysis, ddPCR was tested on chromatin. Three approaches to chromatin extraction were investigated, two of which were based on spheroplasting the *S. cerevisiae* cells with enzymatic breakdown of the cell wall while the other involved formaldehyde crosslinking and mechanical breakdown of the cell wall. Despite their differences they all ended in the following steps: SDS to denature proteins, RNaseA to remove RNA, proteinase K to degrade proteins and phenol-chloroform DNA extraction.

Firstly, wild-type *S. cerevisiae* cells were grown in YPAD, killed with sodium azide and PMSF, pelleted and washed, transformed into spheroplasts in spheroplasting buffer, pelleted, treated with Triton X-100 on ice, added to a sucrose cushion, pelleted, methylated with M.EcoGII in EB buffer and the protocol completed in the standard way (SDS, RNaseA, proteinase K, phenol-chloroform DNA extraction) (see section 2.7.6). Two volumes of M.EcoGII were used (1 and 4 µl) and the genomic sites are the same as in figure 4.8. Figure 4.12 indicates that the extracted DNA in the ScaI-HF region (DB9/10) was not as completely digested as in previous experiments (shown by each black bar being half the height of the corresponding red bar). Additionally, all regions in the PstI-HF digest of unmethylated chromatin condition were incompletely digested suggesting that something is preventing digestion of these samples. This is unlikely to be methylation as ScaI-HF has previously shown high levels of digestion (low DNA concentrations) of M.EcoGII treated DNA at these sites (figure 4.8). Both volumes of M.EcoGII produced similar results to this control. Therefore, methylation does not seem to have occurred, or if it has, it has been masked by the incomplete digestion in the control.

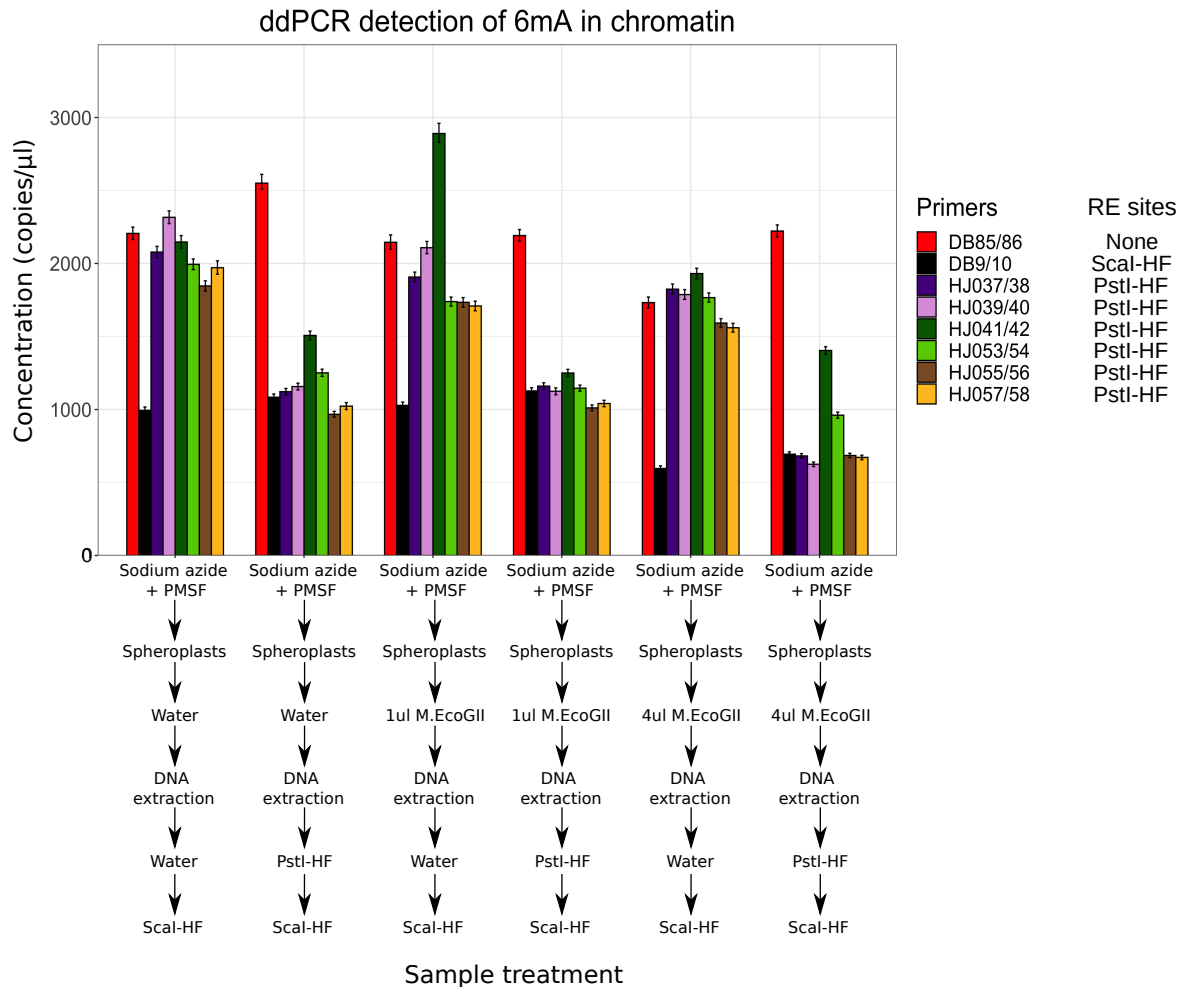


Fig. 4.12 ddPCR quantification of methylation on *S. cerevisiae* chromatin

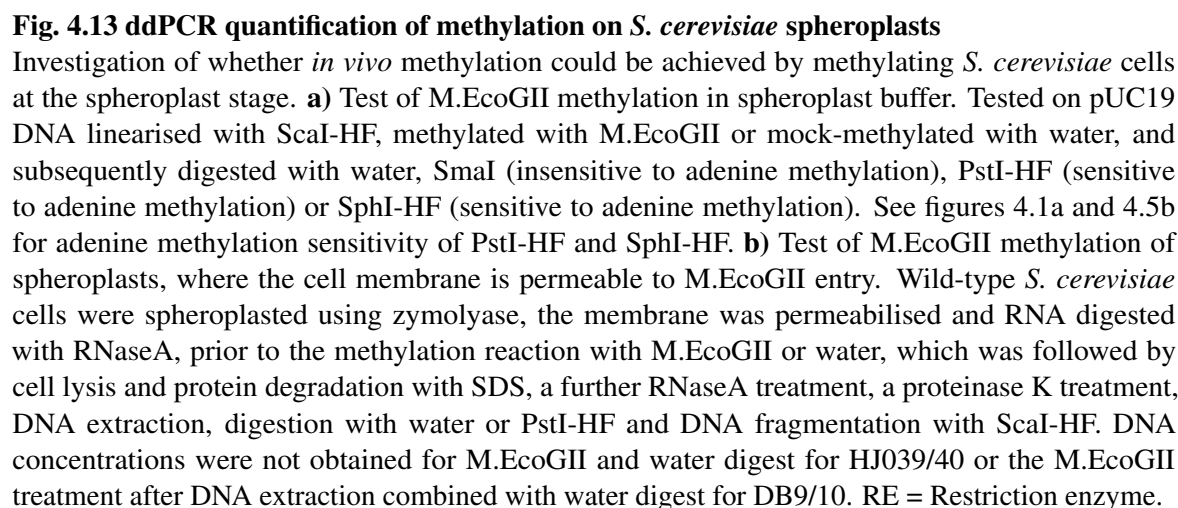
S. cerevisiae cells were grown in YPAD, killed with sodium azide and PMSF, pelleted and washed. Spheroplasts were produced, pelleted, treated with Triton X-100 on ice, added to a sucrose cushion, pelleted, methylated with M.EcoGII in EB buffer, treated with 1% SDS overnight, and treated with RNaseA then proteinase K. DNA was extracted using phenol-chloroform and ethanol precipitation then treated with water or PstI-HF, fragmented with ScaI-HF prior to ddPCR analysis. See section 2.7.6 for more details. RE = Restriction enzyme.

For a variation on this experiment, wild-type *S. cerevisiae* cells were grown in YPAD to mid-log phase, then spheroplasts were generated in spheroplasting buffer, washed in sorbitol, the membrane was permeabilised and RNaseA was added to remove RNA prior to methylation though salt was omitted as this would inhibit M.EcoGII (section 2.7.6). After methylation and the standard end steps, DNA was digested with water or PstI-HF

and fragmented with ScaI-HF followed by ddPCR analysis. M.EcoGII methylation was performed in a buffer containing IGEPAL, DTT, spermidine and sorbitol in addition to the usual Cutsmart buffer and SAM. Therefore, the ability of M.EcoGII to methylate DNA in this new buffer was investigated using ScaI-HF linearised pUC19 DNA. The linearised DNA was incubated with water or M.EcoGII, then treated with water, SmaI (insensitive to adenine methylation), PstI-HF (sensitive to adenine methylation) or SphI-HF (sensitive to adenine methylation). Figure 4.13a shows that M.EcoGII works as effectively in the modified buffer as in previous experiments with NEB Cutsmart buffer, demonstrated by the high purple bars in the M.EcoGII treated PstI-HF or SphI-HF digested conditions compared to these digests in the water treated condition.

For analysing methylation of the spheroplasts, once DNA had been extracted a sample was treated with M.EcoGII to simulate a fully methylated sample. This and the extracted DNA were digested with PstI-HF or water. Figure 4.13b shows that the M.EcoGII treatment of the chromatin did not produce any noticeable methylation, as shown by the low concentration of all sites in the M.EcoGII and PstI-HF condition. The DNA was able to be methylated once extracted, shown by the higher concentrations in the post-methylation and PstI-HF condition. Despite this, the overall finding is that methylating the *S. cerevisiae* cells in the spheroplast stage with this method has not produced any methylated chromatin that could be used for analysing the chromatin structure. The negative control (DB9/10) region with a ScaI-HF site was not consistently low as ScaI is sensitive to adenine methylation [150] therefore would not be able to sufficiently fragment the DNA. However, the differences between the M.EcoGII treatment at spheroplast stage and post-DNA extraction suggest that the methylation did not work for other reasons. For example, M.EcoGII may have been unable to enter the cells due to an inefficient zymolyase digest, or there may have been too many proteins bound to the DNA. The six sites of interest were designed to cover a range of genomic contexts, with three sites (HJ037/38, HJ039/40, HJ041/42) covering well-positioned nucleosomes,

therefore would be expected to be low in methylation, while the other three sites (HJ053/54, HJ055/56, HJ057/58) were designed to be in nucleosome-free regions therefore should have higher levels of methylation. As only a small number of sites were tested the three accessible sites may have been bound by non-nucleosome proteins, therefore protected from methylation. Ideally a greater number of genomic sites would be tested but ddPCR is best suited to studying individual loci and does not scale up well.



An alternative method of methylating chromatin involved formaldehyde crosslinking the proteins to the DNA which also kills the cells. Then the formaldehyde was quenched with glycine and cells were pelleted, washed in ice-cold TBS and resuspended in lysis buffer including EDTA and Triton-X100, with addition of PMSF and an anti-proteolytic cocktail (see section 2.7.6). Cells were subsequently mechanically broken by beating and incubated with water or M.EcoGII for one hour prior to the standard end steps. A pUC19 spike-in was included immediately prior to the methylation step with M.EcoGII at a 1:100 concentration, based on DNA quantification after a previous chromatin extraction. This would identify whether or not M.EcoGII was able to methylate in these conditions and whether sufficient M.EcoGII was added. If pUC19 was successfully methylated while genomic DNA was not, the reason for lack of methylation would be specific to the genomic DNA, for example too many proteins bound.

Firstly the concentration of pUC19 DNA from a chromatin sample was optimised for ddPCR. Compared to previous analysis on isolated DNA (figure 4.3) this time the pUC19 DNA was combined with genomic DNA. It was identified in figure 4.14a that a higher concentration, roughly ten times more (2×10^{-4} ng/ μ l rather than 2×10^{-5} ng/ μ l), of pUC19 DNA was required when starting with a mixed DNA sample with pUC19 DNA to genomic DNA at a 1:100 ratio. Figure 4.14b shows that at both levels of M.EcoGII (1 and 4 μ l) the relatively high concentrations measured in the PstI-HF treated conditions (dark red bars) indicate adenine methylation of pUC19 DNA, though this is lower with 1 μ l M.EcoGII. Therefore, M.EcoGII is able to methylate pUC19 DNA to some extent with either amount of M.EcoGII. Comparing this to the genomic DNA, which had much lower measurements overall, none of the PstI-HF sites showed protection from PstI-HF. This suggests that the M.EcoGII amount was sufficient but it either couldn't access the DNA, which should not have been a problem as the cells had been lysed, or that there were too many proteins bound to DNA, which should not have been a problem as the HJ053/54 and HJ055/56 sites were

designed to be in nucleosome-free regions. However, the sample size is very low therefore this experiment does not rule out either explanation.

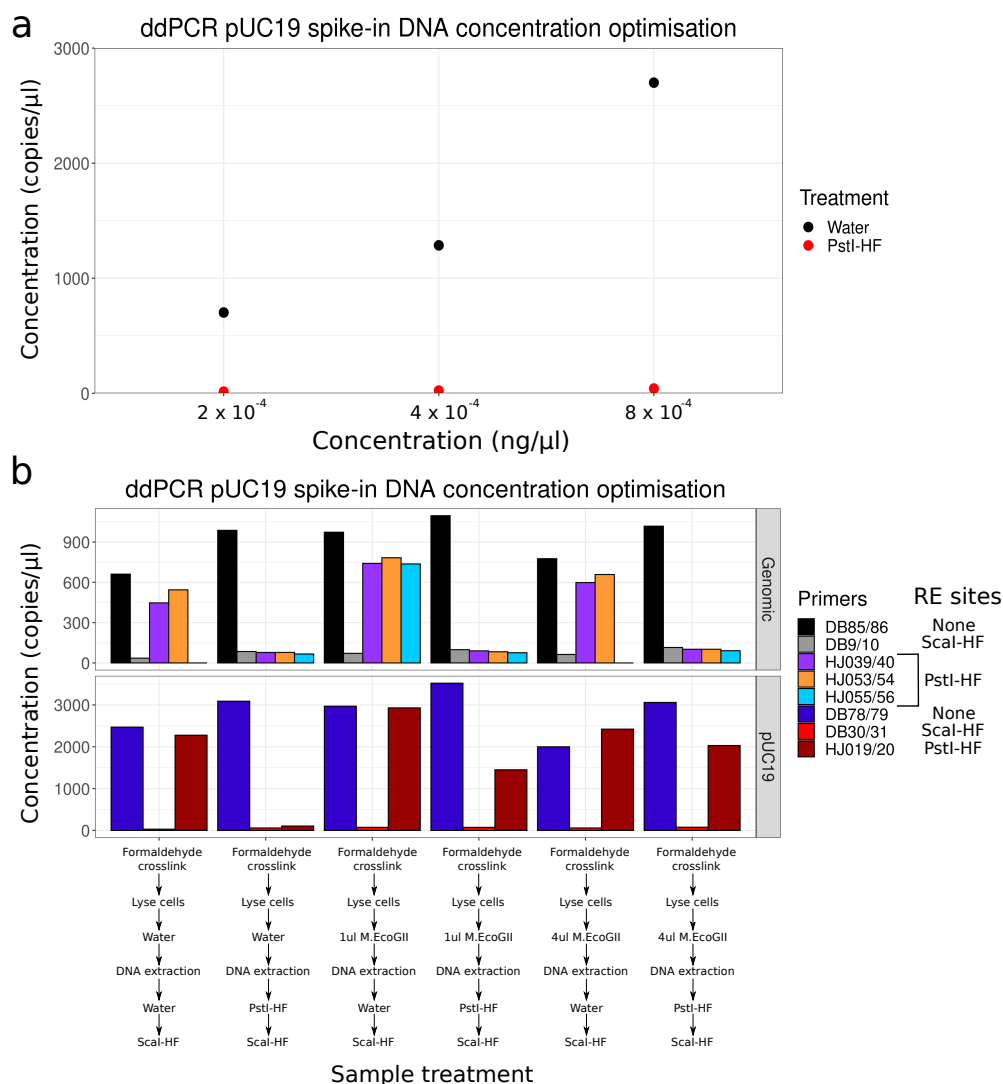


Fig. 4.14 ddPCR quantification of methylation on *S. cerevisiae* chromatin with pUC19 spike-in *S. cerevisiae* wild-type (T7107) cells were crosslinked with formaldehyde and lysed by mechanical means prior to methylation followed by DNA extraction and restriction enzyme digests. A 1:100 pUC19 spike-in was used immediately before the methylation step. **a)** The dilution of pUC19 DNA used as input for ddPCR was optimised as the pUC19 DNA was mixed with genomic DNA compared to in figure 4.3 which was on isolated pUC19 DNA. **b)** ddPCR was performed using primers for the genomic DNA and pUC19 DNA. Both sets of primers included two controls, one with no restriction enzyme (RE) sites, one with a Scal-HF site. All measurements were made in the same ddPCR run. Error with the measurement for water and 4 μ l with water instead of PstI-HF digests for HJ055/56 genomic DNA primers, therefore light blue bar is missing from these conditions.

4.4 *In vitro* assessment of *S. cerevisiae* chromatin structure

Different approaches to investigating chromatin structure include modifying chromatin *in vivo* and extracting the DNA, or extracting the chromatin and modifying it at this stage. Extraction of *S. cerevisiae* chromatin is technically challenging therefore it was important to assess the quality of the chromatin to ensure a chromatin accessibility assay would provide biologically relevant information. Two validation approaches were investigated, fragmenting the DNA with MNase and assessing the DNA fragment sizes, and probing the chromatin and non-chromatin samples with antibodies for nuclear and cytosolic proteins.

4.4.1 Validation of *S. cerevisiae* chromatin preps by MNase treatment

Chromatin predominantly consists of DNA wrapped around histone octamers, forming nucleosomes. Although there are some large nucleosome-free regions, around 80% of the genome is nucleosome-bound [7]. Nucleosomes are a standard size and have around 150 bp of DNA wrapped around them. Therefore, if chromatin is digested with MNase, which cuts in the linker regions between nucleosomes, then a 150 bp DNA fragment would be produced for a single nucleosome. Incomplete MNase digestion, where some linkers are undigested, would allow longer DNA fragments of 300 or 450 bp, corresponding to two or three adjacent nucleosomes respectively, to persist. Therefore, a good chromatin sample would produce a ‘nucleosomal ladder’, with DNA molecules observed at 150 bp intervals.

A wild-type (T7017) *S. cerevisiae* culture was grown in YPAD, crosslinked with 1% formaldehyde to maintain the chromatin structure and subsequently spheroplasted with zymolyase. Samples were then treated with an MNase dilution series after which DNA was extracted. In figure 4.15 there is no band observed in the sample treated with undiluted MNase (1) which may be due to the band having run off the end of the gel. However, a

nucleosomal ladder is observed in the 1:10 dilution where there are bands roughly every 150 bp. For more dilute MNase the nucleosomal fragments are much larger, shown by the more intense bands towards the top of the gel. The two supernatant samples which are discarded in the protocol contain a large amount of protein, most of which is likely to be from the cytosol.

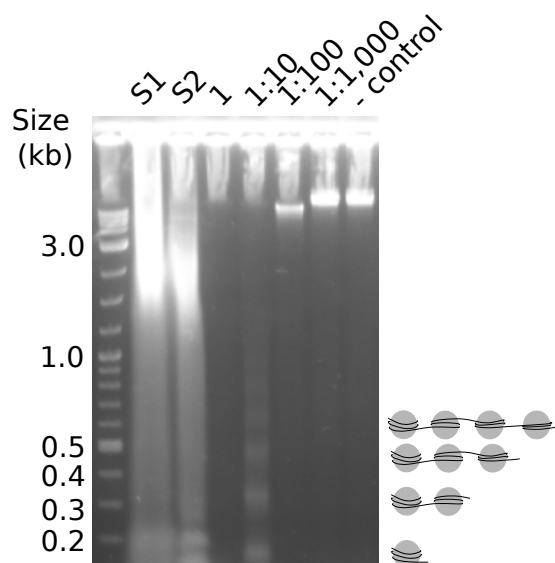


Fig. 4.15 Detection of nucleosome ladder by MNase digest of DNA extracted from *S. cerevisiae* chromatin

Wild-type *S. cerevisiae* (T7107) was crosslinked with 1% formaldehyde, and treated with zymolyase to create spheroplasts which were subsequently treated with MNase to digest accessible DNA. S1 is the supernatant produced after pelleting the spheroplasts, immediately after the zymolyase reaction. S2 is the supernatant collected after the first incubation in MNase buffer with IGEPAL. The 1, 1:10, 1:100 and 1:1000 samples have been treated with the respective dilutions of MNase which digests accessible DNA but cannot digest nucleosome-bound DNA. Dilutions were made with 50% glycerol. The negative (-) control was treated with 50% glycerol instead of MNase. Ladder = Quickload 2-log DNA ladder (NEB). Schematic shows nucleosomes in grey and DNA in black.

4.4.2 Validation of *S. cerevisiae* chromatin preps by protein detection

For chromatin extraction (see section 2.7.6) wild-type *S. cerevisiae* cells were grown in YPAD, killed with sodium azide and PMSF. Subsequent zymolyase treatment generated spheroplasts and the membrane was permeabilised with Triton-X100, after which the whole cell extract sample was collected. Remaining spheroplasts were incubated on a sucrose

cushion and treated with benzonase to digest the DNA and RNA, leaving only proteins that were part of the chromatin. This method of chromatin extraction, though without the benzonase treatment, was also used in the ddPCR analysis (section 4.3.4). For each collected sample, they were mixed with LDS loading buffer and boiled. The ‘total’ protein sample was collected by a different protocol for protein extraction (section 2.7.14), which did not aim to preserve chromatin.

For validation by western blotting, two proteins were probed for: histone H3, a nuclear protein involved in the formation of nucleosomes [214], and tubulin, a cytosolic protein involved in microtubule formation [215]. Figure 4.16, shows that both proteins were present in the total and whole cell extract samples while the chromatin sample contained histone H3 while the soluble did not, and the soluble sample contained tubulin while the chromatin did not. Therefore, the chromatin extraction did separate nuclear and cytosolic proteins and western blotting with this combination of antibodies can be used to assess the quality of the chromatin, using histone H3 as a proxy for chromatin proteins. However, this method is very protein-specific, therefore does not provide a complete view of the chromatin quality.

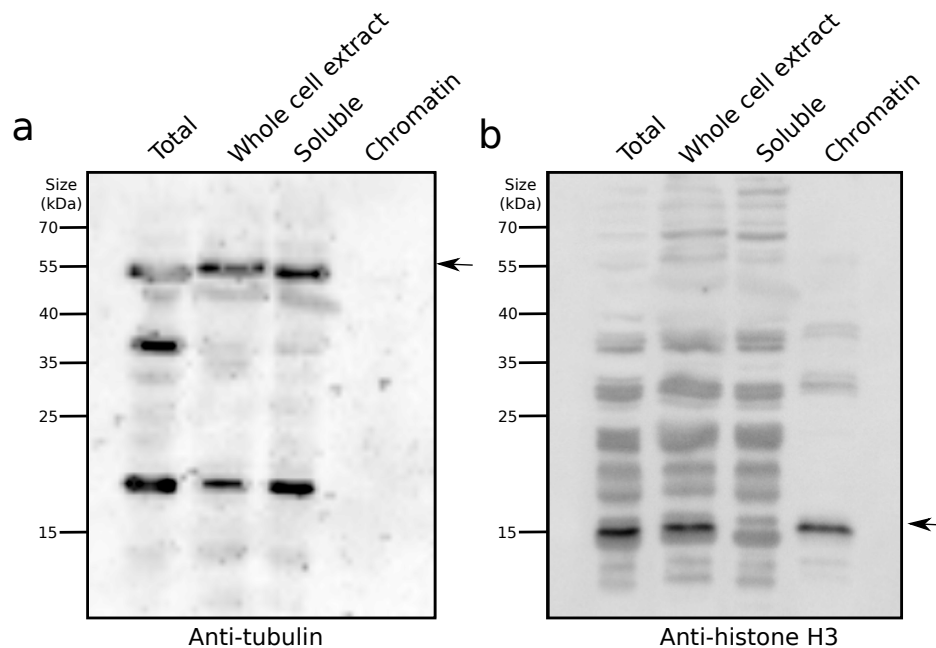


Fig. 4.16 Western blot shows correct separation of proteins in chromatin extraction from *S. cerevisiae*

Chromatin was extracted from wild-type *S. cerevisiae* (T7107), with the whole cell extract, soluble and chromatin fractions collected during this protocol (section 2.7.6). The total sample was from the same *S. cerevisiae* strain (T7107) but was collected using an alternative protocol which extracts all proteins from the cell rather than separating out chromatin (section 2.7.14). For the western blot, a 12% PAGE gel was loaded with 5 μ l PAGE ruler, 5 μ l 'total' sample, and 10 μ l of each of the other samples. The samples were separated using 12% PAGE and transferred to a membrane which was subsequently probed with the indicated primary antibodies using the conditions in section 2.7.15. The membrane from (a) was stored in PBST at 4 °C, then later washed, re-blocked and incubated with anti-histone H3 antibodies as shown in (b). Arrow heads show expected band size for each protein (53 kDa for tubulin and 15 kDa for histone H3).

4.4.3 Mass spectrometry shows very low cytosine methylation of spheroplasts

For *in vitro* methylation of chromatin, cells were treated with a methyltransferase when spheroplasts have been formed to allow the methyltransferase to access the DNA. Wild-type *S. cerevisiae* (T7107) was crosslinked with formaldehyde, spheroplasted with zymolyase and split into three tubes. One was digested with a dilution series of MNase to validate the chromatin (not shown), one was treated with M.CviPI to methylate cytosines in a GC

sequence context and the other was treated with the same buffers and incubation steps as the M.CviPI condition but was treated with 50% glycerol thus forms the negative control. Reactions were stopped either with a Stop buffer for the MNase digest, or heat-inactivation for the other two samples. All samples underwent RNaseA and proteinase K treatments followed by phenol-chloroform DNA extraction and isopropanol DNA precipitation. For a positive control, wild-type (T7107) *S. cerevisiae* genomic DNA was extracted then treated with an excess of M.CviPI (8 units for 3 µg DNA) to maximise DNA methylation. In comparison, the chromatin sample was treated with 40 units of M.CviPI, which NEB recommends is sufficient to methylate 40 µg DNA.

The mass spectrometry experiment involved breaking the DNA down to individual nucleosides and then identifying the concentration of each modified and canonical base present. This method of mass spectrometry is used throughout this chapter. As M.CviPI can only methylate cytosines in a GC sequence context, in the *S. cerevisiae* genome only 452,567 of the 2,313,713 cytosines (approximately 19.5% of all cytosines) are potential targets for this enzyme. Therefore, the positive control would not be able to achieve higher than 19.5%, however only 7.5% was detected (figure 4.17). All sites were accessible in this sample, therefore this suggests that M.CviPI was not as efficient at methylating GC sites as expected. Both the negative and chromatin (spheroplasts treated with M.CviPI) samples had very low levels of cytosine methylation, with 0.015% and 0.070% of total cytosines methylated respectively. A lower level than the positive control was expected due to the presence of proteins bound to DNA and thus providing protection from methylation. However, the M.CviPI treated sample only showed 0.94% of the positive control methylation, when taking the positive control as 100% methylated. Therefore, even if it is assumed that 80% of the genome is bound by nucleosomes [7] and some DNA is occupied by other proteins, for example transcription factors or replication proteins, there may be more accessible DNA than this experiment implies. Conversely, when optimising the level of a methyltransferase,

there is also the concern of over-methylation, where protein-bound sites may be methylated if the methyltransferase is in high excess. Therefore, the optimal level of methylation would need to be investigated by other means, potentially by downstream sequencing of a range of conditions to verify the location in addition to the quantity of methylation. This current experiment was not further analysed due to very low levels of methylation detected (figure 4.17). Future experiments could involve testing a range of higher concentrations of M.CviPI.

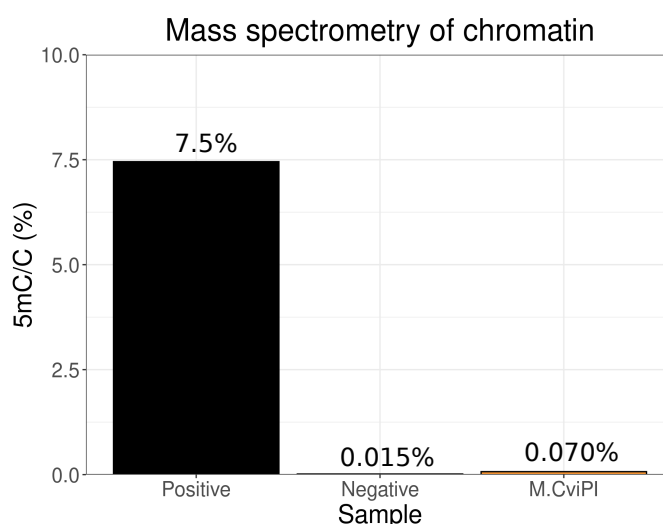


Fig. 4.17 Mass spectrometry of chromatin treated with a GC methyltransferase (M.CviPI)

Chromatin was extracted from wild-type *S. cerevisiae* (T7107) and methylated with M.CviPI, a GC methyltransferase, at the spheroplast stage. Methylation is shown as a percentage of all cytosines, however M.CviPI only targets cytosines in a GC sequence context which makes up approximately 19.5% of all cytosines in the *S. cerevisiae* genome. Mass spectrometry analysis was performed by Paolo Spingardi. Positive control was T7107 genomic DNA methylated *in vitro* with M.CviPI; Negative control was chromatin extracted from T7107 DNA and mock-methylated with water.

4.4.4 Variant of SMAC-seq approach to methylate chromatin *in vitro* does not completely replicate SMAC-seq results

In 2019, a technique called Single Molecule Accessible Chromatin sequencing (SMAC-seq) was published [116]. This involved looking at *in vitro* methylation of *S. cerevisiae* chromatin, therefore an attempt was made to replicate this method. SMAC-seq used a combination of methyltransferases: M.EcoGII, M.CviPI and M.SssI. These methylated adenines, cytosines

in a GC context and cytosines in a CG context respectively. To reproduce this technique, a simplified version with one methyltransferase per condition was performed (section 2.7.9). In this variant, methylation was performed when the cells had been spheroplasted with zymolyase and the membranes permeabilised with IGEPAL, making it possible for the methyltransferase to enter the cells while in SMAC-seq the nuclei were isolated. An RNaseA step was added prior to methylation to reduce off-target effects of the methyltransferases, in addition to just before proteinase K treatment and DNA extraction.

The final readout for SMAC-seq was nanopore sequencing, for which methylation detection was performed with Tombo. Using the scripts Georgi Marinov from the Greenleaf lab (who published the SMAC-seq technique) kindly lent me I was able to produce a figure of the same region with data that had been analysed identically to SMAC-seq. The analysis involved extracting per read data from Tombo, Bayesian aggregation, combining the reads into a single .bgz file which is tabix indexed for fast random access and generating Bismark-like methylation files which are visualised with reads grouped by hierarchical global clustering. Figure 4.18 shows that the data is in a similar format however the area of dense black dots seen in the published SMAC-seq paper [116] in figure 4.18a is not reproduced as clearly in the SMAC-sequencing variation in figure 4.18b. However, in the 6mA condition there is an increase in clustering around the origin, with one cluster slightly downstream (to the right) of the origin in the upper half of reads and another cluster closer to the origin location in the lower half of reads. These may represent different chromatin structures at these sites or forward and reverse strand reads. The differences may be due to small differences in the spheroplasting protocols used, potentially affecting the amount of methyltransferase that was able to reach the chromatin. However, it is more likely that they are due to the difference in methyltransferases used, with 30 units used in this version compared to 200 units in SMAC-seq [116]. In addition, the differences observed may be due to the presence of methylation by one enzyme rather than three, meaning that the overall level of methylation

will be less. As Tombo is used to detect both adenine and cytosine methylation, it is possible that the methylation patterns are only evident when these are combined.

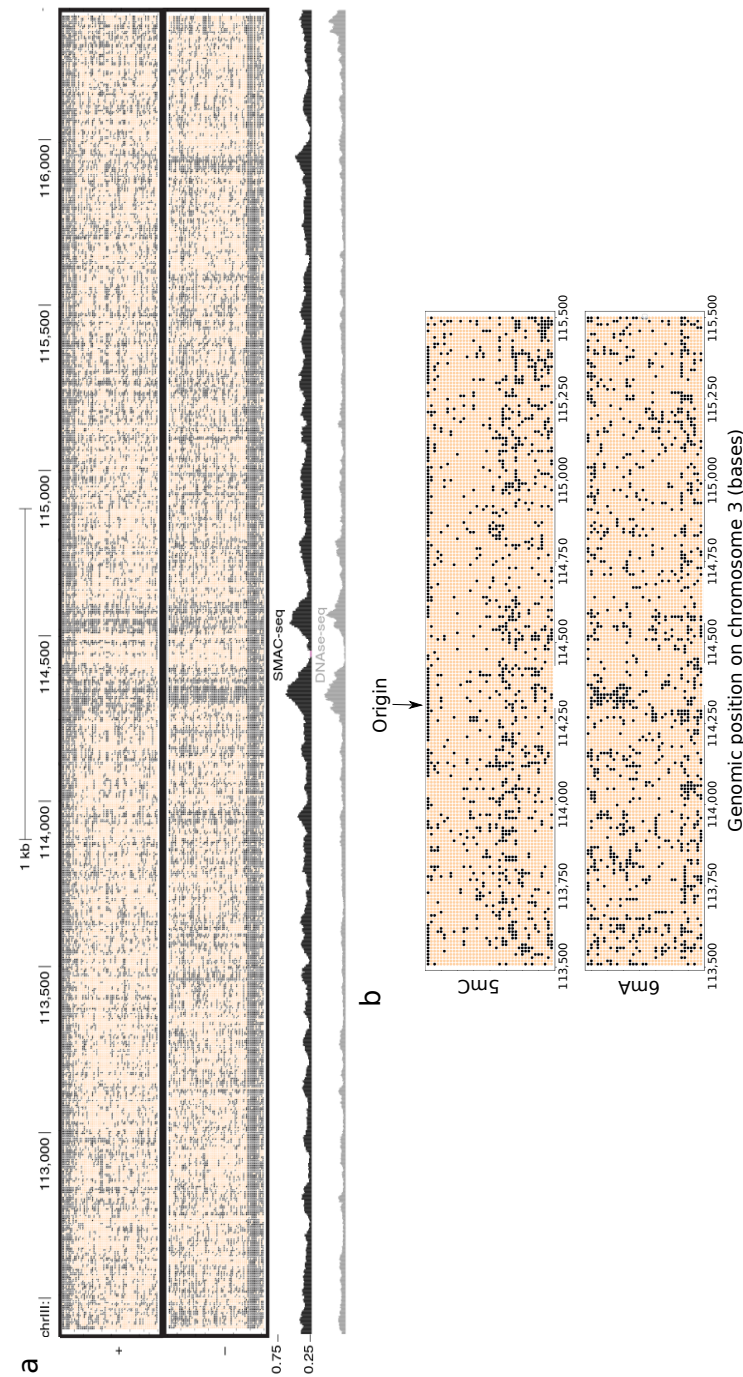


Fig. 4.18 Nanopore sequencing of SMAC-seq variant experiment

Asynchronous wild-type *S. cerevisiae* (T7107) cells were treated with methyltransferases at the spheroplast stage. Nanopore sequencing data was processed with Tombo and SMAC-seq, with the same scripts used for both (a) and (b). Single molecules (individual lines) are shown with data grouped into windows of 10 bases and each site classified as methylated or not based on a 0.5 threshold. DNA replication origin location of *ARS308* is marked. Data is aligned between the (a) and (b) to show the same genomic region. a) Figure 2a from Shipony *et al* [116] where DNA was methylated with M.EcoGII, M.CviPI and M.SssI. An ensemble of single molecules is represented in black below the individual reads to compare to population-based DNase-seq. Reprinted by permission from Springer Nature: Nature Methods. Long-range single-molecule mapping of chromatin accessibility in eukaryotes, Shipony *et al* COPYRIGHT (2019). Figure was cropped to remove the Cse4-ChIP-seq data, H4S47C chemical nucleosome mapping data and genes. A copy of the licence can be found in appendix 2. b) Variant of SMAC-seq using M.EcoGII for 6mA methylation or M.CviPI for 5mC methylation.

4.5 Verification of inducible Dam methyltransferase *S. cerevisiae* strain and assessment of motif-specific adenine methylation detection by nanopore sequencing

Existing strains with an inducible Dam methyltransferase made by Dr Carolin Müller were used to investigate how well *S. cerevisiae* tolerated the presence of adenine methylation, and to investigate options for detecting adenine methylation with nanopore sequencing. The Dam methyltransferase was under the control of a *GALS* promoter, therefore should be repressed in glucose, non-induced in raffinose and activated by the addition of galactose.

4.5.1 Inducible Dam methyltransferase identified by western blot

Firstly, the control of Dam methyltransferase expression by growth in raffinose followed by galactose induction was assessed. Figure 4.19 shows that Dam methyltransferase expression is not detectable in 2% raffinose (0 minutes) with expression slowly starting from 15 minutes post-galactose addition. The overnight sample was incorrectly loaded, hence there are no observable proteins. This figure demonstrates that changing the sugar source can control the expression of Dam methyltransferase in this *S. cerevisiae* strain and a reasonably high level of expression is achieved, shown by the clear bands towards the end of the timecourse.

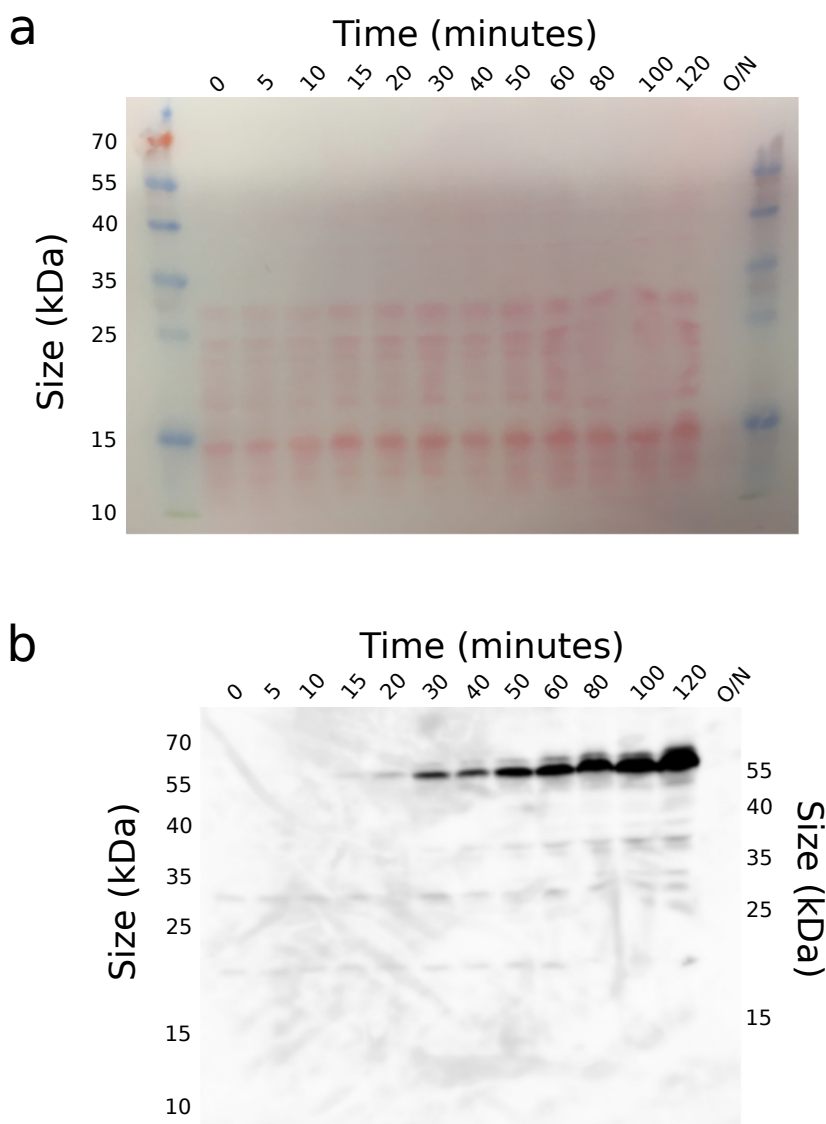


Fig. 4.19 Western blotting detects inducible Dam methyltransferase in timecourse

S. cerevisiae with an inducible Dam methyltransferase under a *GALS* promoter was grown in YPAD, transferred to 2% raffinose and grown for a couple of doubling times. The timecourse then began in raffinose (0 minutes) immediately followed by induction with galactose to a final concentration of 2%. Samples were taken at 5, 10, 15, 20, 30, 40, 50, 60, 80, 100 and 120 minutes, plus an overnight sample. The overnight sample was incorrectly loaded therefore no proteins are seen for this sample. The Dam-AID construct is 57.6kDa. **a)** Ponceau staining of the membrane to identify total protein loaded. **b)** ECL detection of the anti-mini-AID antibody.

4.5.2 Effect of inducible Dam methyltransferase on *S. cerevisiae* growth

As *S. cerevisiae* does not endogenously express any methyltransferases [173, 174] the effect of Dam methylation on growth was investigated. The inducible Dam methyltransferase has a C-terminal mini-AID tag which allows for degradation with auxin (Indole-3-acetic acid; IAA). This provides a mechanism for switching off the expression of Dam methyltransferase which could be used to restrict methylation to a shorter time window. IAA is potentially toxic to cells [216] therefore the effect of IAA was tested using a growth curve with cells grown in 2% raffinose overnight, then split into six cultures. One was transferred into YPAD, one was induced with 2% galactose, and the other four were grown in 2% raffinose. IAA was added to three of the raffinose conditions, one with a single addition at the start, one with additions every hour and one with additions every two hours. Samples were given 100% ethanol at the same time points as IAA in other samples, as a control. Figure 4.20 shows that the cells grow fastest in YPAD, as glucose is the preferred sugar source. Growth is similar between the 2% raffinose culture, the 2% raffinose culture with the initial IAA treatment and the 2% raffinose culture with 2-hourly IAA. Growth slows down slightly with 1-hourly additions of IAA, suggesting that this level of IAA is partially toxic to the cells. The Dam methyltransferase should only have been minimally expressed in 2% raffinose, therefore the 1-hourly IAA was potentially more toxic due to insufficient Dam methyltransferase for it to act on and may be more suitable for use in 2% galactose cultures. The galactose condition grew slightly faster than the 2% raffinose, as it is a preferred sugar. The expression of the Dam methyltransferase does not appear to substantially affect the growth rate, though it was slightly slower than the control strain in all conditions. Therefore further repeats are required to validate any effect of Dam methylation on the growth of *S. cerevisiae*.

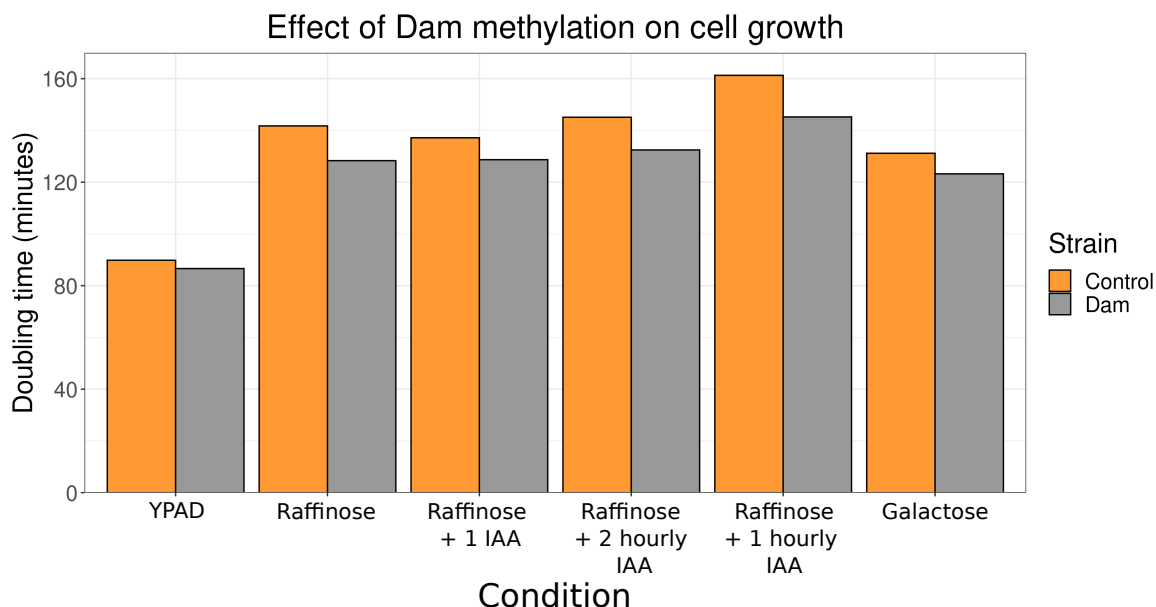


Fig. 4.20 Growth analysis of inducible Dam methyltransferase *S. cerevisiae* strain

S. cerevisiae with an inducible Dam methyltransferase under a *GALS* promoter (strain CAY434) was grown in YPAD, 2% raffinose and 2% galactose to observe differences in growth. There were four raffinose cultures, three of which were treated with IAA to degrade the Dam methyltransferase. One had a single treatment of IAA at the start of the timecourse. One had an addition of IAA every hour, and one every two hours. 100% ethanol was added to all cultures not receiving IAA treatment, at the same times as IAA treatments as a control as IAA was dissolved in 100% ethanol. Control strain was CAY336, a wild-type *S. cerevisiae* strain used to create the inducible Dam methyltransferase strain.

4.5.3 DNA methylation by inducible Dam methyltransferase identified in *S. cerevisiae* by enzyme digests and nanopore sequencing

The next step was to assess the functionality of the Dam methyltransferase. *S. cerevisiae* cells expressing Dam methyltransferase under a *GALS* promoter were grown in YPAD, the repressive condition, transferred to 2% raffinose, the non-inducing condition, for a couple of doubling times after which expression of the Dam methyltransferase was induced with 2% galactose. Two types of analysis were performed on a single timecourse of samples taken post-induction, with a control in 2% raffinose immediately prior to the addition of 2% galactose. The DNA analysis methods were digestion by adenine sensitive restriction enzymes which have a GATC recognition site and nanopore sequencing.

Genomic DNA methylation assessed by restriction enzyme digests

To quickly assess genomic DNA methylation, the DNA was digested with a selection of enzymes which have the same recognition site of GATC, which is shared with the methylation target site of Dam methyltransferase, but different sensitivities to adenine methylation. These were DpnI which cuts at GATC sites containing a methylated adenine, MboI which cuts unmethylated DNA and Sau3AI which cuts regardless of the methylation status. As a mock-digest control, water was added to a set of samples which were exposed to the same buffer and incubation steps as the real samples. Figure 4.21 shows minimal change in any of the conditions in the first 20 minutes, apart from the loss of the slightly more distinct bands just above 2 kb and 5 kb which were observed in the 0 minute timepoint. These bands may be rDNA or another repetitive DNA region. The water condition showed more change than expected, with the DNA fragment length increasing throughout the timecourse. For DpnI, which cuts methylated DNA, the overall fragment length decreases throughout the timecourse and the intensity of short DNA fragments increases. MboI replicates the water pattern for the high molecular weight DNA fragments, and the distinct bands <1.5 kb become fainter and eventually disappear as the timecourse progresses, suggesting a reduction in unmethylated DNA for MboI to digest. The Sau3AI control digests the DNA similarly throughout the timecourse with the exception of the two distinct bands that are present in the raffinose sample (0 minutes). Therefore, DNA methylation occurs throughout the timecourse, predominantly after the first 20 minutes. Nanopore sequencing is a more accurate method therefore will be able to characterise the methylation levels across the timecourse in more detail.

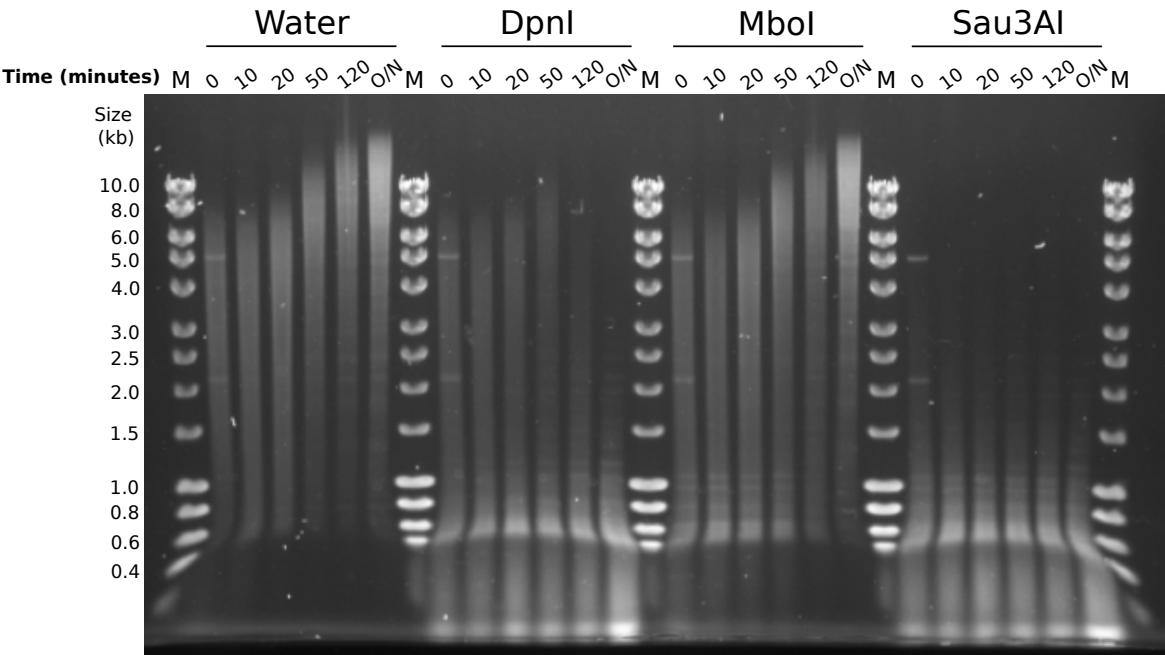


Fig. 4.21 Methylation investigated by restriction enzyme digests of inducible Dam methyltransferase throughout timecourse

S. cerevisiae with an inducible Dam methyltransferase under a *GALS* promoter was grown in YPAD, transferred to 2% raffinose and grown for a couple of doubling times. The timecourse then began in raffinose (0 minutes) immediately followed by induction with galactose to a final concentration of 2%. Samples were taken at 10, 20, 50 and 120 minutes post-galactose induction in addition to overnight (O/N). 0.5% TBE-agarose gel run at 20V for 5 minutes followed by 60V for 2.5 hours and stained with ethidium bromide. M = Bioline Hyperladder I.

Genomic DNA methylation assessed by nanopore sequencing

Considering only the nanopore sequencing reads that mapped to the sacCer3 reference genome, the level of methylation in each sample was calculated. The Dam methyltransferase timecourse samples were basecalled with the first version (v3.3.0) of Guppy that was methylation-aware, with a model that is able to detect adenine methylation in a GATC sequence context. This produces a new set of fast5 files which contain information about each read, including a table of scores of each position in each read being either A, 6mA, C, 5mC, G or T. Each score is a value between 0 and 255. The score of A plus 6mA adds up to 255, likewise for C and 5mC. G and T do not have a modified option therefore always have a value of 255. If an adenine is methylated it will have a higher 6mA score and a lower A score.

Most methylation scores are spread across the possible range of 0 to 255, representing the confidence with which the basecall for each position in each read was made. The probabilities were subsequently normalised to 0-100. Figure 4.22 shows that very low levels of GATC methylation were detected in the first 50 minutes of Dam methyltransferase induction, with an increase in methylation observed at 120 minutes. The methylation level then decreases for the overnight sample, indicating that the majority of GATC sites are protected from the Dam methyltransferase despite the long exposure and that methylation may be removed during longer exposures. The pUC19 DNA shows particularly low methylation considering that all GATC sites were expected to be methylated.

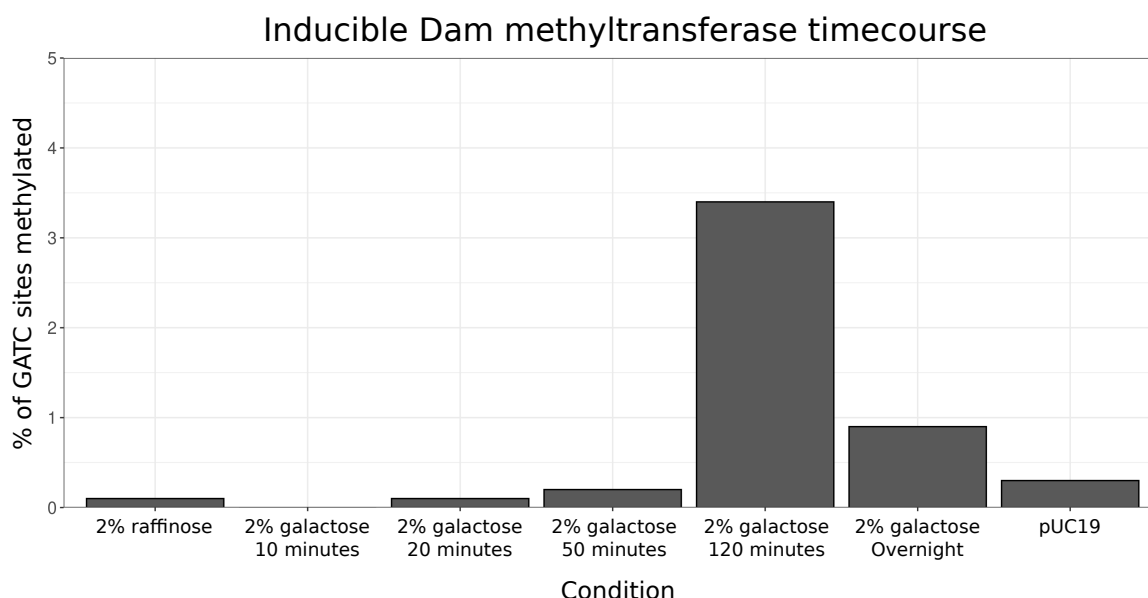


Fig. 4.22 Nanopore sequencing detection of methylation in Dam methyltransferase timecourse Nanopore sequencing data was basecalled and mapped to the sacCer3 or pUC19 reference sequence. The 6mA probability at 1,000 GATC sites was classified as methylated or not based on a 50% threshold. The overall percentage of GATC sites methylated for each sample is plotted.

To verify the Dam methyltransferase sequence specificity, the adenine methylation levels were analysed for the adenine in the second position of GATC and a selection of 4-mers that were one base change away from GATC. These were selected based on reports of Dam promiscuity, particularly at GACC sites [161] and 1,000 sites were analysed for each

4-mer. Basecalling and methylation detection for all sites was performed with Guppy v3.3.0 with a model designed for detecting GATC adenine methylation. Figure 4.23a shows that methylated adenine is detected at GATC sites after 120 minutes induction with 2% galactose but not at 0 minutes (2% raffinose), indicating that the Dam methyltransferase expression is controlled by the absence and presence of 2% galactose. Adenine methylation is not detected at any of the other 4-mers, indicating that either Dam methyltransferase is highly specific, or the model is only capable of detecting adenine methylation in a GATC sequence context, or a combination of both. The model is documented as suitable for Dam methylation detection therefore it is most likely trained on Dam methylated DNA and might exclude other sequence contexts. To test this, the same analysis was performed on SMAC-sequencing published data [116] which was treated with M.EcoGII, therefore the expectation would be methylated adenine in a variety of sequence contexts. Figure 4.23b shows that in SMAC-sequencing data adenine methylation is observed at GATC, GATG and AATC sites, although the highest methylation is detected at GATC sites and no methylation is detected for GACC, GATT, CATC or TATC. Therefore the model can detect methylation in some other sequence contexts, though there is a clear bias towards detection in at GATC sites which means the model is not suitable for detecting M.EcoGII adenine methylation in all sequence contexts.

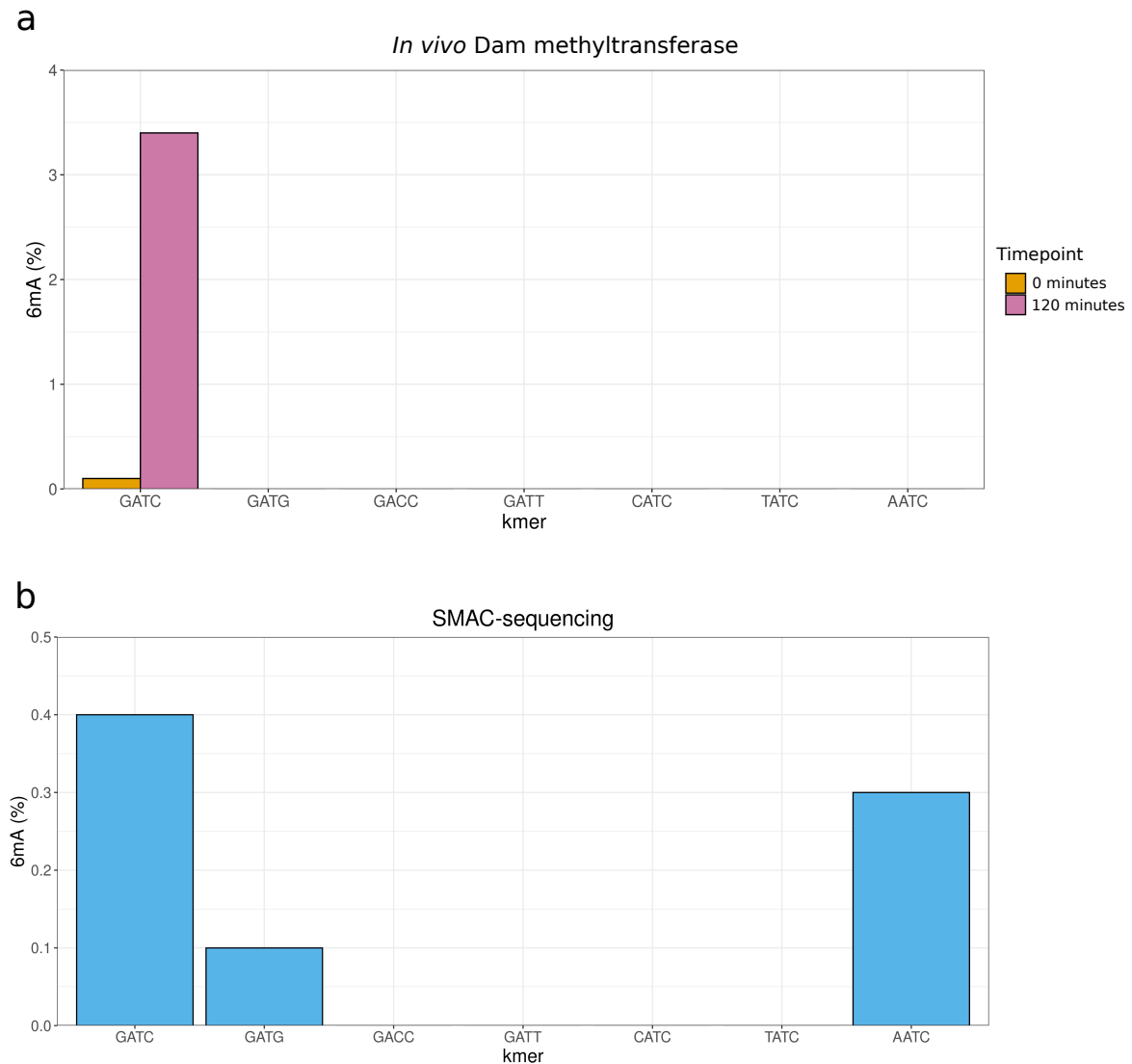


Fig. 4.23 Nanopore sequencing sequence context analysis of Dam methyltransferase methylation
 Nanopore sequencing basecalling and methylation detection performed with Guppy version 3.3.0 and the dna_r9.4.1_450bps_modbases_dam-dcm-cpg_hac.cfg configuration file. Data was mapped to the sacCer3 reference sequence, excluding the mitochondrial genome. The methylation level across a set of 4-mers (1,000 sites each) were analysed. **a)** A timecourse of Dam methyltransferase induction with 2% galactose. 0 minutes = 2% raffinose, sample taken immediately prior to induction of Dam methyltransferase with 2% galactose. **b)** Analysis of SMAC-sequencing data from Shipony *et al* (2019) [116] includes adenine methylation by M.EcoGII (a non-specific adenine methyltransferase). The raw data was analysed in the same way as panel a. For both datasets, the 6mA probability at the second base in a selection of 4-mers, all with one base change from GATC, was investigated. Each instance of each k-mer was classified as methylated or not based on a greater than or equal to 50% threshold. Only the adenine in the second position was analysed for the AATC sequence context.

A later version of Guppy, v3.6.0, was compatible with a Rerio research basecalling model reported to be capable of detecting adenine and cytosine methylation in all contexts. The format of the resulting data was identical which allowed the sequence specificity of this model to be assessed by reproducing the analysis in figure 4.23 with Guppy v3.6.0 and the new model. Figure 4.24 shows that this model produces quite different results. In the Dam methyltransferase data (figure 4.24a) the false positive rate in the 0 minute sample is higher than in figure 4.23a for all analysed 4-mers. At 120 minutes some methylation is detected at all analysed 4-mers, at slightly higher levels for most 4-mers than in the 0 minutes. This is consistent with some Dam promiscuity [161], although the nanopore sequencing data does not show GACC as preferentially modified above the other non-GATC 4-mers. However, there is a clear bias towards detection of methylation at GATC sites and a higher level of methylation is observed at the GATC site than with the previous model. For the SMAC-sequencing data (figure 4.24b), methylated adenine was more evenly detected across the 4-mers though some variation still exists. Therefore, this model can be used to detect M.EcoGII non-specific adenine methylation.

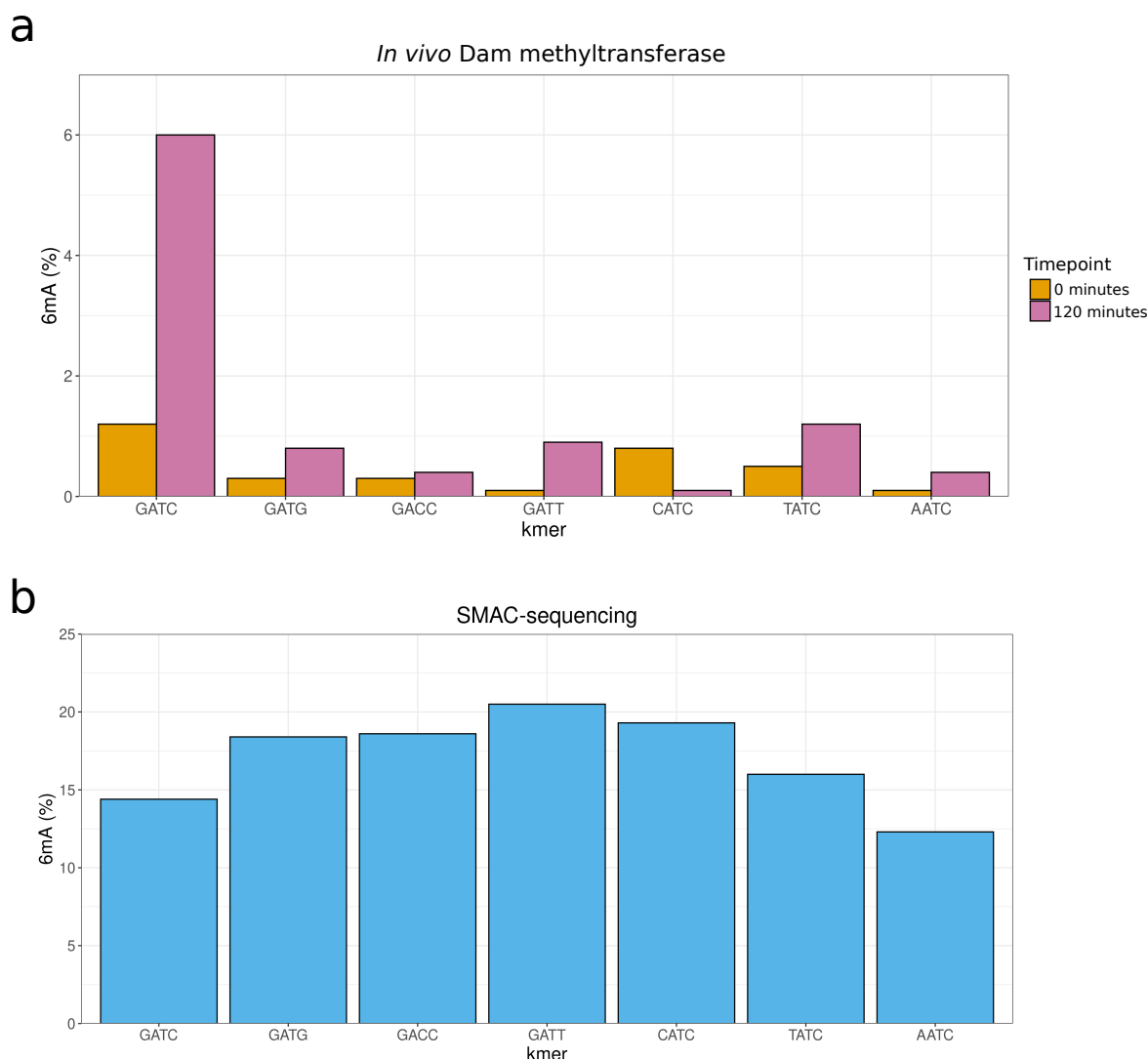


Fig. 4.24 Nanopore sequencing model specificity for adenine methylation

Nanopore sequencing basecalling and methylation detection performed with Guppy version 3.6.0 and a configuration file from Rerio (res_dna_r941_min_modbases-all-context_v001). Data was mapped to the sacCer3 reference sequence, excluding the mitochondrial genome. The methylation level across a set of 4-mers (1,000 sites each) were analysed. **a**) A timecourse of Dam methyltransferase induction with 2% galactose. 0 minutes = 2% raffinose, sample taken immediately prior to induction of Dam methyltransferase with 2% galactose. **b**) Analysis of SMAC-sequencing data from Shipony *et al* (2019) [116] includes adenine methylation by M.EcoGII (a non-specific adenine methyltransferase). The raw data was analysed in the same way as panel a. For both datasets, the 6mA probability at the second base in a selection of 4-mers, all with one base change from GATC, was investigated. Each instance of each k-mer was classified as methylated or not based on a greater than or equal to 50% threshold. Only the adenine in the second position was analysed for the AATC sequence context.

Nanopore sequencing was used to investigate chromatin structure. As the density of GATC sites in the *S. cerevisiae* genome is relatively low a meta-analysis was performed on primary alignments mapped to the forward strand of the sacCer3 reference genome, excluding the mitochondrial genome. Using consensus nucleosome dyad positions [89], GATC sites were mapped to their nearest nucleosome dyad. Figure 4.25 shows that in 2% raffinose, 10 and 20 minutes of 2% galactose induction the methylation level is relatively constant and therefore independent of the distance of the site from the nucleosome dyad. Methylation in these conditions may be false positives as figure 4.19 showed that there is only minimal Dam methyltransferase expression by 20 minutes after 2% galactose induction therefore low DNA methylation would be expected. In the 50 minutes, 2 hours and overnight induction conditions the methylation level increases from the nucleosome dyad towards the edge of the nucleosome (around 73 bp away from the nucleosome dyad). This suggests that the nucleosome dyad is partially protecting the DNA from methylation, while the region of DNA towards the edge and just outside the nucleosome is more accessible to methylation, suggesting that DNA at the edge of the nucleosome is more loosely wrapped. The spikes in the data may be due to an increase in the number of adenines at these positions or increased exposure of these sites to Dam methyltransferase. The spike around 25 bases in figure 4.25c does not correlate with a spike in figure 4.25a or b therefore both factors may play a role for different sites. The same methylation patterns are observed based on mean methylation at each distance (figure 4.25a) and percentage of adenines above a 50% methylation threshold (figure 4.25b), with a very minimal reduction in methylation in the latter. The number of adenines analysed for each condition (figure 4.25c) is relatively flat for each sample throughout the nucleosome-bound region of DNA shown by the grey box. Therefore, the increase in methylation is not an artefact of the number of adenines. The overnight sample had a low number of reads however the methylation level is lower than the two hour sample, suggesting that either the DNA methylation is removed, or the Dam methyltransferase expression is suppressed over

longer time periods. This meta-analysis has shown that Dam methylation can be affected by chromatin structure.

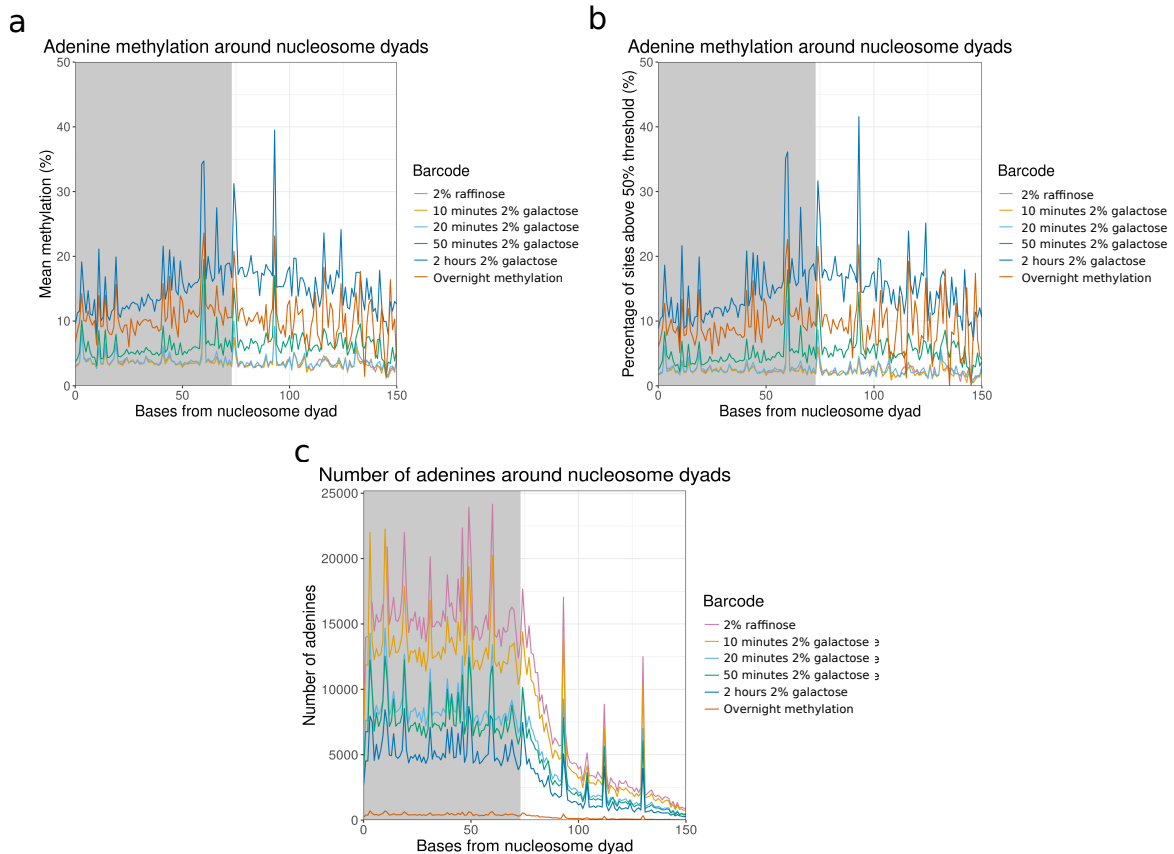


Fig. 4.25 Nanopore sequencing detection of 6mA in GATC sites around nucleosomes

S. cerevisiae with an inducible Dam methyltransferase was grown in YPAD, transferred to 2% raffinose for a couple of doubling times then expression of the Dam methyltransferase was induced with 2% galactose. The 0 minute sample was taken in 2% raffinose, just prior to the addition of galactose, while the other samples are at various points after the start of the 2% galactose induction. Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence, excluding the mitochondrial genome. Each adenine in a GATC sequence context was mapped to the nearest nucleosome dyad (central position of the nucleosome) based on consensus nucleosome positions from Jiang *et al* (2009) [89]. **a)** The mean methylation of all adenines at each distance from the nucleosome dyad is plotted. **b)** The methylation status of each adenine on each read was classified using a 50% threshold. The percentage of methylated sites at each distance from the nucleosome dyad is plotted. **c)** The number of adenines analysed for each distance from the nucleosome dyad for each condition. For all panels, the grey box indicates nucleosome-bound DNA. Distances are relative to the adenine in the GATC site.

Single-molecule analysis was performed to investigate whether chromatin structure could be visualised at specific genomic loci, principally around DNA replication origins. As an

example, single molecules around *ARS1513* were visualised to investigate the application of this system for studying chromatin accessibility during origin licensing. GATC sites within forward primary alignments mapping to a 2 kb region centered on *ARS1513* were classified as methylated or not based on a 50% threshold. The first two hours of galactose induction were analysed in figure 4.26a-e where each black line represents a read, each dot represents an adenine in a GATC site (red if methylated, blue if unmethylated) and nucleosome-binding sites are shown in grey. Throughout the induction period the methylation increases, with almost no sites methylated early on and a large number of adenines at the central GATC site methylated in figure 4.26e. This site is on the edge of nucleosome-bound DNA and a nucleosome-free region encompassing the ACS therefore the methylation at this site and low methylation observed at sites in more densely packed nucleosomes provides support for a methylation-based single-molecule chromatin accessibility assay. However, within a 2 kb region around this ACS there are only seven GATC sites. This does not provide sufficient information to study the binding sites of DNA replication proteins. Therefore, a similar inducible system in *S. cerevisiae* will be developed with a non-specific adenine methyltransferase: M.EcoGII.

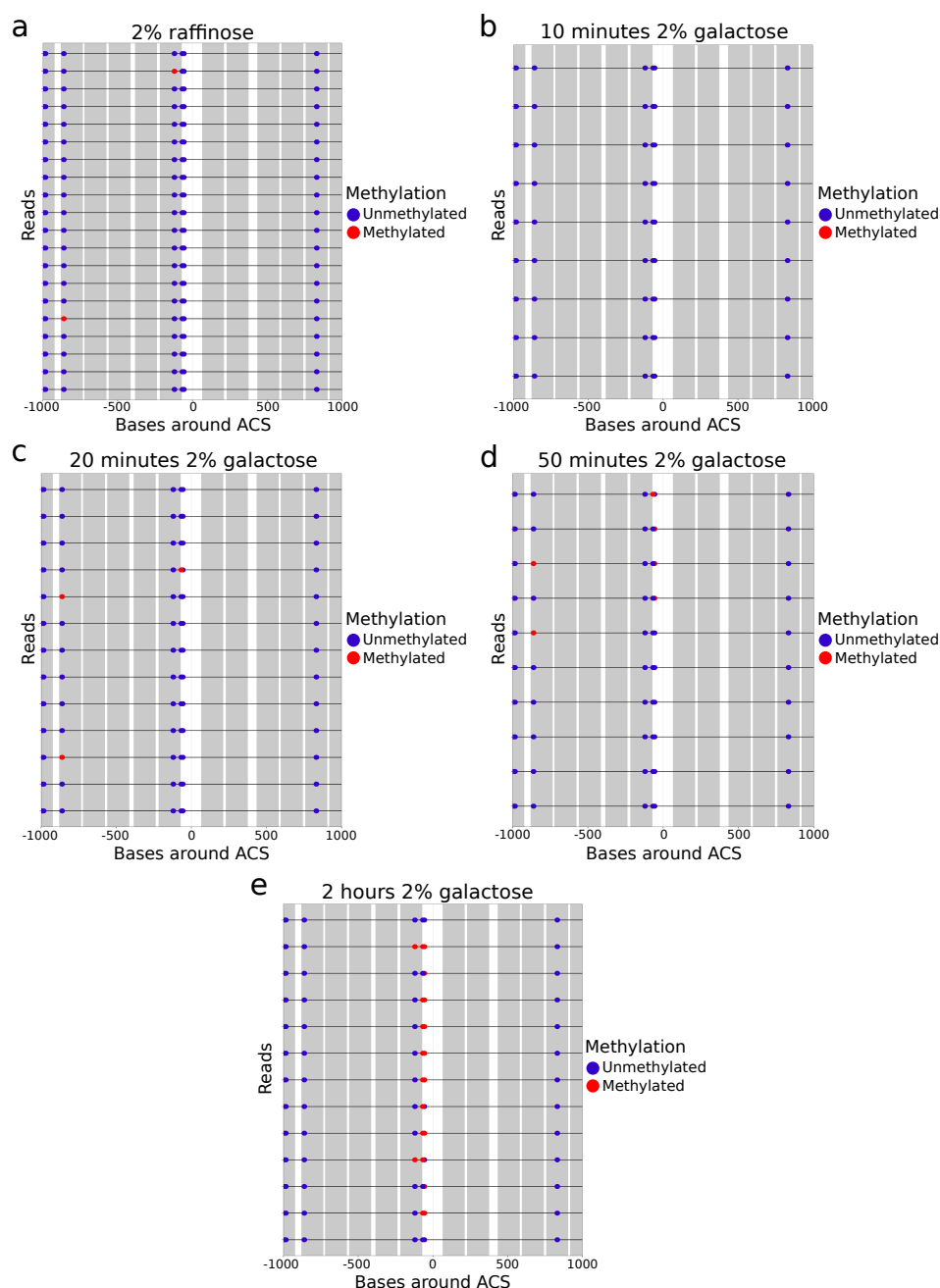


Fig. 4.26 Nanopore sequencing detection of 6mA in GATC sites around ARS1513

S. cerevisiae with an inducible Dam methyltransferase was grown in YPAD, transferred to 2% raffinose for a couple of doubling times then expression of the Dam methyltransferase was induced with 2% galactose. Samples were a) 2% raffinose, and then b) 10 minutes, c) 20 minutes, d) 50 minutes and e) 2 hours after induction of Dam methyltransferase induction with 2% galactose. Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence, excluding the mitochondrial genome. Adenines in GATC sequence contexts in the genomic region of +/- 1 kb from ARS1513 were plotted for each forward primary alignment spanning the entire 2 kb region. Each dot represents an adenine in a GATC site, which is methylated (red) if above a 50% threshold or unmethylated (blue) if below this threshold. For all panels, the grey boxes indicate nucleosome-bound DNA based on consensus nucleosome positions from Jiang *et al* (2009) [89]. The distances around the ACS are relative to the reference sequence.

4.6 *In vitro* M.EcoGII treatment of *S. cerevisiae* genomic DNA

Restriction enzyme protection assays and ddPCR have validated the ability of M.EcoGII to methylate DNA at specific loci. However, they cannot be used to assess genome-wide methylation levels. Therefore, mass spectrometry and nanopore sequencing were used. Both methods provide information on the total level of methylation, but only nanopore sequencing provides positional information. The suitability of Guppy v3.6.0 with the Rerio model for M.EcoGII methylation detection was further tested. The experimental setup involved a 2-fold dilution series of M.EcoGII (1/64x to 2x), with 1x being the NEB recommended amount of M.EcoGII for the amount of DNA used, plus a no M.EcoGII control that was treated with water. This allowed the capture of a range of methylation levels that might be encountered in a biological system.

4.6.1 Mass spectrometry of *in vitro* M.EcoGII methylated genomic DNA

Mass spectrometry involving digestion to individual nucleosides was performed by Paolo Spingardi. As the concentration of M.EcoGII increased, so did the level of adenine methylation (figure 4.27). However, despite all regions of the DNA being accessible, 1x M.EcoGII had only 16% of total adenines methylated. This suggests that the NEB recommended concentration does not lead to complete methylation. Approximately 80% of the genome is covered by nucleosomes [7] and the non-nucleosome bound regions of the genome may be bound by other proteins (e.g. transcription factors, replication proteins, DNA repair proteins) for at least part of the cell cycle. Therefore, the levels of methylation detected in this *in*

vitro experiment cover the methylation range that could be expected from future *in vivo* experiments.

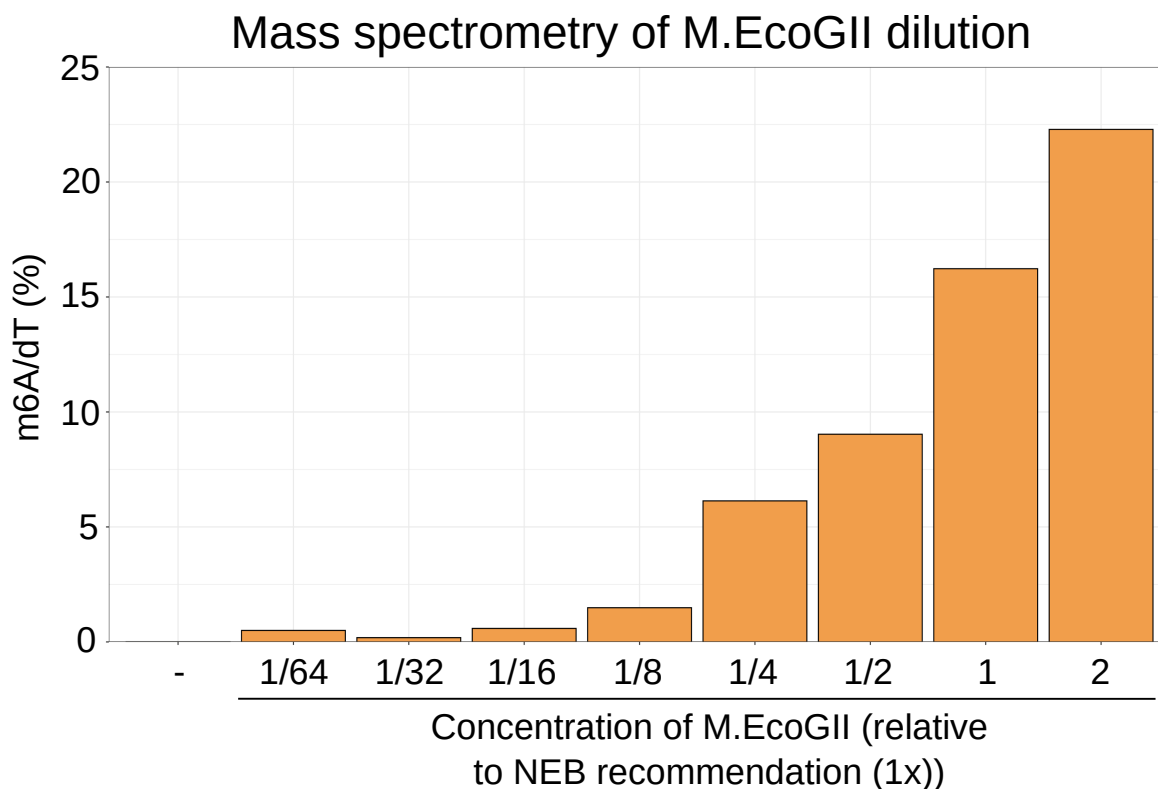


Fig. 4.27 Mass spectrometry of *in vitro* genomic DNA treated with M.EcoGII dilution series
S. cerevisiae wild-type (T7107) genomic DNA was treated with a dilution series of M.EcoGII (1x = NEB recommended concentration for the amount of DNA treated) or water (negative control) and incubated at 37 °C for one hour followed by heat-inactivation of M.EcoGII and DNA clean-up. HPLC mass spectrometry was performed by Paolo Spingardi.

4.6.2 Nanopore sequencing of *in vitro* M.EcoGII methylated genomic DNA

The aims of this experiment were to validate the ability of nanopore sequencing to detect adenine methylation in a variety of contexts and to identify how well this would perform at different levels of adenine methylation.

The nanopore sequencing data was basecalled with Guppy v3.6.0 (as previously used in figure 4.24). The barcoded nanopore data generated from this *in vitro* M.EcoGII 2-fold dilution experiment was used to investigate a variety of thresholds for classifying an adenine as methylated or not. Coverage of more than 60x for each sample (table 4.4) provides sufficient data for determining a threshold.

M.EcoGII dilution	Mean read length (bases)	Coverage*
No M.EcoGII	3,401	63
1/64x	3,509	80
1/32x	3,412	65
1/16x	3,461	76
1/8x	3,439	66
1/4x	3,418	77
1/2x	3,409	72
1x	3,444	85
2x	3,370	82

Table 4.4 M.EcoGII *in vitro* nanopore sequencing overview

Nanopore sequencing run from 13th July 2020. Reads were mapped to the sacCer3 reference genome (excluding mitochondrial genome) with minimap2, filtered for primary alignments and mapping quality greater than zero. Coverage calculated by primary alignments with mapping quality greater than 0 multiplied by mean read length, divided by genome size of 11,209,039 bases excluding the mitochondrial DNA. Mean read length and coverage are both rounded to the nearest integer. * Coverage is genomic coverage, therefore strand-specific coverage would be this coverage divided by 2.

The mass spectrometry provided a reliable quantification of the level of 6mA to which the nanopore data could be compared. Figure 4.28a analysis of 100,000 adenines in any sequence context shows that all thresholds display a similar trend of increasing methylation throughout the dilution series. However, the threshold used for nanopore sequencing 6mA classification greatly affects the level of methylation detected. At low M.EcoGII concentrations many of the

thresholds predict a higher level of methylation than is detected by mass spectrometry, which may lead to a high false positive rate. Conversely, at higher concentrations of M.EcoGII most of the thresholds (30% or 40% and higher) underestimate the total level of methylation, leading to a lower true positive rate. To balance minimising false positives and maximising true positives 70% was selected for future analysis.

The Rerio model used for detecting M.EcoGII methylation is reported to detect both adenine and cytosine methylation in all sequence contexts. Therefore, for this dataset the level of cytosine methylation was also investigated. Figure 4.28b shows that at low levels of adenine methylation there are relatively high levels of methylated cytosine detected despite no endogenous or exogenous cytosine methylation. However, as the level of true adenine methylation increases, with increased M.EcoGII concentrations, the level of methylated cytosine decreases. This suggests that the model expects to find some methylation, and in the absence of any it preferentially detects cytosine methylation.

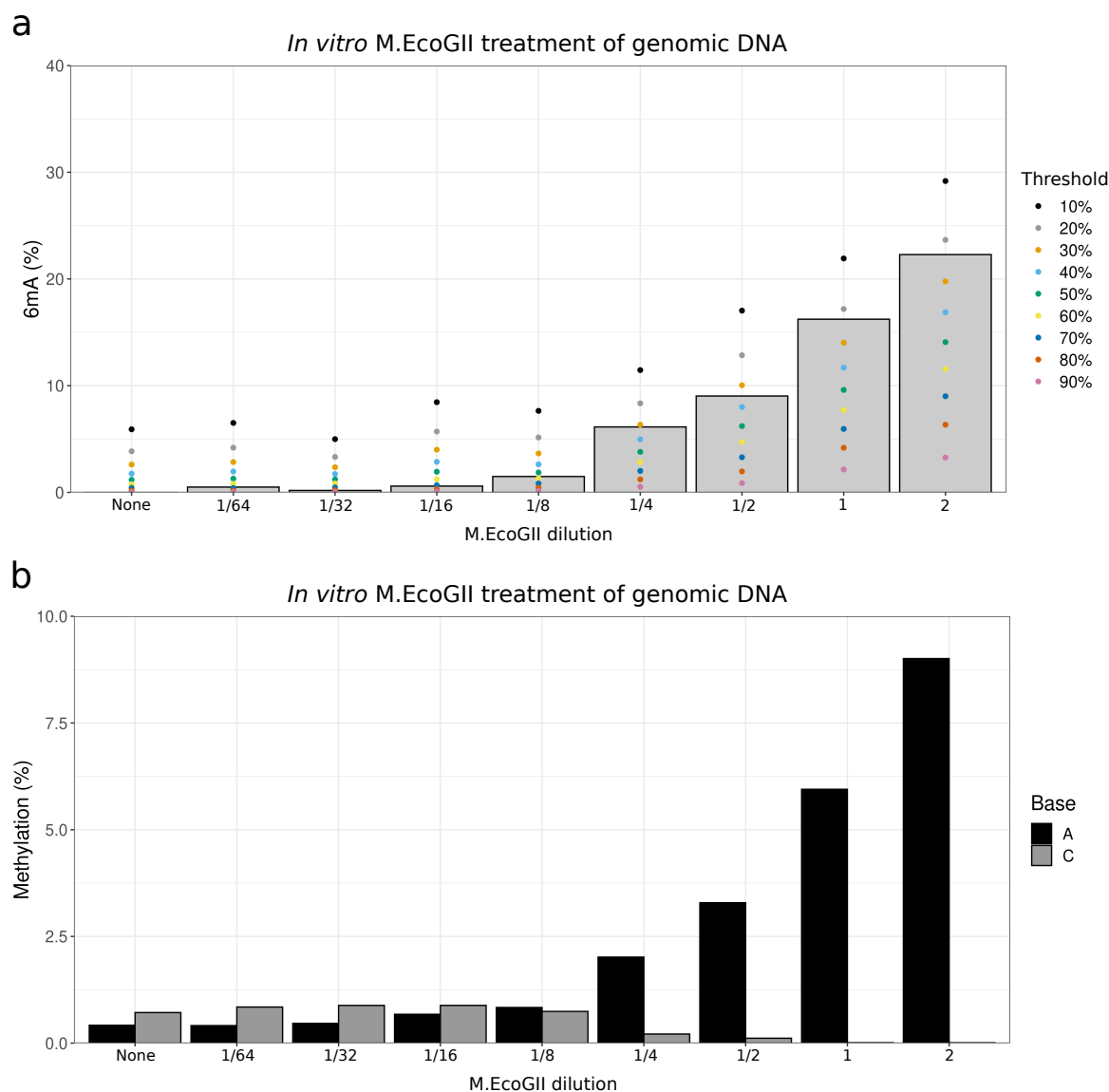


Fig. 4.28 Nanopore sequencing threshold comparisons to mass spectrometry data

S. cerevisiae wild-type genomic DNA was treated with a 2-fold dilution series of M.EcoGII (1x = NEB recommended concentration for the amount of DNA treated) or water (negative control) and incubated at 37 °C for one hour. **a)** Grey bars = HPLC mass spectrometry performed by Paolo Spingardi. For the nanopore sequencing data, a range of thresholds have been used to categorise each adenine as either adenine or 6mA. The mass spectrometry shows the 6mA (%) as 6mA/dT while the nanopore sequencing data shows the 6mA (%) as 6mA/dA. For the nanopore sequencing data, 100,000 adenines were analysed in any sequence context from a selection of random primary alignments mapped to the forward strand with a mapping quality greater than zero. Each threshold has been applied to the same set of 100,000 adenines. **b)** This time, 100,000 adenines (same as in **a)**) and 100,000 cytosines were analysed in any sequence context from a selection of random primary alignments mapped to the forward strand with a mapping quality greater than zero. The methylation probability represents the percentage of either adenines or cytosines that were classified as methylated according to a greater than or equal to 70% threshold.

To build upon the initial test of the sequence specificity of this model (figure 4.24) this analysis was repeated using the *in vitro* M.EcoGII treated nanopore data. Figure 4.29 shows that methylation is observed across all seven 4-mers and there is no clear bias towards detection of 6mA in a GATC context. The presence of variation between 4-mers suggests that either the M.EcoGII methylation, or the detection, contains some biases. Additionally, some false positive methylation was detected in the control sample (black bars) in agreement with figure 4.24a. Similarly to figure 4.28a, there is a substantial difference in the methylation level between samples. Therefore, a high concentration of M.EcoGII leads to increased methylation which is clearly distinguishable from the background level. In summary, M.EcoGII methylation can be detected in a variety of sequence contexts by nanopore sequencing and thus used to assess chromatin accessibility.

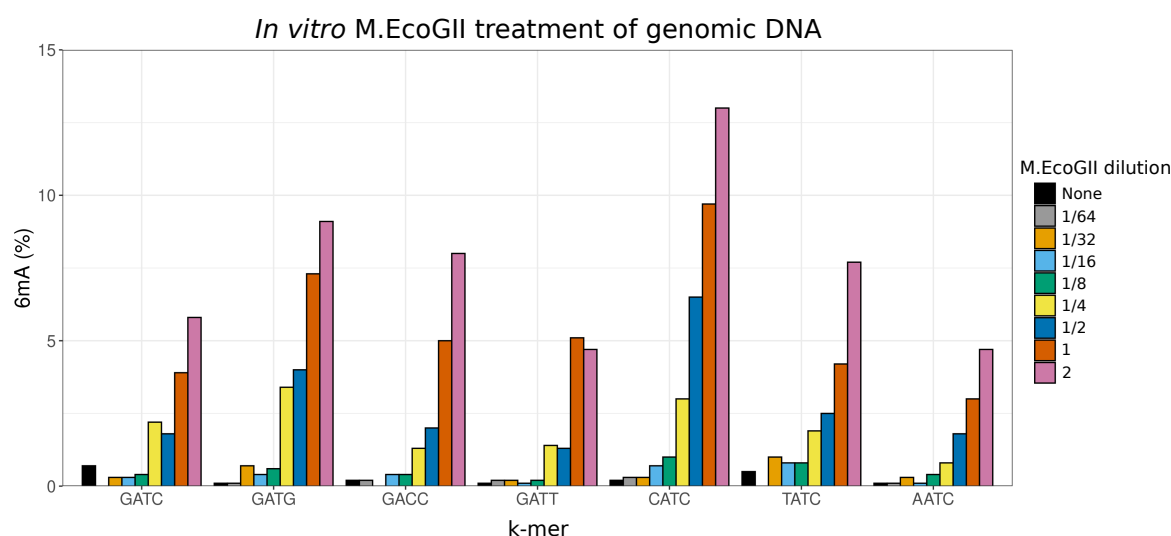


Fig. 4.29 Nanopore sequencing 4-mer analysis of *in vitro* M.EcoGII-treated DNA

Yeast wild-type (T7107) genomic DNA was treated with a 2-fold dilution series of M.EcoGII and incubated at 37 °C overnight. Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence, excluding the mitochondrial genome. The 6mA probability at the second base in a selection of 4-mers, all with one base change from GATC, was investigated in 1,000 sites for each k-mer, from a random selection of 50,000 reads. These were classified as methylated or not based on a greater than or equal to 70% threshold. Only the adenine in the second position was analysed for the AATC sequence context. None = Negative control where water was added in place of M.EcoGII; 1 = NEB recommended amount of M.EcoGII (other dilutions are relative to this).

To investigate the sequence specificity in wider contexts, all 3,367 6-mers containing at least one adenine were each analysed in 1,000 sites of primary alignments with a mapping quality above zero. To reduce the likelihood of the sites being the same genomic position but different reads, 50,000 reads were randomly selected. If a 6-mer contained more than one adenine then for every instance of the 6-mer a random adenine was selected for analysis. The methylation level of these 6-mers were analysed in the 'No M.EcoGII' control and the '2x M.EcoGII' (0 and 22% methylation respectively according to mass spectrometry (figure 4.27)) to facilitate a comparison between the two extremes of methylation levels. In figure 4.30 each 6-mer is represented as a dot, with colour coding by number of adenines. The position of the dots indicates the percentage methylated when treated with M.EcoGII versus the background percentage methylation (in the No M.EcoGII condition). Figure 4.30 indicates that the '2x M.EcoGII' sample had higher levels of methylation detected than the 'No M.EcoGII' sample since all the dots are above the black line. There are no clear groups of each colour therefore the number of adenines in each 6-mer is not biasing the results. This also showed that methylation is observed across different 6-mers, providing further evidence of the non-sequence specificity of both M.EcoGII methylation and detection with Guppy v3.6.0 and the Rerio model for 6mA in all sequence contexts. As the methylation threshold increased from 30% to 70% the level of background methylation was substantially reduced (shown by the shift of the dots towards the left of the graph) and the data points became more tightly clustered together, with a reduction in the level of M.EcoGII methylation observed. Higher thresholds had lower false positives (shown by a reduction in the methylation in the 'No M.EcoGII' condition) but also had lowered measurements in the '2x M.EcoGII' condition compared to the same condition in the mass spectrometry data (table 4.5). At a 70% threshold, about 8.9% of adenines were methylated in the '2x M.EcoGII' condition, compared to 22% by mass spectrometry (figure 4.28) and around 0.3% of adenines were

methyated in the unmethylated negative control, which acts as a false positive rate. Therefore the selection of 70% was selected as a compromise between these two aspects.

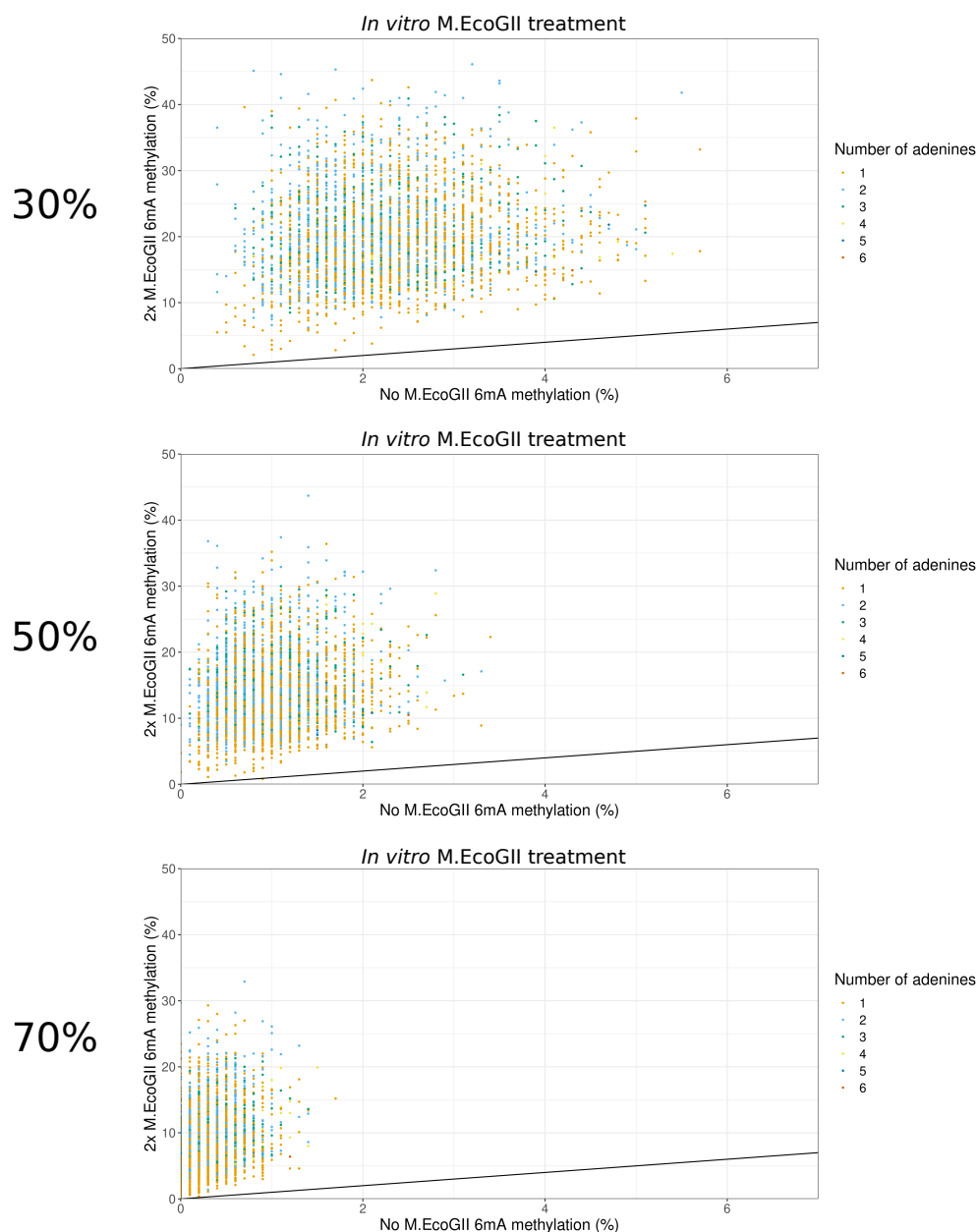


Fig. 4.30 Nanopore sequencing 6-mer analysis across 6mA classification thresholds

Analysis of the background level of methylation across 6-mers. For this, 50,000 primary alignments mapped to the forward strand with mapping quality above zero were randomly selected for each condition ('No M.EcoGII' and '2x M.EcoGII'). From this selection of reads, the methylation state of each 6-mer was identified in 1,000 sites. Methylation was classified using either a 30%, 50% or 70% threshold. The same 1,000 instances of the 6-mers were analysed for each of the methylation thresholds. The black line is where the methylation is equal in the two conditions.

Threshold (%)	Mean for '2x M.EcoGII' (%)	'2x M.EcoGII' relative to 22.3% mass spectrometry measurement (%)	Mean for 'No M.EcoGII' (FP (%))
10	30.0	-	5.3
20	24.1	-	3.4
30	20.1	90.1	2.3
40	16.9	75.8	1.6
50	14.1	63.2	1.0
60	11.5	51.6	0.64
70	8.9	39.9	0.33
80	6.2	27.8	0.13
90	3.2	13.5	0.027

Table 4.5 M.EcoGII *in vitro* 6-mer methylation across 6mA classification thresholds

Mean = Mean percentage of sites methylated with each threshold for each condition. The proportion of the mass spectrometry value for '2x M.EcoGII' (22.3%) is only shown where it would be in the range of 0 to 100. The mean for 'No M.EcoGII' is used as the false positive rate. Total number of 6-mers with at least one adenine is 3,367. 1,000 sites investigated for each 6-mer.

4.7 Creation and validation of inducible M.EcoGII *S. cerevisiae* strain and assessment of non-specific adenine methylation detection by nanopore sequencing

4.7.1 Inducible M.EcoGII *S. cerevisiae* strain generation and verification

An inducible M.EcoGII strain was produced by transforming a vector containing M.EcoGII under the control of a *GALI* promoter, into wild-type *S. cerevisiae* (T7107). This vector (figure 4.31) was produced by Dr Dzmitry Batrakou and contains FLAG and histidine tags at the C-terminus of the M.EcoGII gene, allowing the M.EcoGII protein to be detected with antibodies by western blotting. The *URA3* homology regions facilitated integration into

the *S. cerevisiae* genome at the *URA3* locus. Transformation was performed with pDB81 (M.EcoGII) DNA pre-linearised with NotI, pEDH3 (gRNA against *URA3*) and a plasmid expressing Cas9 under tryptophan (Trp) selection. After transformation eight clones were analysed by PCR (figure 4.32) and Sanger sequencing (not shown) to check for the presence of the M.EcoGII gene and whether or not it contained any errors as previous attempts by Dr Dzmitry Batrakov all resulted in early stop codons or deletions.

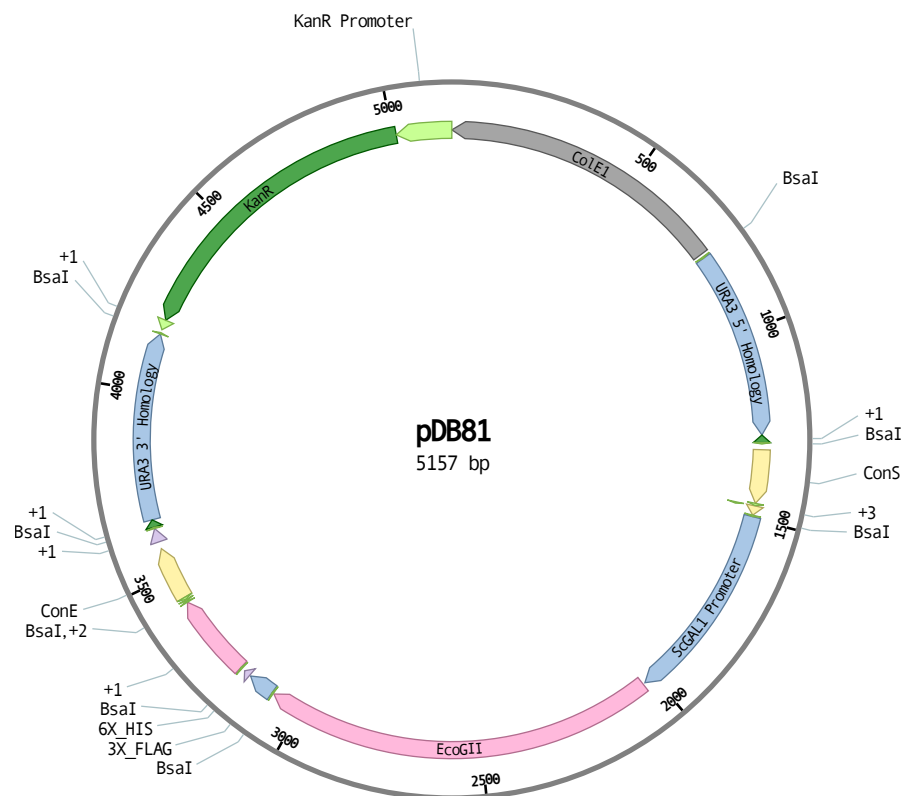


Fig. 4.31 Plasmid containing M.EcoGII for transformation

The plasmid (pDB81) was used in the transformation of T7107 to create an inducible M.EcoGII strain. This plasmid was generated by Dr Dzmitry Batrakov and has the M.EcoGII gene under a *GAL1* promoter, with both FLAG and histidine tags at the end of the gene. Image produced using Benchling. KanR = Kanamycin resistance gene.

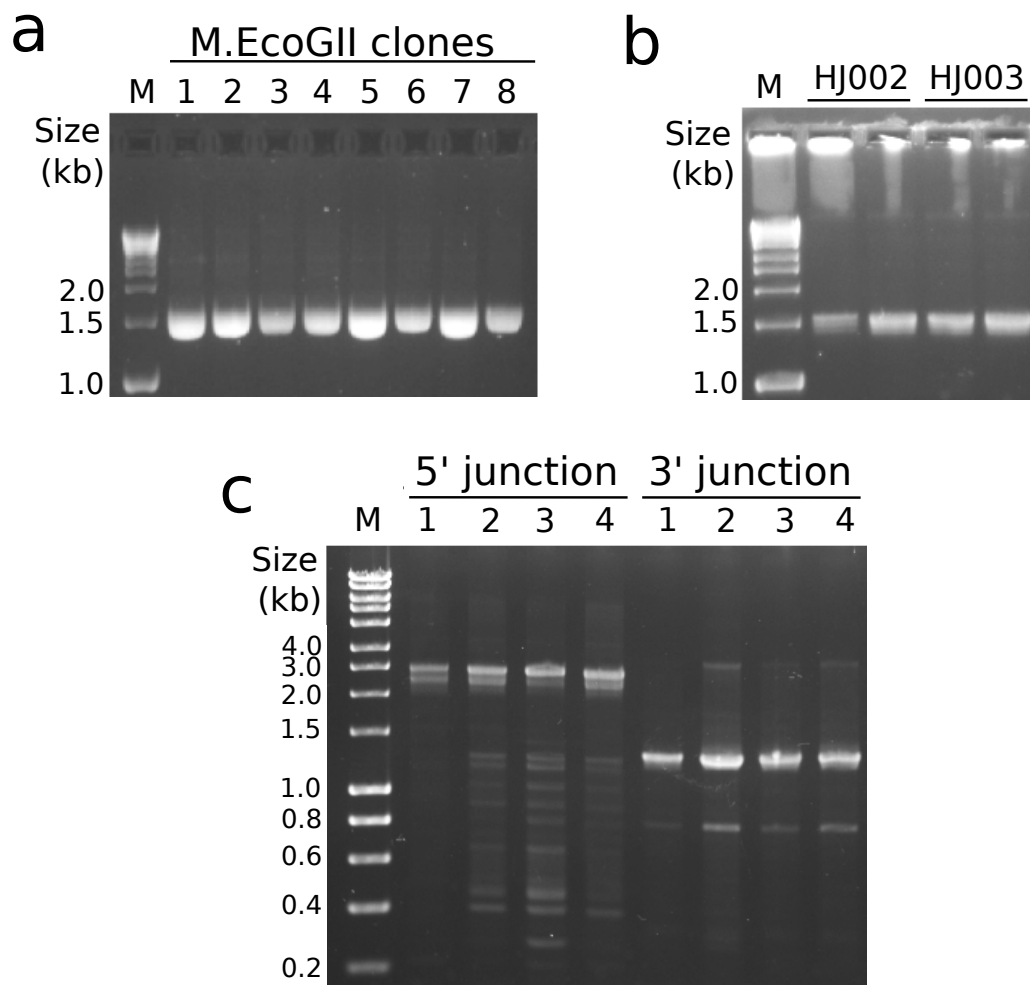


Fig. 4.32 Generation and verification of an inducible M.EcoGII *S. cerevisiae* strain

The plasmid containing the M.EcoGII gene (pDB81 as shown in figure 4.31) was transformed into wild-type *S. cerevisiae* (T7107). **a)** PCR of the 5' junction of the *URA3* locus in 8 clones grown on the Ura- Trp- plate after transformation. Primers include CA033 which binds in the *URA3* locus within the *S. cerevisiae* genome, and DB559 which binds in the M.EcoGII gene in the pDB81 plasmid, creating a 1,560 bp amplicon. **b)** PCR of the 5' junction of the *URA3* locus of the two strains that were frozen at -80 °C and reawoken on to YPAD. Two isolates were tested for each strain. Same primers used as in (a). **c)** PCR of the 5' junction (with CA033 and DB508, creating a 2,447 bp amplicon) and 3' junction (with DB526 and CA038, creating a 1,341 bp amplicon) of the *URA3* locus. 1 and 2 are clones from HJ002 freshly woken up from -80 °C; 3 and 4 are clones from HJ003 freshly woken up from -80 °C. All gels were 1% TBE-agarose, run at 90V for one hour and stained with ethidium bromide. Bands will only be present if the M.EcoGII gene is present at the *URA3* locus. M = Bioline Hyperladder I.

4.7.2 Inducible M.EcoGII identified by western blotting

To test whether the M.EcoGII protein was present after induction with 2% galactose, a timecourse of samples were collected post-galactose induction. The inducible M.EcoGII strain was grown in YPAD, pelleted, washed and resuspended in 2% raffinose and grown in this for two or more doubling times, then induced with galactose to a final concentration of 2%. Samples were collected post-galactose induction, with 0 minutes being immediately before addition of galactose. Proteins were extracted and probed with either an anti-histone H3 antibody, or anti-FLAG antibody, to detect the histone H3 (as a control) or FLAG-tagged M.EcoGII respectively. These were subsequently probed with anti-rabbit HRP and anti-mouse HRP respectively to allow visualisation by ECL chemiluminescence. Figure 4.33 shows that the galactose induction induces expression of M.EcoGII, shown as a band at 42 kDa which is absent at the start, appears at 10 minutes and then increases in intensity throughout the timecourse, appearing relatively stable (or saturated on this western blot) beyond half an hour. Therefore, high levels of M.EcoGII are being expressed relatively quickly post induction.

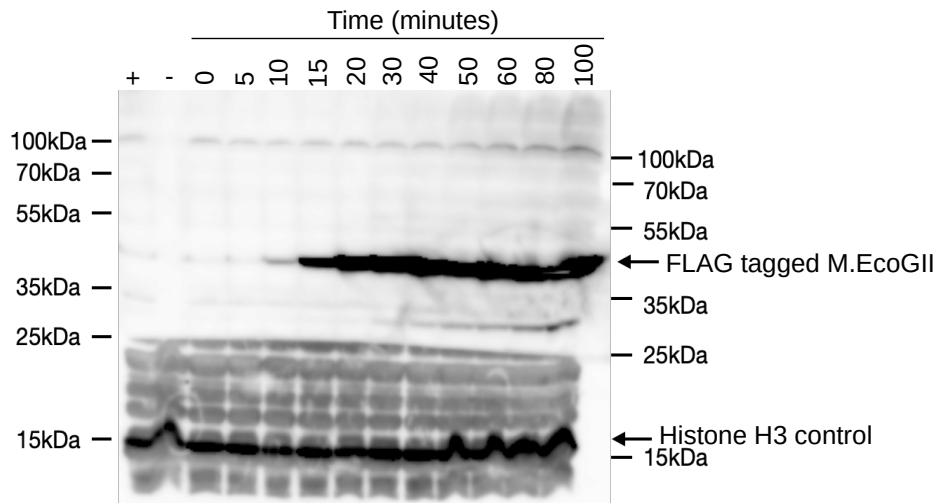


Fig. 4.33 Detection of inducible M.EcoGII by western blotting

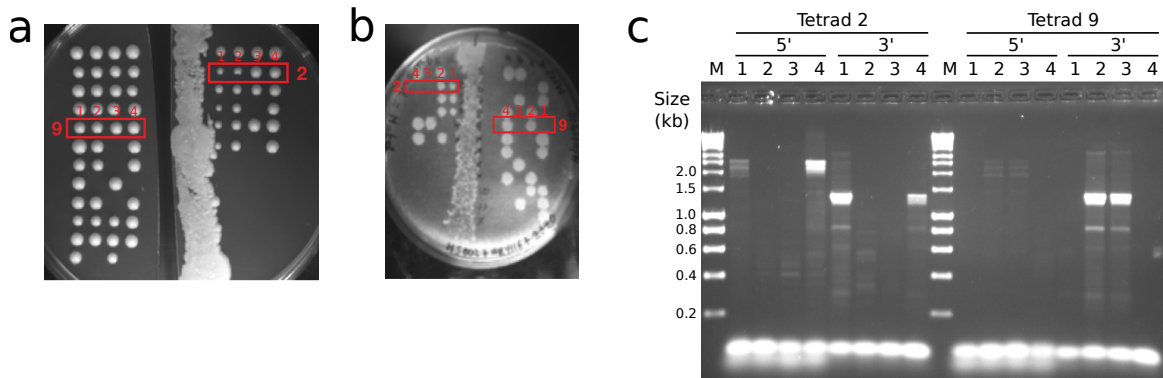
S. cerevisiae cells were grown in YPAD, transferred to 2% raffinose for a couple of doubling times and then the expression of M.EcoGII methyltransferase was induced with 2% galactose. 0 minutes = 2% raffinose, sample taken immediately prior to galactose induction. Inducible M.EcoGII strain (HJ002) has not been backcrossed. M.EcoGII is expected to be seen at 42 kDa, as the molecular weight of M.EcoGII is 40 kDa, the FLAG tag is 1 kDa and the His tag is 1 kDa. Histone H3 is expected to be seen at 17 kDa. + = Proteins extracted from POY001 *S. cerevisiae* strain with a FLAG-tagged PolII (turned out that the size was too large to be resolved on this gel); - = Proteins extracted from T7107 (wild-type *S. cerevisiae* strain).

4.7.3 Backcrossing of inducible M.EcoGII *S. cerevisiae* strain

Once expression of M.EcoGII had been confirmed, the new *S. cerevisiae* strain was backcrossed with a wild-type strain with a W303 background (AUY116) to correct for any mutations that may have occurred across the genome during the transformation step. Backcrossing involved mating with AUY116 on YPAD, sporulation and tetrad dissection to isolate the four spores (figure 4.34a and d). To assess which tetrads showed a 2:2 segregation for mating type and which spores were of which mating type, the tetrads were mated with two control strains of known α or a mating type on YPAD plates and replica-plated on to minimal media which only facilitates the growth of diploid colonies. Growth on minimal media indicated the spore was of the opposite mating type, thus was able to mate and form diploids (figure 4.34b and e). Tetrads showing growth for four spores and 2:2 segregation of mating

type were tested for the presence of M.EcoGII by PCR across the 5' and 3' junctions of the integration site at the *URA3* locus (figure 4.34c and f). The presence of the M.EcoGII gene was also expected to display 2:2 segregation. The integrated M.EcoGII was also PCR amplified and Sanger sequenced to verify that the integrated copy was error-free. By testing two complete tetrads, a spore was identified that contained the desired characteristics (mating type a to allow future α -factor arrest, and the integration of the M.EcoGII gene).

Backcross 1: HJ002 + AUY116



Backcross 2: HJ004 + AUY116

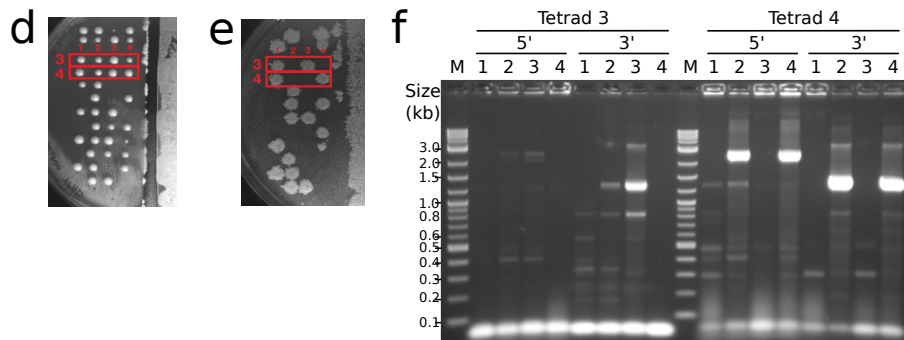


Fig. 4.34 Backcrossing of an inducible M.EcoGII *S. cerevisiae* strain

The *S. cerevisiae* strain containing the M.EcoGII gene (HJ002) was backcrossed with AUY116 (wild-type W303 background) twice to correct for mutations that may have occurred throughout the genome during the transformation. **a-c)** The first backcross was between the T7107 strain transformed with the M.EcoGII gene (HJ002) and a wild-type *S. cerevisiae* strain (AUY116). Tetrad 2 spore 4 was frozen as a new strain called HJ004. **d-f)** The second backcross was between HJ004 (generated in the first backcross) and a wild-type *S. cerevisiae* strain (AUY116). Tetrad 4 spore 2 was frozen as a new strain called HJ005, while tetrad 4 spore 4 was frozen as HJ006. **a and d)** Growth of tetrads on a YPAD plate with no selection performed. Tetrad numbers are given to the side of the red boxes, spore numbers above the red boxes. **b and e)** Growth of tetrads which have been replica-plated on to a YPAD plate with strain 2370 of mating type α , and subsequently replica-plated on to minimal media. Growth will only be observed for spores that are of mating type α . **c and f)** PCR of the 5' and 3' junctions of the region of the *URA3* locus where the M.EcoGII gene may be integrated (see figure 4.32 for expected sizes). For panel c, M = Bioline Hyperladder I; for panel f, M = Quickload 2-log ladder.

4.7.4 Effect of inducible M.EcoGII on *S. cerevisiae* growth

As *S. cerevisiae* does not naturally methylate its DNA [173, 174], the effect of the presence of this methyltransferase was assessed. The strain with the inducible M.EcoGII and backcrossed twice (HJ005) was compared to the wild-type strain used for backcrossing (AUY116). These were grown in YPAD and 2% raffinose, when M.EcoGII should be repressed and non-induced respectively. There may be very low levels of M.EcoGII protein expressed in these conditions. Figure 4.35 shows that in YPAD, the repressive condition, the two strains grow at a comparable rate. The doubling time for both strains is longer in 2% raffinose than in YPAD, which is to be expected as raffinose is a poorer sugar source than glucose. However, in 2% raffinose HJ005 (the strain with the inducible M.EcoGII) has a longer doubling time than AUY116 (figure 4.35) which suggests that 2% raffinose may permit some M.EcoGII expression which may have a negative effect on growth. M.EcoGII has the potential to perform a high level of methylation as it can target every accessible adenine, creating a high number of target sites throughout the genome. To assess whether higher levels of DNA methylation affects the growth of the strain the growth curve should be repeated in galactose. It is hypothesised that growth in 2% galactose would be slower for a strain with an inducible M.EcoGII, despite the preferential sugar source, as high levels of exogenous protein expression and DNA methylation would be an extra burden on the cell. For experiments involving induction of M.EcoGII with galactose, samples are taken within the first couple of hours. It takes time for the M.EcoGII to build up and 2% galactose is a richer sugar source than 2% raffinose, therefore it is unlikely that the growth will be substantially altered between galactose induction and sample collection.

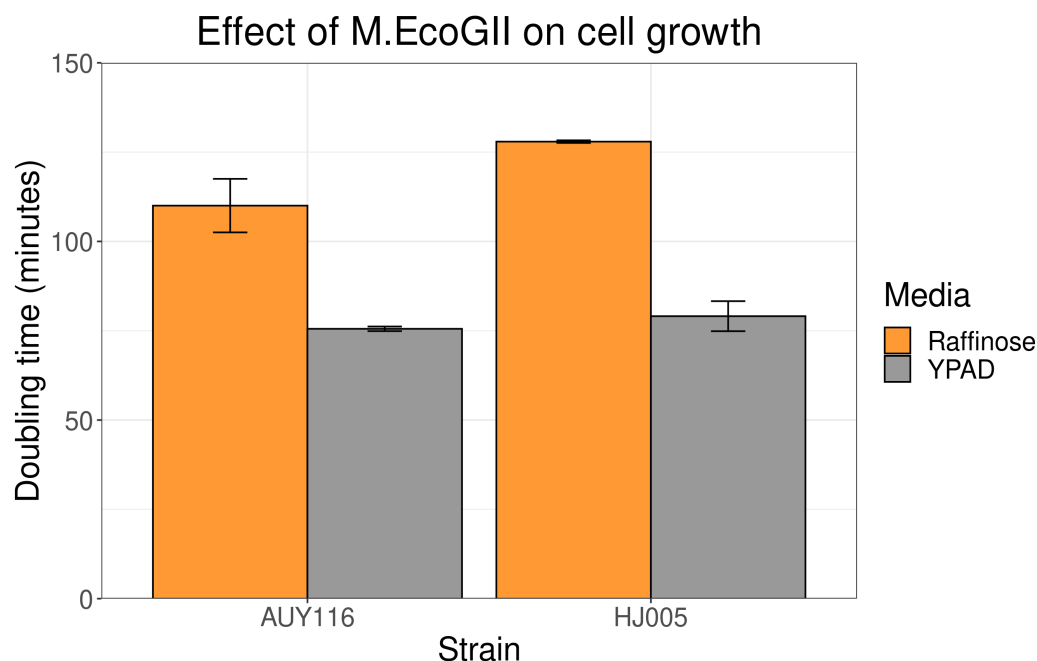


Fig. 4.35 Growth of an *S. cerevisiae* strain with an inducible M.EcoGII

Growth curves were performed on the twice-backcrossed strain with an inducible M.EcoGII (HJ005) and the wild-type strain used for backcrossing (AUY116). Concentration of raffinose in the medium was 2%. Strains were grown in the media for at least two doubling times prior to measurements. Three repeats were measured for all conditions, except for HJ005 in raffinose where two replicates were measured as the third culture did not get to a suitable OD₆₀₀. Error bars are standard deviation.

4.7.5 DNA methylation by inducible M.EcoGII methyltransferase identified in *S. cerevisiae* by mass spectrometry and nanopore sequencing

To test whether or not the M.EcoGII was an active methyltransferase a similar timecourse to that for the protein detection experiment was performed, though this time with DNA extractions rather than protein extractions. The DNA was then split in half, one half for mass spectrometry analysis, and the other half for nanopore sequencing.

Genomic DNA methylation assessed by mass spectrometry

Mass spectrometry, performed by Paolo Spingardi, provided validation of the methylation level detected by nanopore sequencing, which is useful as the nanopore sequencing data can vary based on how the analysis is performed (for example by altering the 6mA threshold) while the mass spectrometry data is a static number of total 6mA detected per sample. Three timecourses with subsequent DNA extraction have been performed on *S. cerevisiae* strains with an inducible M.EcoGII. The first two were to test the ability of M.EcoGII to methylate DNA, after either one or two backcrosses. These had more early timepoints after galactose induction as figure 4.33 indicated that M.EcoGII was first present around 10 minutes after galactose induction. However, mass spectrometry showed that although M.EcoGII did methylate DNA and very consistently between the two strains, methylation did not occur until later in the timecourse, particularly until around one hour after galactose induction (figure 4.36a). Therefore, the third timecourse, performed on the twice-backcrossed strain of inducible M.EcoGII (HJ005), was performed with samples taken every 15 minutes in the first two hours after galactose induction. The results in figure 4.36b reinforce the observation from figure 4.36a that observable methylation does not occur much before one hour post-galactose induction. This suggests that although the protein is produced earlier in the timecourse, there is a delay between the production of the protein and the methyltransferase function becoming active. This delay may be caused by the protein moving from the cytosol where it is produced to the nucleus where it performs its methyltransferase function, as the protein samples were whole cell extracts therefore the protein localisation was not investigated. Beyond one hour after galactose induction the methylation continues to increase until potentially starting to tail off after 105 minutes (figure 4.36b). This may be due to reaching a maximal amount of methylation as although the methylation is occurring in living cells with dynamic protein loading and unloading on the DNA, the majority of the DNA will be protein bound [7] thus presumably inaccessible to M.EcoGII. Further repeats would be required to confirm this.

The three biological replicates all agree reasonably well for the YPAD, 2% raffinose and the 30 and 60 minute samples in 2% galactose. However, there is some variation for the 120 minutes post-galactose induction. Both replicates of the twice-backcrossed strain (dark green in figure 4.36a and blue in b) show around 15% methylation at this time point, thus are in good agreement. The single backcross biological replicate (light green in figure 4.36a) shows reduced methylation, around 8%, at this timepoint, which indicates that the M.EcoGII was less functional in this strain. There may have been other mutations in the genome, that were lost in the second backcross, that suppressed the M.EcoGII methyltransferase functionality over time. The mass spectrometry has shown the single and double backcrossed M.EcoGII strains to contain a functional M.EcoGII, with results being consistent across biological replicates of the same strain.

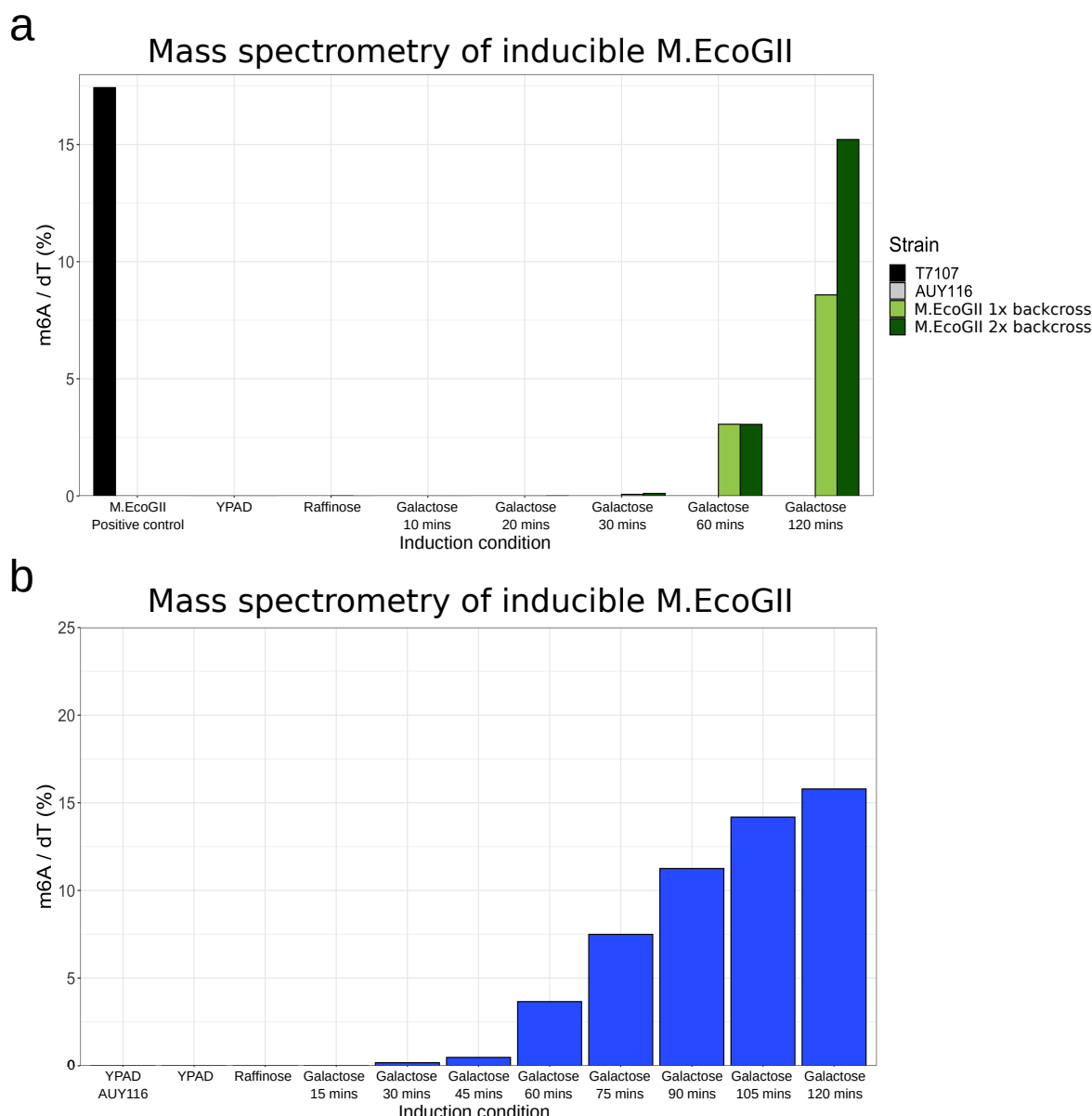


Fig. 4.36 Mass spectrometry of *in vivo* inducible M.EcoGII timecourse

S. cerevisiae were grown in YPAD at 23 °C to OD 0.1 then pelleted and resuspended in 2% raffinose for 2 doublings (until around OD 0.5), at which point galactose was added to a final concentration of 2%. This induced the expression of the M.EcoGII under the *GALI* promoter. Sample collection, DNA extraction and digestion to nucleosides were performed by Heather Jeffery prior to HPLC mass spectrometry by Paolo Spingardi. **a)** Mass spectrometry of the first inducible M.EcoGII timecourse. T7107 and AUY116 are wild-type strains which do not express any methyltransferases. T7107 was treated with and without M.EcoGII *in vitro* in the M.EcoGII condition. A lack of methylation is observed in AUY116 by the absence of a visible grey bar in the YPAD condition. HJ004 = M.EcoGII 1 x backcross; HJ005 = M.EcoGII 2 x backcross. **b)** An additional inducible M.EcoGII timecourse was performed with samples taken every 15 minutes to further characterise the level of methylation at later timepoints. This was using strain HJ005 (EcoGII 2 x backcross from panel (a)) and shows good agreement with panel (a). For both timecourses raffinose and galactose were both 2% final concentrations; mins = minutes.

Genomic DNA methylation assessed by nanopore sequencing

The methylation level of these timecourses using the twice-backcrossed inducible M.EcoGII strain were also analysed by nanopore sequencing which will be used later to investigate chromatin structure. Coverage was more than 150x for the initial timecourse and generally more than 400x for the second timecourse with more evenly spaced timepoints (table 4.6). The particularly low coverage observed for some samples may be due to inefficient adapter ligation.

Strain	Media	Timepoint	Mean read length (bases)	Coverage*
AUY116	YPAD	-	4,454	162
HJ005	YPAD	-	3,839	189
HJ005	2% Raffinose	0 minutes	3,798	163
HJ005	2% Galactose	10 minutes	3,919	152
HJ005	2% Galactose	20 minutes	3,891	187
HJ005	2% Galactose	30 minutes	4,017	209
HJ005	2% Galactose	60 minutes	3,947	175
HJ005	2% Galactose	120 minutes	4,259	156
AUY116	YPAD	-	4,051	3
HJ005	YPAD	-	3,986	563
HJ005	2% Raffinose	-	3,744	497
HJ005	2% Galactose	15 minutes	4,078	3
HJ005	2% Galactose	30 minutes	3,995	492
HJ005	2% Galactose	45 minutes	3,895	576
HJ005	2% Galactose	60 minutes	3,749	397
HJ005	2% Galactose	75 minutes	3,815	470
HJ005	2% Galactose	90 minutes	3,918	423
HJ005	2% Galactose	105 minutes	3,528	422
HJ005	2% Galactose	120 minutes	3,384	423

Table 4.6 M.EcoGII *in vivo* nanopore sequencing overview

Nanopore sequencing runs from 15th, 24th and 30th June 2020. Reads were mapped to the sacCer3 reference genome (excluding mitochondrial genome) with minimap2, filtered for primary alignments and mapping quality greater than zero. Coverage calculated by primary alignments with mapping quality greater than 0 multiplied by mean read length, divided by genome size of 11,209,039 bases excluding the mitochondrial DNA. Mean read length and coverage are both rounded to the nearest integer. * Coverage is genomic coverage, therefore strand-specific coverage would be this coverage divided by 2.

Figure 4.37 shows that nanopore sequencing detects adenine methylation in a similar pattern as observed in the mass spectrometry data (figure 4.36). A low level of methylation was detected in the control samples (YPAD and 2% raffinose) and in agreement with the mass spectrometry data, noticeable methylation was not observed until 60 minutes post-galactose induction of M.EcoGII. The overall quantity of methylation detected by nanopore sequencing was about half of that detected by mass spectrometry due to the use of a 70% threshold to reduce detection of background methylation in samples with lower or no methylation. This was a similar level as was seen with the 70% threshold in the *in vitro* experiment (figure 4.28). Importantly, as in the mass spectrometry analysis, the methylation level appeared to be reasonably consistent for the timepoints that appeared in both timecourses.

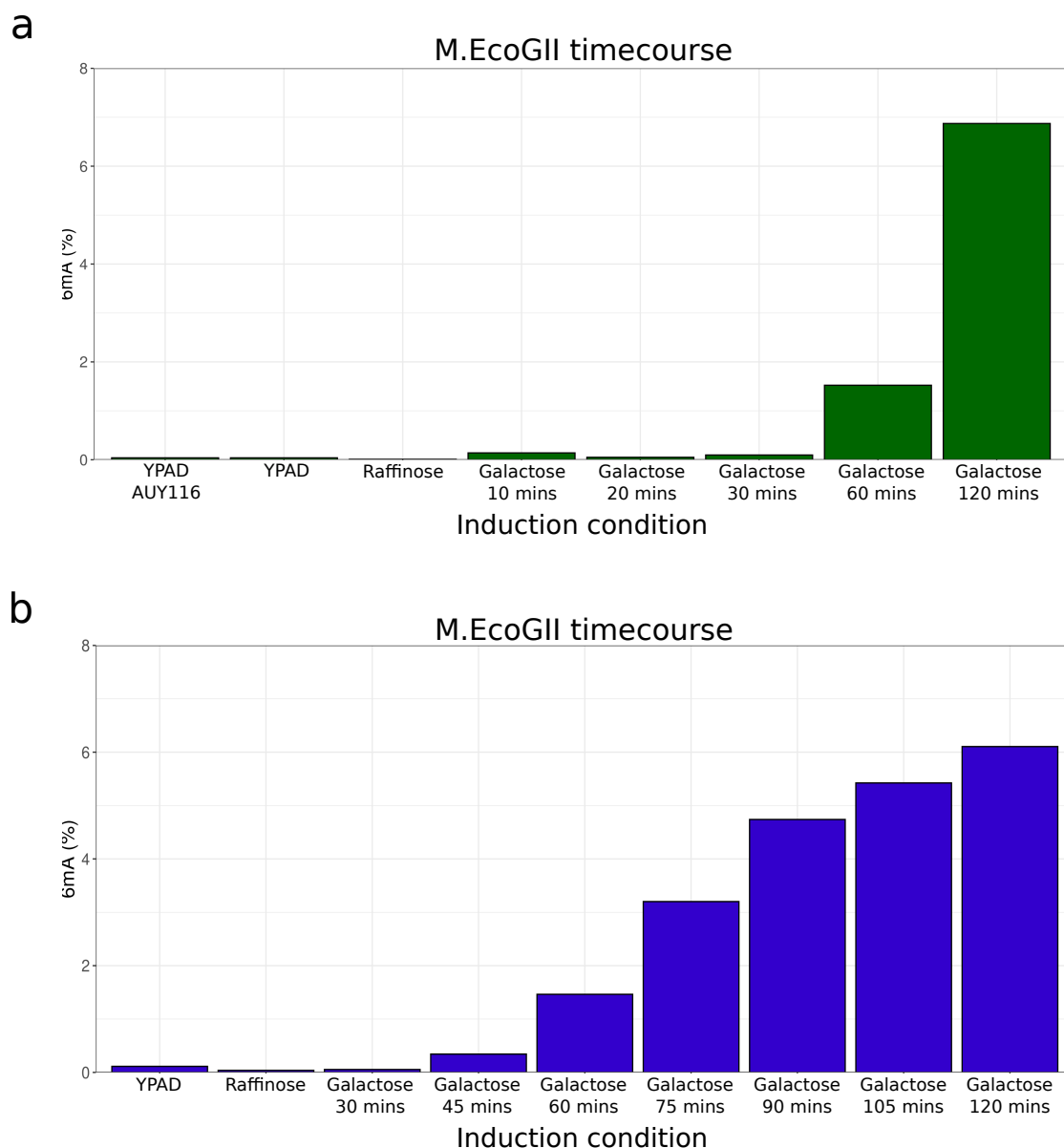


Fig. 4.37 Nanopore sequencing 6mA detection of *in vivo* inducible M.EcoGII timecourses

S. cerevisiae were grown in YPAD at 23 °C to OD₆₀₀ 0.1 then pelleted and resuspended in 2% raffinose for 2 doublings (until around OD₆₀₀ 0.5), at which point galactose was added to a final concentration of 2%. This induced the expression of the M.EcoGII under the *GAL1* promoter. Nanopore sequencing data was basecalled and methylation detected with Guppy v3.6.0 and the Rerio model used in figure 4.24 then mapped to the sacCer3 reference sequence excluding mitochondrial DNA. For both graphs adenines were evaluated in any sequence context at 100,000 sites from a random selection of primary alignments mapped to the forward strand with a mapping quality of greater than zero. Each site was classified as methylated or not based on a 70% threshold (greater than or equal to was classed as methylated). **a)** Initial M.EcoGII timecourse. **b)** 15 minute M.EcoGII timecourse. Samples for AUY116 in YPAD and Galactose 15 minutes had low coverage so were not assessed. AUY116 = control wild-type *S. cerevisiae* strain with no methyltransferase; all other samples are based on HJ005 which has an inducible M.EcoGII; mins = minutes.

4.8 Inducible M.EcoGII *S. cerevisiae* strains with replication protein mutations

The first proteins involved in origin licensing are Orc1-6 and Cdc6. There are mutations which can prevent these proteins from binding to the DNA [90, 217]. Two yeast strains containing mutations of this nature were selected to be combined with the inducible M.EcoGII to allow the chromatin structure around origins to be analysed in the absence of these protein binding events [90, 217].

4.8.1 Strain generation and verification

The *orc1-161* strain [90] was kindly sent to me from the MacAlpine lab. The *cdc6-1* strain [217] was made *RAD5* and *BUD4* wild-type by Stephanie Barker. To combine these mutations with an inducible M.EcoGII, the HJ006 inducible M.EcoGII strain was used. HJ006 was generated at the same time as HJ005 and also verified to possess an integrated M.EcoGII gene at the *URA3* locus by PCR (figure 4.34) and Sanger sequencing (not shown), although the functional activity of M.EcoGII in this strain was not assessed by mass spectrometry. HJ006 was of mating type α so could mate with *orc1-161* and *cdc6-1* mutant strains which were both mating type a. Spores from tetrad dissection were assessed for mating type (ideally type a desired to allow future α -factor arrest), presence of the *orc1-161* or *cdc6-1* mutation and incorporation of the M.EcoGII gene at the *URA3* locus (figures 4.38 and 4.39). Mating type and M.EcoGII verification were performed as in sections 4.7.3 and 2.7.10 with the mating tests performed at 23 °C to prevent selection against the isolates containing the *orc1-161* and *cdc6-1* mutations which both cause the strains to be temperature sensitive. The presence of the *orc1-161* or *cdc6-1* mutant was verified by a temperature sensitivity assay which involved replica plating the tetrad dissection plate on to a YPAD plate and growing it at 30 °C.

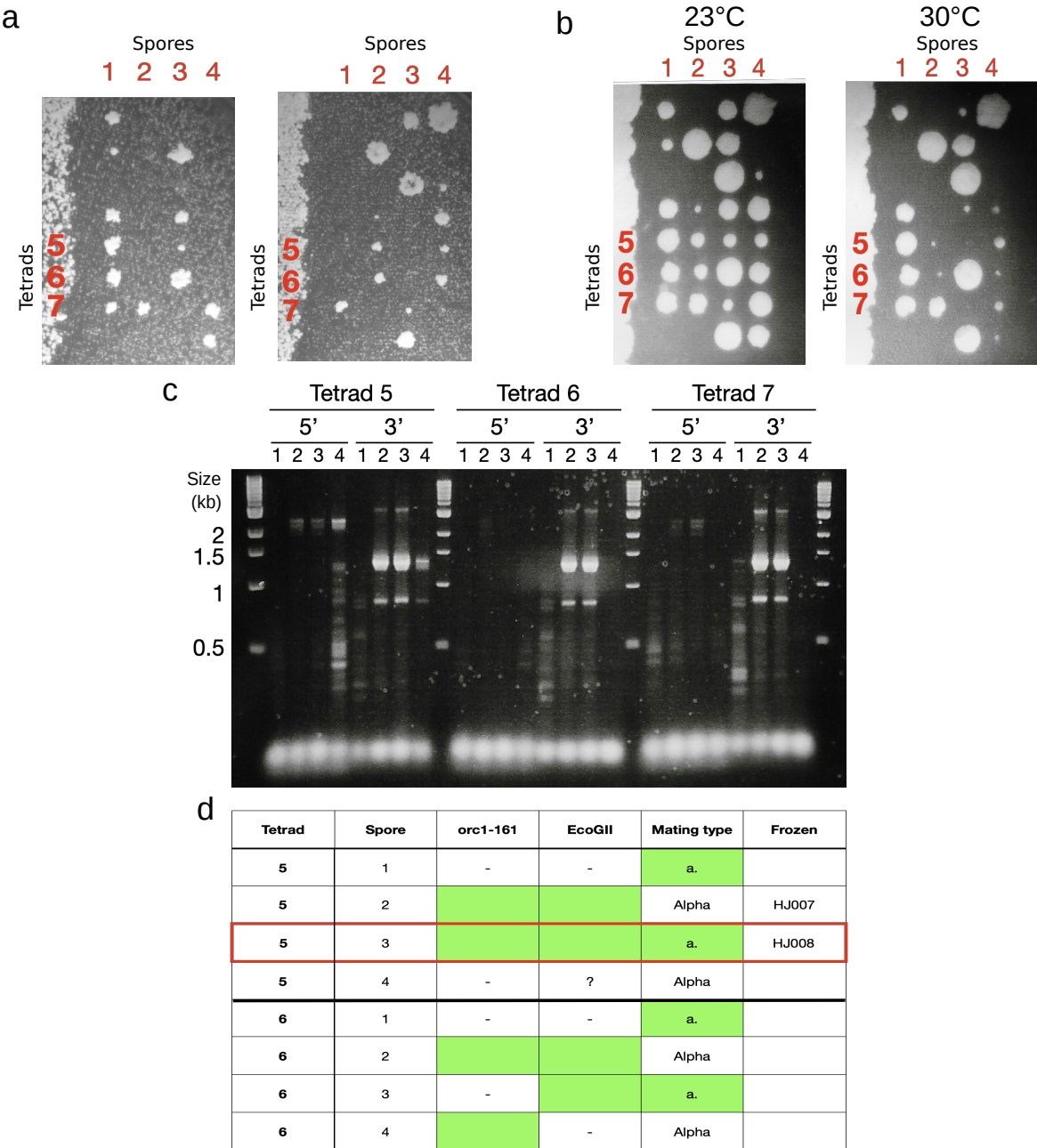


Fig. 4.38 Generation and verification of *orc1-161* strain with an inducible M.EcoGII

Verification of **a)** mating type **b)** presence of *orc1-161* and **c)** presence of M.EcoGII gene at *URA3* locus. **a)** Mating type was verified by replica plating tetrads on to a plate covered in a strain of known mating type α (shown on the left) or *a* (shown on the right). Once grown at 23 °C, the plate was replica plated on to minimal media. Growth was seen for spores that have formed diploids after successful mating. Mating can only occur between spores of opposite mating types, therefore growth of a colony on a plate mated with known type *a* would indicate the spore is type α and vice versa. **b)** Presence of *orc1-161* mutation was verified by temperature sensitivity. A spore containing the *orc1-161* mutation will be less able to grow at 30 °C. **c)** Presence of M.EcoGII gene at *URA3* locus was verified by PCR at the 5' and 3' ends of the *URA3* inserted region. Spores within each tetrad are represented by the numbers 1 to 4. The outer primers bind in the *sacCer3* *S. cerevisiae* genome, while the inner bind in the inserted M.EcoGII gene. PCR was performed with CA033 + DB508 for the 5' end and DB526 + CA038 for the 3' end. **d)** Summary of the characteristics in a-e, highlighting the spores of interest (mating type *a*, with *orc1-161* and M.EcoGII).

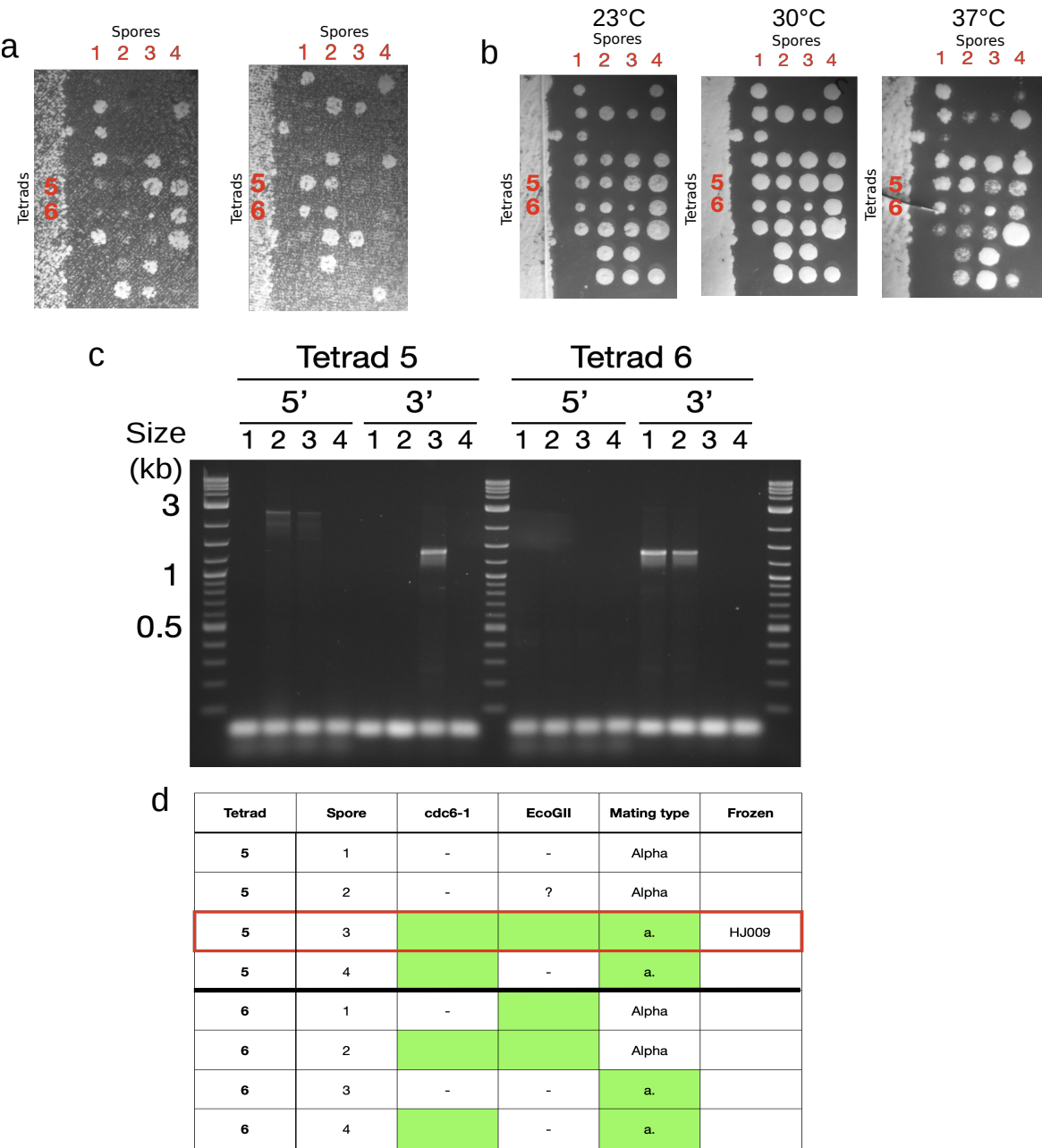


Fig. 4.39 Generation and verification of *cdc6-1* strain with an inducible M.EcoGII

Verification of **a)** mating type, **b)** presence of *cdc6-1* and **c)** presence of M.EcoGII gene at *URA3* locus. **a)** Mating type was verified by replica plating tetrads on to a plate covered in a strain of known mating type α (shown on the left) or a (shown on the right). Once grown at 23 °C, the plate was replica plated on to minimal media. Growth was seen for spores that have formed diploids after successful mating. Mating can only occur between spores of opposite mating types, therefore growth of a colony on a plate mated with known type a would indicate the spore is type α and vice versa. **b)** Presence of *cdc6-1* mutation was verified by temperature sensitivity. A spore containing the *cdc6-1* mutation will be less able to grow at higher temperatures, where the clearest effect of the *cdc6-1* mutation on spores was observed by the lack of growth at 37 °C. **c)** Presence of M.EcoGII gene at *URA3* locus was verified by PCR at the 5' and 3' ends of the *URA3* inserted region. Spores within each tetrad are represented by the numbers 1 to 4. The outer primers bind in the yeast genome, while the inner bind in the inserted M.EcoGII gene. PCR was performed with CA033 + DB508 for the 5' end and DB526 + CA038 for the 3' end. **d)** Summary of the characteristics in a-e, highlighting the spores of interest (mating type a, with *cdc6-1* and M.EcoGII).

4.8.2 Effect of replication protein mutations combined with inducible M.EcoGII on *S. cerevisiae* growth

There were three genomic changes that could have perturbed the ability of the *S. cerevisiae* strains to perform normal cell cycle behaviour. These were the insertion of an M.EcoGII gene at the *URA3* locus, and two mutations in DNA replication proteins: *orc1-161* and *cdc6-1*. The two replication protein mutations cause the strains to be temperature sensitive. This means that the replication proteins (Orc1 and Cdc6) can bind to DNA at 23 °C but as the temperature is increased the mutation either prevents the protein from binding to the DNA or makes the protein more unstable leading to degradation at higher temperatures, thus affecting growth of *S. cerevisiae* strains with these mutations (figure 4.40). This figure shows that 37 °C inactivates both Orc1 and Cdc6 while 30 °C only inactivates Orc1. There is also an indication of slower growth of the *orc1-161* strain at 23 °C, suggesting the mutation may have a minor role even at this low temperature.

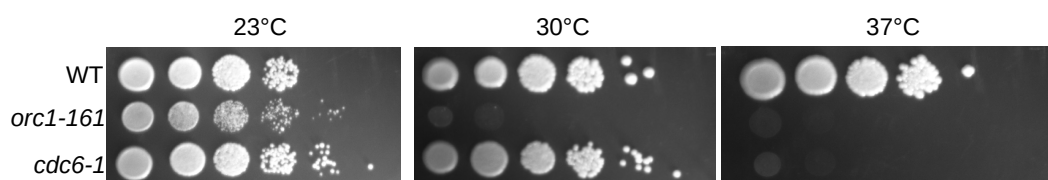


Fig. 4.40 Spot assay assesses temperature sensitive replication protein mutant strains

The *orc1-161* and *cdc6-1* mutations are temperature sensitive, thus should not be activated at 23 °C. From this spot assay, where the further right the dots are in each picture, the more diluted they are, it can be seen that 30 °C prevents most *orc1-161* growth, caused by activation of the *orc1-161* mutation, while the *cdc6-1* mutation is not activated until 37 °C. WT = Wild-type (T7107).

The ability of the new strains, combining *orc1-161* or *cdc6-1* with M.EcoGII, to grow in YPAD at 23 °C in liquid cultures was investigated. Figure 4.41 shows that overall the solo presence of each of these genome alterations don't affect the growth, although there is some variation in the growth of *orc1-161* between isolates (panel a). However, when the *orc1-161* mutation is combined with the M.EcoGII gene slower growth (shown by an

increase in doubling time of approximately 30 minutes) is consistently observed, while the growth of strains with the *cdc6-1* mutation are not affected by the M.EcoGII gene. The slower growth of strains with both *orc1-161* and M.EcoGII suggests that these two genetic manipulations cause the cell too much stress together, as Orc1 is present throughout the cell cycle, while Cdc6 is only expressed in G1. However, the expression of M.EcoGII is repressed in YPAD (figures 4.33 and 4.36) and an inducible M.EcoGII strain has been shown to not have reduced growth in YPAD 4.35. Conversely, the *orc1-161* strain showed large variation between replicates in figure 4.41a, with two replicates having a longer doubling time than the replicates for the combination of *orc1-161* and M.EcoGII. Therefore any reduced growth is most likely primarily due to the *orc1-161* mutation.

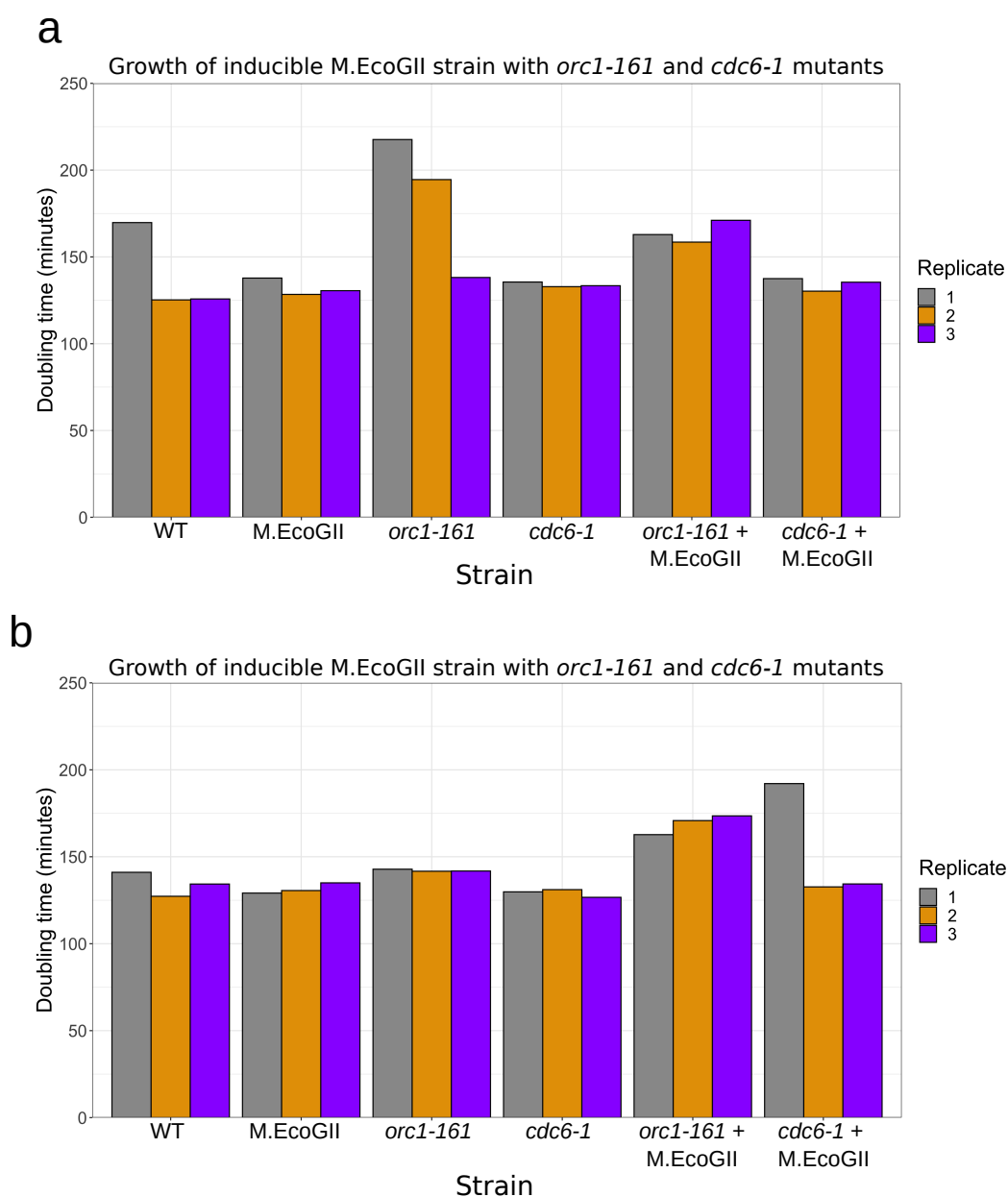


Fig. 4.41 Doubling times of replication protein mutant and inducible M.EcoGII strains

Yeast were grown in YPAD at 23 °C. The *orc1-161* and *cdc6-1* mutations are temperature sensitive, thus should not be activated at 23 °C. **a)** and **b)** were two growth curves performed on different days, each with 3 replicates. The isolates in the two growth curves are independent of each other, each freshly picked from a plate streaked from the -80 °C glycerol stock. For **a)** the wild-type (WT) strain was T7107, while for **b)** it was AUY116. These two strains are both W303 background, but of opposite mating types.

4.8.3 Cell cycle progression of strains with replication protein mutations and inducible M.EcoGII

The replication protein mutations should prevent the respective proteins (Orc1 and Cdc6) from binding DNA at high temperatures, thus prevent the cells from performing DNA replication due to incomplete origin licensing, leading to a cell cycle arrest in G1 phase. To identify the efficiency of this cell cycle arrest, flow cytometry was performed. The proteins should be able to bind the DNA at 23 °C as these strains predominantly show similar growth to wild-type strains at this temperature (figure 4.41). However, growth was impacted at 30 and 37 °C (figure 4.40) therefore these temperatures were investigated. Each strain was grown at 23 °C and then split in half with one culture transferred to 30 °C and the other to 37 °C, with samples taken every half an hour post temperature shift. Figure 4.42 shows that the inducible M.EcoGII strain (HJ006) is mostly not affected by the temperature shift to 30 °C and still appears relatively normal but with a higher G1 peak at 37 °C. Conversely, the *orc1-161* and *cdc6-1* mutations, either on their own or in combination with M.EcoGII, show distorted cell cycle progression, with an accumulation of cells in G1 at 37 °C. None of the strains manage to completely eliminate the G2 peak, however for all four strains containing a replication protein mutation the collection of cells in G1 is more effective at 37 °C than 30 °C. This suggests that at 37 °C the majority of cells have an inability for either Orc1 or Cdc6 to bind the DNA. The cells in G2 may be those where sufficient Orc1 or Cdc6 has bound to allow replication at a selection of origins. The doubling time of these strains was approximately two hours therefore three hours should have allowed time for all cells to enter G1, particularly given growth rates typically increase for yeast at 30 °C and 37 °C.

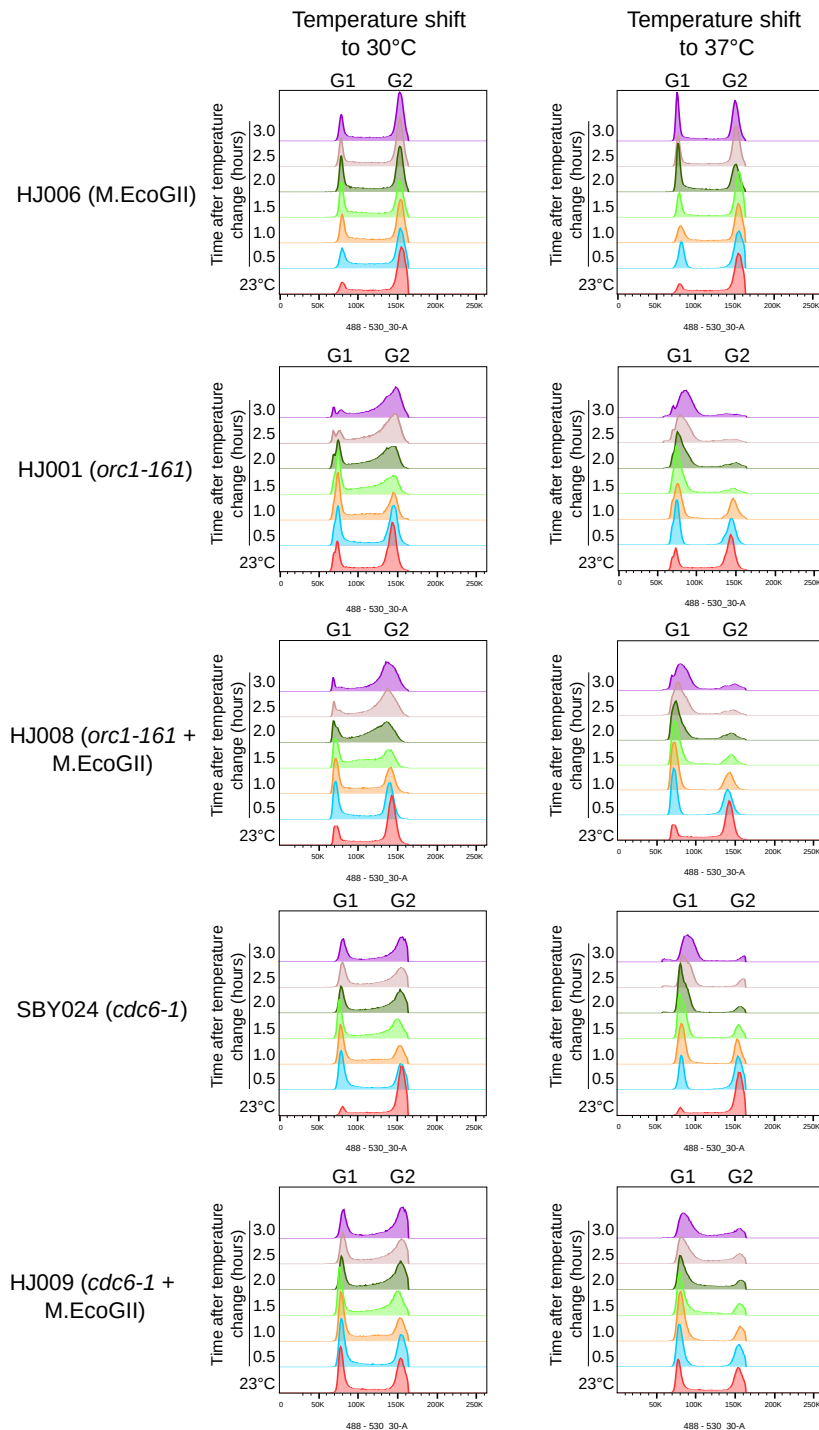


Fig. 4.42 Flow cytometry of replication protein mutant and inducible M.EcoGII strains

Yeast were grown in YPAD at 23 °C to OD 0.3, then the cultures were moved to 30 or 37 °C. Samples were taken every half an hour for 3 hours. The *orc1-161* and *cdc6-1* mutations are temperature sensitive and required for cell cycle progression, thus showed cell cycle arrest at higher temperatures. Conversely, inducible M.EcoGII is not temperature sensitive and is not required for cell cycle progression.

4.8.4 Mass spectrometry of strains with replication protein mutations and inducible M.EcoGII

To test the functionality of the inducible M.EcoGII in these new strains, mass spectrometry was performed. Figure 4.43 shows a substantial difference between the two strains, with the *orc1-161* inducible M.EcoGII strain showing very low levels of methylation (panel a). For this strain, methylation increased slightly throughout the timecourse however it appears that the production of M.EcoGII, and subsequent DNA methylation by this enzyme, is severely restricted. This corresponds to the slightly slower growth of this strain in YPAD (figure 4.41) and the observation that during this experiment the strain grew particularly slowly in raffinose, with a doubling time of more than six hours. Conversely, the *cdc6-1* and inducible M.EcoGII strain showed similar levels of methylation to those observed in wild-type strains with an inducible M.EcoGII (figures 4.43b and 4.36).

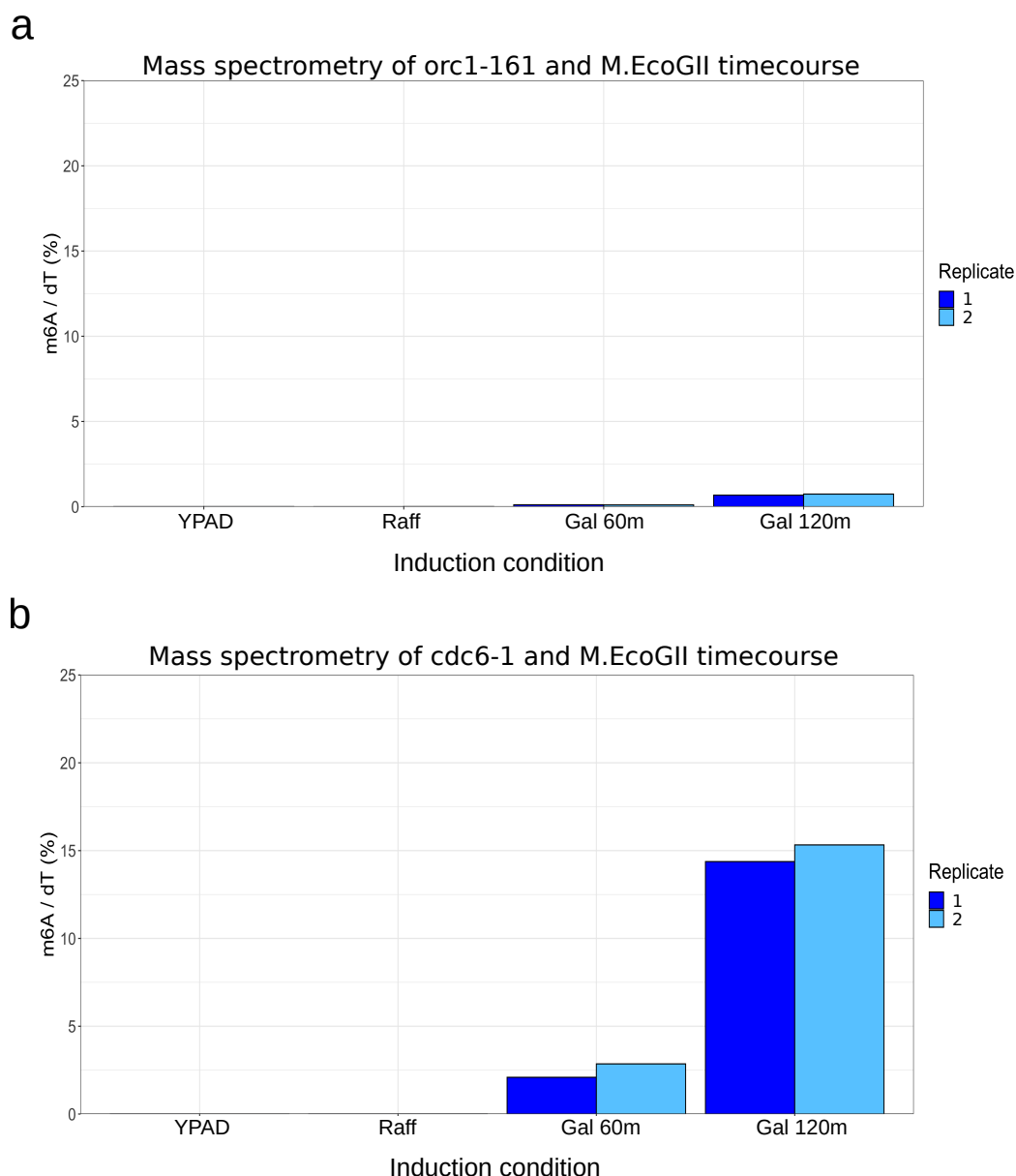


Fig. 4.43 Mass spectrometry of replication protein mutant and inducible M.EcoGII strains

Yeast were grown in YPAD at 23 °C to OD 0.1 then pelleted and resuspended in 2% raffinose for 2 doublings (until around OD₆₀₀ 0.4), at which point galactose was added to a final concentration of 2%. This induced the expression of the M.EcoGII under the *GAL1* promoter. DNA was extracted from the new strains prior to induction (the raffinose condition) and at 60 and 120 minutes post-galactose induction to verify the methyltransferase ability of the M.EcoGII in each strain. The two replicates were from different isolates when streaked directly from the -80 °C glycerol stock. Sample preparation performed by Heather Jeffery, HPLC mass spectrometry performed by Paolo Spingardi.

4.9 Inducible M.EcoGII *S. cerevisiae* strain arrested in G1 phase

4.9.1 Experimental procedure and cell cycle analysis

To look at DNA replication initiation in more detail the number of *S. cerevisiae* cells in G1 phase of the cell cycle, when origin licensing occurs, was increased. For this experiment *S. cerevisiae* cells with an inducible M.EcoGII were grown in YPAD to mid-log phase, at which point they were pelleted, washed with water and resuspended in 2% raffinose for a few doubling times. Then α -factor was added at regular intervals and the cells visually inspected by light microscopy (section 2.7.12). Galactose was added to a final concentration of 2% at 430 minutes after the initial α -factor treatment, when the cells appeared to be predominantly arrested in G1 phase. In G1 phase, *S. cerevisiae* cells are expected to ‘shmoo’ (as seen in figure 4.44), viewed as the formation of an elongation protruding from the main circular cell. This also means that they would not be budded, with the formation of two distinct cells, as this would indicate that the cells were undergoing cell division. Final samples for DNA analysis by nanopore sequencing were taken at 60 and 90 minutes post-galactose induction.

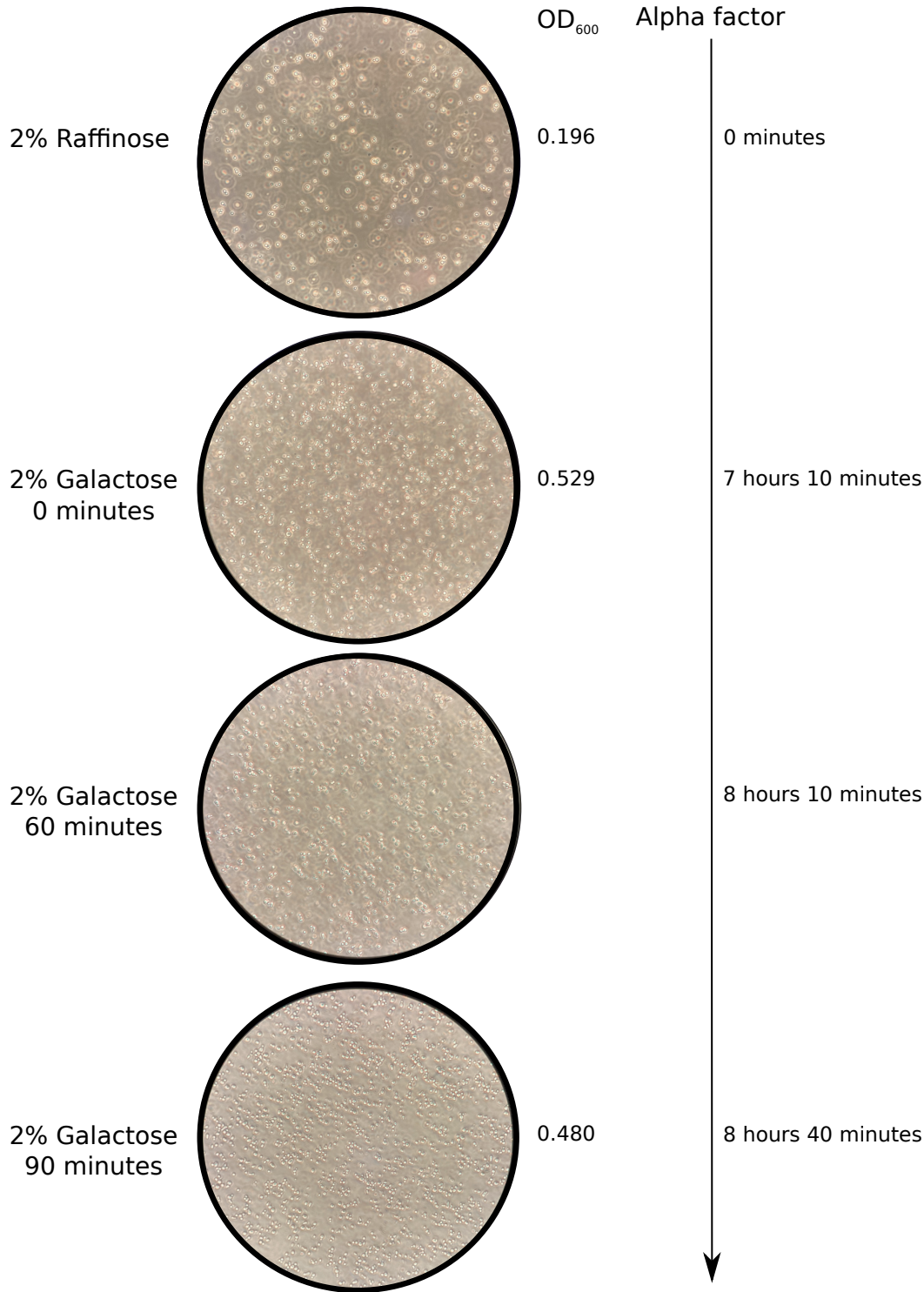


Fig. 4.44 Light microscopy images of *S. cerevisiae* G1 α -factor arrest

S. cerevisiae were grown in YPAD, pelleted and washed and grown in 2% raffinose. α -factor was added at OD₆₀₀ 0.2, and galactose was added to a final concentration of 2% when the cells reached a G1 arrest (7 hours 10 minutes after α -factor addition). This induced the expression of the M.EcoGII under the *GAL1* promoter. DNA was extracted at 60 and 90 minutes post-galactose addition. The OD₆₀₀ measurement for 0 minutes 2% galactose was taken immediately before addition of galactose. The addition of galactose increases the total volume, therefore reducing the OD₆₀₀ for the 90 minutes measurement. Images taken through the eyepiece of a light microscope.

4.9.2 Nanopore sequencing of inducible M.EcoGII *S. cerevisiae* strain arrested in G1 phase

Nanopore sequencing analysis was performed as in section 4.5.3, with Guppy version 4.4.0 used for basecalling. The two timepoints of 60 and 90 minutes post-galactose induction were selected as in figure 4.36b these two timepoints had a substantially different total level of methylation, both of which were clearly above the low levels previously detected, thus suggesting the mass spectrometry had detected true 6mA rather than background signal. In addition both timepoints were below the saturation point which potentially appeared by 120 minutes, at which point the information on the chromatin structure may be partially lost as the methyltransferase may have succeeded in methylating some previously accessible DNA. Nanopore sequencing produced a high coverage of 1,350x and 2,019x for the 60 and 90 minute timepoints respectively (table 4.7).

Condition	Reads	Mapped reads	Mapping quality > 0	Mean read length (bases)	Coverage*
Galactose 60 minutes	8,983,016	5,339,415	4,747,963	3,188	1,350
Galactose 90 minutes	11,734,362	7,371,551	6,590,682	3,433	2,019

Table 4.7 EcoGII *in vivo* nanopore sequencing overview of G1 arrest experiment

Two MinION flow cells run, one per sample, on 6th November 2020. Mapped reads are the primary alignments that have mapped to the sacCer3 reference genome (excluding mitochondrial genome). Mapping quality was only assessed on reads which were primary alignments. Mean read length is calculated using the read sequence from the bam file from primary alignments with a mapping quality > 0. Mean read length rounded to nearest integer. Coverage calculated by reads with mapping quality > 0 multiplied by mean read length, divided by genome size of 11,209,039 bases excluding the mitochondrial DNA. Coverage rounded to nearest integer. * Coverage is genomic coverage, therefore strand-specific coverage would be this coverage divided by 2.

As previously discussed, the level of methylation for each position is generated during basecalling and is stored as a value (0-255) which is subsequently normalised to 0-100. Figure 4.45 shows that across 100,000 adenines and 100,000 cytosines, methylation (classified on a 70% threshold) from induction in α -factor arrested *S. cerevisiae* cells is observed at an equal level for adenines and cytosines at 60 minutes with methylation at just over 1% of all adenines or cytosines analysed. In the 90 minute timepoint the level of adenine methylation has slightly increased while the level of cytosine methylation has decreased, suggesting that some ‘true’ adenine methylation is present. Mass spectrometry was not performed for these samples however these methylation levels correspond to the region of false positive methylation observed in figure 4.28. These α -arrested cultures have shown a lower level of methylation than the same timepoints in the asynchronous M.EcoGII induced experiments (figure 4.37). The reduction in methylation meant that the chromatin structure was not analysed in these samples.

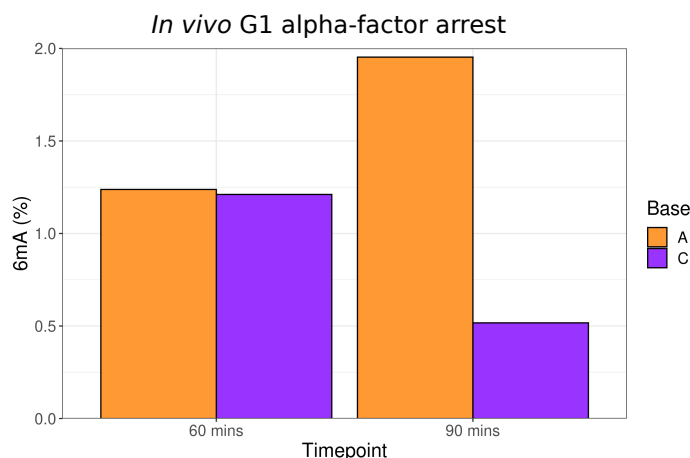


Fig. 4.45 Overall methylation detected in nanopore sequencing of G1 arrested cells with an inducible M.EcoGII

S. cerevisiae with an inducible M.EcoGII was arrested in G1 with α -factor treatment, and subsequently the expression of M.EcoGII was induced with 2% galactose. Nanopore sequencing data was basecalled with Guppy v4.4.0 and mapped to the sacCer3 reference sequence, excluding the mitochondrial genome. 100,000 adenines and 100,000 cytosines were analysed in any sequence context from a selection of 10,000 random primary alignments mapping to the forward strand with a mapping quality greater than zero. Each site was classified as methylated or not based on a 70% threshold.

4.10 Detecting *S. cerevisiae* chromatin structure with adenine methylation performed *in vivo* by an inducible M.EcoGII

Finally, the ability of induced M.EcoGII adenine methylation to identify chromatin structure was investigated in a variety of ways using the nanopore sequencing data of the asynchronous M.EcoGII timecourse with samples taken every fifteen minutes. Firstly, by investigating methylation within nucleosome binding sites as nucleosomes are thought to protect DNA from methylation [212]. Secondly, by analysing methylation patterns around transcription start sites as transcription factors predominantly bind in regions of accessible DNA [218]. Thirdly, by performing a meta-analysis of methylation around two sets of DNA replication origins [7, 76]. Fourthly, by performing a cumulative count of methylation across regions containing a DNA replication origin. In this scenario, the presence of nucleosomes may be visible as a plateau, due to protection of the nucleosome-bound DNA from methylation. Another approach that was used to investigate replication origins was the clustering of origins by chromatin structure using the same clusters observed by Berbenetz *et al* [76] to see how well the data recreates figure 1.2. For the analysis mentioned so far, all forward strand primary alignments with a mapping quality greater than zero were analysed and every adenine in every analysed read was classified as methylated or not based on a 70% threshold. Adenines were then grouped by distance to a region of interest (nucleosome dyads, transcription start sites or replication origins) and the percentage of adenines methylated at each position calculated (see Appendix 1). The final analysis was single-molecule visualisation across a replication origin where the aggregation step was omitted.

The majority of DNA is nucleosome-bound [7] therefore, as in figure 4.25, nucleosomes were used to assess whether inducible M.EcoGII can be used to detect chromatin structure.

Nucleosome dyads are the central position within the symmetrical nucleosome, therefore upstream and downstream distances were combined. Figure 4.46a shows that most adenines are within nucleosome-bound DNA. Figure 4.46b shows that adenine methylation increases throughout the *in vivo* timecourse. The nucleosome-bound DNA is not completely protected from methylation, as previously suggested by figure 4.11. However, the methylation increases from the dyad towards the edge of the nucleosome, suggesting that the centre of the nucleosome provides some protection while the DNA at the nucleosome edge is more accessible.

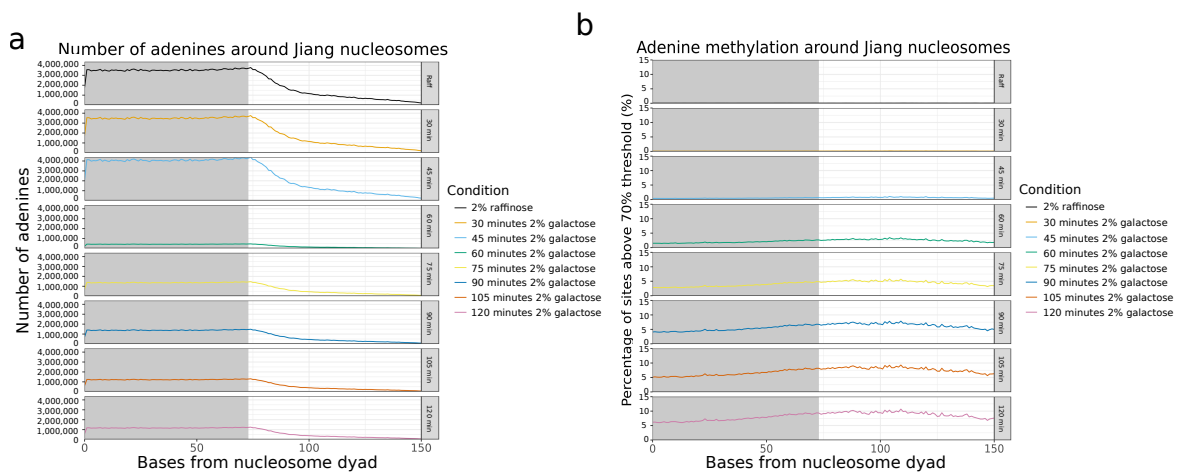


Fig. 4.46 *In vivo* M.EcoGII adenine methylation around nucleosome dyads

Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence excluding mitochondrial DNA. Consensus nucleosome dyad positions are from Jiang *et al* [89]. For this meta-analysis the adenines around each nucleosome dyad were analysed in all forward primary alignments with a mapping quality above zero. Each adenine in each read was classified as methylated or not based on a 70% threshold. Adenines were then aggregated based on distance from the nucleosome dyad and the percentage of methylated sites at each position was calculated. As nucleosomes are symmetrical around the dyad, the upstream and downstream distances were combined. Grey box represents nucleosome-bound DNA.

A typical way of validating chromatin accessibility assays is by analysing the accessibility around transcription start sites [100, 116, 194]. At these sites regions of accessibility are important for transcription factor binding [218]. Figure 4.47 shows a meta-analysis of adenine methylation around 5,979 transcription start sites [219] through a two hour timecourse of M.EcoGII induction. In figure 4.47b there is a pattern in the methylation around these

transcription start sites, starting to be visible from one hour post-galactose induction of M.EcoGII. This pattern is not replicated in the number of adenines around the transcription start sites (figure 4.47a) therefore is not confounded by this. Interestingly, this pattern appears to be reasonably symmetrical, with lower methylation indicating reduced accessibility at the actual transcription start sites, with regions of accessibility on either side. Other studies have shown increased accessibility upstream of the transcription start site without a corresponding increase downstream [116, 194] however this was not observed with the current approach. Additionally, figures 4.47c and d indicate that the observed pattern from all transcription start sites is clearer in the sites not containing a TATA motif. However, this may also be caused by the larger number of TSS without a TATA motif (4,981) compared to the 998 sites with one. Overall, the transcription start site meta-analysis shows that these sites contain conserved regions of accessible DNA which can be detected by this new chromatin accessibility method.

To assess methylation around replication origins, which are typically regions of accessible DNA, a similar approach was used as for the nucleosomes and transcription start sites. Two origin datasets were used, from Eaton *et al* [7] and Berbenetz *et al* [76], of which the latter is predominantly based on Nieduszynski *et al* [5]. Most origins are common between the two datasets, though different ACS sites may be reported. The size of the nanopore sequencing datasets of the M.EcoGII timecourse varied (figures 4.48a and 4.49a). The methylation analysis of both origin datasets (figures 4.48b and 4.49b) show that the overall methylation levels increase throughout the timecourse. As the methylation increases, the variation in methylation also appears to increase, with smaller changes around the ACS in the Eaton dataset, with peaks of methylation further away, with a potential nucleosome-binding site with reduced methylation around 150-220 bases downstream of the ACS. The Berbenetz origins show greater variation closer to the ACS, with a reduction at the ACS, potentially indicating ORC binding, with some peaks around 30-60 bases downstream of the ACS, which may indicate accessible DNA prior to a nucleosome binding site. The peak around 160 bases

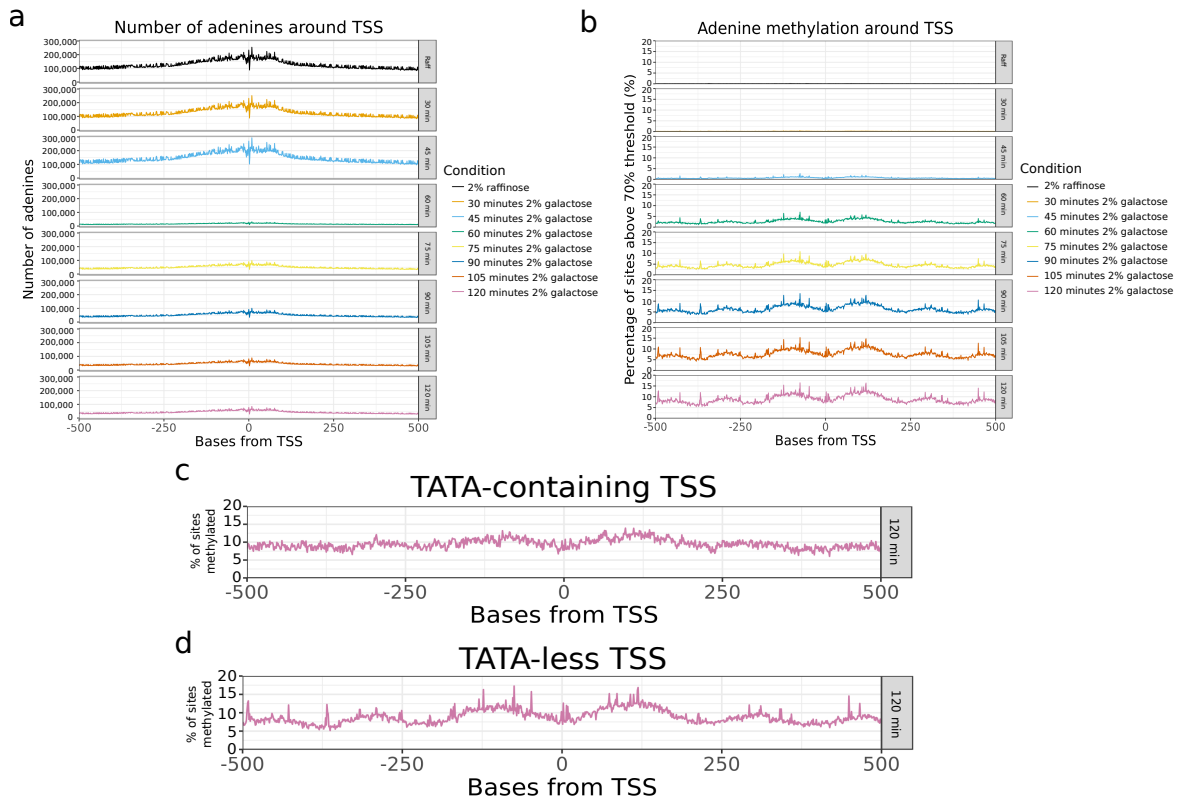


Fig. 4.47 *In vivo* M.EcoGII adenine methylation around transcription start sites

Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence excluding mitochondrial DNA. For this meta-analysis the adenines around each transcription start site (TSS) were analysed in all forward primary alignments with a mapping quality above zero. Each adenine in each read was classified as methylated or not based on a 70% threshold. Adenines were then aggregated based on distance from the TSS and the percentage of methylated sites at each position was calculated. The median TSS from Qiu *et al* [219] were used. For all panels of this figure analysis was performed relative to the strand on which the TSS is located. **a)** Number of adenines around all 5,979 transcription start sites for a range of samples across the M.EcoGII timecourse. **b)** Adenine methylation around all 5,979 transcription start sites for a range of samples across the M.EcoGII timecourse. **c)** Adenine methylation around 998 TATA-containing transcription start sites for 120 minutes post-galactose induction of M.EcoGII. **d)** Adenine methylation around 4,981 TATA-less transcription start sites for 120 minutes post-galactose induction of M.EcoGII.

downstream may indicate a linker between two nucleosomes. As this is a meta-analysis and nucleosome positions vary between origins, a 147 bp region of a nucleosome would be unlikely to be seen, thus shorter regions may indicate nucleosome binding.

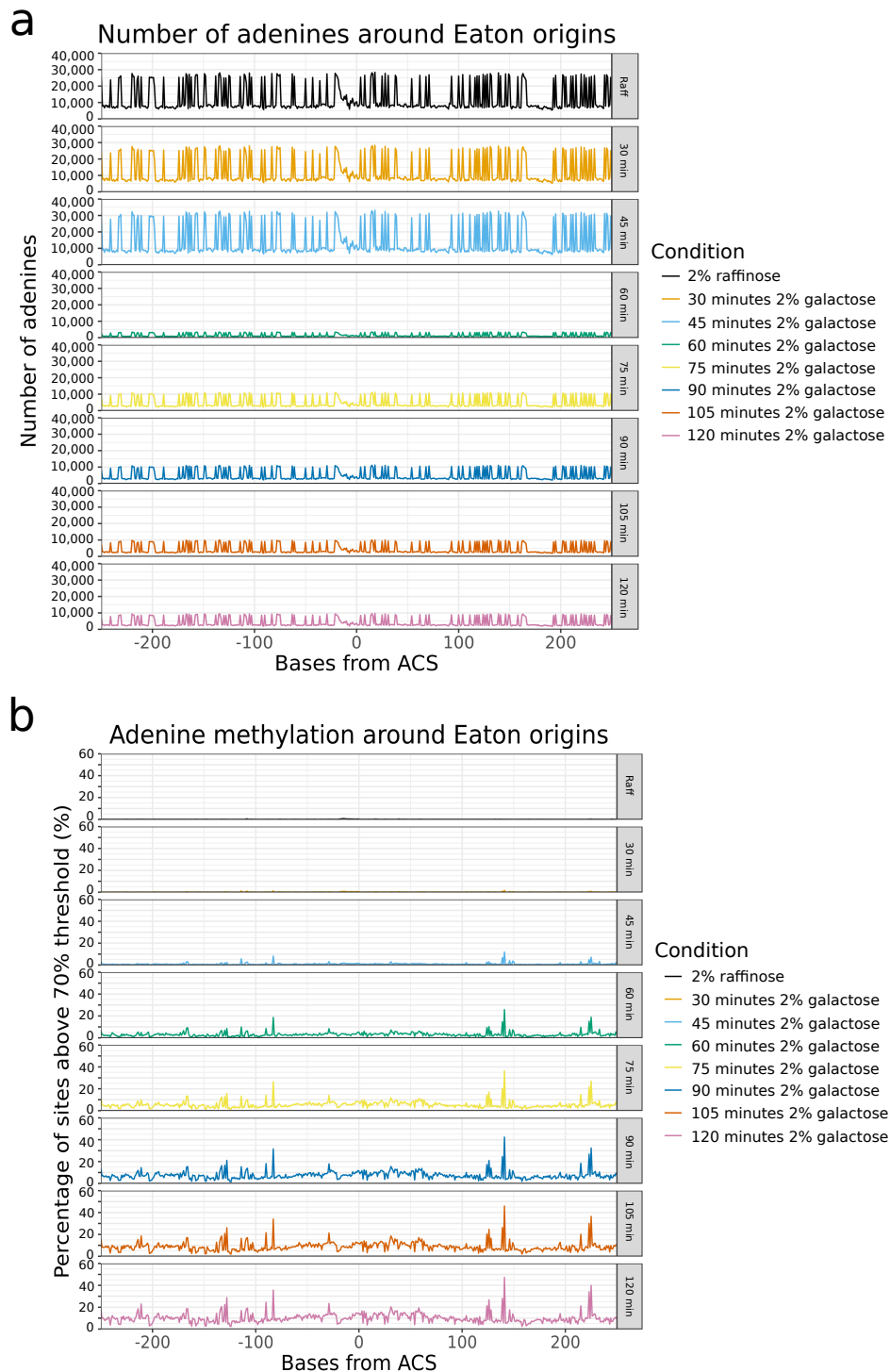


Fig. 4.48 *In vivo* M.EcoGII adenine methylation around Eaton replication origins

Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence excluding mitochondrial DNA. Reads mapping to ACS locations representing DNA replication origins from Eaton *et al* [7] were analysed. For the meta-analysis shown here, the adenines around each of the ACS sites were analysed in all forward primary alignments with a mapping quality above zero. Each adenine in each read was classified as methylated or not based on a 70% threshold. Adenines were then aggregated based on distance from the ACS (-250 to 250 bases) and the percentage of methylated sites at each position was calculated. The negative numbers represent upstream with the ACS oriented by the T-rich strand. **a)** The number of adenines analysed at each position +/- 250 bases around Eaton origins in each condition. **b)** The percentage of methylated sites that are above a 70% threshold at each position around Eaton origins.

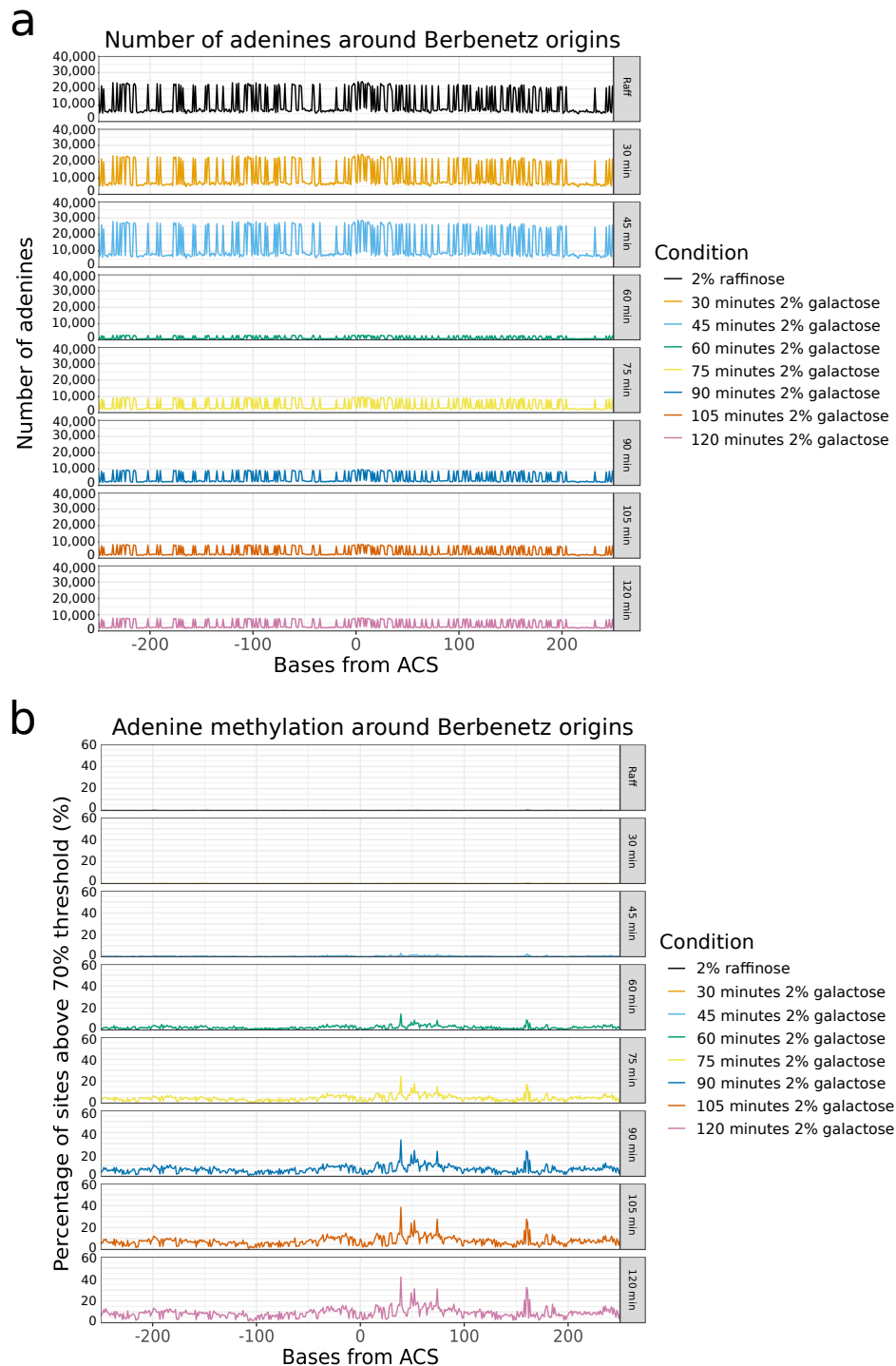


Fig. 4.49 *In vivo* M.EcoGII adenine methylation around Berbenetz replication origins

Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence excluding mitochondrial DNA. Reads mapping to ACS locations representing DNA replication origins from Berbenetz *et al* [76] were analysed. For the meta-analysis shown here, the adenines around each of the ACS sites were analysed in all forward primary alignments with a mapping quality above zero. Each adenine in each read was classified as methylated or not based on a 70% threshold. Adenines were then aggregated based on distance from the ACS (-250 to 250 bases) and the percentage of methylated sites at each position was calculated. The negative numbers represent upstream with the ACS oriented by the T-rich strand. **a)** The number of adenines analysed at each position +/- 250 bases around Berbenetz origins in each condition. **b)** The percentage of methylated sites that are above a 70% threshold at each position around Berbenetz origins.

For the cumulative sum approach the percentage of adenines methylated at each position were added across a 1 kb region centered on an ACS. Figure 4.50a-c shows three replication origins on chromosome four. Accessible DNA is represented by a steeper gradient (a large increase in methylation) while protein-bound regions are represented by a shallower gradient (a small increase in methylation). In figure 4.50a-c the nucleosome-bound DNA (grey boxes) generally have flatter lines, therefore lower methylation, than the regions of accessible DNA, indicating that nucleosomes partially protect DNA from methylation. The regions of accessible DNA contain a variety in the level of methylation, with figure 4.50a showing a region in the central accessible DNA that might be bound by another protein, shown by the flat line, followed by a sharp increase in methylation until the DNA reaches the next nucleosome. Figure 4.50b, the well studied *ARS1*, shows much less distinction between the regions, making it harder to determine where nucleosomes could be bound. *ARS1* partially overlaps the *TRP1* gene which may affect the accessibility of this region. Figure 4.50c shows that smaller regions of accessible DNA around an ACS can be highly methylated. To consider the ability of each time point to provide information on chromatin structure the cumulative data was normalised to 0-100% (figure 4.51). The timepoints with low overall methylation (2% raffinose and 30 minutes 2% galactose) show big steps in methylation therefore are not suitable for assessing chromatin structure. However, the timepoints from 75 minutes onwards, and sometimes including 45 and 60 minutes, group together across these regions. Therefore, although in figure 4.50 the longer incubation steps may have looked more suitable for chromatin structure analysis due to larger increases, 75-120 minutes produce similar information on the chromatin structure when normalised. Figure 4.52 shows that *ARS1513*, which was used to visualise single-molecule Dam methylation in nanopore sequencing data (figure 4.26), has increased methylation over the ACS compared to the nucleosomes on either side. This further supports the requirement for high frequency modification sites as this was not visible with the Dam methyltransferase data.

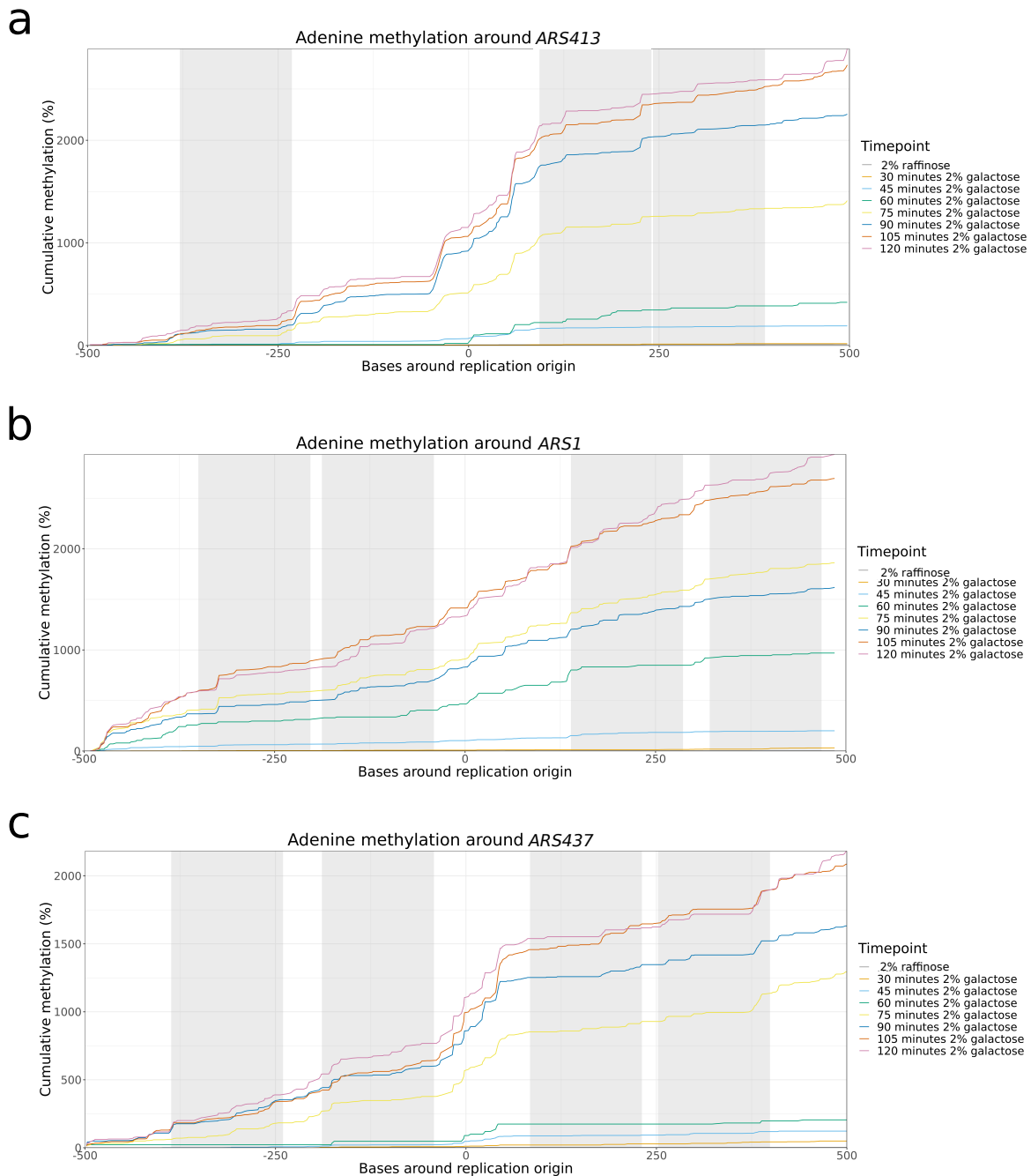


Fig. 4.50 Cumulative *in vivo* adenine methylation around three replication origins

Nanopore sequencing data was basecalled and mapped to the sacCer3 reference sequence excluding mitochondrial DNA. Reads mapping to ACS locations representing DNA replication origins from Eaton *et al* [7] were analysed. Three origins from chromosome four are shown here (a-c) and others can be viewed at https://github.com/HJeffery/Thesis_supplementary. For each ACS, all the forward primary alignments with a mapping quality above zero mapping to this region contribute to the results. Each adenine in each read was classified as methylated or not based on a 70% threshold. Then for each adenine around the ACS of interest, the percentage of methylated sites across the reads was calculated. The plots show a cumulative count of methylation, starting with zero at 500 bases upstream of the ACS. Therefore going across the origin left to right, the lines show the change in methylation across the region, with flat areas representing low methylation and steeper gradients representing higher methylation. The negative numbers represent upstream with the ACS oriented by the T-rich strand. Grey boxes represent consensus nucleosome positions [89].

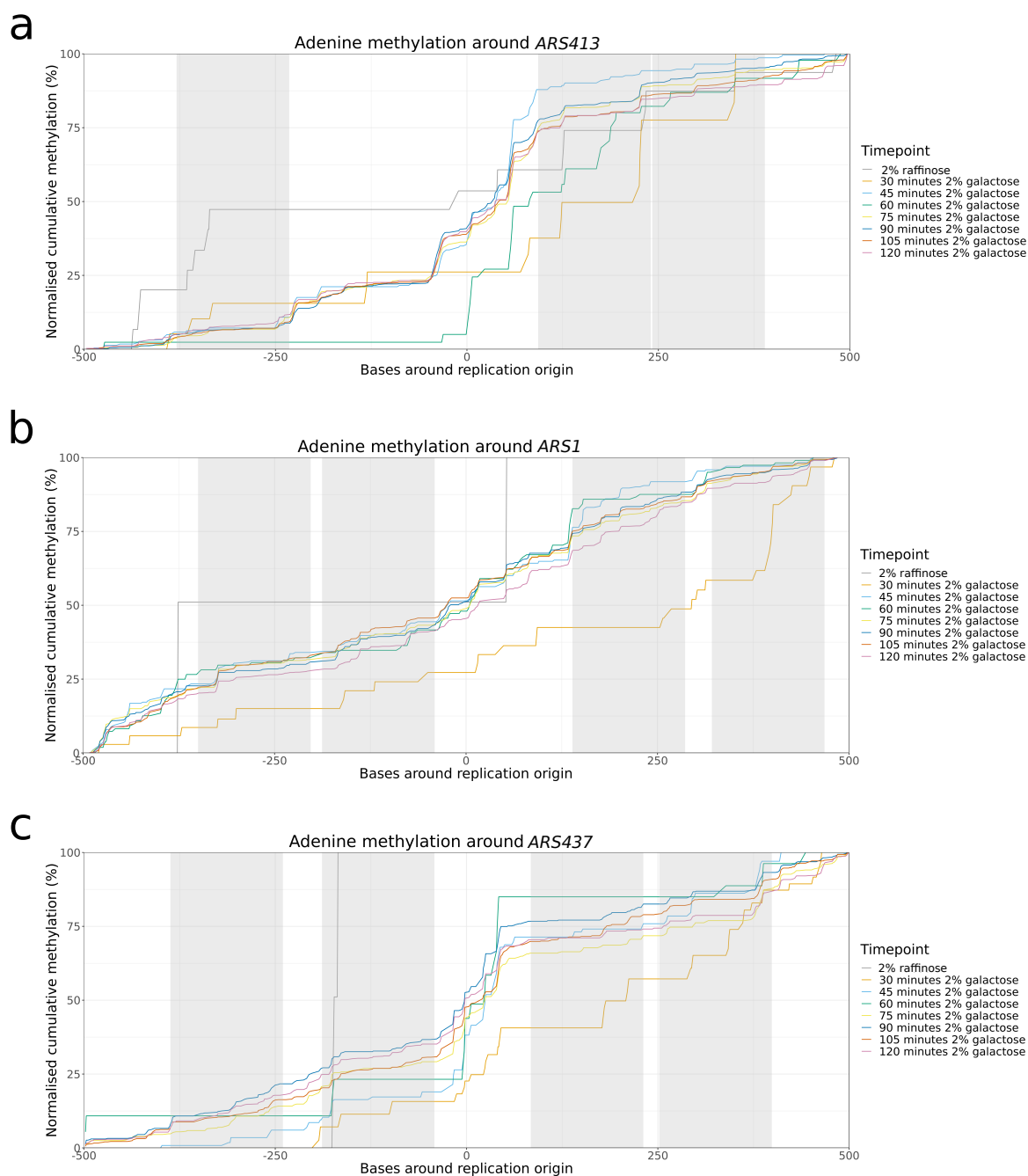


Fig. 4.51 Cumulative *in vivo* adenine methylation around three replication origins

The same data and analysis was performed as for figure 4.50. However, this time each time point has been normalised so that the cumulative range is from 0 to 100%. The negative numbers represent upstream with the ACS oriented by the T-rich strand. Grey boxes represent consensus nucleosome positions [89]. Three origins from chromosome four are shown here and others can be viewed at https://github.com/HJeffery/Thesis_supplementary.

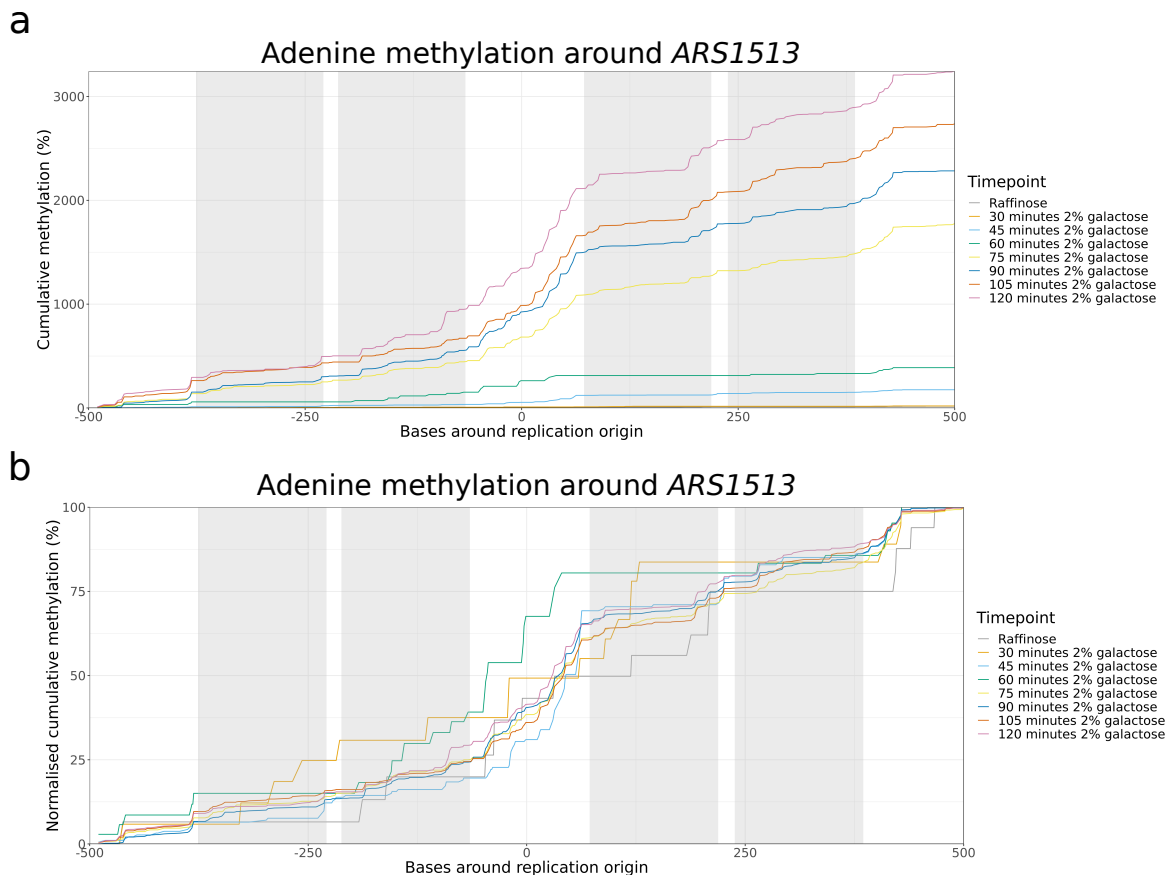


Fig. 4.52 Cumulative *in vivo* adenine methylation around *ARS1513*

Nanopore sequencing data was basecalled and mapped to the *sacCer3* reference sequence excluding mitochondrial DNA. Reads mapping to *ARS1513* from Eaton *et al* [7] were analysed. All the forward primary alignments with a mapping quality above zero mapping to this region contribute to the results. Each adenine in each read was classified as methylated or not based on a 70% threshold. Then for each adenine around the ACS of interest, the percentage of methylated sites across the reads was calculated. **a)** A cumulative count of methylation, starting with zero at 500 bases upstream of the ACS. Therefore going across the origin left to right, the lines show the change in methylation across the region, with flat areas representing low methylation and steeper gradients representing higher methylation. **b)** For the normalisation the cumulative range has been adjusted to 0 to 100%. The negative numbers represent upstream with the ACS oriented by the T-rich strand. Grey boxes represent consensus nucleosome positions [89].

For the clustering approach, 228 replication origins were split into four clusters based on the published Berbenetz *et al* study (figure 1.2) [76]. Figure 4.53a-d shows the percentage of methylated sites at each adenine around each ACS (one row per ACS). There may be slight patterns around the ACS in the clusters however there is nothing as clear as in figure 1.2.

Therefore, this new methylation nanopore sequencing chromatin accessibility assay requires further optimisation to be able to confidently detect chromatin structure.

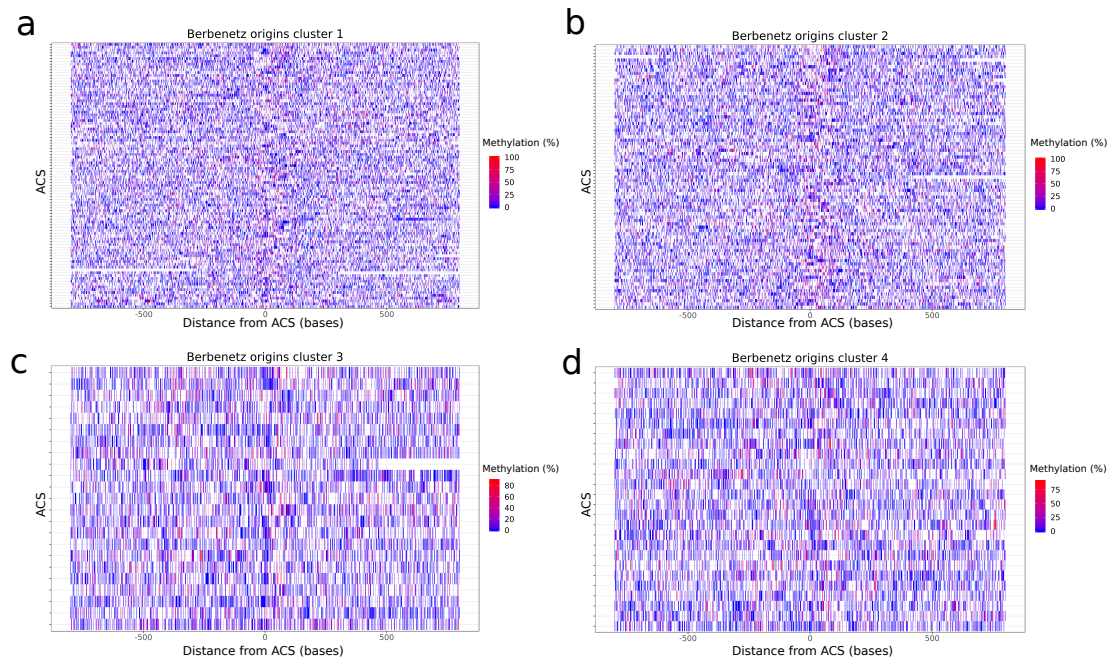


Fig. 4.53 *In vivo* adenine methylation around clustered replication origins

The same data is analysed as was shown in figure 4.49. This time the replication origins (ACS) have been split into four groups based on the clusters that were observed by Berbenetz *et al* [76] (figure 1.2). Each line in the heatmaps represents an ACS. Each adenine in each read was classified as methylated or not based on a 70% threshold. The percentage of adenines methylated at each distance from the ACS where an adenine is present were then calculated. The negative numbers represent upstream with the ACS oriented by the T-rich strand.

For the single-molecule analysis, M.EcoGII methylation provides much more information across *ARS1513* than Dam methylation did (compare figures 4.26 and 4.54). Most adenines around *ARS1513* are unmethylated in the initial conditions (figure 4.54a-d) but later in the timecourse (e-h) higher methylation is observed, particularly around 30-60 bases downstream of the ACS. This indicates that there is a non-nucleosome protein binding event over the ACS and extending upstream. Despite the high occurrence of adenines, small regions with no information can exist, for example around 150 bp downstream of the ACS. Single-molecule analysis of the three origins from chromosome four analysed by the cumulative sum approach

(figures 4.50 and 4.51) shows that methylation in accessible regions is increased with longer time in 2% galactose (figure 4.55). Additionally, as shown by the cumulative sum analysis, the single-molecule analysis also shows that regions of high methylation are localised to specific parts of the accessible regions rather than the entirety, indicating the presence of other protein binding events in these regions. The sections of the accessible regions with higher methylation in figure 4.55 are more tightly grouped for *ARS413* and *ARS437* while they are more evenly spaced in *ARS1* which supports the lack of definition between nucleosome-bound and accessible DNA observed in figure 4.50. In summary, this ensemble data shows regions to be generally accessible or not. However, at high methylation levels methylation within the nucleosome-bound DNA is observed. Therefore, this data may not be particularly informative if a single read was investigated in isolation.

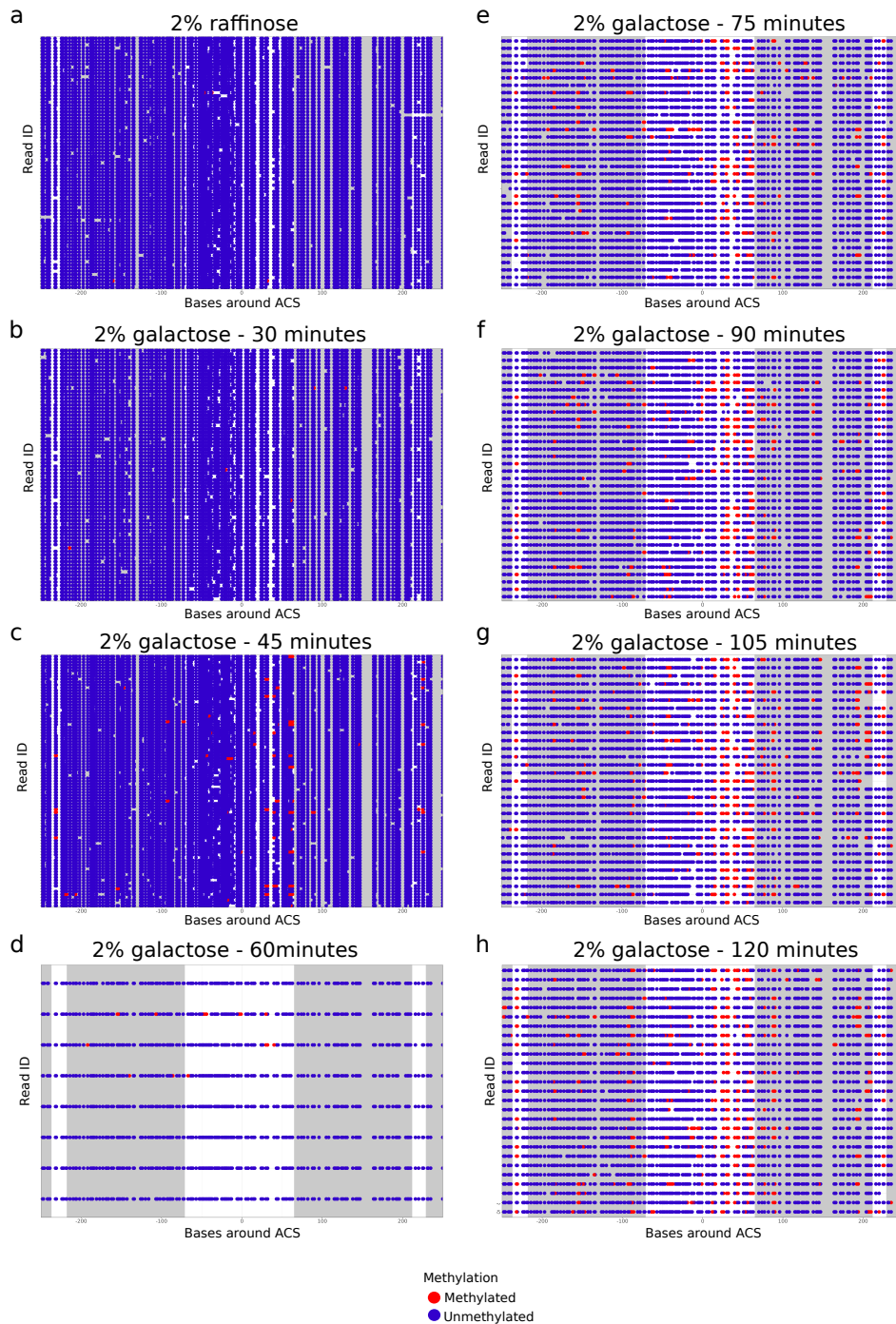


Fig. 4.54 M.EcoGII adenine methylation around ARS1513

The nanopore sequencing data has been analysed in the same way as for figure 4.48 minus the aggregation step. Instead each read is plotted as a line, with each adenine shown as methylated (red) or not (blue) based on a binary 70% threshold. Grey boxes represent nucleosomes (based on consensus positions [89]). Reads are oriented to the T-rich strand of the ACS. Reads shown contain at least 90% of the adenines within this region. Unlike with the Dam methyltransferase data (figure 4.26) the high frequency of adenines means that some are likely to be in deletions therefore filtering for 100% of sites produced very low numbers of reads for each sample and was not appropriate.

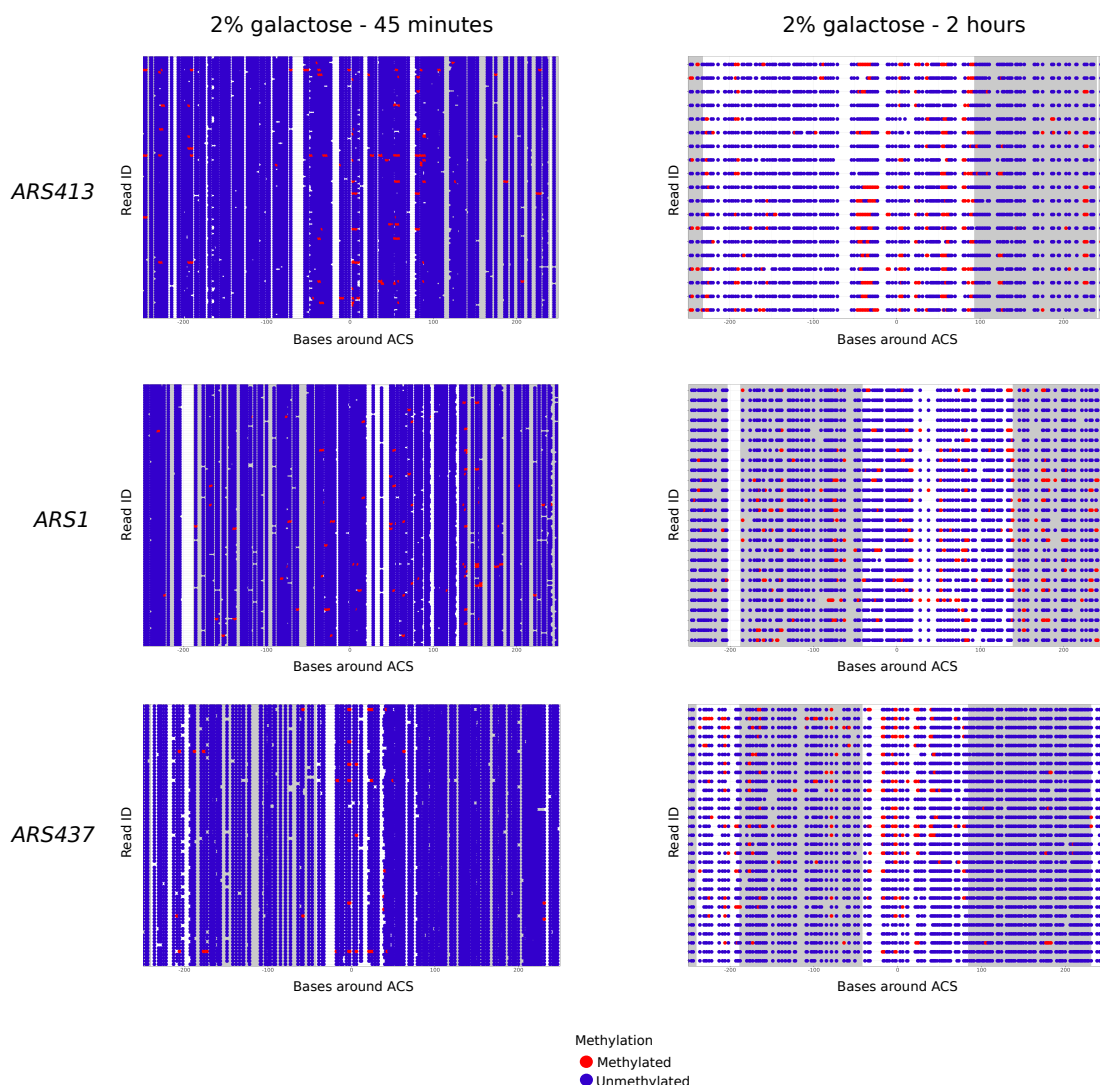


Fig. 4.55 M.EcoGII adenine methylation around *ARS413*, *ARS1* and *ARS437*

The nanopore sequencing data has been analysed and plotted in the same way as for figure 4.54. Each read is plotted as a line, with each adenine shown as methylated (red) or not (blue) based on a binary 70% threshold. Grey boxes represent nucleosomes (based on consensus positions [89]). Reads are oriented to the T-rich strand of the ACS. Reads shown contain at least 90% of the adenines within this region.

4.11 Discussion

In this chapter methylation of adenine, and occasionally cytosine, has been performed to investigate different methods of detection and potential methods for methylating DNA while preserving biological chromatin structures.

Firstly, the methylation abilities of M.EcoGII were characterised. This involved testing the adenine methylation sensitivity of a selection of restriction enzymes (figure (figure 4.1). This confirmed that PstI-HF and SphI-HF were sensitive to adenine methylation, meaning that they are unable to digest DNA where the adenines in their recognition site are methylated. This information was used to quantitatively measure the level of methylation at a specific recognition site by combining the enzyme digests with a Droplet Digital PCR readout. It was identified that close to 100% methylation could be achieved in both plasmid and genomic DNA. Therefore, these enzymes provide a mechanism for confirming methylation at specific loci.

To apply methylation to the investigation of the *S. cerevisiae* chromatin structure, an *in vitro* approach was attempted. For this, either chromatin needed to be extracted while maintaining its biological state, or the methyltransferase needed to enter the nuclei, requiring destruction of the cell wall and permeabilisation of the cell membrane. For the main chromatin experiments, with validation by western blotting or MNase digests, the chromatin structure was preserved by formaldehyde crosslinking as in FAIRE-seq [95] and ChIP-seq [93]. In yeast 1% formaldehyde immediately stops cell growth and results in 50% lethality within 100 seconds, with 99% lethality achieved by 360 seconds [95]. This suggests that formaldehyde crosslinking can be used to preserve the chromatin structure without substantially perturbing it. ddPCR analysis of formaldehyde-crosslinked chromatin did not show any adenine methylation despite methylation of the pUC19 spiked-in DNA. Using a spheroplast-based approach, cytosine methylation of chromatin with M.CviPI showed very low levels of methylation (figure 4.17) suggesting that a much higher methyltransferase concentration, similar to that used by SMAC-seq [116], would be required to methylate chromatin. SMAC-seq treated isolated nuclei obtained from the spheroplasts rather than the spheroplasts themselves, providing more direct access to the DNA. However, M.CviPI DNA methylation at the spheroplast stage has been performed successfully [220] therefore

there were indications it could work as the protocol in this thesis was largely based on this prior publication. Methylation of chromatin could be repeated with a much higher amount of methyltransferase, or with an additional nuclei isolation step.

Variability in the mass spectrometry results was observed for the samples in the *in vitro* dilution series with very low levels of methylation. This aligns with speculation that HPLC tandem mass spectrometry can be inaccurate at low levels as it is susceptible to potential contamination from a variety of sources including RNA and bacterial contamination of reagents [180, 181]. At this point in time the *in vivo* approach was looking more promising so this was pursued.

An *in vivo* approach was theoretically more likely to capture the true biological chromatin state than an *in vitro* method as the DNA is being modified while the cells are still alive. Conversely, the protein-DNA binding events are dynamic and transient interactions are less likely to be seen. DNA-protein interactions involved in pre-RC formation predominantly persist for the duration of G1 therefore are likely to be detected. The *in vivo* approach involved transforming wild-type *S. cerevisiae* with the M.EcoGII gene under a *GALI* promoter at the *URA3* locus. Integration of the gene into genomic DNA was shown to be successful (figures 4.32 and 4.34), and the M.EcoGII protein was confirmed to be both expressed (figure 4.33) and functional (figure 4.36). This provided a model system in which biological insights into chromatin structure could theoretically be gained although reasonably long incubation periods were required to achieve noticeable levels of DNA methylation.

As *S. cerevisiae* doesn't normally express functional methyltransferases the growth of the strain with the inducible M.EcoGII was monitored. In figure 4.35 it was observed that the growth of *S. cerevisiae* strains are minimally affected by low levels of M.EcoGII in 2% raffinose. A greater reduction in growth rate may have been seen if M.EcoGII was more

highly expressed, as M.EcoGII expression has been associated with slower growth in *E. coli* [150]. M.EcoGII is not expressed in *E. coli* during normal growth [183] and may only be expressed as the prophage is ready to excise and initiate a new infection [150]. Therefore, *E. coli* may have developed a particular sensitivity to the presence of even low levels of M.EcoGII and the resulting adenine methylation. Also, adenine methylation has known biological roles in *E. coli* whereas there are no known roles of adenine methylation in *S. cerevisiae*.

Slightly slower growth was also observed for *S. cerevisiae* strains with an *orc1-161* mutation while growth was not affected by the *cdc6-1* mutation (figure 4.41). In *S. cerevisiae* ORC binds throughout the cell cycle [13] whereas Cdc6 only binds in G1 [11]. Therefore ORC can have a larger effect on the entire cell cycle. The binding of Cdc6 potentially stabilises ORC DNA-binding [11]. Therefore, ORC may be relatively unstable throughout the rest of the cell cycle and in the absence of Cdc6 ORC-DNA binding may also be more dynamic during G1.

Nanopore sequencing experiments showed that 6mA methylation by M.EcoGII could be detected in a variety of contexts (figures 4.29 and 4.30) due to recent technological developments. This opens up many opportunities for looking at modifications at substantially higher resolutions than was previously possible. Furthermore, figure 4.28 shows that the total level of methylation detected can vary considerably depending on which threshold is used. Analysis of nanopore sequencing data identified that nucleosomes partially protected DNA from methylation (figures 4.25 and 4.46). However, nucleosomes did not completely protect the DNA from methylation as previously suggested [212] therefore the boundary between nucleosome-bound and accessible DNA was less clear (figure 4.46). The gradual gradient rather than distinct boundary makes it more challenging to assign regions of a DNA molecule to be either nucleosome-bound or accessible. This might also explain why the

regions of low methylation in the transcription start site meta-analysis (figure 4.47) were smaller than the typical 147 bases of a nucleosome-binding site. In this analysis it is likely that the central reduction in methylation is due to transcription factor binding [218] although a collection of transcription factors would be required for the size of the methylation-depleted region. With regards to validating the method for the detection of chromatin structure, it is promising that accessible regions with high methylation were found on either side of this central reduction in methylation. This meta-analysis also highlights that despite changes in methylation between regions, greater than 5% methylation was observed at all positions around the transcription start sites. Therefore, background methylation of protein-binding sites is likely to be occurring although this may also be a result of the transcription start sites selected for analysis if they do not all behave in a similar fashion. Patterns of methylation, and thus chromatin structure, around replication origins were also observed (figures 4.48, 4.49 and 4.50). These were more challenging to interpret than the transcription start-site meta-analysis, which may be due to either a greater diversity of chromatin structures at these sites, or a result of the reduced number of sites (252 compared to 5,979). Distinct chromatin structures were not observed on a per-ACS basis as previously reported [76]. Single-molecule visualisation shows that when viewed together, nanopore sequencing reads can show general regions of chromatin accessibility (figure 4.54). It also highlights that, in some cases, only certain regions not covered by consensus nucleosome positions were accessible to M.EcoGII methylation, indicating either that the nucleosomes were in different positions, or that other proteins were bound. Therefore, there are indications of chromatin structure information in these nanopore sequencing datasets. However, the presence of low levels of methylation within protein-binding sites reduces the accuracy with which protein-binding events can be confidently detected.

Most of the *in vivo* experiments were performed in asynchronous cultures as the aim of these was to investigate the ability of nanopore sequencing to detect methylated adenine, and

to identify the level of DNA methylation that occurred at different time points post-induction of M.EcoGII. During the M.EcoGII induction period some cells would have performed DNA replication, creating new DNA. Therefore, it was impossible for M.EcoGII to methylate every single adenine as some may be very newly replicated prior to sample collection. Ongoing DNA replication also means that the ratio of methylated to unmethylated DNA in a particular cell at any point is not entirely due to the methylation ability of M.EcoGII. If DNA is replicated while M.EcoGII levels are high then there is a greater chance of the DNA being methylated prior to assembly of the chromatin structure, or during replication while single-stranded DNA is accessible (as M.EcoGII can also methylate single-stranded DNA [150]). This would likely have averaged out over the entire population and the mass spectrometry data did show an increase in DNA methylation throughout the M.EcoGII induction timecourse (figure 4.36). The presence of DNA replication indicates that slightly more methylation may be possible to achieve at the same time point in an arrested *S. cerevisiae* culture. However, when inducing M.EcoGII in G1-arrested cells, the level of methylation detected at the same timepoints was substantially lower than previously seen (figure 4.45 compared to figure 4.37). This may be due to a reduction of protein production in the arrested cells.

Overall, this chapter has shown that DNA methylation can be detected by different methods, nanopore sequencing can be used to detect methylated adenine in all sequence contexts, and an inducible M.EcoGII *S. cerevisiae* model can be used to provide some insights into chromatin structure.

Chapter 5

Discussion and Conclusion

5.1 Discussion

The principle aim of this thesis was to develop a new method for probing chromatin structure in *S. cerevisiae*. Focusing on the main concept of genomic footprinting, where DNA bound by proteins can be protected from modification, a variety of modifications, both chemical (chapter 3) and enzymatic (chapter 4), have been trialled. In addition, both *in vitro* and *in vivo* systems have been investigated in combination with DNA modification by methylation (chapter 4).

Genomic footprinting was first used in 1992 [13] where it was used to look at specific loci. More recent development of sequencing technologies that collect data on a genomic scale at base resolution, combined with their ability to detect modified bases without amplification, has facilitated the application of the footprinting concept on a much larger scale.

The genomic footprinting method developed in this thesis has shown that it is possible to get a single-molecule view of the *in vivo* chromatin structure in *S. cerevisiae* by combining

adenine methylation with nanopore sequencing. The read length achieved in this thesis was approximately 4 kb on average, which was limited by the mechanical methods used to break open the *S. cerevisiae* cells to extract the DNA. Longer read lengths could be achieved by extracting the DNA by enzymatic means with zymolyase.

The chemical approach involved testing two furanocoumarins: TMP and angelicin which modify thymines in a TA sequence context. These provide an advantage of fast modification as the addition of these chemicals immediately kills the cells and the period of UV activation to modify the DNA is not particularly long. However, the UV activation requires optimisation to provide sufficient activation of the chemicals while limiting DNA damage, particularly the formation of pyrimidine dimers [129]. As the addition of TMP kills the cells it would be unable to identify chromatin dynamics or transient interactions. TMP was shown in chapter 3 to be able to modify DNA at a TA site on oligonucleotides when analysed by mass spectrometry however no modification was observed by nanopore sequencing. Conversely, angelicin, which formed monoadducts rather than DNA crosslinks, did not appear to modify the DNA at all in mass spectrometry analysis.

The enzymatic approach was first tested *in vitro* where M.EcoGII, a non-specific adenine methyltransferase, was further characterised and a quantitative method involving Droplet Digital PCR was developed for measuring methylation at specific loci. For the *in vitro* methylation there is some potential chromatin preparation variability and the risk that the proteins, even if crosslinked, will not be in their true biological binding positions. There was also the possibility that nucleosomes did not fully protect the DNA from methylation [116]. Methylation of *S. cerevisiae* cells at the spheroplast stage was achieved at a very minimal level, thus indicating that methylation of spheroplasts requires a much high concentration of methyltransferase. This may be due to insufficient spheroplasting meaning the cell membrane

blocks the methyltransferase from entering the cell, or due to a large number of proteins and other cell debris getting in the way between the methyltransferase and the DNA.

The enzymatic approach was then tested *in vivo* where M.EcoGII was inserted into the *S. cerevisiae* genome under the control of a *GALI* promoter. This approach involved inducing expression of M.EcoGII in the cells by addition of galactose. Minimal methylation was observed in the first forty-five minutes of induction, thus indicating that relatively long time frames are required. Most of the studies in this thesis used asynchronous cultures in which the cells were still growing during the induction period. This also means there is ongoing DNA replication and mitosis which may lead to the presence of methylation in protein binding sites due to protein movement. Transient binding events would also be lost in the long induction period. This could be simplified by arresting cells in G1, and inducing expression of M.EcoGII upon arrest, as performed at the end of chapter 4. In this case, the cells are mostly in the same cell cycle phase and there should be no DNA replication to confound the results.

Another modifying agent that could have been investigated is formaldehyde (HCHO) which is the most common crosslinking agent. It can perform protein-protein, protein-lipid and lipid-lipid crosslinking in addition to protein-DNA crosslinking, thus allowing it to be used to identify protein interactions and complex formation. In addition, formaldehyde can form interstrand crosslinks at adenines in AT sites in synthetic annealed oligonucleotides [4] or modify guanine and cytosine on mononucleotides or single-stranded DNA [221, 222]. Conversely, there are competing ideas that formaldehyde is unable to modify double-stranded DNA [223, 222]. The large variety of interactions that formaldehyde makes in the cell, and the potential inability to modify double-stranded DNA render it unsuitable for acting as a marker of accessible DNA in the way intended in this project. Conversely, TMP has minimal interactions with proteins [142] and has consistently been shown able to modify

double-stranded DNA [141, 132, 125, 127, 113] therefore was a more suitable candidate. Either the TMP, angelicin or methylation approaches could potentially be combined with formaldehyde crosslinking to preserve the chromatin state during modification.

In theory, formaldehyde could be used in another way, as a marker of protein-bound DNA by retaining the modification from the DNA-protein crosslink. Solomon and Varshavsky applied this concept to Simian virus 40 (SV40) chromosomes by treating them with formaldehyde either *in vivo* or *in vitro* [224]. Subsequent protein digestion with pronase and SDS left pronase-resistant DNA-peptide adducts. Using these, they found that histones were efficiently crosslinked to DNA therefore allowing the identification of nucleosome locations [224]. Other proteins, including the lac repressor were not crosslinked to DNA [224] meaning that this method was not able to determine the DNA binding sites of all proteins. However, it could be a good marker of *in vivo* nucleosome-bound DNA. This would pose a significant challenge for nanopore sequencing, as nanopore sequencing detection of a formaldehyde-modification on DNA has not been reported and the DNA-peptide adducts may be variable depending on the protein and level of digestion, meaning that there wouldn't be a consistent signal to look for in the nanopore data. Furthermore, modifications of that size may affect the mapping of the read to the reference sequence.

The chromatin accessibility methods discussed in this thesis rely on DNA-binding proteins protecting DNA from modification. It has been shown that nucleosomes can protect DNA from cytosine methylation both *in vitro* and *in vivo* [212, 225]. Conversely, it was identified in SMAC-seq that nucleosomes did not offer complete protection from methylation by M.CviPI, M.SssI and M.EcoGII [116]. In this case the nucleosome dyads were particularly highly methylated relative to the rest of the nucleosome-bound DNA [116]. This supports the data in this thesis that showed that there was sometimes a reduction in methylation in nucleosome-bound sites based on a consensus dataset [89] but the nucleosomes did not

offer complete protection. However, the peak in methylation at the nucleosome-dyad from SMAC-seq was not observed in this thesis. Further validation could involve aligning the data to other nucleosome datasets for comparison. Although an experiment could be designed where the cells are split into two fractions during the sample collection, one for MNase-seq and one for methylation chromatin accessibility investigation, this would still yield variation in the nucleosome positions as MNase-seq remains a population-based method. Ideally the exact positioning of a nucleosome on a specific DNA molecule would be able to be detected by two methods for verification.

In addition to nucleosomes, this chromatin accessibility assay was aiming to investigate protein-binding sites of proteins involved in replication initiation. ORC and MCM both bind DNA in ring structures [11, 20] which theoretically protect the DNA from modification. SMAC-seq has shown that Orc1, a component of the ORC ring, can protect DNA from methylation which is shown by a reduction in methylation at the binding site [116]. From the meta-analysis and cumulative sum representations (figures 4.49 and 4.50) there are potential indications of ORC binding to the DNA but these were far from conclusive.

A long timeframe introduces the opportunity for adenine methylation to affect DNA-binding events as, like restriction enzymes, DNA-binding proteins may have different sensitivities to adenine methylation. Despite being a small addition to the base, DNA methylation can affect the curvature, particularly if methylated sites are correctly phased, as shown at GATC sites [176], although this appears to vary between sequences [175]. Adenine methylation preferentially occurs at the N6 position of adenine in the *cis* conformation, where it interferes with Watson-Crick base pairing [226] and can lead to context-dependent destabilisation of the DNA [227]. This is not sufficient to lead to DNA unwinding or mismatch base pairing [227] but it may affect protein binding. For example, in *E. coli* methylation of the GATC sites to which Integration Host Factor (IHF) binds can slightly increase IHF binding

[228]. Therefore, it is possible for adenine methylation to affect the binding of proteins to DNA and may act to support DNA-bending roles of DNA-binding proteins. If the binding of transcription or regulatory factors are affected then adenine methylation may have an indirect effect on gene expression. To measure this RNA-seq could be performed on both the inducible M.EcoGII strain and a control wild-type strain that was used to generate the inducible M.EcoGII strain.

In this study, the focus of the genomic footprinting method has been nucleosomes and DNA replication. However, it can be applied to any DNA-binding proteins. The requirements are that the protein needs to bind the DNA in a way that will protect the DNA from modification and the binding needs to be relatively stable over the period of M.EcoGII induction. In addition, this method has been developed in *S. cerevisiae* but the principles can be applied to other organisms which lack endogenous adenine methylation if using 6mA, or cytosine methylation if using 5mC. The nanopore sequencing analysis would remain similar across a variety of organisms.

One of the major challenges encountered in this thesis was the ability of nanopore sequencing to detect DNA base modifications. An alternative approach would have been to independently train a model for detecting adenine methylation in all sequence contexts. This would have required knowledge of exactly which positions in a substrate are methylated, in order to train the model accurately, thus ideally requiring a synthetic substrate. In nanopore sequencing, although the main signal arises from the six bases currently in the narrowest part of the pore, the flanking sequences also affect the signal. Therefore there is an enormous range of possible sequence contexts that would be needed for training to reduce context bias. A model could also be optimised for a specific organism based on the frequency of each sequence context in the genome. Therefore, a model could have varying accuracies based on how well the train and test datasets align. It is really important not to overtrain a

model in order to allow detection in unseen sequence contexts. A DNA modification with high frequency in the genome was required for this project, therefore model training was not suitable.

Future extensions of the nanopore sequencing analysis could involve performing detection with new alternative software, for example Megalodon, or adding a different weighting to each k-mer based on the likelihood of it being methylated by M.EcoGII in the absence of proteins. K-mers with a particularly high modification likelihood in control conditions could also be excluded from analysis. It is also possible that in the future nanopore sequencing may be able to capture information for both strands of an original dsDNA molecule, which would increase the validity of modified base calls on single reads. A further extension would be to create an algorithm to classify regions as accessible or protein-bound. A starting point could be to exclude isolated methylated bases, as they are likely to be false-positives. In addition, a reference sequence based on the W303 background of the inducible M.EcoGII strain could be used as this may have slight differences to the *sacCer3* reference that is based on S288C. However, the S288C reference is well annotated and widely used as a standard.

Future work could also include nanopore sequencing the M.EcoGII inducible strains with *orc1-161* and *cdc6-1* mutations, as created in chapter 4, arrested in G1 to identify any changes in chromatin structure that could be caused by the formation of the pre-RC. The *orc1-161* mutation is in the BAH domain which is responsible for the stable interaction between ORC and DNA [87]. This would provide further validation of the method in addition to potential insights into the interactions of replication initiation proteins on single molecules.

In the future this approach could be extended to use multiple DNA modifications. This is currently limited by the ability of nanopore sequencing software to accurately distinguish between modifications on a single DNA molecule. However, there have been many devel-

opments in the ability of nanopore sequencing to detect base modifications in recent years [106–108, 110] therefore this is likely to become a possibility in the future. Based on this thesis, the best combinations would include multiple types of methylation as combined in SMAC-seq [116] as there are methods for detecting these with nanopore sequencing. In the future, if angelicin could be shown to modify DNA it may provide a good alternative modification as the chemical structure is substantially different to methylated or canonical bases and it does not crosslink the two DNA strands together as TMP does. However, angelicin only modifies at TA and potentially AT sites so the resolution is reasonably low, roughly equivalent to that of a GC or CG methyltransferase depending on the genomic base composition. A combination of sequencing methods could also be investigated, for example with EM-seq (Enzymatic Methyl-seq) or TAPS (TET-assisted pyridine borane sequencing) to detect cytosine methylation, which preserve the DNA better than bisulfite sequencing which uses high temperatures and long incubation times. By avoiding these harsh treatments, they facilitate the generation of longer sequencing reads. Involving these techniques would however prevent the modifications being investigated on the same DNA molecules. Even more ambitiously, it might in the future be possible to combine protein peptide marks from formaldehyde crosslinking proteins to DNA with the modifications shown in this thesis for modifying accessible DNA. An alternative would be combination of adenine methylation with nanopore sequencing with a DamID-style method with one type of methylation providing accessible chromatin information and another protein-specific information. This would be time-consuming to generate the fusion proteins but could provide unique insights and act as an additional validation of the chromatin accessibility assay. This would provide protein identification at binding sites, thereby conceptually combining the information gained from ChIP-seq with the accessible DNA information gained by the method developed in this thesis.

5.2 Conclusion

In conclusion this thesis has tested chemical modification of DNA with psoralen and angelicin and enzymatic modification of DNA by methylation as potential markers of chromatin accessibility, with detection by nanopore sequencing, further characterised the M.EcoGII non-specific adenine methyltransferase and developed an *in vivo* single-molecule chromatin accessibility assay for *S. cerevisiae*.

References

- [1] Seetha V. Srinivasan, David Dominguez-Sola, Lily C. Wang, Olivier Hyrien, and Jean Gautier. Cdc45 is a critical effector of Myc-dependent DNA replication stress. *Cell Reports*, 3:1629–1639, 5 2013.
- [2] Martin F. Arlt, Jennifer G. Mulle, Valerie M. Schaibley, Ryan L. Ragland, Sandra G. Durkin, Stephen T. Warren, and Thomas W. Glover. Replication stress induces genome-wide copy number changes in human cells that resemble polymorphic and pathogenic variants. *The American Journal of Human Genetics*, 84:339–350, 3 2009.
- [3] Susanna Ekholm-Reed, Juan Méndez, Donato Tedesco, Anders Zetterberg, Bruce Stillman, and Steven I. Reed. Dereglulation of cyclin E in human cells interferes with prereplication complex assembly. *Journal of Cell Biology*, 165:789–800, 6 2004.
- [4] Huifang Huang and Paul B. Hopkins. DNA interstrand cross-linking by formaldehyde: nucleotide sequence preference and covalent structure of the predominant cross-link formed in synthetic oligonucleotides. *Journal of the American Chemical Society*, 115:9402–9408, 1993.
- [5] Conrad A. Nieduszynski, Yvonne Knox, and Anne D. Donaldson. Genome-wide identification of replication origins in yeast by comparative genomics. *Genes and Development*, 20:1874–1879, 2006.
- [6] Adam M. Breier, Sourav Chatterji, and Nicholas R. Cozzarelli. Prediction of *Saccharomyces cerevisiae* replication origins. *Genome Biology*, 5, 2004.
- [7] Matthew L. Eaton, Kyriaki Galani, Sukhyun Kang, Stephen P. Bell, and David M. MacAlpine. Conserved nucleosome positioning defines replication origins. *Genes and Development*, 24:748–753, 2010.
- [8] Vinod Kumar Singh and Annangarachari Krishnamachari. Context based computational analysis and characterization of ARS consensus sequences (ACS) of *Saccharomyces cerevisiae* genome. *Genomics Data*, 9:130–136, 9 2016.
- [9] Stephen P. Bell and Bruce Stillman. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature*, 357:128–134, 1992.
- [10] York Marahrens and Bruce Stillman. A yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science*, 255:817–823, 1992.

- [11] Jingchuan Sun, Hironori Kawakami, Juergen Zech, Christian Speck, Bruce Stillman, and Huilin Li. Cdc6-induced conformational changes in ORC bound to origin DNA revealed by cryo-electron microscopy. *Structure*, 20:534–544, 3 2012.
- [12] Margit Foss, Francis J. McNally, Patricia Laurenson, and Jasper Rine. Origin Recognition Complex (ORC) in transcriptional silencing and DNA replication in *S. cerevisiae*. *Science*, 262:1838–1844, 1993.
- [13] John F. X. Diffley and Julie H. Cocker. Protein-DNA interactions at a yeast replication origin. *Nature*, 357:169–172, 1992.
- [14] John F. X. Diffley, Julie H. Cocker, Simon J. Dowell, and Adele Rowley. Two steps in the assembly of complexes at yeast replication origins in vivo. *Cell*, 78:303–316, 1994.
- [15] Daniel Duzdevich, Megan D. Warner, Simina Ticau, Nikola A. Ivica, Stephen P. Bell, and Eric C. Greene. The dynamics of eukaryotic replication initiation: Origin specificity, licensing, and firing at the single-molecule level. *Molecular Cell*, 58:483–494, 5 2015.
- [16] Shuyan Chen, Milan A. De Vries, and Stephen P. Bell. Orc6 is required for dynamic recruitment of Cdt1 during repeated Mcm2-7 loading. *Genes and Development*, 21:2897–2907, 11 2007.
- [17] Alejandra Fernández-Cid, Alberto Riera, Silvia Tognetti, M. Carmen Herrera, Stefan Samel, Cecile Evrin, Christian Winkler, Emanuela Gardenal, Stefan Uhle, and Christian Speck. An ORC/Cdc6/MCM2-7 complex is formed in a multistep reaction to serve as a platform for MCM double-hexamer assembly. *Molecular Cell*, 50:577–588, 5 2013.
- [18] Jayson L. Bowers, John C.W. Randell, Shuyan Chen, and Stephen P. Bell. ATP hydrolysis by ORC catalyzes reiterative Mcm2-7 assembly at a defined origin of replication. *Molecular Cell*, 16:967–978, 12 2004.
- [19] Simina Ticau, Larry J. Friedman, Kanokwan Champasa, Ivan R. Corrêa, Jeff Gelles, and Stephen P. Bell. Mechanism and timing of Mcm2-7 ring closure during DNA replication origin licensing. *Nature Structural and Molecular Biology*, 24:309–315, 3 2017.
- [20] Dirk Remus, Fabienne Beuron, Gökhan Tolun, Jack D. Griffith, Edward P. Morris, and John F.X. Diffley. Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell*, 139:719–730, 11 2009.
- [21] Simina Ticau, Larry J. Friedman, Nikola A. Ivica, Jeff Gelles, and Stephen P. Bell. Single-molecule studies of origin licensing reveal mechanisms ensuring bidirectional helicase loading. *Cell*, 161:513–525, 4 2015.
- [22] Irina Bruck and Daniel L. Kaplan. Origin single-stranded DNA releases Sld3 protein from the Mcm2-7 complex, allowing the GINS tetramer to bind the Mcm2-7 complex. *Journal of Biological Chemistry*, 286:18602–18613, 5 2011.

- [23] Seiji Tanaka, Toshiko Umemori, Kazuyuki Hirai, Sachiko Muramatsu, Yoichiro Kamimura, and Hiroyuki Araki. CDK-dependent phosphorylation of Sld2 and Sld3 initiates DNA replication in budding yeast. *Nature*, 445:328–332, 1 2007.
- [24] Philip Zegerman and John F.X. Diffley. Phosphorylation of Sld2 and Sld3 by cyclin-dependent kinases promotes DNA replication in budding yeast. *Nature*, 445:281–285, 1 2007.
- [25] Yoichiro Kamimura, Hiroshi Masumoto, Akio Sugino, and Hiroyuki Araki. Sld2, which interacts with Dpb11 in *Saccharomyces cerevisiae*, is required for chromosomal DNA replication. *Molecular and Cellular Biology*, 18:6102–6109, 1998.
- [26] Yon-Soo Tak, Yoshimi Tanaka, Shizuko Endo, Yoichiro Kamimura, and Hiroyuki Araki. A CDK-catalysed regulatory phosphorylation for formation of the DNA replication complex Sld2-Dpb11. *EMBO Journal*, 25:1987–1996, 5 2006.
- [27] Yoichiro Kamimura, Yon-Soo Tak, Akio Sugino, and Hiroyuki Araki. Sld3, which interacts with Cdc45 (Sld4), functions for chromosomal DNA replication in *Saccharomyces cerevisiae*. *The EMBO journal*, 20:2097–2107, 2001.
- [28] Hiroshi Masumoto, Akio Sugino, and Hiroyuki Araki. Dpb11 controls the association between DNA polymerases alpha and epsilon and the autonomously replicating sequence region of budding yeast. *Molecular and Cellular Biology*, 20:2809–2817, 2000.
- [29] Navin K Sinha, Charles F Morris, and Bruce M Alberts. Efficient in vitro replication of double-stranded dna templates by a purified t4 bacteriophage replication system. *The Journal of Biological Chemistry*, 255:4290–4303, 1980.
- [30] Toshiki Tsurimoto, Thomas Melendy, and Bruce Stillman. Sequential initiation of lagging and leading strand synthesis by two different polymerase complexes at the sv40 dna replication origin. *Nature*, 346, 8 1990.
- [31] Cheng-Keat Tan, C. Castillo, A. G. So, and K. M. Downey. An auxiliary protein for dna polymerase- δ from fetal calf thymus. *Journal of Biological Chemistry*, 261:12310–12316, 1986.
- [32] Federico Focher, Max Gassmann, Peter Hafkemeyer, Elena Ferrari, Silvio Spadari, and Ulrich Hiibscher. Calf thymus dna polymerase delta independent of proliferating cell nuclear antigen (pcna). *Nucleic Acids Research*, 17:1805–1821, 1989.
- [33] F Zachary Pursell, Isabelle Isoz, Else-Britt Lundström, Erik Johansson, and Thomas A Kunkel. Yeast dna polymerase ϵ participates in leading-strand dna replication. *Science*, 317:127–130, 2007.
- [34] Robert E. Johnson, Roland Klassen, Louise Prakash, and Satya Prakash. A major role of dna polymerase δ in replication of both the leading and lagging dna strands. *Molecular Cell*, 59:163–175, 7 2015.
- [35] Yi-Jun Sheu and Bruce Stillman. Cdc7-Dbf4 phosphorylates MCM proteins via a docking site-mediated mechanism to promote S phase progression. *Molecular Cell*, 24:101–113, 10 2006.

- [36] Shuyan Chen and Stephen P. Bell. CDK prevents Mcm2-7 helicase loading by inhibiting Cdt1 interaction with Orc6. *Genes and Development*, 25:363–372, 2 2011.
- [37] Shankar P. Das, Tyler Borrmann, Victor W.T. Liu, Scott C.-H. Yang, John Bechhoefer, and Nicholas Rhind. Replication timing is regulated by the number of MCMs loaded at origins. *Genome Research*, 25:1886–1892, 12 2015.
- [38] Julien Gros, Charanya Kumar, Gerard Lynch, Tejas Yadav, Iestyn Whitehouse, and Dirk Remus. Post-licensing specification of eukaryotic replication origins by facilitated Mcm2-7 sliding along DNA. *Molecular Cell*, 60:797–807, 12 2015.
- [39] Reiji Okazaki, Tuneko Okazaki, Kiwako Sakabe, Kazunori Sugimoto, and Akio Sugino. Mechanism of DNA chain growth I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Biochemistry*, 59:598–605, 1968.
- [40] Jean-Michel Louarn and Robert E. Bird. Size distribution and molecular polarity of newly replicated DNA in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 71:329–333, 1974.
- [41] Jia Xu, Yoshimi Yanagisawa, Alexander M. Tsankov, Christopher Hart, Keita Aoki, Naveen Kommajosyula, Kathleen E. Steinmann, James Bochicchio, Carsten Russ, Aviv Regev, Oliver J. Rando, Chad Nusbaum, Hironori Niki, Patrice Milos, Zhiping Weng, and Nicholas Rhind. Genome-wide identification and characterization of replication origins by deep sequencing. *Genome Biology*, 13:1–14, 2012.
- [42] Jianli Dai, Ray-Yuan Chuang, and Thomas J. Kelly. DNA replication origins in the *Schizosaccharomyces pombe* genome. *Proceedings of the National Academy of Sciences of the United States of America*, 102:337–342, 2005.
- [43] Ivan Liachko, Anand Bhaskar, Chanmi Lee, Shau Chee Claire Chung, Bik-Kwoon Tye, and Uri Keich. A comprehensive genome-wide map of autonomously replicating sequences in a naive genome. *PLoS Genetics*, 6:22, 5 2010.
- [44] Satoshi Tabata, Atsuhiko Oka, Kazunori Sugimoto, Mituru Takanami, Seiichi Yasuda, and Yukinori Hirota. The 245 base-pair *oriC* sequence of the *E. coli* chromosome directs bidirectional replication at an adjacent region. *Nucleic Acids Research*, 11:2617–2626, 1983.
- [45] Robert E. Bird, Jacqueline Louarn, Jaime Martuscelli, and Lucien Caro. Origin and sequence of chromosome replication in *Escherichia coli*. *Journal of Molecular Biology*, 70:549–566, 1972.
- [46] David Bramhill and Arthur Kornberg. Duplex opening by *dnaA* protein at novel sequences in initiation of replication at the origin of the *E. coli* chromosome. *Cell*, 52:743–755, 1988.
- [47] Jasmine Arora, Kasturi Goswami, and Swati Saha. Characterization of the replication initiator Orc1/Cdc6 from the archaeon *Picrophilus torridus*. *Journal of Bacteriology*, 196:276–286, 1 2014.

- [48] Fujihiko Matsunaga, Patrick Forterre, Yoshizumi Ishino, and Hannu Myllykallio. In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. *Proceedings of the National Academy of Sciences of the United States of America*, 98:11152–11157, 2001.
- [49] Nicholas P. Robinson and Stephen D. Bell. Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 104:5806–5811, 2007.
- [50] Magnus Lundgren, Anders Andersson, Lanming Chen, Peter Nilsson, and Rolf Bernander. Three replication origins in *Sulfolobus* species: Synchronous initiation of chromosome replication and asynchronous termination. *Proceedings of the National Academy of Sciences of the United States of America*, 101:7046–7051, 2004.
- [51] Erik A. Pelve, Ann-Christin Lindås, Anna Knöppel, Alex Mira, and Rolf Bernander. Four chromosome replication origins in the archaeon *Pyrobaculum calidifontis*. *Molecular Microbiology*, 85:986–995, 9 2012.
- [52] Hannu Myllykallio, Philippe Lopez, Purificación López-García, Roland Heilig, William Saurin, Yvan Zivanovic, Hervé Philippe, and Patrick Forterre. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science*, 288:2212–2215, 2000.
- [53] Cymon J Cox, Peter G Foster, Robert P Hirt, Simon R Harris, and T Martin Embley. The archaeobacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 105:20356–20361, 2008.
- [54] Alison Rowles, James P. J. Chong, Lamorna Brown, Mike Howell, Gerard I. Evan, and J. Julian Blow. Interaction between the origin recognition complex and the replication licensing system in *Xenopus*. *Cell*, 87:287–296, 1996.
- [55] Piotr Romanowski, Mark A. Madine, Alison Rowles, J. Julian Blow, and Ronald A. Laskey. The *Xenopus* origin recognition complex is essential for DNA replication and MCM binding to chromatin. *Current Biology*, 6:1416–1425, 1996.
- [56] Melissa C. Edwards, Antonin V. Tutter, Christin Cvetic, Catherine H. Gilbert, Tatyana A. Prokhorova, and Johannes C. Walter. Mcm2-7 complexes bind chromatin in a distributed pattern surrounding the origin recognition complex in *Xenopus* egg extracts. *Journal of Biological Chemistry*, 277:33049–33057, 9 2002.
- [57] Dirk Remus, Eileen L. Beall, and Michael R. Botchan. DNA topology, not DNA sequence, is a critical determinant for *Drosophila* ORC-DNA binding. *EMBO Journal*, 23:897–907, 2 2004.
- [58] Federico Comoglio, Tommy Schlumpf, Virginia Schmid, Remo Rohs, Christian Beisel, and Renato Paro. High-resolution profiling of *Drosophila* replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Reports*, 11:821–834, 2015.

- [59] Stephen E. Moyer, Peter W. Lewis, and Michael R. Botchan. Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proceedings of the National Academy of Sciences of the United States of America*, 103:10236–10241, 2006.
- [60] James F. Theis and Carol S. Newlon. Two compound replication origins in *Saccharomyces cerevisiae* contain redundant origin recognition complex binding sites. *Molecular and Cellular Biology*, 21:2790–2801, 4 2001.
- [61] Nataliya Petryk, Malik Kahli, Yves d’Aubenton Carafa, Yan Jaszczyzyn, Yimin Shen, Maud Silvain, Claude Thermes, Chun-Long Chen, and Olivier Hyrien. Replication landscape of the human genome. *Nature Communications*, 7, 1 2016.
- [62] Peiyao A. Zhao, Takayo Sasaki, and David M. Gilbert. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biology*, 21, 3 2020.
- [63] Kai Stoeber, Thea D. Tlsty, Lisa Happerfield, Geraldine A. Thomas, Sergei Romanov, Lynda Bobrow, W. Dillwyn Williams, and Gareth H. Williams. DNA replication licensing and human cell proliferation. *Journal of Cell Science*, 114:2027–2041, 2001.
- [64] Benoit Miotto, Zhe Ji, and Kevin Struhl. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 113:E4810–E4819, 8 2016.
- [65] Sanjay Vashee, Christin Cvetic, Wenyan Lu, Pamela Simancek, Thomas J. Kelly, and Johannes C. Walter. Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes and Development*, 17:1894–1908, 8 2003.
- [66] Larry D. Mesner, Veena Valsakumar, Neerja Karnani, Anindya Dutta, Joyce L. Hamlin, and Stefan Bekiranov. Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. *Genome Research*, 21:377–389, 3 2011.
- [67] Gaetano Ivan Dellino, Davide Cittaro, Rossana Piccioni, Lucilla Luzi, Stefania Banfi, Simona Segalla, Matteo Cesaroni, Ramiro Mendoza-Maldonado, Mauro Giacca, and Pier Giuseppe Pelicci. Genome-wide mapping of human DNA-replication origins: Levels of transcription at Orc1 sites regulate origin selection and replication timing. *Genome Research*, 23:1–11, 1 2013.
- [68] Arkaitz Ibarra, Etienne Schwob, and Juan Méndez. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proceedings of the National Academy of Sciences of the United States of America*, 105:8956–8961, 2008.
- [69] Romain Drissi, Anaïs Chauvin, Alyson McKenna, Dominique Lévesque, Simon Blais-Brochu, Dominique Jean, and François Michel Boisvert. Destabilization of the MiniChromosome Maintenance (MCM) complex modulates the cellular response to DNA double strand breaks. *Cell Cycle*, 17:2593–2609, 12 2018.

- [70] Teresa Casar Tena, Lars D. Maerz, Karol Szafranski, Marco Groth, Tamara J. Blätte, Cornelia Donow, Sabrina Matysik, Paul Walther, Penelope A. Jeggo, Martin D. Burkhalter, and Melanie Philipp. Resting cells rely on the DNA helicase component MCM2 to build cilia. *Nucleic Acids Research*, 47:134–151, 1 2019.
- [71] Louise S Bicknell, Ernie M.H.F. Bongers, Andrea Leitch, Stephen Brown, Jeroen Schoots, Margaret E. Harley, Salim Aftimos, Jumana Y. Al-Aama, Michael Bober, Paul A. J. Brown, Hans van Bokhoven, John Dean, Alaa Y. Edrees, Murray Feingold, Alan Fryer, Lies H. Hoefsloot, Nikolaus Kau, Nine V.A.M. Knoers, James MacKenzie, John M. Opitz, Pierre Sarda, Alison Ross, I. Karen Temple, Annick Toutain, Carol A. Wise, Michael Wright, and Andrew P. Jackson. Mutations in the pre-replication complex cause Meier-Gorlin syndrome. *Nature Genetics*, 43:356–359, 2011.
- [72] Stephen L. McDaniel, Allison J. Hollatz, Anna M. Branstad, Marissa M. Gaskill, Catherine A. Fox, and Melissa M. Harrison. Tissue-specific DNA replication defects in *Drosophila melanogaster* caused by a Meier-Gorlin syndrome mutation in Orc4. *Genetics*, 214:355–367, 2020.
- [73] Louise S. Bicknell, Sarah Walker, Anna Klingseisen, Tom Stiff, Andrea Leitch, Claudia Kerzendorfer, Carol-Anne Martin, Patricia Yeyati, Nouriya Al Sanna, Michael Bober, Diana Johnson, Carol Wise, Andrew P. Jackson, Mark O’Driscoll, and Penny A. Jeggo. Mutations in ORC1, encoding the largest subunit of the origin recognition complex, cause microcephalic primordial dwarfism resembling Meier-Gorlin syndrome. *Nature Genetics*, 43:350–356, 2 2011.
- [74] Likun Yao, Jing Chen, Xiaotong Wu, Shunji Jia, and Anming Meng. Zebrafish cdc6 hypomorphic mutation causes Meier-Gorlin syndrome-like phenotype. *Human Molecular Genetics*, 26:4168–4180, 2017.
- [75] Darren A. Natale, Robert M. Umek, and David Kowalski. Ease of DNA unwinding is a conserved property of yeast replication origins. *Nucleic Acids Research*, 21:555–560, 1993.
- [76] Nicolas M. Berbenetz, Corey Nislow, and Grant W. Brown. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genetics*, 6, 9 2010.
- [77] J. Russell Lipford and Stephen P. Bell. Nucleosomes positioned by ORC facilitate the initiation of DNA replication. *Molecular Cell*, 7:21–30, 2001.
- [78] David M. MacAlpine, Heather K. Rodríguez, and Stephen P. Bell. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes and Development*, 18:3094–3105, 12 2004.
- [79] Xiao-Qin Cao, Jia Zeng, and Hong Yan. Structural properties of replication origins in yeast DNA sequences. *Physical Biology*, 5, 9 2008.
- [80] Andreas W. Thomae, Dagmar Pich, Jan Brocher, Mark-Peter Spindler, Christian Berens, Robert Hock, Wolfgang Hammerschmidt, and Aloys Schepers. Interaction between HMGA1a and the origin recognition complex creates site-specific replication origins. *Proceedings of the National Academy of Sciences of the United States of America*, 105:1692–1697, 2008.

- [81] Julie Norseen, Andreas Thomaе, Venkatesh Sridharan, Ashok Aiyar, Aloys Schepers, and Paul M. Lieberman. RNA-dependent recruitment of the origin recognition complex. *EMBO Journal*, 27:3024–3035, 11 2008.
- [82] Till Bartke, Michiel Vermeulen, Blerta Xhemalce, Samuel C. Robson, Matthias Mann, and Tony Kouzarides. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell*, 143:470–484, 10 2010.
- [83] Zhen Shen, Kizhakke M. Sathyan, Yijie Geng, Ruiping Zheng, Arindam Chakraborty, Brian Freeman, Fei Wang, Kannanganattu V. Prasanth, and Supriya G. Prasanth. A WD-repeat protein stabilizes ORC binding to chromatin. *Molecular Cell*, 40:99–111, 10 2010.
- [84] Daniel T. S. Pak, Michelle Pflumm, Igor Chesnokov, Da Wei Huang, Rebecca Kellum, Jacqueline Marr, Piotr Romanowski, and Michael R. Botchan. Association of the origin recognition complex with heterochromatin and HP1 in higher eukaryotes. *Cell*, 91:311–323, 1997.
- [85] Heather K. MacAlpine, Raluca Gordân, Sara K. Powell, Alexander J. Hartemink, and David M. MacAlpine. Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome Research*, 20:201–211, 2 2010.
- [86] Philipp Muller, Sookhee Park, Erika Shor, Dana J. Huebert, Christopher L. Warren, Aseem Z. Ansari, Michael Weinreich, Matthew L. Eaton, David M. MacAlpine, and Catherine A. Fox. The conserved bromo-adjacent homology domain of yeast Orc1 functions in the selection of DNA replication origins within chromatin. *Genes and Development*, 24:1418–1433, 7 2010.
- [87] Kohji Noguchi, Alex Vassilev, Soma Ghosh, John L. Yates, and Melvin L. DePamphilis. The BAH domain facilitates the ability of human Orc1 protein to activate replication origins in vivo. *EMBO Journal*, 25:5372–5382, 11 2006.
- [88] Alex J. Kuo, Jikui Song, Peggie Cheung, Satoko Ishibe-Murakami, Sayumi Yamazoe, James K. Chen, Dinshaw J. Patel, and Or Gozani. The BAH domain of Orc1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature*, 484:115–119, 4 2012.
- [89] Cizhong Jiang and B. Franklin Pugh. A compiled and systematic reference map of nucleosome positions across the *Saccharomyces cerevisiae* genome. *Genome Biology*, 10:–undefined, 2009.
- [90] Jason A. Belsky, Heather K. MacAlpine, Yoav Lubelsky, Alexander J. Hartemink, and David M. MacAlpine. Genome-wide chromatin footprinting reveals changes in replication origin architecture induced by pre-RC assembly. *Genes and Development*, 29:212–224, 2014.
- [91] David S. Gilmour and John T. Lis. Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America*, 81:4275–4279, 1984.
- [92] Debasish Raha, Miyoung Hong, and Michael Snyder. ChIP-Seq: a method for global identification of regulatory elements in the genome, 2010.

- [93] Feinan Wu, Brennan G. Olson, and Jie Yao. DamID-seq: Genome-wide mapping of protein-DNA interactions by high throughput sequencing of adenine-methylated DNA fragments. *Journal of Visualized Experiments*, 2016:1–11, 2016.
- [94] Bas van Steensel and Steven Henikoff. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nature Biotechnology*, 18:424–428, 2000.
- [95] Paul G. Giresi, Jonghwan Kim, Ryan M. McDaniell, Vishwanath R. Iyer, and Jason D. Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17:877–885, 2007.
- [96] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10:1213–1218, 2013.
- [97] Lingyun Song and Gregory E. Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 5, 2 2010.
- [98] Kairong Cui and Keji Zhao. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-seq. *Methods in Molecular Biology*, 833:413–419, 2012.
- [99] Muhammad Shoaib, David Walter, Peter J. Gillespie, Fanny Izard, Birthe Fahrenkrog, David Lleres, Mads Lerdrup, Jens Vilstrup Johansen, Klaus Hansen, Eric Julien, J. Julian Blow, and Claus S. Sørensen. Histone H4K20 methylation mediated chromatin compaction threshold ensures genome integrity by limiting DNA replication licensing. *Nature Communications*, 9, 12 2018.
- [100] Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzénburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523:486–490, 7 2015.
- [101] Jacob Peter Matson, Raluca Dumitru, Philip Coryell, Ryan M. Baxley, Weili Chen, Kirk Twaroski, Beau R. Webber, Jakub Tolar, Anja-Katrin Bielinsky, Jeremy E. Purvis, and Jeanette Gowen Cook. Rapid DNA replication origin licensing protects stem cell pluripotency. *eLIFE*, 6, 2017.
- [102] A. Bensimon, A. Simon, A. Chiffaudel, V. Croquette, F. Heslot, and D. Bensimon. Alignment and sensitive detection of DNA by a moving interface. *Science*, 265:2096–2098, 1994.
- [103] Xavier Michalet, Rosemary Ekong, Francoise Fougereousse, Sophie Rousseaux, Catherine Schurra, Nick Hornigold, Marjon van Slegtenhorst, Jonathan Wolfe, Sue Povey, Jacques S. Beckmann, and Aaron Bensimon. Dynamic molecular combing: Stretching the whole human genome for high-resolution studies. *Science*, 277:1518–1523, 1997.

- [104] Atanas Kaykov, Thibaud TAILLEFUMIER, Aaron Bensimon, and Paul Nurse. Molecular combing of single DNA molecules on the 10 megabase scale. *Scientific Reports*, 6, 1 2016.
- [105] Teresa Fazio, Mari-Liis Visnapuu, Shalom Wind, and Eric C. Greene. DNA curtains and nanoscale curtain rods: High-throughput tools for single molecule imaging. *Langmuir*, 24:10524–10531, 9 2008.
- [106] Carolin A. Müller, Michael A. Boemo, Paolo Spingardi, Benedikt M. Kessler, Skirmantas Kriaucionis, Jared T. Simpson, and Conrad A. Nieduszynski. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nature Methods*, 16:429–436, 2019.
- [107] Jared T Simpson, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14:407–410, 2017.
- [108] Arthur C. Rand, Miten Jain, Jordan M. Eizenga, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson, and Benedict Paten. Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods*, 14:411–413, 2017.
- [109] Magali Hennion, Jean-Michel Arbona, Laurent Lacroix, Corinne Cruaud, Bertrand Theulot, Benoît Le Tallec, Florence Proux, Xia Wu, Elizaveta Novikova, Stefan Engelen, Arnaud Lemainque, Benjamin Audit, and Olivier Hyrien. FORK-seq: Replication landscape of the *Saccharomyces cerevisiae* genome by nanopore sequencing. *Genome Biology*, 21, 5 2020.
- [110] Daniela Georgieva, Qian Liu, Kai Wang, and Dieter Egli. Detection of base analogs incorporated during DNA replication by nanopore sequencing. *Nucleic Acids Research*, 48:1–12, 2020.
- [111] Christian Speck, Zhiqiang Chen, Huilin Li, and Bruce Stillman. ATPase-dependent cooperative binding of ORC and Cdc6p to origin DNA. *Nature Structural and Molecular Biology*, 12:965–971, 11 2005.
- [112] Jieqiong Zhang, James M Dewar, Magda Budzowska, Anna Motnenko, Martin A. Cohn, and Johannes C Walter. DNA interstrand cross-link repair requires replication fork convergence. *Nature Structural and Molecular Biology*, 22:242–247, 2015.
- [113] Jun ichiro Komura, Hironobu Ikehata, Yoshio Hosoi, Arthur D. Riggs, and Tetsuya Ono. Mapping psoralen cross-links at the nucleotide level in mammalian cells: Suppression of cross-linking at transcription factor- or nucleosome-binding sites. *Biochemistry*, 40:4096–4105, 2001.
- [114] Joseph Schacherer, Douglas M. Ruderfer, David Gresham, Kara Dolinski, David Botstein, and Leonid Kruglyak. Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. *PLoS ONE*, 2, 3 2007.
- [115] M.D. Rose, F. Winston, and P. Hieter. *Methods in Yeast Genetics - A Laboratory Course Manual* . 1990.

- [116] Zohar Shipony, Georgi K. Marinov, Matthew P. Swaffer, Nicholas A. Sinnott-Armstrong, Jan M. Skotheim, Anshul Kundaje, and William J. Greenleaf. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nature Methods*, 17:319–327, 2019.
- [117] Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094–3100, 9 2018.
- [118] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079, 8 2009.
- [119] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 1 2010.
- [120] Stephen F. Altschup, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [121] Audray Dugrand-Judek, Alexandre Olry, Alain Hehn, Gilles Costantino, Patrick Ollitrault, Yann Froelicher, and Frédéric Bourgaud. The distribution of coumarins and furanocoumarins in Citrus species closely matches Citrus phylogeny and reflects the organization of biosynthetic pathways. *PLoS ONE*, 10:1–25, 2015.
- [122] Radek Peroutka, Vera Schulzova, Petr Botek, and Jana Hajslova. Analysis of furanocoumarins in vegetables (Apiaceae) and citrus fruits (Rutaceae). *Journal of the science of food and agriculture*, 87:2152–2163, 2007.
- [123] Nguyen Thanh Dong, KiHwan Bae, Young Ho Kim, Gwi Seo Hwang, Ok Soon Heo, Se Eun Kim, and Jong Seong Kang. Quantitative determination of psoralen and angelicin from some medicinal herbs by high performance liquid chromatography. *Archives of Pharmacal Research*, 26:516–20, 2003.
- [124] Anthony T. Yeung, Beverly K. Jones, and C. Timothy Chu. Photoreactivities and thermal properties of psoralen cross-links. *Biochemistry*, 27:3204–3210, 1988.
- [125] Franca Esposito, Robert G. Brankamp, and Richard R. Sinden. DNA sequence specificity of 4,5',8-trimethylpsoralen cross-linking. *The Journal of Biological Chemistry*, 263:11466–11472, 1988.
- [126] R. M. Widmer, Th. Koller, and J. M. Sogo. Analysis of the psoralen-crosslinking pattern in chromatin DNA by exonuclease digestion. *Nucleic Acids Research*, 16:7013–7024, 1988.
- [127] Ralf Erik Wellinger, Renzo Lucchini, Reinhard Dammann, and José M Sogo. In vivo mapping of nucleosomes using psoralen-DNA crosslinking and primer extension. *Methods in Molecular Biology*, 119:161–173, 1999.

- [128] David Kanne, Kenneth Straub, Henry Rapoport, and John E. Hearst. Psoralen-deoxyribonucleic acid photoreaction. Characterization of the monoaddition products from 8-methoxypsoralen and 4,5',8-trimethylpsoralen. *Biochemistry*, 21:861–871, 1982.
- [129] R. B. Setlow, W. L. Carrier, and F. J. Bollum. Pyrimidine dimers in UV-irradiated poly dI:dC. *Proceedings of the National Academy of Sciences of the United States of America*, 53:1111–1118, 1965.
- [130] Yun bo Shi, H. Peter Spielmann, and John E. Hearst. Base-catalyzed reversal of a psoralen-DNA cross-link. *Biochemistry*, 27:5174–5178, 1988.
- [131] William R. Kobertz and John M. Essigmann. Solid-phase synthesis of oligonucleotides containing a site-specific psoralen derivative. *Journal of the American Chemical Society*, 119:5960–5961, 1997.
- [132] Richard R. Sinden and Paul J. Hagerman. Interstrand psoralen cross-links do not introduce appreciable bends in DNA. *Biochemistry*, 23:6299–6303, 1984.
- [133] Isaac Willis and Harris David R. Resistant psoriasis: combined methoxsalen-anthralin therapy. *Arch Dermatol.*, 107:358–62, 1973.
- [134] D. Vella Briffa, A. P. Warin, and C. D. Calnan. Parakeratosis variegata: a report of two cases and their treatment with PUVA. *Clinical and Experimental Dermatology*, 4:537–541, 1979.
- [135] Aaron Bunsen Lerner, Cleveland R. Denton, and Thomas B. Fitzpatrick. Clinical and experimental studies with 8-methoxypsoralen in vitiligo. *The Journal of investigative dermatology*, 20:299–314, 1953.
- [136] Walter Liszewski, David Gram Naym, Edyta Biskup, and Robert Gniadecki. Psoralen with ultraviolet A-induced apoptosis of cutaneous lymphoma cell lines is augmented by type I interferons via the JAK1–STAT1 pathway. *Photodermatology Photoimmunology and Photomedicine*, 33:164–171, 2017.
- [137] Frederick A. Derheimer, J. Kevin Hicks, Michelle T. Paulsen, Christine E. Canman, and Mats Ljungman. Psoralen-induced DNA interstrand cross-links block transcription and induce p53 in an ataxia-telangiectasia and rad3-related-dependent manner. *Molecular Pharmacology*, 75:599–607, 2009.
- [138] S.-H. Kim, S. Peckler, B. Graves, D. Kanne, H. Rapoport, and J. E. Hearst. Sharp kink of DNA at psoralen-cross-link site deduced from crystal structure of psoralen-thymine monoadduct. *Cold Spring Harbor Symposia on Quantitative Biology*, 47:361–365, 1983.
- [139] Seth Peckler, Bradford Graves, David Kanne, Henry Rapoport, John E. Hearst, and Sung-Hou Kim. Structure of a psoralen-thymine monoadduct formed in photoreaction with DNA. *Journal of Molecular Biology*, 162:157–172, 1982.

- [140] Gary Wiesehahn and John E. Hearst. DNA unwinding induced by photoaddition of psoralen derivatives and determination of dark-binding equilibrium constants by gel electrophoresis. *Proceedings of the National Academy of Sciences of the United States of America*, 75:2703–2707, 1978.
- [141] Carl Veith Hanson, Che-Kun James Shen, and John E. Hearst. Cross-linking of DNA in situ as a probe for chromatin structure. *Science*, 193:62 – 64, 1976.
- [142] M. A. Pathak and D. M. Krämer. Photosensitization of skin in vivo by furocoumarins (psoralens). *Biochimica et Biophysica Acta - Nucleic Acids And Protein Synthesis*, 195:197–206, 1969.
- [143] Shiv Kumar, Chuanjuan Tao, Minchen Chien, Brittney Hellner, Arvind Balijepalli, Joseph W.F. Robertson, Zengmin Li, James J. Russo, Joseph E. Reiner, John J. Kasianowicz, and Jingyue Ju. PEG-labeled nucleotides and nanopore detection for single molecule DNA sequencing by synthesis. *Scientific Reports*, 2:1–8, 2012.
- [144] Kristin Brogaard, Liquan Xi, Ji-Ping Wang, and Jonathan Widom. A base pair resolution map of nucleosome positions in yeast. *Nature*, 486:496–501, 2012.
- [145] Thierry Douki, Anne Reynaud-Angelin, Jean Cadet, and Evelyne Sage. Bipyrimidine photoproducts rather than oxidative lesions are the main type of DNA damage involved in the genotoxic effect of solar UVA radiation. *Biochemistry*, 42:9221–9226, 2003.
- [146] Raymond D. Miller, Louise Prakash, and Satya Prakash. Genetic control of excision of *Saccharomyces cerevisiae* interstrand DNA cross-links induced by psoralen plus near-UV light. *Molecular and Cellular Biology*, 2:939–948, 1982.
- [147] M. Dardalhon and D. Averbeck. Pulsed-field gel electrophoresis analysis of the repair of psoralen plus UVA induced DNA photoadducts in *Saccharomyces cerevisiae*. *Mutation Research-DNA Repair*, 336:49–60, 1995.
- [148] Ross B. Greenberg, Marie Alberti, John E. Hearst, Mark A. Chua, and Wilma A. Saffran. Recombinational and mutagenic repair of psoralen interstrand cross-links in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 276:31551–31560, 2001.
- [149] J.A.P. Henriques and E. Moustacchi. Isolation and characterization of pso mutants sensitive to photo-addition of psoralen derivatives in *Saccharomyces cerevisiae*. *Genetics*, 95:273–288, 1980.
- [150] Iain A. Murray, Richard D. Morgan, Yvette Luyten, Alexey Fomenkov, Ivan R. Corrêa, Jr, Nan Dai, Mohammed B. Allaw, Xing Zhang, Xiaodong Cheng, and Richard J. Roberts. The non-specific adenine DNA methyltransferase M.EcoGII. *Nucleic Acids Research*, 46:840–848, 2017.
- [151] Masaki Okano, Daphne W. Bell, Daniel A. Haber, and En Li. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99:247–257, 1999.
- [152] Mehrnaz Fatemi, Andrea Hermann, Humaira Gowher, and Albert Jeltsch. Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. *European Journal of Biochemistry*, 269:4981–4984, 2002.

- [153] Paula M. Vertino, Jennifer A. Sekowski, Jennifer M. Coll, Nancy Applegren, Suhua Han, Robert J. Hickey, and Linda H. Malkas. DNMT1 is a component of a multiprotein DNA replication complex. *Cell cycle (Georgetown, Tex.)*, 1:416–423, 2002.
- [154] Susan E. Goelz, Bert Vogelstein, Stanley R. Hamilton, and Andrew P. Feinberg. Hypomethylation of DNA from benign and malignant human colon neoplasms. *Science*, 228:187–190, 1985.
- [155] Marie-France Robert, Steves Morin, Normand Beaulieu, France Gauthier, Ian C. Chute, Annie Barsalou, and A. Robert MacLeod. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nature Genetics*, 33:61–65, 2003.
- [156] Paul A. Ginno, Paul L. Lott, Holly C. Christensen, Ian Korf, and Frédéric Chédin. R-Loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Molecular Cell*, 45:814–825, 2012.
- [157] Takashi Sado, Martin H. Fenner, Seong-Seng Tan, Patrick Tam, Toshihiro Shioda, and En Li. X inactivation in the mouse embryo deficient for Dnmt1: Distinct effect of hypomethylation on imprinted and random X inactivation. *Developmental Biology*, 225:294–303, 2000.
- [158] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K. Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, Kazuhiro R. Nitta, Minna Taipale, Alexander Popov, Paul A. Ginno, Silvia Domcke, Jian Yan, Dirk Schübeler, Charles Vinson, and Jussi Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356, 2017.
- [159] Simcha Urieli-Shoval, Yosef Gruenbaum, and Aharon Razin. Sequence and substrate specificity of isolated DNA methylases from *Escherichia coli* C. *Journal of Bacteriology*, 153:274–280, 1983.
- [160] Gail E. Geier and Paul Modrich. Recognition sequence of the Dam methylase of *Escherichia coli* K12 and mode of cleavage of DpnI endonuclease. *Journal of Biological Chemistry*, 254:1408–1413, 1979.
- [161] Tyson A. Clark, Iain A. Murray, Richard D. Morgan, Andrey O. Kislyuk, Kristi E. Spittle, Matthew Boitano, Alexey Fomenkov, Richard J. Roberts, and Jonas Korlach. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Research*, 40:e29–undefined, 2012.
- [162] Gail E. Herman and Paul Modrich. *Escherichia coli* K-12 clones that overproduce Dam methylase are hypermutable. *Journal of Bacteriology*, 145:644–646, 1981.
- [163] M. G. Marinus, Anthony Poteete, and Judy A. Arraj. Correlation of DNA adenine methylase activity with spontaneous mutability in *Escherichia coli* K-12. *Gene*, 28:123–125, 1984.
- [164] Erik Boye and Anders Løbner-Olesen. The role of Dam methyltransferase in the control of DNA replication in *E. coli*. *Cell*, 62:981–989, 1990.

- [165] Joseph L. Campbell and Nancy Kleckner. *E. coli* oriC and the dnaA gene promoter are sequestered from Dam methyltransferase following the passage of the chromosomal replication fork. *Cell*, 62:967–979, 1990.
- [166] Ming X. Wang and George M. Church. A whole genome approach to in vivo DNA-protein interactions in *E. coli*. *Letters to Nature*, 360:606 – 610, 1992.
- [167] Cynthia Lark. Studies on the in vivo methylation of DNA in *Escherichia coli* 15T-. *Journal of Molecular Biology*, 31:389–399, 1968.
- [168] Joan E Brooks, Robert M Blumenthal, and Thomas R Gingeras. The isolation and characterization of the *Escherichia coli* dna adenine methylase (dam) gene. *Nucleic Acids Research*, 11:837–851, 1983.
- [169] Guoqiang Zhang, Hua Huang, Di Liu, Ying Cheng, Xiaoling Liu, Wenxin Zhang, Ruichuan Yin, Dapeng Zhang, Peng Zhang, Jianzhao Liu, Chaoyi Li, Baodong Liu, Yuewan Luo, Yuanxiang Zhu, Ning Zhang, Shunmin He, Chuan He, Hailin Wang, and Dahua Chen. N6-methyladenine DNA modification in *Drosophila*. *Cell*, 161:893–906, 2015.
- [170] Eric Lieberman Greer, Mario Andres Blanco, Lei Gu, Erdem Sendinc, Jianzhao Liu, David Aristizábal-Corrales, Chih Hung Hsu, L. Aravind, Chuan He, and Yang Shi. DNA methylation on N6-adenine in *C. elegans*. *Cell*, 161:868–878, 2015.
- [171] Ye Fu, Guan Zheng Luo, Kai Chen, Xin Deng, Miao Yu, Dali Han, Ziyang Hao, Jianzhao Liu, Xingyu Lu, Louis C. Doré, Xiaocheng Weng, Quanjiang Ji, Laurens Mets, and Chuan He. N6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell*, 161:879–892, 2015.
- [172] Stephen J. Mondo, Richard O. Dannebaum, Rita C. Kuo, Katherine B. Louie, Adam J. Bewick, Kurt LaButti, Sajeet Haridas, Alan Kuo, Asaf Salamov, Steven R. Ahrendt, Rebecca Lau, Benjamin P. Bowen, Anna Lipzen, William Sullivan, Bill B. Andreopoulos, Alicia Clum, Erika Lindquist, Christopher Daum, Trent R. Northen, Govindarajan Kunde-Ramamoorthy, Robert J. Schmitz, Andrii Gryganskyi, David Culley, Jon Magnuson, Timothy Y. James, Michelle A. O’Malley, Jason E. Stajich, Joseph W. Spatafora, Axel Visel, and Igor V. Grigoriev. Widespread adenine N6-methylation of active genes in fungi. *Nature Genetics*, 49:964–968, 2017.
- [173] Stanley Hattman, Carol Kenny, Lisbeth Berger, and Karen Pratt. Comparative study of DNA methylation in three unicellular eucaryotes. *Journal of Bacteriology*, 135:1156–1157, 1978.
- [174] Floriana Capuano, Michael Mülleder, Robert Kok, Henk J. Blom, and Markus Ralser. Cytosine DNA methylation is found in *Drosophila melanogaster* but absent in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and other yeast species. *Analytical Chemistry*, 86:3697–3702, 2014.
- [175] Stephan Diekmann. DNA methylation can enhance or induce DNA curvature. *The EMBO Journal*, 6:4213–4217, 1987.

- [176] P. Polaczek, K. Kwan, and J. L. Campbell. GATC motifs may alter the conformation of DNA depending on sequence context and N6-adenine methylation status: possible implications for DNA-protein recognition. *Molecular and General Genetics*, 258:488–493, 1998.
- [177] Chhaya W. Achwal, Chitra A. Iyer, and H. Sharat Chandra. Immunochemical evidence for the presence of 5mC, 6mA and 7mG in human, *Drosophila* and mealybug DNA. *FEBS Letters*, 158:353–358, 1983.
- [178] Chuan-Le Xiao, Song Zhu, Minghui He, De Chen, Qian Zhang, Ying Chen, Guoliang Yu, Jinbao Liu, Shang-Qian Xie, Feng Luo, Zhe Liang, De-Peng Wang, Xiao-Chen Bo, Xiao-Feng Gu, Kai Wang, and Guang-Rong Yan. N6-methyladenine DNA modification in the human genome. *Molecular Cell*, 71:306–318.e7, 2018.
- [179] F. Zheng, D. Tang, H. Xu, Y. Xu, W. Dai, X. Zhang, X. Hong, D. Liu, and Y. Dai. Genome-wide analysis of 6-methyladenine DNA in peripheral blood mononuclear cells of systemic lupus erythematosus. *Lupus*, 28:359–364, 2019.
- [180] Karolos Douvlataniotis, Maïke Bensberg, Antonio Lentini, Björn Gylemo, and Colm E. Nestor. No evidence for DNA N6-methyladenine in mammals. *Science Advances*, 6:1–10, 2020.
- [181] Zach K. O’Brown, Konstantinos Boulias, Jie Wang, Simon Yuan Wang, Natasha M. O’Brown, Ziyang Hao, Hiroki Shibuya, Paul-Enguerrand Fady, Yang Shi, Chuan He, Sean G. Megason, Tao Liu, and Eric L. Greer. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics*, 20:1–15, 2019.
- [182] Ziyang Hao, Tong Wu, Xiaolong Cui, Pingping Zhu, Caiping Tan, Xiaoyang Dou, Kai-Wen Hsu, Yueh-Te Lin, Pei-Hua Peng, Li-Sheng Zhang, Yawei Gao, Lulu Hu, Hui-Lung Sun, Allen Zhu, Jianzhao Liu, Kou-Juey Wu, and Chuan He. N6-deoxyadenosine methylation in mammalian mitochondrial DNA. *Molecular Cell*, 78:382–395.e8, 2020.
- [183] Gang Fang, Diana Munera, David I. Friedman, Anjali Mandlik, Michael C. Chao, Onureena Banerjee, Zhixing Feng, Bojan Losic, Milind C. Mahajan, Omar J. Jabado, Gintaras Deikus, Tyson A. Clark, Khai Luong, Iain A. Murray, Brigid M. Davis, Alona Keren-Paz, Andrew Chess, Richard J. Roberts, Jonas Korlach, Steve W. Turner, Vipin Kumar, Matthew K. Waldor, and Eric E. Schadt. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nature Biotechnology*, 30:1232–1239, 2012.
- [184] Theresa K. Kelly, Yaping Liu, Fides D. Lay, Gangning Liang, Benjamin P. Berman, and Peter A. Jones. Genome-wide mapping of nucleosome positioning and DNA methylation within individual dna molecules. *Genome Research*, 22:2497–2506, 2012.
- [185] Andrew B. Stergachis, Brian M. Debo, Eric Haugen, L. Stirling Churchman, and John A. Stamatoyannopoulos. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368:1449–1454, 2020.

- [186] Nour J. Abdulhay, Colin P. McNally, Laura J. Hsieh, Sivakanthan Kasinathan, Aidan Keith, Laurel S. Estes, Mehran Karimzadeh, Jason G. Underwood, Hani Goodarzi, Geeta J. Narlikar, and Vijay Ramani. Massively multiplex single-molecule oligonucleosome footprinting. *eLife*, 9:1–23, 12 2020.
- [187] Michael Weber, Jonathan J. Davies, David Wittig, Edward J. Oakeley, Michael Haase, Wan L. Lam, and Dirk Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37:853–862, 2005.
- [188] Thomas A. Down, Vardhman K. Rakyen, Daniel J. Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Gräf, Nathan Johnson, Javier Herrero, Eleni M. Tomazou, Natalie P. Thorne, Liselotte Bäckdahl, Marlis Herberth, Kevin L. Howe, David K. Jackson, Marcos M. Miretti, John C. Marioni, Ewan Birney, Tim J. P. Hubbard, Richard Durbin, Simon Tavaré, and Stephan Beck. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*, 26:779–785, 2008.
- [189] Nancy H. Nabili, Loic P. Deleyrolle, Russell P. Darst, Alberto Riva, Brent A. Reynolds, and Michael P. Kladde. Multiplex mapping of chromatin accessibility and DNA methylation within targeted single molecules identifies epigenetic heterogeneity in neural stem cells and glioblastoma. *Genome Research*, 24:329–339, 2 2014.
- [190] Arnaud R. Krebs, Dilek Imanci, Leslie Hoerner, Dimos Gaidatzis, Lukas Burger, and Dirk Schübeler. Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Molecular Cell*, 67:411–422.e4, 2017.
- [191] Stephen J. Clark, Ricard Argelaguet, Chantierint-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle, and Wolf Reik. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9:1–9, 2018.
- [192] Yunhao Wang, Anqi Wang, Zujun Liu, Andrew L. Thurman, Linda S. Powers, Meng Zou, Yue Zhao, Adam Hefel, Yunyi Li, Joseph Zabner, and Kin Fai Au. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Research*, 2019.
- [193] Elisa Oberbeckmann, Michael Wolff, Nils Krietenstein, Mark Heron, Jessica L. Ellins, Andrea Schmid, Stefan Krebs, Helmut Blum, Ulrich Gerland, and Philipp Korber. Absolute nucleosome occupancy map for the *Saccharomyces cerevisiae* genome. *Genome Research*, 29:1996–2009, 2019.
- [194] Isac Lee, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J. Sedlazeck, Kasper D. Hansen, Jared T. Simpson, and Winston Timp. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods*, 17:1191–1199, 12 2020.
- [195] Peng Ni, Neng Huang, Zhi Zhang, De-Peng Wang, Fan Liang, Yu Miao, Chuan-Le Xiao, Feng Luo, and Jianxin Wang. DeepSignal: detecting DNA methylation state

- from Nanopore sequencing reads using deep-learning. *Bioinformatics*, 35:4586–4595, 2019.
- [196] Alexa B. R. McIntyre, Noah Alexander, Kirill Grigorev, Daniela Bezdan, Heike Sichtig, Charles Y. Chiu, and Christopher E. Mason. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nature Communications*, 10:1–11, 2019.
- [197] Qian Liu, Daniela C. Georgieva, Dieter Egli, and Kai Wang. NanoMod: A computational tool to detect DNA modifications using Nanopore long-read sequencing data. *BMC Genomics*, 20(Suppl 1), 2019.
- [198] Huanle Liu, Oguzhan Begik, Morghan C. Lucas, Jose Miguel Ramirez, Christopher E. Mason, David Wiener, Schraga Schwartz, John S. Mattick, Martin A. Smith, and Eva Maria Novoa. Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications*, 10:1–9, 2019.
- [199] Qian Liu, Li Fang, Guoliang Yu, Depeng Wang, Chuan-Le Xiao, and Kai Wang. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications*, 10, 2019.
- [200] Piroon Jenjaroenpun, Thidathip Wongsurawat, Taylor D. Wadley, Trudy M. Wassenaar, Jun Liu, Qing Dai, Visanu Wanchai, Nisreen S. Akel, Azemat Jamshidi-Parsian, Aime T. Franco, Gunnar Boysen, Michael L. Jennings, David W. Ussery, Chuan He, and Intawat Nookaew. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Research*, 49:1–13, 2020.
- [201] Liu Xu and Masahide Seki. Recent advances in the detection of base modifications using the Nanopore sequencer. *Journal of Human Genetics*, 2019.
- [202] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex DeWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, 2009.
- [203] Benjamin A Flusberg, Dale Webster, Jessa Lee, Kevin Travers, Eric Olivares, Tyson A. Clark, Jonas Korlach, and Stephen W Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7:461–465, 2010.
- [204] Matthew Ouellette, J. Peter Gogarten, Jessica Lajoie, Andrea M. Makkay, and R. Thane Papke. Characterizing the DNA methyltransferases of *Haloferax volcanii* via bioinformatics, gene deletion, and SMRT sequencing. *Genes*, 9, 2018.

- [205] Andreas E. Zautner, Anne-Marie Goldschmidt, Andrea Thürmer, Jörg Schuldes, Oliver Bader, Raimond Lugert, Uwe Groß, Kerstin Stingl, Gabriela Salinas, and Thomas Lingner. SMRT sequencing of the *Campylobacter coli* BfR-CA-9557 genome sequence reveals unique methylation motifs. *BMC Genomics*, 16:1–12, 2015.
- [206] Samantha R. Coy, Eric R. Gann, Spiridon E. Papoulis, Michael E. Holder, Nadim J. Ajami, Joseph F. Petrosino, Erik R. Zinser, James L. Van Etten, and Steven W. Wilhelm. SMRT sequencing of *Paramecium Bursaria* Chlorella Virus-1 reveals diverse methylation stability in adenines targeted by restriction modification systems. *Frontiers in Microbiology*, 11:1–15, 2020.
- [207] Mohamad R. Abdul Sater, Araceli Lamelas, Guilin Wang, Tyson A. Clark, Katharina Röltgen, Shrikant Mane, Jonas Korlach, Gerd Pluschke, and Christoph D. Schmid. DNA methylation assessed by SMRT sequencing is linked to mutations in *Neisseria meningitidis* isolates. *PLoS ONE*, 10:1–19, 2015.
- [208] Brian M. Forde, Lauren J. McAllister, James C. Paton, Adrienne W. Paton, and Scott A. Beatson. SMRT sequencing reveals differential patterns of methylation in two O111:H-*STEC* isolates from a hemolytic uremic syndrome outbreak in Australia. *Scientific Reports*, 9:1–11, 2019.
- [209] Benjamin J. Hindson, Kevin D. Ness, Donald A. Masquelier, Phillip Belgrader, Nicholas J. Heredia, Anthony J. Makarewicz, Isaac J. Bright, Michael Y. Lucero, Amy L. Hiddessen, Tina C. Legler, Tyler K. Kitano, Michael R. Hodel, Jonathan F. Petersen, Paul W. Wyatt, Erin R. Steenblock, Pallavi H. Shah, Luc J. Bousse, Camille B. Troup, Jeffrey C. Mellen, Dean K. Wittmann, Nicholas G. Erndt, Thomas H. Cauley, Ryan T. Koehler, Austin P. So, Simant Dube, Klint A. Rose, Luz Montesclaros, Shenglong Wang, David P. Stumbo, Shawn P. Hodges, Steven Romine, Fred P. Milanovich, Helen E. White, John F. Regan, George A. Karlin-Neumann, Christopher M. Hindson, Serge Saxonov, and Bill W. Colston. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, 83:8604–8610, 2011.
- [210] Dzmitry G. Batrakou, Emma D. Heron, and Conrad A. Nieduszynski. Rapid high-resolution measurement of DNA replication timing by droplet digital PCR. *Nucleic Acids Research*, 46, 2018.
- [211] John K. Wiencke, Paige M. Bracci, George Hsuang, Shichun Zheng, Helen Hansen, Margaret R. Wrensch, Terri Rice, Melissa Eliot, and Karl T. Kelsey. A comparison of DNA methylation specific droplet digital PCR (ddPCR) and real time qPCR with flow cytometry in characterizing human T cells in peripheral blood. *Epigenetics*, 9:1360–1365, 2014.
- [212] Max Felle, Helen Hoffmeister, Julia Rothhammer, Andreas Fuchs, Josef H. Exler, and Gernot Längst. Nucleosomes protect DNA from DNA methylation in vivo and in vitro. *Nucleic Acids Research*, 39:6956–6969, 2011.
- [213] Robert T. Simpson and Darrell W. Stafford. Structural features of a phased nucleosome core particle. *Proceedings of the National Academy of Sciences of the United States of America*, 80:51–55, 1983.

- [214] W. F. Brandt and C. von Holt. The occurrence of histone h3 and h4 in yeast. *FEBS Letters*, 65:386–390, 1976.
- [215] Peter Baum, Jeremy Thorner, and Lawrence Honig. Identification of tubulin from the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, 75:4962–4966, 1978.
- [216] Katarzyna Hac-Wydro and Michał Flasiński. The studies on the toxicity mechanism of environmentally hazardous natural (IAA) and synthetic (NAA) auxin - The experiments on model *Arabidopsis thaliana* and rat liver plasma membranes. *Colloids and Surfaces B: Biointerfaces*, 130:53–60, 6 2015.
- [217] Leland H. Hartwell. Sequential function of gene products relative to DNA synthesis in the yeast cell cycle. *Journal of Molecular Biology*, 104:803–817, 1976.
- [218] ENCODE. Chromatin patterns at transcription factor binding sites. *Nature*, 2019.
- [219] Chenxi Qiu, Chenxi Qiu, Huiyan Jin, Irina Vvedenskaya, Irina Vvedenskaya, Jordi Abante Llenas, Jordi Abante Llenas, Tingting Zhao, Indranil Malik, Indranil Malik, Alex M. Visbisky, Scott L. Schwartz, Ping Cui, Pavel Čabart, Pavel Čabart, Kang Hoo Han, William K.M. Lai, William K.M. Lai, Richard P. Metz, Charles D. Johnson, Sing Hoi Sze, Sing Hoi Sze, B. Franklin Pugh, B. Franklin Pugh, Bryce E. Nickels, Bryce E. Nickels, and Craig D. Kaplan. Universal promoter scanning by Pol II during transcription initiation in *Saccharomyces cerevisiae*. *Genome Biology*, 21, 6 2020.
- [220] Eliza C. Small, Liqun Xi, Ji-Ping Wang, Jonathan Widom, and Jonathan D. Licht. Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, 111:E2462–71, 2014.
- [221] Robert Haselkorn and Paul Doty. The reaction of formaldehyde with polynucleotides. *Journal of Biological Chemistry*, 236:2738–2745, 1961.
- [222] Hikoya Hayatsu, Yasuhiro Yamashita, Seiko Yui, Yuriko Yamagata, Ken ichi Tomita, and Kazuo Negishi. N-sulfomethylation of guanine, adenine and cytosine with formaldehyde-bisulfite. A selective modification of guanine in DNA. *Nucleic Acids Research*, 10:6281–6294, 1982.
- [223] H. Fraenkel-Conrat. Reaction of nucleic acid with formaldehyde. *Biochimica et Biophysica Acta*, 15:307–309, 1954.
- [224] Mark J. Solomon and Alexander Varshavsky. Formaldehyde-mediated DNA-protein crosslinking: A probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences of the United States of America*, 82:6470–6474, 1985.
- [225] Hideyuki Takeshima, Isao Suetake, and Shoji Tajima. Mouse Dnmt3a preferentially methylates linker DNA and is inhibited by histone H1. *Journal of Molecular Biology*, 383:810–821, 11 2008.

-
- [226] James D. Engel and Peter H. von Hippel. Effects of methylation on the stability of nucleic acid conformations: Studies at the monomer level. *Biochemistry*, 13:4143–4158, 1974.
- [227] James Douglas Engel and Peter H. von Hippel. Effects of methylation on the stability of nucleic acid conformations: Studies at the polymer level. *The Journal of Biological Chemistry*, 253:927–934, 1978.
- [228] Piotr Polaczek, Kelvin Kwan, David A. Liberles, and Judith L. Campbell. Role of architectural elements in combinatorial regulation of initiation of DNA replication in *Escherichia coli*. *Molecular Microbiology*, 26:261–275, 1997.

Appendix A

Code for producing figures

In this appendix the commands to produce the graphs are shown. Custom scripts can be found at https://github.com/HJeffery/Thesis_supplementary where the scripts are organised based on the chapter and topic they were used for. All scripts have been tested with python version 3.8.6 or R version 4.0.2. Figures were compiled and formatted in Inkscape v1.0.1.

For any analysis using ACS data from Eaton *et al* (2010) [7] the data used was the mid-points of each ACS which was calculated using the ‘acs_mid_points.py’ script.

A.1 Chapter 3

A.1.1 Tables

Table 3.2

The number of dinucleotide sites was calculated with ‘dinucleotide_prevalence.py’. The number of nucleosomes was calculated by the number of lines in the nucleosome bed file. The mean linker length and number of linkers were calculated in the ‘linker_lengths.r’ script. The mean length of linkers with an ACS was calculated with ‘linker_lengths_with_acs.py’.

The ‘violin_plots_dinucleotide_nucleosomes.r’ script was used to calculate the mean number of sites per nucleosome. The mean number of dinucleotide sites per linker was calculated with ‘violin_plots_dinucleotide_linkers.r’. The number of linkers with specific TA or AT sites were extracted from the file produced from the ‘sites_in_linkers.py’ script.

Table 3.4

The nanopore data was basecalled with Albacore v2.0.2. The data was barcoded, therefore it was demultiplexed using ‘demultiplex.py’. To assess different levels of coverage the ‘ref_cover’ variable in the ‘demultiplex.py’ script was edited to either 0.15, 0.3 or 0.8 for 15%, 30% and 80% coverage of the reference sequence respectively. Read counts were generated using the following command.

```
samtools view file.bam | wc -l
```

Table 3.5

Data was processed as in the section for figure 3.20.

A.1.2 Figures

Figures 3.1 and 3.3

The chemical structures in these figures were drawn using ChemDraw v19.0.

Figure 3.4

Nucleosome sites of 147 bases and linker regions were generated based on nucleosome dyad positions from Jiang *et al* (2009) [89] using the ‘make_nucleosomes_and_linkers.py’ script. Linkers of length 0 (where the nucleosomes touch) were removed with the following command.

```
awk '{if ($2 != $3) {print}}' 2009_Jiang_linkers_sacCer3_initial.bed  
> 2009_Jiang_linkers_sacCer3.bed
```

Bed files containing dinucleotide sites were generated using ‘dinucleotide_prevalence.py’ and subsequently sorted using bedtools sort. For panel a the prevalence of each dinucleotide in either nucleosome or linker regions was identified by using bedtools intersect to identify which genomic context each dinucleotide was in followed by counting the number of lines in each file and taking this as a percentage of the total number of sites in the original dinucleotide file. These percentages were plotted with ‘dinucleotide_distribution_genomic_regions.r’.

```
bedtools sort -i TA_positions_sacCer3_noM.bed  
> TA_positions_sacCer3_noM.sorted.bed
```

```
bedtools intersect -a TA_positions_sacCer3_noM.sorted.bed  
-b 2009_Jiang_nucleosomes_sacCer3.bed -wo  
> TA_in_nucleosomes.bedtools_intersect
```

```
bedtools intersect -a TA_positions_sacCer3_noM.sorted.bed  
-b 2009_Jiang_linkers_sacCer3.bed -wo -v  
> TA_in_linkers.bedtools_intersect
```

```
cat TA_in_nucleosomes.bedtools_intersect | wc -l
```

```
cat TA_in_linkers.bedtools_intersect | wc -l
```

-v = Only report those entries in A that have no overlap in B.

-wo = Write the original A and B entries plus the number of base pairs of overlap between the two features. Only A features with overlap are reported.

For panel b the bedtools intersect files produced below were used as inputs for plotting with ‘violin_plot_dinucleotide_nucleosomes.r’.

```
bedtools intersect -a 2009_Jiang_nucleosomes_sacCer3.bed
-b TA_positions_sacCer3_noM.sorted.bed -wo
> TA_in_nucleosomes.bedtools_intersect
```

```
bedtools intersect -a 2009_Jiang_linkers_sacCer3.bed
-b TA_positions_sacCer3_noM.sorted.bed -wao
> TA_in_linkers.bedtools_intersect
```

-wao = Write the original A and B entries plus the number of base pairs of overlap between the two features. Write all features in A to file even if no overlap is found.

For panel c the dinucleotide combination (TA and AT, or TA, AT, TG and GT) bed files needed to be filtered to prevent T’s being analysed in multiple sequence contexts as for example a T can be in both a TA and an AT site. They were subsequently re-sorted, and the distance to the nearest two nucleosome dyads calculated (although only distances to the nearest nucleosome dyad were plotted in panel c). The data was plotted with ‘violin_plot_dinucleotide_to_nucleosome_dyad_distances.r’.

```
uniq -u TA_AT_positions_sacCer3_noM.bed
> TA_AT_unique_positions_sacCer3_noM.bed
```

```
bedtools sort -i TA_AT_unique_positions_sacCer3_noM.bed
> TA_AT_unique_positions_sacCer3_noM.sorted.bed
```

```
bedtools closest -a TA_AT_unique_positions_sacCer3_noM.sorted.bed
-b 2009_Jiang_nucleosome_dyad_positions_sacCer3.sorted.bed -d
-t first -k 2 > TA_AT_distance_nucleosome-dyads.bedtools_closest
```

Figure 3.5

For panel a the linker lengths were calculated using ‘linker_sizes.py’ and plotted using ‘linker_lengths.r’. The prevalence of TA and AT sites in linkers was calculated using ‘sites_in_linkers.py’. Then ‘bubble_plots_linkers_nucleosomes.r’ was used to plot panel b.

The linkers containing an ACS were identified using the following bedtools intersect command. The graph in panel c was then generated using ‘ACS_in_linkers.r’

```
bedtools intersect
-a GSM424494_Eaton_2010_acs_locations_midpoints_sacCer3_G2_23degrees_WT.
sorted.bed
-b 2009_Jiang_linkers_sacCer3.bed -wao
> ACS_Eaton_in_linkers_Jiang.bedtools_intersect
```

Figure 3.6

For identifying ACS sites in linker regions, the bed file containing ACS positions from Eaton *et al* (2010) [7] were sorted and the distance from each ACS to the closest nucleosome dyad upstream (-fu argument) and downstream (-fd argument) was calculated using bedtools closest.

```
bedtools sort -i GSM424494_Eaton_2010_acs_locations_sacCer3_G2_23degrees
_WT.bed
> GSM424494_Eaton_2010_acs_locations_sacCer3_G2_23degrees_WT.sorted.bed

bedtools closest -a GSM424494_Eaton_2010_acs_locations_midpoints_sacCer3
_G2_23degrees_WT.sorted.bed
-b 2009_Jiang_nucleosome_dyad_positions_sacCer3.sorted.bed -D a -fu
> ACS_Eaton_nucleosome_dyads_Jiang_upstream_distances.bedtools_closest
```

```
bedtools closest -a GSM424494_Eaton_2010_acs_locations_midpoints_sacCer3
_G2_23degrees_WT.sorted.bed
-b 2009_Jiang_nucleosome_dyad_positions_sacCer3.sorted.bed -D a -fd
> ACS_Eaton_nucleosome_dyads_Jiang_downstream_distances.bedtools_closest
```

-D a = Signed distance taking into consideration strand of ACS (negative will be upstream of ACS, positive will be downstream of ACS).

A bed file containing linkers with an ACS was generated with 'get_linkers_with_acs.py'. A bedtools intersect command was then used to find the dinucleotide positions within the linkers.

```
bedtools intersect -a linkers_containing_ACS_Eaton.bed
-b GSM424494_Eaton_2010_acs_locations_midpoints_sacCer3_G2_23degrees
_WT.sorted.bed
TA_positions_sacCer3_noM.sorted.bed
AT_positions_sacCer3_noM.sorted.bed
-wao -names ACS TA AT
> TA_AT_sites_in_linkers_with_ACS_Eaton.bedtools_intersect
```

-names = Allows identification of the file from which each match comes from by making a new column of ACS/TA/AT in the bed file created.

-wao = Add information on positions of each a or b file in the output file.

The ACS positions were extracted from the ACS ID's using 'get_acs_pos.py'. This figure was plotted with 'TA_AT_in_linkers_with_ACS_plot.r'

Figure 3.19

The nanopore data was basecalled with Albacore v2.0.2. The data was barcoded, therefore it was demultiplexed using 'demultiplex.py'.

To calculate and plot coverage, indels and mismatches the ‘mpileup_nanopore_qc_analysis.r’ script was used.

Figure 3.20

The reference sequences (forward and reverse) were indexed with BWA v0.7.12.

```
bwa index TA_crosslink_TMP.fasta
bwa index TA_crosslink_TMP_revComp.fasta
```

A fastq file was generated from the basecalled nanopore sequencing data using the following command.

```
sudo read_fast5_basecaller.py -i F_reads -t 1 -s F_reads_fastq
-c r95_450bps_linear.cfg -r -o fastq
```

The fastq was then indexed with nanopolish v0.13.2 and both the fastq sequences and the events were aligned to the reference, with BWA and nanopolish eventalign respectively.

```
nanopolish index -d F_reads F_reads_combined_pass+fail.fastq

bwa mem TA_crosslink_TMP.fasta F_reads_combined_pass+fail.fastq
> F_reads_bwa_mem_aligned.bam

nanopolish eventalign -r F_reads_combined_pass+fail.fastq
-b F_reads_bwa_mem_aligned_sorted.bam
-g TA_crosslink_TMP.fasta
--scale-events
--print-read-names > nanopolish_F_reads.txt
```

Rows with a k-mer that could not be determined were represented by a ‘NNNNNN’ as the ‘ONT model predicted k-mer’ and were removed from the data.

```
awk '$10 !~ /'NNNNNN'/' nanopolish_F_reads.txt > nanopolish_F_reads  
_noNNNNNN.txt
```

```
awk '$10 !~ /'NNNNNN'/' nanopolish_R_reads.txt > nanopolish_R_reads  
_noNNNNNN.txt
```

The R script used to plot panels a, b, d and e was ‘nanopolish_graph.r’. The R script used to plot panels c and f was ‘NNNNNN_distribution.r’.

A.2 Chapter 4

All scripts used to analyse nanopore sequencing data in this chapter filter out reads that are unmapped, secondary alignments or supplementary alignments. Reads must also have a mapping quality above 0. For all analysis only the forward strands were analysed.

A.2.1 Tables

Table 4.4

Read numbers were calculated as follows. Total number of reads per barcode:

```
samtools view -c sample.sorted.bam
```

Number of primary alignment mapped reads:

```
samtools view -F 0x904 -c sample.sorted.bam
```

Excluded flags are 800 = supplementary alignment, 100 = secondary alignment, 4 = segment unmapped.

Number of primary alignment mapped reads with quality > 0:

```
samtools view -F 0x904 sample.sorted.bam | awk -F"\t" '\$5>0' | wc -l
```

Column 5 of SAM file = MAPQ (mapping quality). This describes the uniqueness of the alignment; 0 = non-unique.

Mean read length (used length of sequence (column 10 of bam file) for length of each read, analysed only primary alignments with mapping quality > 0):

```
samtools view -F 0x904 sample.sorted.bam | awk -F"\t" '\$5>0' |  
awk 'BEGIN{FS="\t"; sum=0; count=0} {sum += length(\$10); count += 1}  
END {print sum/count}'
```

Table 4.6

See section for table 4.4.

Table 4.7

See section for table 4.4.

A.2.2 Figures

Figure 4.3

Panel a and b were generated with the BioRad Quantasoft software. The R script used to plot panel c was 'ddPCR_pUC19_optimisation.r'.

Figure 4.4

Panel a was generated with the BioRad Quantasoft software. The R script used to plot panel b was 'ddPCR_pUC19_digestion_tests.r'

Figure 4.5

The R script used to plot panel a was ‘ddPCR_pUC19_MboI.r’. The R script used to plot panel b was ‘ddPCR_pUC19_SphI-HF.r’.

Figure 4.6

The R script used to plot this figure was ‘gDNA_replicates_effect_of_methylation.r’.

Figure 4.7

The R script used to plot this figure was ‘ddPCR_gDNA_test.r’.

Figure 4.8

The R script used to plot panel a and b was ‘ddPCR_gDNA_graph.r’

Figure 4.9

Benchling was used to produce this plasmid representation.

Figure 4.10

The R script used to plot this figure was ‘ddPCR_pUC19_AluI.r’.

Figure 4.11

The R script used to plot panel a and b was ‘ddPCR_recombinant_nucleosomes_graph.r’.

Figure 4.12

The R script used to plot this figure was ‘ddPCR_chromatin.r’.

Figure 4.13

The R script used to plot panel a and b was ‘ddPCR_spheroplasts.r’.

Figure 4.14

The R script used to plot panel a was ‘ddPCR_pUC19_chromatin_spike-in.r’. The R script used to plot panel b was ‘ddPCR_chromatin_pUC19_spike-in.r’.

Figure 4.17

The R script used to plot this figure was ‘5mC_mass_spec.r’.

Figure 4.18

The data was analysed with Tombo using three commands (tombo preprocess, tombo resquiggle and tombo detect_modifications_de_novo). These were run from a bash script which can be found in the GitHub repository, for which one example is the tombo_6mA.sh file. This outputs a ‘barcodeX_per_read_stats.de_novo.tombo.per_read_stats’ file which contains single-molecule information. This needed to be run on each sample separately.

For post-processing and plotting SMAC-sequencing scripts from the Greenleaf lab were used. These can be found here: <http://mitra.stanford.edu/kundaje/marinovg/oak/code/bisulfite/>

A few alterations were required to run with python3. The ‘from sets import Set’ line from methylation-reads-tsv-to-coverage.py script was removed, and some type errors were corrected within the same script. The updated script can be found in the GitHub repository.

The commands to run this analysis were as follows and these were run once for the 6mA data and once for the 5mC data.

Step 1 - This step extracts the tombo results for Tombo v1.5 or greater. In this step barcodeX is used as the output file prefix.

```
python /path/to/TomboSingleReadsExtract-tombo_de_novo-1.5.py
/path/to/6mA_barcodeX_per_read_stats.de_novo.tombo.per_read_stats
sacCer3.fa
barcodeX
-m6A-only or -GC-only
```

Step 2 - This step performs Bayesian aggregation and the barcodeX.0.tombo.m6A-GC-CG-only.reads.BI_w10_a10_b10.tsv will have the reads with the aggregated base calls.

```
python3 /path/to/methylation-reads-tsv-to_coverage.py
barcodeX.reads.tsv 0.5
barcodeX.reads.tsv.min_p_val0.4.cutoff_0.5.coverage.BI_w10_a10_b10
-minAbsPValue 0.4
-BayesianIntegration 10 1 10 10 50
-saveNewSingleMoleculeFile barcodeX.reads.BI_w10_a10_b10.tsv
```

Step 3 - This step sorts and compresses the tsv files and tidies up the workspace by deleting files that are no longer needed. It also indexes the compressed tsv files for fast access.

```
cat barcodeX*.m6A-GC-CG-only.reads.BI_w10_a10_b10.tsv |
sort -k1,1 -k2,2n -k3,3n |
bgzip > barcodeX.m6A-GC-CG-only.BI_w10_a10_b10.reads.tsv.bgz &

cat barcodeX*.m6A-GC-CG-only.reads.tsv |
sort -k1,1 -k2,2n -k3,3n |
bgzip > barcodeX.m6A-GC-CG-only.reads.tsv.bgz &

rm barcodeX*.m6A-GC-CG-only.reads.tsv
```

```
rm *.m6A-GC-CG-only.reads.BI_w10_a10_b10.tsv

tabix -s 1 -b 2 -e 3 barcodeX.m6A-GC-CG-only.reads.tsv.bgz &
tabix -s 1 -b 2 -e 3 barcodeX.m6A-GC-CG-only.BI_w10_a10_b10.reads.tsv
.bgz &
```

Step 4 - This step produces single molecule plots using hierarchical global clustering. It requires a bed file for the regions of interest.

- window = Number of bases combined into a single dot
- resize = The size of the plot (which will need to be changed for smaller/bigger regions)
- binarise = The threshold for classifying a site as methylated or not.

```
python3 /path/to/SMAC-footprints-from-methylation-reads-tsv-tabix.py
barcodeX.reads.BI_w10_a10_b10.tsv.bgz
regions.bed 0 1 2 3
/path/to/nanopolish-10.17/htslib/tabix
barcodeX.reads.BI_w10_a10_b10.10bp.resize5.window10.binarise0.5.HGC
-window 10 -minCov 1 -deleteMatrix -binarize 0.5
-scatterPlot gist_heat 0 1.1 w -resize 2 -unstranded
```

Figure 4.22

To calculate the overall methylation level at GATC sites, the analysis was run in the same way as 4-mer analysis for figure 4.23, with the GATC overall methylation level being plotted at a 50% methylation threshold.

Figure 4.23

To calculate 6mA probabilities for seven 4-mers (GATC, GACC, AATC, CATC, GATG, GATT, TATC) the following command was run. The reference genome was sacCer3. For

4-mer analysis, 1,000 sites of each kmer were analysed in 50,000 randomly selected reads for each sample to increase the variety of genomic contexts analysed.

```
python3 m6A_EcoGII_4-mer_analysis_pysam.py
-p /path/to/protobufs/
-g /path/to/genome.fasta
-o /path/to/outfile.tsv
-a /path/to/barcodeX_alignments.sorted.bam
-b barcodeX
```

If non-barcoded the `-b` argument was replaced by `-n True`.

To run this on multiple barcodes a bash script can be used as in the following example:

```
#!/bin/bash

set -e

for i in {X..Y};
do
echo "Starting barcode $i"

python3 m6A_EcoGII_4-mer_analysis_pysam.py
-p /path/to/protobufs/
-g /path/to/genome.fasta
-o /path/to/outfile_barcode"$i".tsv
-a /path/to/barcode"$i"_alignments.sorted.bam
-b barcode"$i"
done
```

X and Y are integers representing the first and last barcodes to be analysed. If a barcode (e.g. Z) is missing within the continuous list the following can be inserted between the 'do' and 'echo' lines of the above command.

```
if [ $i -eq Z ]  
    then  
        continue  
fi
```

Figure 4.24

Same analysis as section for figure 4.23 but with basecalling performed with Guppy v3.6.0 and using a Rerio model: `res_dna_r941_min_modbases-all-context_v001.cfg`

Figures 4.25 and 4.26

The methylation probability for every adenine in a GATC site on primary alignments mapped to the forward strand was extracted using a custom python script. Reads were also filtered on a mapping quality above zero. A bed file of chr, start, end, name (read ID), score (6mA value normalised to between 0 and 100), strand is produced, in which each row will be an adenine at a specific position in a specific read.

```
python3 analyse_A_GATC_forward_reads.py  
-p /path/to/protobufs/  
-b /path/to/barcodeX_alignments.sorted.bam  
-o /path/to/outfile.bed  
-e barcodeX  
-g /path/to/genome.fasta  
-s GATC
```

The resulting bed file was sorted.

```
bedtools sort -i input.bed > output.sorted.bed
```

The distances to nucleosome dyads (where the direction does not matter as nucleosomes are symmetrical around the dyad positions) [89] were calculated using bedtools closest.

```
bedtools closest  
-a output.sorted.bed  
-b 2009_Jiang_nucleosome_dyad_positions_sacCer3.sorted.bed  
-d -t first  
> barcodeX_GATC_Jiang_nucleosomes.bedtools_closest
```

Awk commands were subsequently used to filter the files based on distance prior to plotting in R. For nucleosomes the distances are not directional, therefore only one filter (less than 150 bp) is required. For single molecule analysis of Dam methyltransferase data 1 kb upstream and 1 kb downstream was used.

```
awk '{ if ($10 <= 150) { print } }'  
barcodeX_GATC_Jiang_nucleosomes.bedtools_closest >  
barcodeX_GATC_Jiang_nucleosomes_filtered_150bp_distance.bedtools_closest
```

In the bedtools closest command, the ‘-d’ indicates that the distance between sites in the two files will be calculated and ‘-t first’ means that, in the case of a tie, the first one will be chosen.

Meta-analysis, performed within R, involved grouping by distance and barcode, and plotting either the mean methylation at each point or the percentage of sites above a 50% methylation threshold. The R script used for this was ‘nucleosomes_meta-analysis.r’.

Single-molecule analysis, performed within R, involved filtering the data for a specific origin, calculating the maximum number of GATC sites within the region for each barcode,

filtering out reads that did not include all sites, and plotting these. The R script used for this was ‘acs_single_molecule_plots.r’

Figure 4.28

For overall methylation analysis, 100,000 adenines and 100,000 cytosines on the forward strand were analysed from 10,000 randomly selected reads. This analysis outputs the methylation level at a variety of thresholds (10 - 90% with 10% steps). The same python script was run for analysis of overall methylation in all M.EcoGII datasets (*in vitro* and *in vivo*).

```
python3 m6A_EcoGII_overall_analysis_pysam.py
-p /path/to/protobufs/
-g /path/to/genome.fasta
-a /path/to/barcodeX_alignments.sorted.bam
-b barcodeX
-o /path/to/outfile.tsv
-s 100000
```

s = Number of sites to be analysed

Figure 4.29

See section for figure 4.23.

Figure 4.30

To analyse 6-mers, any 6-mer with at least one adenine was analysed. If the 6-mer had more than one adenine, a random adenine was selected each time this 6-mer was encountered so that for each 6-mer the methylation status of one adenine was evaluated. 6-mers were

analysed in forward strands only in a random selection of 50,000 reads. A bash script like the following was used to analyse multiple barcodes efficiently. The basis of the script could be used for running any of the nanopore sequencing data analysis on multiple barcodes.

```
#!/bin/bash

set -e

for i in {13..21};
do
echo "Starting barcode $i"
python3 m6A_EcoGII_6-mer_analysis_pysam.py
-p /path/to/protobufs/
-g /path/to/sacCer3.fa
-a /path/to/barcode"$i"_alignments.sorted.bam
-b barcode"$i"
-o /path/to/in_vitro_EcoGII_6-mer_analysis_pysam_barcode"$i".tsv
done
```

Figure 4.37

See section for figure 4.28.

Figure 4.45

See section for figure 4.28.

Figures 4.46, 4.48 and 4.49

The analysis is similar to that performed for figures 4.25 and 4.26.

All primary alignments mapped to the forward strand with mapping quality greater than zero were analysed. Each adenine in each read was classified as methylated or not based on a 70% threshold. To reduce computational time the methylation of adenines was split into per-chromosome files with BedOps bedextract v2.4.39.

```
# Barcoded
```

```
python3 analyse_A_GATC_forward_reads.py
-p /path/to/protobuf/folder/
-b /path/to/barcodeX_alignments.sorted.bam
-g /path/to/sacCer3.fa
-o /path/to/barcodeX_output_file.bed
-e barcodeX
-s A
```

```
# Non-barcoded:
```

```
python3 analyse_A_GATC_forward_reads.py
-p /path/to/protobuf/folder/
-b /path/to/barcodeX_alignments.sorted.bam
-g /path/to/sacCer3.fa
-o /path/to/barcodeX_output_file.bed
-s A
-n True
```

```
for chr in `bedextract --list-chr barcodeX_A_methylation.sorted.bed`;
do bedextract $chr barcodeX_A_methylation.sorted.bed
> barcodeX/barcodeX.$chr.bed; done
```

The following bash script was used to calculate the distances between the adenines for which methylation data has just been produced, and the nucleosome dyads or replication origins. Binding events around a DNA replication origin do not occur symmetrically [7, 76] therefore a minimum and maximum value are required for filtering and the orientation information is maintained through use of the ‘-D b’ flag. Conversely, nucleosomes are symmetrical around the dyad therefore the upstream and downstream distances were combined by using the ‘-d’ flag.

```
#!/bin/bash

chr='chrI chrII chrIII chrIV chrV chrVI chrVII chrVIII chrIX chrX chrXI
chrXII chrXIII chrXIV chrXV chrXVI'

for i in {19..23}
do
  for chrom in $chr
  do
    echo "$i - Eaton"
    bedtools closest -a barcode"$i"/barcode"$i"."$chrom".bed
    -b GSM424494_Eaton_2010_acs_locations_midpoints_sacCer3_
    G2_23degrees_WT.sorted.bed
    -D b -t first
    > barcode"$i"/barcode"$i"_EcoGII_Eaton_origins."$chrom"
    .bedtools_closest;
    echo "$i - Berbenetz"
    bedtools closest -a barcode"$i"/barcode"$i"."$chrom".bed
    -b ACS_S228C_summary_formatted.sorted.bed
```

```

-D b -t first

> barcode"$i"/barcode"$i"_EcoGII_Berbenetz_origins.
"$chrom".bedtools_closest;

echo "$i - Jiang"

bedtools closest -a barcode"$i"/barcode"$i"."$chrom".bed
-b 2009_Jiang_nucleosome_dyad_positions_sacCer3.sorted.bed
-d -t first

> barcode"$i"/barcode"$i"_EcoGII_Jiang_nucleosomes.
"$chrom".bedtools_closest;

done

done

```

These files were then filtered to reduce file size (150 bp for nucleosome dyad meta-analysis, 250 bp for Eaton replication origin meta-analysis, 1000 bp for Eaton replication origin single-molecule analysis and 800 bp for Berbenetz meta-analysis and heatmap).

```

#!/bin/bash

chr='chrI chrII chrIII chrIV chrV chrVI chrVII chrVIII chrIX chrX
chrXI chrXII chrXIII chrXIV chrXV chrXVI'

for i in {19..23}
do
  for chrom in $chr
  do
    echo "$i - Eaton"

    awk '{ if ($13 <= 250 && $13 >= -250) { print } }'

    barcode"$i"/barcode"$i"_EcoGII_Eaton_origins."$chrom"
  done
done

```

```

        .bedtools_closest
> barcode"$i"/barcode"$i"_EcoGII_Eaton_origins_250bp_
distance."$chrom".bedtools_closest
awk '{ if ($13 <= 1000 && $13 >= -1000) { print } }'
barcode"$i"/barcode"$i"_EcoGII_Eaton_origins."$chrom"
.bedtools_closest
> barcode"$i"/barcode"$i"_EcoGII_Eaton_origins_1000bp_
distance."$chrom".bedtools_closest
echo "$i - Berbenetz"
awk '{ if ($13 <= 800 && $13 >= -800) { print } }'
barcode"$i"/barcode"$i"_EcoGII_Berbenetz_origins.
"$chrom".bedtools_closest
> barcode"$i"/barcode"$i"_EcoGII_Berbenetz_origins_
800bp_distance."$chrom".bedtools_closest
echo "$i - Jiang"
awk '{ if ($10 <= 150) { print } }'
barcode"$i"/barcode"$i"_EcoGII_Jiang_nucleosomes.
"$chrom".bedtools_closest
> barcode"$i"/barcode"$i"_EcoGII_Jiang_nucleosomes_
150bp_distance."$chrom".bedtools_closest

done

done

```

The chromosome files were then merged to create one file per barcode for each of the biological sets of regions of interest.

```
#!/bin/bash
```

```
for i in {19..23}
do
echo "$i - Eaton"
cat barcode"$i"/barcode"$i"_EcoGII_Eaton_origins_250bp_
distance.*.bedtools_closest
> barcode"$i"/barcode"$i"_Eaton_origins_250bp_distance_
_combined.bedtools_closest
cat barcode"$i"/barcode"$i"_EcoGII_Eaton_origins_1000bp_
distance.*.bedtools_closest
> barcode"$i"/barcode"$i"_Eaton_origins_1000bp_distance_
combined.bedtools_closest
echo "$i - Berbenetz"
cat barcode"$i"/barcode"$i"_EcoGII_Berbenetz_origins_800bp_
distance.*.bedtools_closest
> barcode"$i"/barcode"$i"_EcoGII_Berbenetz_origins_800bp_
distance_combined.bedtools_closest
echo "$i - Jiang"
cat barcode"$i"/barcode"$i"_EcoGII_Jiang_nucleosomes_150bp_
distance.*.bedtools_closest
> barcode"$i"/barcode"$i"_EcoGII_Jiang_nucleosomes_150bp_
distance_combined.bedtools_closest
done
```

The nucleosome meta-analysis was plotted with ‘nucleosomes_meta-analysis.r’. The Eaton replication origin meta-analysis was plotted with ‘origins_meta-analysis.r’. The Berbenetz replication origin meta-analysis was plotted with ‘Berbenetz_meta-analysis.r’.

Figure 4.47

The transcription start site (TSS) data were median positions from Qiu *et al* [219]. This data was saved as a csv and processed as follows to separate out TSS with and without TATA motifs.

```
# Convert csv to bed file
awk -F "," -v OFS="\t" '{print $1, $2, $3, $10}' TSS.csv > TSS_all.bed

# Filter for TATA-less or TATA-containing TSS
grep "TATA-less" TSS.csv |
awk -F "," -v OFS="\t" '{print $1, $2, $3, $10}' -
> TSS_TATA-less.bed

grep "TATA-containing" TSS.csv |
awk -F "," -v OFS="\t" '{print $1, $2, $3, $10}' -
> TSS_TATA-containing.bed

# Sort each of the above bed files
bedtools sort -i TSS_all.bed > TSS_all.sorted.bed
bedtools sort -i TSS_TATA-less.bed > TSS_TATA-less.sorted.bed
bedtools sort -i TSS_TATA-containing.bed
> TSS_TATA-containing.sorted.bed
```

For the meta-analysis of transcription start sites, a similar approach was taken as in the previous section for analysis of nucleosomes and replication origins. The bash scripts used were TSS_meta-analysis.sh, TSS_filter_by_distance.sh and TSS_merge_files.sh. This

analysis was plotted with `TSS_meta_analysis.r` for all transcription start sites, `TSS_TATA-containing.r` for TSS with TATA motifs and `TSS_TATA-less.r` for those without TATA motifs.

Figures 4.50 and 4.51

For the cumulative sum, the data files used for analysis were the same as for figures 4.48 and 4.49 but were plotted with `'cumulative_sum.r'`.

For plotting the nucleosome positions in relation to the ACS sites the distances were calculated using the following `bedtools` `closest` command.

```
bedtools closest -a 2009_Jiang_nucleosomes_sacCer3.sorted.bed
-b GSM424494_Eaton_2010_acs_locations_midpoints_sacCer3_G2_23degrees
_WT.sorted.bed
-D b >
2009_Jiang_nucleosomes_2010_Eaton_acs_midpoints_sacCer3_distances
.bedtools_closest
```

Figure 4.53

The origins were grouped into four clusters in the supplementary data from Berbenetz *et al* [76]. This data was then added to the ACS by formatting it into a bed file which was subsequently read in and combined with the adenine methylation data during the plotting stage with `'Berbenetz_heatmap.r'`.

```
sed '1d' ACS_S228C_4_clusters_k-means_clustering.txt | awk '{print $2}' -
| paste Berbenetz_origins.bed - > Berbenetz_clusters.bed
```

Figures 4.54 and 4.55

Using the `bedtools` `closest` data generated for figure 4.48 the plots were generated using `'EcoGII_single_molecule.r'` for the origin on chromosome 15 and `'EcoGII_single_molecule_acs.r'`

for the origins on chromosome 4. This filtered the reads based on inclusion of a minimum of 90% of the adenines within the 500 base region centered on the ACS and added nucleosome positions.

Appendix B

Licence for reproducing figure 4.18a

The licence from Springer Nature for reproduction of figure 2a from [116] as figure 4.18a in this thesis is included on the next page.

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Jul 26, 2021

This Agreement between Miss. Heather Jeffery ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number 5116460093910

License date Jul 26, 2021

Licensed Content
Publisher Springer Nature

Licensed Content
Publication Nature Methods

Licensed Content Title Long-range single-molecule mapping of chromatin accessibility in eukaryotes

Licensed Content Author Zohar Shipony et al

Licensed Content Date Feb 10, 2020

Type of Use Thesis/Dissertation

Requestor type academic/university or research institute

Format print and electronic

Portion figures/tables/illustrations

Number of
figures/tables/illustrations

1

High-res required

no

Will you be translating?

no

Circulation/distribution

1 - 29

Author of this Springer
Nature content

no

Title

Genomic single molecule analysis of chromatin structure around
DNA replication origins in *Saccharomyces cerevisiae*

Institution name

University of Oxford

Expected presentation
date

Aug 2021

Portions

Figure 2a single molecule, SMAC-seq ensemble and DNase-seq
data

Miss. Heather Jeffery
98 Crescent Road

Requestor Location

Oxford, OX4 2PD
United Kingdom
Attn: Miss. Heather Jeffery

Total

0.00 GBP

Terms and Conditions

Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the **Licence**) between you

and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

1. Grant of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

1. 2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

1. 3. If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Scope of Licence

2. 1. You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

2. 2. A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

2. 3. Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

2. 4. Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

2. 5. An alternative scope of licence may apply to signatories of the [STM Permissions Guidelines](#), as amended from time to time.

3. Duration of Licence

3. 1. A licence for is valid from the date of purchase ('Licence Date') at the end of the

relevant period in the below table:

Scope of Licence	Duration of Licence
Post on a website	12 months
Presentations	12 months
Books and journals	Lifetime of the edition in the language purchased

4. Acknowledgement

4. 1. The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

5. Restrictions on use

5. 1. Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

5. 2. You must not use any Licensed Material as part of any design or trademark.

5. 3. Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

6. Ownership of Rights

6. 1. Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

7. Warranty

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON

LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

8. Limitations

8.1. *BOOKS ONLY*: Where 'reuse in a dissertation/thesis' has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

8.2. For content reuse requests that qualify for permission under the [STM Permissions Guidelines](#), which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

9. Termination and Cancellation

9.1. Licences will expire after the period shown in Clause 3 (above).

9.2. Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

Appendix 1 — Acknowledgements:

For Journal Content:

Reprinted by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication)]

For Advance Online Publication papers:

Reprinted by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)]

For Adaptations/Translations:

Adapted/Translated by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication)]

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:

Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj. [JOURNAL ACRONYM])

For Book content:

Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author(s)**] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.3

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.