Methods to Probe the Function of Modified Bases in DNA

Robyn Elizabeth Hardisty

Fitzwilliam College

September 2017

This dissertation is submitted for the degree of Doctor of Philosophy

Methods to Probe the Function of Modified Bases in DNA

Robyn Elizabeth Hardisty

This thesis is focused on the development and utilisation of chemical and biological tools to probe the function of modified bases in DNA with specific exploration of the less well-studied T-modifications: 5-hmU, 5-fU and Base J.

LCMS/MS techniques are first utilised to enable the accurate global quantification of T-modifications (5-hmU, 5-fU and Base J) in both trypanosomatids and mammalian DNA.

A chemical affinity-enrichment sequencing method for the T-modifications is next described, which allows their chemoselective tagging over their C-modification counterparts. DNA fragments containing 5-fU are selectively tagged and enriched via oxime, hydrazine or benzimidazole formation using a biotinylated probe, and DNA fragments containing 5-hmU can be first chemically oxidised to 5-fU using $KRuO_4$. .Proof-of-principle T-modification enrichment is demonstrated by DNA sequencing.

In the following chapter, sequencing methods are employed to investigate the role of T-modifications in both trypanosomatids and mammalian samples. In *T.Brucei*, Base J formation is probed by artificial incorporation of 5-hmU and subsequent Base J chemical sequencing. Base J is preferentially formed or depleted at certain genomic loci; suggesting that Base J formation is sequence-specific. This may imply a distinct role for the 5-hmU sites which are not further glucosylated. Next, 5-hmU enrichment sequencing is performed in SMUG1 knockdown HEK293T cells to determine the genomic location of 5-hmU in mammals. An increase in 5-hmU loci is observed upon SMUG1 knockdown. 5-hmU enriched regions are found to be T-rich and depleted in exons and promoters. Furthermore, 5-hmU sites show poor overlap with known TET-enzyme binding sites, indicating that 5-hmU is formed via a TET-independent mechanism in HEK293T cells.

Next, mass spectrometry-based proteomics studies are utilised to determine 5-fU protein-binders in mammals. Pulldown of proteins using biotinylated baits enables the identification of proteins which are enriched or suppressed in the presence of the 5-fU modification compared to a non-modified control. Enriched proteins include those associated with DNA-damage, consistent with the current understanding that 5-fU is a product of oxidative damage in mammalian DNA.

Finally, a mechanistic insight into the effect of formylated bases on nucleosomal structure is described. Schiff base formation between formylated nucleobases and histone protein lysine side-chains is demonstrated. This provides a molecular mechanism for the association of 5-fC with increased nucleosomal occupancy *in vivo*.

Acknowledgements

I would firstly like to acknowledge Shankar for providing me with the opportunity to study for a PhD within his research group. I have learnt so much throughout this project under his guidance. The next thanks goes to those who provided funding for both my research and stipend: the University of Cambridge, Herchel Smith Fund, the Wellcome Trust, Cancer Research UK and the Leathersellers'.

The research group is a lovely environment to work in. There are a number of postdoctoral researchers who are incredibly supportive, namely Fumiko Kawasaki who has been a great mentor and influence throughout my PhD, and Euni Raiber for her advice and encyclopaedic knowledge of the literature. A huge thank you to: Sergio for the large and copious amount of bioinformatics analysis (still more to come!), Pieter for his help and expertise regarding LCMS/MS spectroscopy, and Sabrina for teaching me how to cell-culture and for HEK293T nuclear extract. From the Carrington group, I would like to acknowledge both Mark, for useful discussion, and Janaina who helped me with trypanosome culture. Also, much gratitude to both the Cambridge Institute Proteomics core facility, especially Valar, and to the Cambridge Proteomics Centre for LCMS/MS support and proteomics analysis respectively. Finally, thanks to Shankar, Fumiko, Euni, Chris, Kim, Alex and Marco for providing comments and careful proofreading for all or parts of this thesis, and Jo for her support while writing up. I take responsibility for any further errors!

I have made some great friends in the lab. Big love to the schwesters - Vicki, Judith and Darcie, along with Kim, Sabrina, Marco, Areeb, Jane, Jess, Louis, Nico & Mini.

Finally, I would like to acknowledge Michael and my family for lots of love and constant encouragement. I have also made some amazing friends while living in Cambridge, with a big shout-out to my Fitz besties Callum and Richard.

Abbreviations

5-caC	5-carboxycytosine
5-fC	5-formylcytsoine
5-fC$_m$	*N*-methyl-5-formylcytosine
5-fU	5-formyluracil
5-fU$_m$	*N*-methyl-5-formyluracil
5-glchmc	5-glucosylhydroxymethylcytosine
5-hmC	5-hydroxymethylcytosine
5-hmU	5-hydroxymethyluracil
5-mC	5-methylcytosine
8-oxoG	8-oxoguanine
A	Adenine
AID	Activation-induced cytidine deaminase
AP	Apurinic/apyrmidinic
APE1	Apurinic/apyrmidinic endonuclease 1
APOBEC	Apolipoprotein B mRNA editing enzyme
aq.	Aqueous
ARP	Aldehyde Reactive Probe
ASH2	ASH2 like Histone Lysine Methyltransferase complex subunit
BER	Base Excision Repair
BH	(+)-biotinamidohexanonic acid hydrazide
bp	Base-pair
BS	Bisulfite
BSA	Bovine Serum Albumin
BSF	Bloodstream form
C	Cytosine
cDNA	Complimentary DNA
CHAF	Chromatin Assembly Factor
ChIP	Chromatin Immunoprecipitation
CpG	Cytosine-phosphate-guanine
CpGi	Cytosine-phosphate-guanine islands
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDX1	Dead box 1
DIP	DNA Immunoprecipitation
DMAP1	DNA Methyltransferase 1-Associated Protein 1
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
DMT-MM	4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methyl-morpholinium chloride
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
DNMT	DNA Methyltransferases
dNTP	Deoxynucleotide triphosphate
DREME	Discriminative Regular Expression Motif Elicitation
ds	Double stranded
DTM	Differentiating Trypanosome Medium

EDC	1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide
ELISA	Enzyme Linked Immunosorbent Assay
EMSA	Electrophoretic Mobility Shift Assay
ES	Electrospray
ESI	Electrospray Ionisation
Et	Ethyl
FEN	Flap Endonuclease
Fw	Forward
G	Guanine
GADPH	Glyceraldehyde 3-phosphate dehydrogenase
GAT	Genomics Annotation Tester
GLIB	Glucosylation, periodate oxidation and biotinylation
GPx	Glutathione Peroxidase
H3K4	Lysine 4 of histone 3 subunit
H3K9	Lysine 9 or histone 3 subunit
HDAC	Histone Deacetylase
HDLBP	High Density Lipoprotein-binding protein
HEK	Human Embryonic Kidney
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonoic acid
HOBt	Hydroxybenzotriazole
HOMO	Highest Occupied Molecular Orbital
HPLC	High Performance Liquid Chromatography
HRMS	High Resolution Mass Spectrometry
J	Coupling constant
J-GT	J-glycosyltransferase
JBP	J-Binding protein
K	Lysine
k.d.	Knockdown
k.o.	Knockout
L.major	Leishmania major
LC	Liquid Chromotography
LC-MS	Liquid Chromotography-Mass Spectrometry
LC-MS/MS	Liquid Chromotography tandem mass Spectrometry
LDS	Lithium dodecyl sulfate
LUMO	Lowest unoccupied molecular orbital
m	Multiplet
m/z	Mass to charge ratio
Me	Methyl
MeCP	Methyl CpG Binding Protein
MEME	Multiple EM for Motif Elicitation
MES	2-(*N*-morpholino)ethanesulfonic acid or minimum energy structure
mESC	Mouse Embryonic Stem Cells
MK	Merz-Singh Kollmann
MNase	Micrococcal Nuclease
MOPS	(3-(N-morpholino)propanesulfonic acid)
MTHFD2	Bifunctional methylenetetrahydrofolate dehydrogenase/cyclohydrolase

Mull	Mulliken
n.d.	Not detected
NAP1	Nuclesome Assembly Protein 1
NBO	Natural Bond Orbital
NGS	Next-generation sequencing
NMR	Nuclear Magnetic Resonance
o	Ortho
ODN	Deoxynucleotide oligmer
OGG1	8-oxoguanine DNA glycosylase 1
ox-BS	Oxidative Bisulfate
p	Para
PAGE	Polyacrylamine Gel Electrophoresis
PBS	Phosphate Buffered Saline
PCF	Procyclic form
PCR	Polymerase Chain Reaction
PGC	Porous Graphitic Column
PMSF	Phenyl methyl sulfonyl fluoride
ppm	Parts per million
QE	Q-Exactive
qn	Quintet
qPCR	Quantitative PCR
red-BS	Reductive Bisulfite
Rev	Reverse
Ribo A	Ribonulcease A
RNA	Ribonucleic acid
ROS	Reactive Oxygen Species
RPKM	Reads Per Kilobase Million
RSLC	Rapid seperation liquid chromtography
RT	Room temperature
RTCB	tRNA-splicing ligase RtcB homolog
SAGA	Spt-Ada-Gcn5-acetyltransferase complex
SAM	*S*-adenosyl Methionine
SAP130	Histone Deacetylase Complex Subunit SAP130
SDS	Sodium dodecyl sulfate
seq	Sequencing
SGF29	SAGA Complex Associated Factor 29
SIL	Isotopically labelled standard
SILAC	Stable isotope labelling by/with amino acids in cell culture
SIN3a	Paired Amphipathic Helix Protein Sin3a
SMRT	Single Molecule Real-Time sequencing
SMUG1	Single-stranded Monofunctional Uracil Glycosylase 1
SND1	Staphylococcal Nuclease Domain-containing Protein 1
ss	Single-standed
T	Thymine
T.Brucei	Trypanosoma Brucei
T.cruzi	Trypanosoma Cruzi

TAB	TET-assisted bisulfite
TARDBP	TAR DNA binding protein
TBE	Tris base, boric acid and EDTA
TCOF1	Treacle protein
TDG	Thymine-DNA glycosylase
TET	Ten-eleven translocation
TFA	Trifluoroacetic acid
TIC	Total Ion Count
TLC	Thin Layer Chromotography
U	Uracil
UNG	Uracil-DNA Glycosylase
UV	Ultraviolet
w.t.	Wildtype
YY1	Yin-yang Transcription Factor 1
β-GT	β-glycosyltransferase

Publications

The following publications have resulted from work described in this thesis:

- Hardisty, R. E., Kawasaki, F., Sahakyan, A. B. & Balasubramanian, S. Selective Chemical Labeling of Natural T Modifications in DNA. *J. Am. Chem. Soc.* **137,** 9270–9272 (2015).
- Kawasaki, F., Beraldi, D., Hardisty, R. E., McInroy, G. M., van Delft., P., Balasubramanian.S. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biol.* **18,** 23 (2017).
- Raiber, E.-A., Hardisty, R., van Delft, P. & Balasubramanian, S. Mapping and elucidating the function of modified bases in DNA. *Nat. Rev. Chem.* **1,** 69 (2017).
- Raiber, E.A., Portella, G. Martinez Cuesta, S., Hardisty, R., Murat, P., Li, Z., Iurlaro, M., Dean, W., Beraldi, D., Dawson, M. A., Reik, W., Balasubramanian, S. 5-formylcytosine is a determinant of nucleosomal organisation. *Manuscript in preparation.*

Table of contents

1. Introduction

1.1. Structure of DNA

DNA is a remarkable molecule which forms the basis of our genetic code; it is composed of four nucleobases, adenine (A), guanine (G), cytosine (C), thymine (T), which are attached to a deoxyribose sugar and linked together via a phosphodiester backbone. Nucleobases form a specific recognition pattern, via hydrogen-bonding, in which G hybridises with C, and A hybridises with T (Figure 1 - A). This cognate base-pair recognition is essential for DNA replication, mRNA synthesis during transcription, and translation for protein synthesis.[1] All our genetic information is stored in the linear sequence of these four DNA nucleobases, and is responsible for the passage of information from one generation to the next.

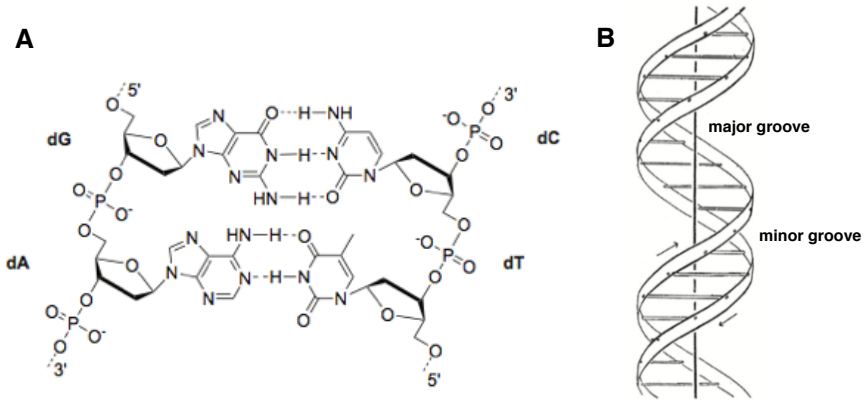


Figure 1: A) Cognate base-pairing within the structure of DNA. B) The asymmetric double helical structure of DNA. The figure is adapted from the original text by Watson and Crick who first described the structure of DNA.[1]

DNA most commonly folds into a double helical structure, B-DNA, composed of two complementary DNA strands that are asymmetric and run in opposite directions (Figure 1 - B). H-bonding is responsible for holding the two strands together, and the overall structure is further stabilised by Van der Waals and π-π interactions between stacks of aromatic nucleobases. The double helical structure enables large amounts of information to be stored in a compact volume, whilst genetic information is still accessible through its major and minor grooves. Proteins and small molecules bind to distinct sequences in these grooves due to differences in the patterns of H-bond donors or acceptors (Figure 2). This is important for sequence-based activation of DNA function; e.g. the recruitment and binding of transcription factors and DNA helicases in the major groove.[2] Furthermore, polyamide small molecules have been designed to bind in the minor groove in a sequence specific manner.[3]

In higher organisms, DNA is condensed with histone proteins to form chromatin. This further stabilises the DNA structure, protects genomic integrity and compacts the DNA into a small volume for storage.[4]

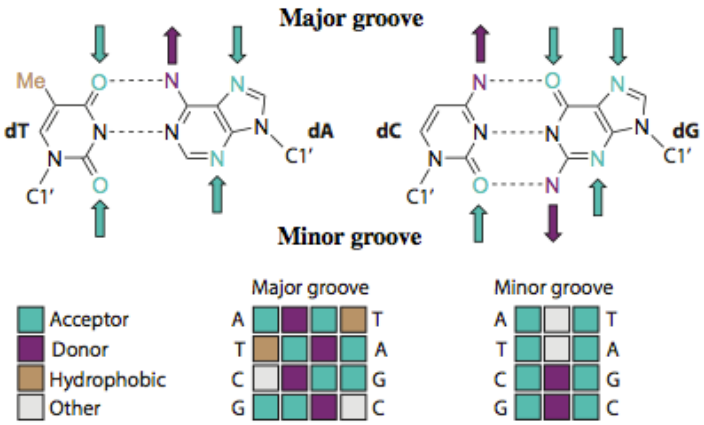


Figure 2: Cognate base-pairs and H-bonding donor and acceptor patterns in the major and minor groove of DNA. This is essential for sequence-specific recognition of proteins and small molecules.

1.2. Epigenetics

Nature has evolved mechanisms to control the expression of genes via an additional layer of information, based on the same primary code. The term "epigenetics" was first coined in 1942 and defined as "changes in phenotype without changes in genotype".[5] Epigenetics can explain the vast phenotypic changes that occur throughout an organism's lifespan, as well as the ability of tissue to differentiate; e.g. heart and lung tissue share the same genetic code, despite having vastly different functions.

1.2.1. Histone Modifications

Certain modifications which occur on histone protein side-chains are established epigenetic marks. These include lysine acetylation; lysine (mono-, di- and tri-) methylation; arginine methylation and serine, threonine and tyrosine phosphorylation, among others.[6] Such modifications have a contributory role in determining whether the DNA within chromatin is in an accessible (euchromatin) or non-accessible and repressed (heterochromatin) state.[5] Lysine acetylation, for example, is a marker for transcriptionally active genes. The resultant loss of positive charge associated with acetylated lysine results in weakened interactions between histone proteins and the negatively charged DNA phosphate backbone, meaning chromatin is no longer packed as tightly (Figure 3).

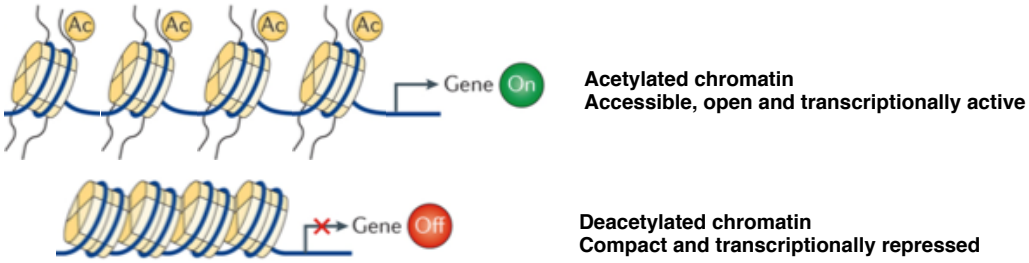


Figure 3: Histone modifications, shown here the effect of lysine acetylation, can alter chromatin accessibility and gene expression. Figure adapted from [7].

DNA base modifications exist in addition to histone modifications[8,9], which expand the scope of the genetic alphabet beyond the four canonical bases A, G, C and T (Figure 4). These modifications do not alter the base-pairing pattern of the DNA primary code, but have a unique chemical moiety that protrudes into the major groove. As such, DNA base modifications have the propensity to alter the way in which proteins, including transcription factors, can bind or recognise DNA; these bases can modulate gene expression and are epigenetic as a result.

1.2.2. DNA Modifications

DNA base modifications occur in all forms of life, and the chemical scope of these modifications are extensive, ranging from methylation to the introduction of amino acids or sugar substituents (Figure 4). DNA modifications are particularly prevalent in bacteria and bacteriophages;[10] in these systems modifications are used as a protective mechanism from host-endonuclease digestion, and are usually incorporated into the nucleobase pool (Table 1). In certain phages, hypermodified bases are generated by DNA modifying enzymes in a post-replicative manner, e.g. T-even phages further modify 5-hmC derived from the nucleobase pool to form α and β-D-Glucosyl-5-hydroxymethylcytosine (5-glchmC).

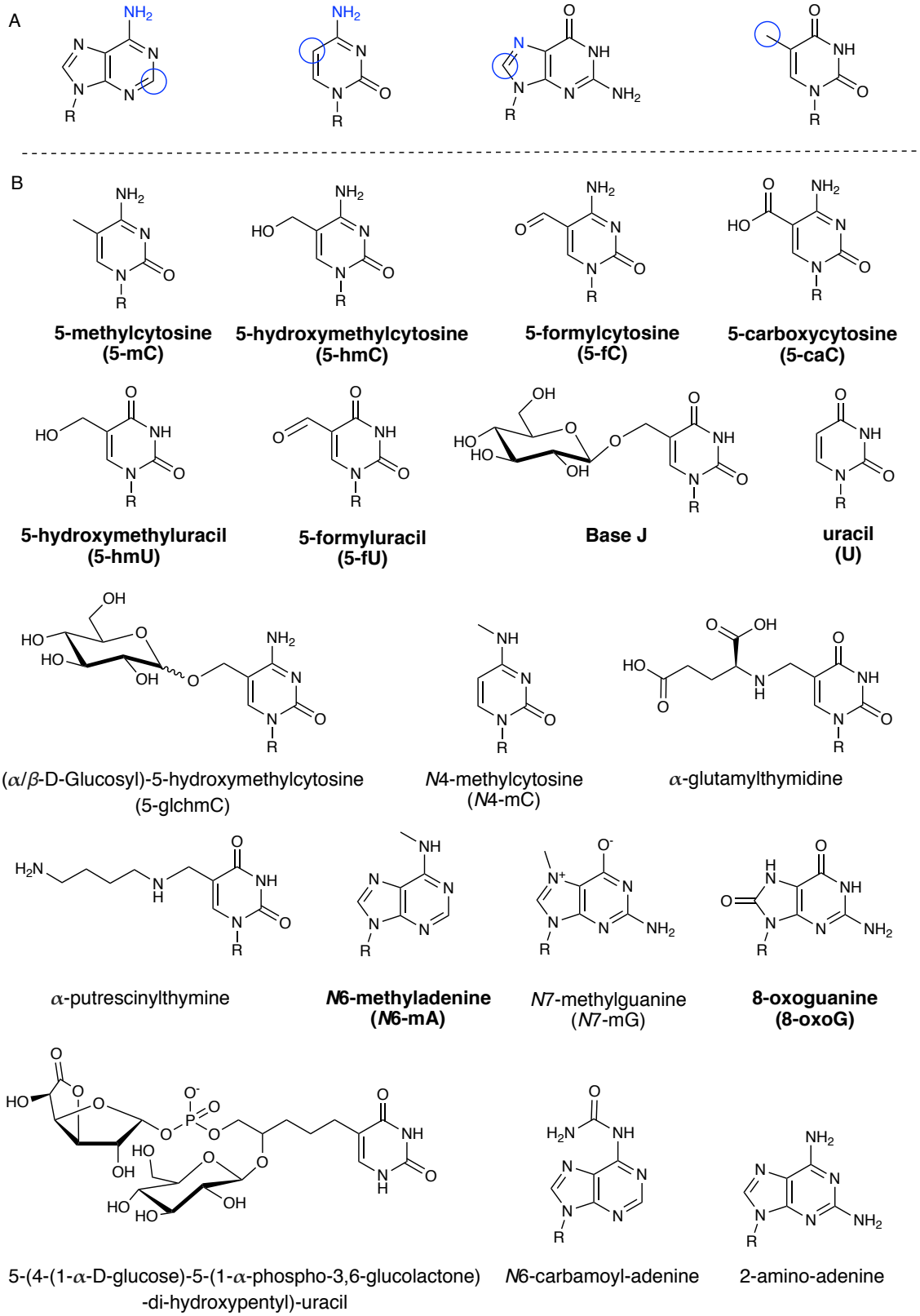


Figure 4: Top) Canonical bases with sites of modification labelled in blue. Bottom) Base modifications known to occur in the genomic DNA of various organisms. Those in bold are those reported to exist in eukaryotic DNA. Modified bases can be detected and quantified by a number of methods, such as LCMS/MS (Section 1.5.1).

Base	Phage	Abundance (% of canonical nucleoside)
5-methylcytosine (5-mC)	Xaruhomonas orzae	100%
	Φ174	~0.5%
Uracil (U)	B subtilis PBS2	100%
5-hydroxymethyluracil (5-hmU)	SP8, φe, SPO1, H1, SP82G, 2C, φ25	100%
2-aminoadenine	S elongazus S-2L	100%
5-hydroxymethylcytosine (5-hmC) + glycosylated derivatives (5-glchmC)	E.coli T2, T4, T6	100% (5-hmC + 5-glchmC)
*N*6-carbamoylmethyladenine	E. Coli phage Mu	15%
7-methylguanine (7-mG)	Singella sonnei phage DDV1	1%
α -glutamylthymine	B subitilis SP10	15-20%
α -putrescinylthymine	Pseudomonas acidovorans phage φW14	50%
*N*4-methylcytosine (4-mC)	Bacteria	~0.5-2%
*N*6-methyladenine (6-mA)	Bacteria	~0.3-3%
	T2, T4	~0.5-2%

Table 1: The large chemical scope and abundance of modified bases in bacteria and bacteriophage.[10]

1.3. Eukaryotic DNA Base Modifications

DNA base modifications are also prevalent in eukaryotes, and a number of these are 'epigenetic' and implicated in gene regulation.[9] Our knowledge of the eukaryotic DNA alphabet is rapidly expanding, and the development of chemical and biological tools is therefore crucial to probe and understand their functional role.

1.3.1. 5-methylcytosine

The best known and most studied eukaryotic modification is 5-methylcytosine (5-mC); commonly referred to as the "fifth" base of DNA. 5-mC is the most abundant DNA modification in mammalian systems, accounting for ~1% of all bases in humans.[11] 5-mC is formed enzymatically in a post-replicative manner via DNA methyltransferase (DNMT) enzymes. Their mode of action involves a key cysteine residue which attacks at the C6 position of cytosine. Methylation at the C5 position of the base then follows, with the co-factor S-adenosyl methionine (SAM) acting as a methyl donor, followed by rearomatisation of the base (Figure 5). There are several classes of DNMT enzymes: DNMT1 acts as a maintenance methylase, which restores the methylation marks on newly replicated DNA, whilst DNMT3a, DNMT3b and DNMT3c act as *de novo* methylases leading to the methylation of new sites.[12,13,14]

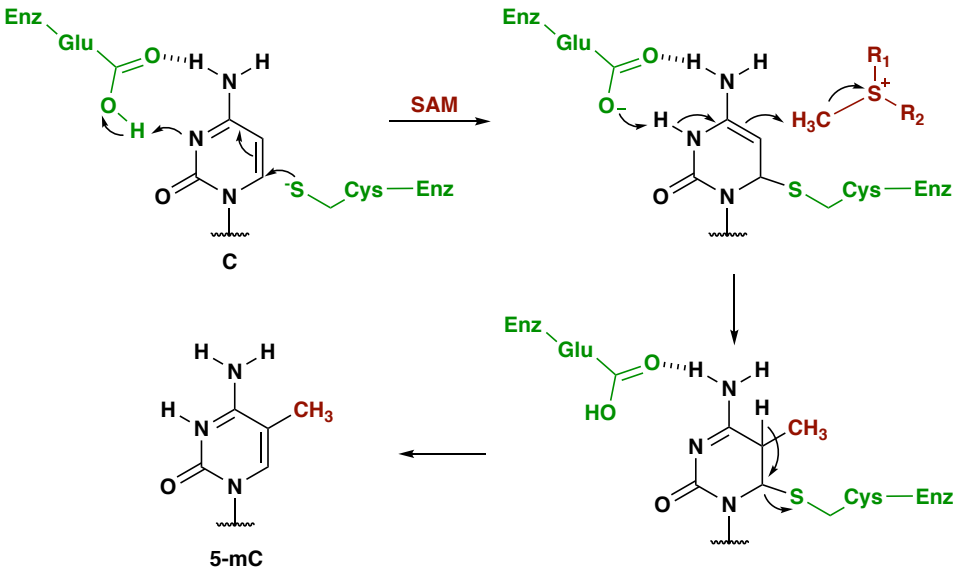


Figure 5: Mechanism of C methylation via the DNMT family of enzymes

60-80% of 5-mC occurs mainly in a cytosine-phosphate-guanosine (CpG) dinucleotide context,[15] since CpG sites are the preferred substrate of the DNMT enzymes. The palindromic nature of these sites provides a mechanism for inheritance, and DNMT1 has been shown to exhibit a preference towards hemimethylated rather than unmethylated sites.[16] DNA methylation is mainly associated with gene silencing and transcriptional repression, and is responsible for a number of biological processes such as X chromosome inactivation, genetic imprinting and cell-differentiation.[17] CpG dinucleotides depleted of 5-mC marks strongly correlate with the gene promoters/transcriptional start-sites of active genes. These genes can be dynamically regulated, and certain tissue-specific CpG sites can become methylated and thus transcriptionally blocked during differentiation or early development.[18] The importance of this mark for normal development is evident from the abnormal cytosine methylation patterns that are associated with disease states. Aberrant methylation is observed in cancerous cells; genomes generally become globally hypomethylated (reduced methylation), whilst specific regions including tumour processor genes are hypermethylated (increased methylation),[19,20] leading to tumorgenesis.[20] DNMT1 knockdown and the associated loss of methylation is lethal in mice and somatic human cells[21,22], and mutations of methyltransferases are associated with developmental disease.[23]

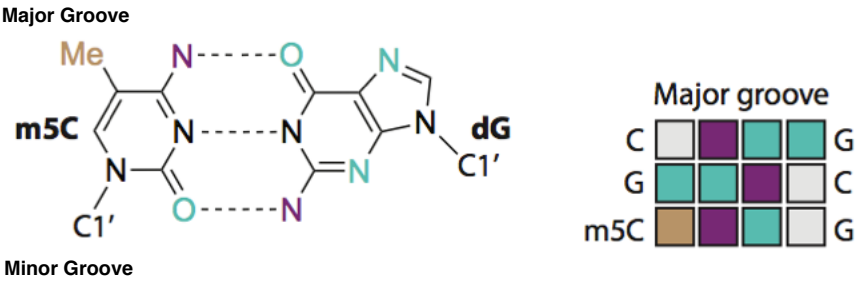


Figure 5: The hydrophobic methyl group of 5-mC affects duplex stability and protein recognition.

5-mC does not interfere with the base-pairing properties of C, yet the addition of a hydrophobic methyl group in the DNA major groove (Figure 5) has been shown to affect DNA recognition and double helix stability.[24] The presence of a methyl molecular handle leads to recruitment of specific 5-mC binding proteins, including the methyl CpG binding class of protein (MeCPs), which further recruit chromatin remodellers and repressive complexes to methylated DNA.[25] As a result, it is evident that the methylation mark does not work independently, and a large amount of epigenetic crosstalk between DNA C-methylation and histone modifications exist.[26] In mammals, C-methylation and histone methylation are strongly correlated at certain sites (H3K9), while de-novo DNMT3-mediated methylation is linked with unmethylated H3K4.

The study of 5-mC clearly demonstrates that DNA modifications have the propensity to alter and direct gene expression. However, within the last decade, our knowledge of the eukaryotic DNA alphabet has expanded even further.

1.3.2. The Base Excision Repair (BER) Pathway

Modified bases can be written or erased from the genome by DNA modifying enzymes. The absence or presence of 5-mC can, for example, determine whether a gene is transcriptionally active or repressed respectively. During early mammalian development, substantial changes in DNA methylation, notably demethylation, are essential for the differentiation of pluripotent stem-cells into distinct tissues.[27,28]

Removal of the methylation mark or modified bases can either occur passively, e.g. in the case of 5-mC, when newly replicated cells fail to be remethylated by DNMT1,[29] or can be excised and replaced in a DNA replication independent manner via the base-excision repair (BER) pathway (Figure 6). In BER, DNA glycosylases first recognise and bind a specific modified base, and flip it out from the double helix. Modified bases can then be excised via cleavage of the *N*-glycosidic bond to generate an

apurinic/apyrimidinic (AP) site. This in turn is removed by an AP endonuclease/lyase which creates a nick in the DNA backbone, followed by either short-patch (replacement of a single-nucleotide) or long-patch (> 2 nucleotides) repair. In short-patch repair, a DNA polymerase inserts the canonical base at the single excision site via Watson-Crick base-pairing and a DNA ligase is recruited to seal the nick. In long-patch repair, the DNA polymerase continues to displace DNA downstream of the excision site (2-13 bases); a Flap endonuclease (FEN1) subsequently removes the overhang, and a DNA ligase subsequently seals the nick between the newly-synthesised and initial DNA strand.[30,31]

In plants, a specific 5-mC glycosylase can excise the methylation mark directly, however no such ortholog exists in mammals.[32] Further investigation into the dynamic removal of 5-mC in mammals led to the discovery of other oxidised C-modifications in the mammalian genome.

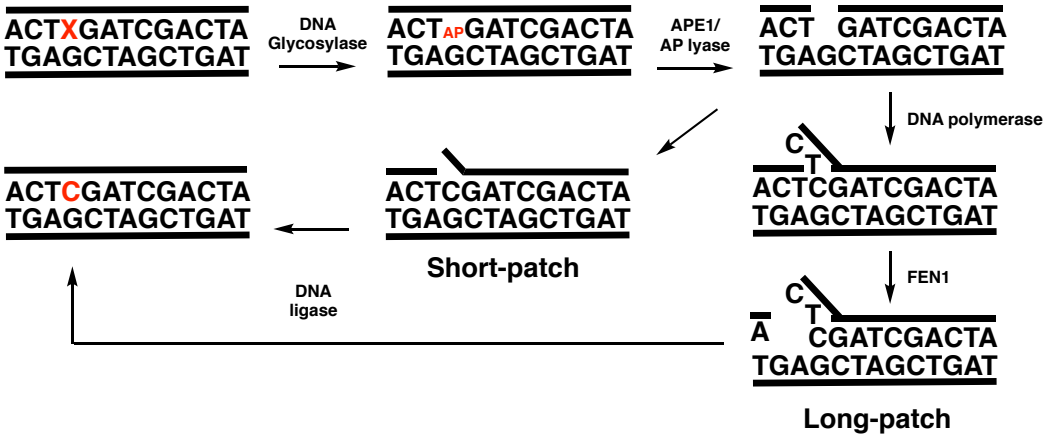


Figure 6: Short and long-patch base excision repair pathway of modified or damaged bases, leading to their replacement with a canonical nucleobase.

1.3.3. Oxidised Cytosine Derivatives and their Role in Active Demethylation

The oxidised C derivative 5-hydroxymethylcytosine (5-hmC) was first reported to occur in mammals in 1972[33], but renewed interest in this mark arose due to the discovery of the ten-eleven-translocation (TET) family of enzymes 1, 2 and 3. The TET enzymes are Fe(II)/2-oxoglutarate-dependent dioxygenases which utilise molecular oxygen for the sequential oxidation of 5-mC.[34] TET1-dependent oxidation of 5-mC to form 5-hmC was first reported by Talihiani *et al* in 2009,[35] and further investigation revealed that the TET enzymes can successively oxidise 5-hmC to both 5-formylcytosine (5-fC) and 5-carboxycytosine (5-caC).[36] Within somatic cells, the 5-hmC mark occurs approximately once every 10^3 bases[37,38], and the levels of 5-fC and 5-caC are three and four fold lower respectively.[39,40] Whilst 5-hmC is stable to excision by mammalian DNA glycosylases, both 5-fC and 5-caC can be excised by thymine DNA glycosylase (TDG).[41,42] This enables 5-mC active demethylation after iterative TET oxidation and subsequent BER (Figure 7). TDG is found to be essential for epigenetic regulation, and TDG knockout is embryonically lethal in mice due to the altered regulation of developmental genes.[43]

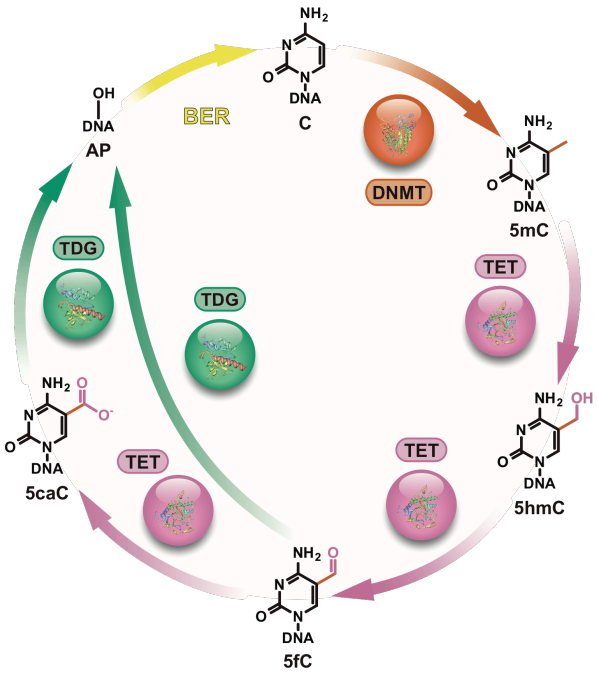


Figure 7: Demethylation of 5-mC can occur via iterative oxidation of 5-mC to 5-fC and 5-caC followed by TDG excision via BER.[44]

1.3.4. Oxidised Cytosine Derivatives as Independent Regulatory Marks

In addition to being intermediates of the active demethylation pathway, a wealth of evidence suggests that the oxidised C derivatives are epigenetic marks in their own right. Rather than being transient intermediates, 5-hmC and 5-fC are found to be stable marks in mouse embryonic stem cell (mESC) DNA, shown via isotopic labelling studies and subsequent tandem liquid chromatography mass spectrometry (LC-MS/MS) analysis (Introduction – 1.5.1).[45,46] Genomic maps of 5-hmC reveal the mark is mainly enriched in poised enhancers of developmental genes.[47] Furthermore, LC-MS/MS analysis of 5-hmC finds this mark, unlike 5-mC, to be tissue specific and age-dependent; for example, higher 5-hmC levels occur in the central nervous system[38] compared to other murine tissues. Low levels of 5-hmC are also associated with cancer,[48] however this may be explained by the fact TET-mediated 5-hmC formation fails to be properly maintained in fast-proliferating tissue.[46]

5-fC, surprisingly, is reported to recruit more unique protein readers compared to all the other C-modifications, including a number of transcription factors (e.g. FOXK1) and chromatin remodelling proteins (e.g. EHMT1).[49,50] 5-fC and 5-caC genomic maps showed these marks to be enriched in gene bodies, enhancers and promoters;[42,51,52] while recent profiling in different murine tissues revealed a tissue-specific role for 5-fC, where unique 5-fC signals were present in the active enhancers of developmental genes.[53] Furthermore, He and co-workers suggest the oxidised cytosine modifications, 5-fC and 5-caC, may lead to increased RNA Polymerase II (Pol II) stalling during transcription. Specific H-bonding interactions between 5-caC and the Pol II recognition loop were identified via crystallography, in support of this mechanism,[54] indicating these marks can affect gene expression. As such, the oxidised cytosine derivatives appear to have a distinct epigenetic function, aside from solely being intermediate products of active demethylation.

1.3.5. 8-oxoguanine - Oxidative Stress and Gene Regulation

There are a number of alternative DNA base modifications aside from the well-studied C marks that also have a unique role in gene regulation. 8-oxoguanine (8-oxoG) arises due to radical oxygen species (ROS) damage to G (Figure 4), and is both mutagenic (due to its propensity to base-pair with A) and a marker of oxidative damage.[55] 8-oxoG is excised from the genome by the 8-oxoguanine glycosylase 1 (OGG1) and BER, where steady-state levels estimate its occurrence to be one 8-oxoG per million nucleosides in mammalian tissue.[40] 8-oxoG is mainly found to occur in gene deserts.[56]

Several studies have suggested a link between the repair of this oxidative-damage mark and altered gene expression. Promoters containing 8-oxoG were found to markedly reduce the transcription of reporter genes as a result of 8-oxoG excision.[57] However, other studies suggest that 8-oxoG excision upregulates gene expression at certain genes; directed DNA-oxidative damage, resulting from H3K9 demethylation, was shown to recruit BER enzymes leading to facilitated transcription.[58] Furthermore, Burrows and co-workers suggest that the excision of 8-oxoG leads to gene activation within promoters capable of forming G-quadruplex structures.[59] In addition, global transcriptome analysis of OGG1 knockout mice indicated a link between OGG1 BER and cellular signalling.[60] Thus, these reports are the first to suggest that oxidative damage marks may also affect gene expression, and are potentially 'epigenetic' in nature.

1.4. T-modifications (5-hmU, 5-fU and Base J)

1.4.1. Mammalian T-modifications

Mammalian tissues also contain the oxidised T analogues 5-hmU and 5-fU (Figure 4), however the functional role of these bases is less well-explored. These modifications are demonstrated to occur at a level of 0.5-5 per 10^6 nucleosides[40], where 5-hmU levels are highest in mESCs and sperm.[40,61] Both 5-hmU and 5-fU modifications have traditionally been considered to be oxidative damage products of T, caused by ROS radical oxidation of the 5'-methyl group.[55] These marks can be excised from the mammalian genome via the BER pathway. In their cognate base-pair with A, both 5-hmU and 5-fU are excised by Single-stranded monofunctional uracil glycosylase 1 (SMUG1),[62] whilst a whole host of DNA glycosylases are found to repair these T-modifications when mispaired with G, including SMUG1, TDG and Methyl-CpG-binding domain 4 (MBD4).[63] hmU:G excision is much more efficient - even when only comparing the relative rates of SMUG1, it is found that hmU:G is excised 60-fold faster than hmU:A.[64] There are also single *in vitro* reports that 5-fU:A can be excised by TDG and NTH1.[65,66]

1.4.1.i 5-hmU - Association with Disease

5-hmU is associated with disease and aging; levels of 5-hmU autoantibodies were increased in individuals that suffered from cancer, and remarkably in those that were diagnosed with breast, colon and rectal cancers 0.5-6 years after sampling.[67] Djuric *et al.* also found increased 5-hmU in the blood of individuals suffering from breast cancer,[68] whilst further work highlighted a correlation between 5-hmU and both cancer invasiveness and tissue age.[69] Such studies suggest 5-hmU could be a potential biomarker for disease, potentially due to BER dysfunction.

1.4.1.ii 5-hmU - Beyond a Random Oxidative Damage Mark

There are now suggestions that 5-hmU may have a unique regulatory role aside from being the result of random oxidative damage in mammalian systems. It was initially hypothesised that 5-hmU may derive from enzymatic 5-hmC deamination in mammalian DNA, via the activation induced cytidine deaminases (AID) and Apolipoprotein B mRNA editing enzyme (APOBEC) family of enzymes. This is proposed to be an alternative TET-dependent active demethylation pathway, in which unmethylated C can be reinstated after efficient excision of the hmU:G mispair by BER (Figure 8 - A). AID is required for the demethylation of silenced promoters in heterokaryons[70] and for paternal demethylation after fertilisation in mice.[71] In addition, Cortellini *et al.* found that AID forms a complex with TDG in HEK239T cells, supporting this mechanism further.[72] Since

AID/APOBEC deaminases are overexpressed in a number of cancer types[73], enhanced 5-hmU from a deamination mechanism could be an alternative explanation for the low levels of 5-hmC found in cancer tissue.[73] However, the 5-hmC deamination proposal remains unsupported since cytidine deaminases are found to show no activity towards 5-hmC *in vitro;* deamination activity of the AID/APOBEC enzymes was found to be inversely proportional to the size of the substituent at the 5' position (e.g. C > mC >> hmC).[74] Thus, AID-related demethylation is instead likely to occur via C → U, or mC → T deamination, followed by BER excision with TDG.[71] Nevertheless, Guo *et al.* found that AID or APOBEC overexpression led to much increased demethylation of 5-hmC-transfected DNA, compared to 5-mC-transfected DNA in HEK293T cells.[75] When AID was overexpressed in the adult dentate gyrus (part of the hippocampus in the brain), 5-hmC levels were reduced by 59%. Further still, shRNA knockdown of APOBEC3 in these systems also led to reduced 5-hmC demethylation in neuronal promoters, suggesting such a 5-hmC deamination mechanism exists.

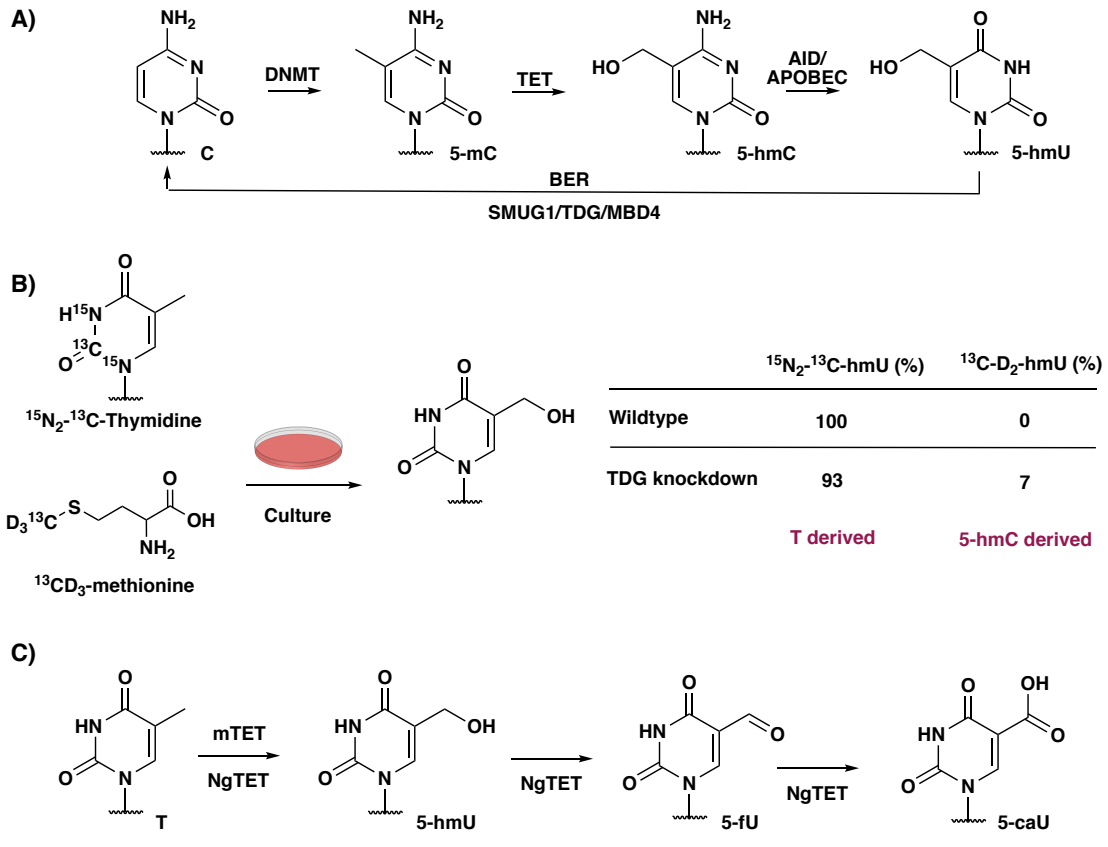


Figure 8: A) Proposed alternative active demethylation pathway where 5-hmC is enzymatically deaminated to form a 5-hmU:G mispair, which is removed by BER. B) Isotopic labelling and LC-MS/MS analysis determined the origin of 5-hmU in mESCs. C) The TET family of enzymes are shown to oxidise T to 5-hmU, 5-fU and 5-caU.

To definitively determine the origin of 5-hmU in mammalian cells, Carrell and co-workers utilised isotopic labelling and LC-MS/MS measurements to trace the formation of 5-hmU.[40] This was achieved by growing mESCs in culture medium supplemented with $^{15}N_2$-^{13}C-thymidine to label T, or $^{13}CD_3$-methionine to label 5-mC and its oxidised derivatives (Figure 8 – B). In steady-state mESCs, all 5-hmU was found to derive from T, questioning the significance of a 5-mC-oxidation-deamination based mechanism. However, in TDG knockdown mESCs, ~7% of 5-hmU modifications arose from 5-hmC deamination. This suggests that although 5-hmC deamination can occur, it is coupled with rapid and efficient excision in the steady-state. Furthermore, the 5-hmC deamination mechanism may be more prominent in certain tissues, such as in the brain.

Two independent groups have now suggested that 5-hmU is generated enzymatically from T, via the TET family of enzymes.[40,76] Carell and co-workers demonstrated the ability of the recombinant mTET1 enzyme to oxidise T to form 5-hmU *in vitro.* Furthermore, the levels of 5-hmU in mESCs are found to vary with TET expression: 5-hmU is depleted by ~3-fold as a result of TET1 and TET2 knockdown in mESCs; whilst ectopic overexpression of the catalytic domain of TET1 in HEK293T cells led to a 65-fold increase in both 5-hmU and 5-hmC respectively. By assessing the dynamic changes in response to mESC differentiation, changes in 5-hmU levels mostly resembled those observed for 5-hmC, peaking 8-16 hours after differentiation.[40] Pais *et al* also demonstrated *in vitro* T oxidation with *Naegleria* TET led to the formation of 5-hmU and other oxidised derivatives, including 5-fU (Figure 8 - B).[76]

5-hmU-protein interactions have also been assessed to probe the function of this mark.[40] 5-hmU binding proteins included 1) regulatory proteins Uhrf2 and HIVEP3, 2) chromatin modellers Chd1 and Chd9, and 3) DNA methyltransferases DNMT3a and DNMT3b. This suggests a regulatory role for 5-hmU in the mammalian genome. However, thus far, no methods have been developed to determine the genomic location of 5-hmU in mammalian DNA, which will be crucial for further insight into its biological function.

1.4.1.iii 5-fU in Mammalian DNA

5-fU is also present in mammalian DNA, however any potential function beyond a marker of oxidative damage has not been explored. Isotopic labelling studies to determine the origin of 5-fU in mESCs revealed the majority of 5-fU to derive from T, although ~3% of all 5-fU arose from deamination both in wild-type and TDG knockdown cells.[40] LC-MS/MS analysis during mESC differentiation revealed that changes in 5-fU levels were more likely to cluster with the oxidative damage marker 8-oxoG. Furthermore, 5-fU levels did not significantly vary with TET expression in mESCs,[40] although other enzymes in the TET family (*Naegleria*) were found to have the capability to oxidise 5-hmU further to 5-fU and 5-caU (Figure 8 - C).[76]

The occurrence of 5-fU in DNA is slightly mutagenic, likely due to its higher propensity to mispair with G during replication.[77,78] The presence of the electron withdrawing formyl group increases the acidity of the NH proton (pKa = 8.12) compared to thymine (pKa = 9.69), facilitating formation of the enol tautomer, especially upon ionisation (Figure 9). As such, the presence of 5-fU has the propensity to cause T→C transitions in the cell, where this potential source of mutation (if 5-fU fails to be repaired efficiently) may be implicated in the onset of disease. Whilst *in vitro* studies demonstrate that the presence of 5-fU in DNA can impede transcription factor binding (shown both with AP-1 and NFκB)[79,80] a global proteomics study to determine 5-fU protein-binders in mammals has yet to be reported. Similarly to 5-hmU, a method to generate a genome-wide map of 5-fU is also yet to be developed.

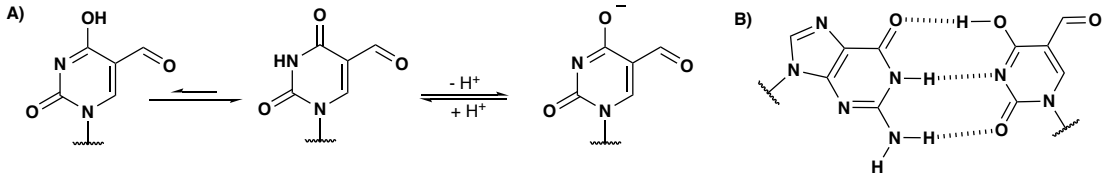


Figure 9: A) 5-formyluracil is more likely to form the enol tautomer. B) fU:G mispair with the enol tautomer of 5-fU

1.4.2. SMUG1

SMUG1 is the dominant DNA glycosylase responsible for both 5-hmU and 5-fU excision, and is the only glycosylase that is consistently reported to excise these bases in their natural base-pair context.[62,63] In addition to its role repairing oxidized T derivatives, SMUG1 is also a significant back-up enzyme for excising genomic uracil (U)[81], which occurs due to misincorporation of dUTP, or chemical or enzymatic deamination of cytosine in DNA.[82] SMUG1 is only present in higher eukaryotes, including humans, and is found to be absent in lower organisms including yeast and *E.coli*. Since a correlation exists between organisms that contain 5-mC and those that express SMUG1, it was initially hypothesized that the enzyme may be important in the processing of cytosine methylation.[83]

In contrast to TDG knockout mice which are found to be embryonically lethal, SMUG1 knockout mice are viable for >1 year of age.[81] In these mice, all notable 5-hmU excision capacity was lost, confirming that SMUG1 is the sole mammalian glycosylase for removal of this mark. Whilst the feeding of 5-hmU mononucleoside is found to be toxic to human cell-lines and mice[81,84], this has been attributed to excessive SMUG1 repair and cells fed with 5-hmU remain viable when SMUG1 is knocked down.[85]

SMUG1 deficiency is found to be linked with breast cancer invasiveness,[86] and a single nucleotide polymorophism (SNP) in the SMUG1 gene is associated with the most significant increased risk (1.42 fold) in bladder cancer compared to all other BER enzymes.[87] Levels of SMUG1 are also reduced in Werner's syndrome cells, a disease associated with accelerated aging.[88] Furthermore, SMUG1 knockdown in mouse embryonic fibroblasts (MEF) show a mutator phenotype, with increased mutation occurrence (~ 2.4 fold) compared to wild-type cells.[89] Analysis revealed that >90% of mutational sites occurred at non-CpG sequences. An enhanced rate of mutations can potentially be explained by a corresponding elevation of mutagenic 5-fU, or the inability to efficiently process U:G, hmU:G or fU:G mispairs, leading to C→T transitions.

These studies imply that an inability to excise the T-modifications via the BER pathway may have implications in disease. Thus the specific study of SMUG1 may also infer the function, role or consequence of oxidized T-modifications in DNA.

1.4.3. T-modifications in Trypanosomatids and Base J

The T-modifications are also prominently observed in the genomes of trypanosomatids.[90] Trypanosomatids are a class of eukaryotic flagellated protists which are mainly parasitic, and are responsible for a number of tropical diseases including African sleeping sickness, Chagas disease and Leishmaniasis. Species include *Trypanosoma* and *Leishmania*, which have multiple life-forms depending on their host. Both their life-cycles include a proliferative procyclic form (PCF - insect host) and a proliferative bloodstream form (BSF - vertebrate/mammalian host) (Figure 10 - A), separated by non-proliferative life-stages where trypanosomatids undergo differentiation. Transcription in trypanosomes varies from mammalian systems (polycistronic vs monocistronic), since transcribed mRNA codes for several genes in one cluster.[91] (Figure 10 - B)

Hypermodified β-glucosylated hydroxymethyluracil, named Base J, occurs in trypanosomatids alongside 5-hmU (Figure 10 – C). T-modifications are more abundant than those in mammals, and occur at a level of 0.5% and 0.04% of all T nucleosides for Base J and 5-hmU respectively.[90] In these systems, T is enzymatically oxidized to 5-hmU by the sequence-specific J-binding proteins (JBP) 1 and 2, which are Fe(II) and 2-oxoglutarate-dependent dioxygenases homologous to the TET enzymes found in mammals[92,93] (Figure 10- C). Whilst JBP1 is mainly involved in T-modification maintenance and binds to Base J containing DNA directly, JBP2 is considered to be the major *de-novo* regulator of 5-hmU, and binds to chromatin in a Base J-independent manner. No identified 5-hmU glycosylase (e.g. SMUG1 ortholog) exists in trypanosomes, and 5-hmU is instead enzymatically glucosylated by J-glucosyltransferase (J-GT)[94] to form Base J, reported to occur in a non-sequence specific manner.[95] Base J is heavily implicated in gene regulation and transcriptional regulation, and is enriched at sites of Pol II initiation and termination. In *L. major*, 98% of Base J occurs in telomeric or repetitive elements. This number falls to ~75% in *T. brucei*, whilst Base J is found to mainly occur outside of telomeric regions in other trypanosomes and kinetoplastids.[96] Knockdown of JBP enzymes and subsequent depletion of Base J in *L. major* is lethal; in *T. brucei*, this leads to increased transcriptional read-through or transcriptional defects and altered expression of downstream genes.[97,98,99] In *T. brucei,* JBP enzymes are developmentally regulated, and found to be downregulated in PCF trypanosomes,[100] leading to the absence of detectable Base J in this life-form.[101]

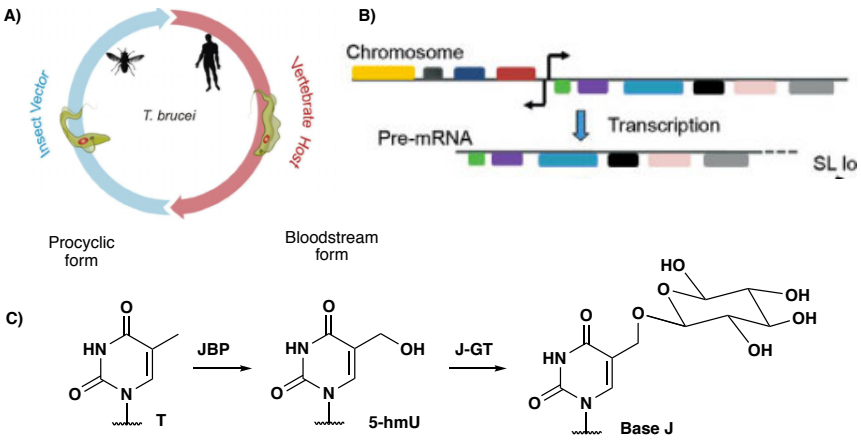


Figure 10: A) A life-cycle of a trypanosomatid passes through a BSF and PCF depending on the host species[102] B) Schematic of polycistronic gene regulation where several genes are transcribed into one mRNA.[103] C) Biosynthesis of T-modifications in trypanosomatids.

Base J loci are also found to be associated with modified histone H3 variants in certain trypanosomes,[104] which gives further evidence to the role of Base J as a regulator of transcription, and provides evidence of epigenetic crosstalk between DNA and histones in chromatin.

There have been many studies into the role of Base J, and its overall functional role is still under some scrutiny. However, investigations into the 5-hmU intermediate have been limited. J-GT knockdown and JBP overexpression is shown to cause an associated decrease in 5-hmU levels, and an increase in detectable 5-fU; this suggests that 5-hmU may be dynamically regulated in trypanosomes[94], and a possible distinct role for 5-hmU requires further exploration. Furthermore, a greater understanding of the dynamics and regulatory roles of T-modifications in these systems may highlight potential druggable targets, specific for trypansomatids (e.g. JBP1 and J-GT enzymes are not expressed in mammals) for the novel treatment of parasitic diseases.

1.5. Methods to Determine the Function of Modified Bases in DNA

The role and presence of T-modifications is less explored and understood compared to the analogous C-modifications. In mammals, preliminary studies suggest 5-hmU may be formed enzymatically, and have a role in gene regulation. Since 5-hmU and 5-fU are also products of ROS, they may also have a similar biological role to 8-oxoG. Thus, the origin and genomic location of these marks should be determined. Furthermore, the T-modifications are heavily implicated in the gene regulation of trypanosomatids. Therefore a unique role for 5-hmU, aside from being an intermediate in Base J synthesis, requires investigation.

To further understand the role of DNA modifications in biology, the development of chemical and biological tools is essential. Several techniques were explored throughout this PhD thesis: 1) the global quantification of modified bases in DNA; 2) DNA sequencing of modified bases to determine their genomic loci; 3) the study of modified-base-protein interactions using proteomics; and 4) the use of chemical biology to probe the effect of DNA modifications on nucleosomal structure. These methods will now be reviewed in some detail, with regards to how they have been utilised previously in the rapidly expanding modified bases field. These strategies were exploited to further explore the role of modified bases in DNA, with a specific focus on the development and utilisation of methods to probe T-modification function.

1.5.1. Detection and Global Quantification Methods

Detection and subsequent global quantitation of modified bases is a powerful tool for the discovery and elucidation of modified bases in organisms. Such methods can be used to assess the level and dynamics of modifications in different biological systems: e.g. in a variety of organisms, tissues, cell-types or phenotypes.[37,105] Quantification methods, particularly those for low abundance modifications, require high resolution and sensitivity due to the high background signal caused by the canonical bases.

Novel modified bases have, in the past, been detected and quantified via thin layer chromatography (TLC). This method enables separation of constituent DNA bases via their unique polar properties in two dimensions,[106] where sensitivity can be improved via the use of radioactively labelled ^{32}P-DNA. TLC was used for the discovery of 5-hmC in mammals[107] and Base J in trypanosomes.[108] Whilst this method has been valuable in discovery, it requires the laborious and sometimes impractical use of radioisotopes, and is limited in sensitivity. Furthermore, modified nucleobases may co-elute under certain

conditions, highlighting potential issues with specificity.[109] Alternative quantitation methods include: enzyme-linked immunosorbant assays (ELISA) or 'dot-blot' techniques, used to quantify the levels of a modified base by using specific antibodies,[110] and Forster-resonance energy transfer (FRET)-based assays which quantifies modified bases that have been fluorescently tagged.[111] These methods, although useful, have typically higher detection limits and also have issues with specificity;[112] e.g. fluorescent tagging of modified bases requires a strictly chemoselective method.

The gold-standard for modified base quantitation uses LC-MS/MS techniques,[113] discussed further in Chapter 2. The advantage of LC-MS/MS is its superior discrimination between different modified bases which are simultaneously resolved via both chromatography and mass. This method is extremely sensitive, can detect down to femtomolar quantities, and has been used for the detection and quantification of low abundance modifications such as 5-caC and 5-hmU in mammals.[40] LC-MS/MS is crucial for probing the dynamics of modifications in biological systems; thus methods will be developed to enable accurate LC-MS/MS measurements of T-modifications, and other low abundance modifications, in this PhD.

1.5.2. Sequencing of modified bases

1.5.2.i. Next-Generation DNA Sequencing Technology

Next-generation sequencing (NGS), e.g. Solexa/Illumina, has enabled scientists to routinely decode the genome of any organism in a rapid and inexpensive manner.[114] Using a sequencing by synthesis approach, the DNA sequence is decoded stepwise by complementary reversible-terminator fluorescent deoxyribonucleoside triphosphates (dNTPs). These dNTPs, labelled with a specific fluorophore to distinguish each base, are inserted opposite the template strand via Watson-Crick base pairing. Protecting groups, present on the 3'-OH group of each nucleotide, temporarily stall polymerisation until they are chemically removed, and this process is then repeated for the remainder of the DNA strand. Before sequencing, genomic DNA must be first sonicated into smaller fragments before the ligation of DNA sequencing adaptors. These adaptors bind to complementary primers on the DNA sequencer flow-cell, and each DNA template is subsequently amplified via solid-phase PCR. This generates a cluster of the initial DNA fragment which enhances the fluorescent signal (Figure 11). The sequences obtained (known as 'reads') can be aligned to the reference genome of the organism of interest.

Modified bases do not give a unique readout via NGS, as modified bases share the same Watson-Crick base-pairing pattern as the canonical bases (e.g. both 5-hmU and T would base-pair with A; and hence would be 'read' as a 'T').[115] However, NGS in conjunction with other methods can still be used to determine the genomic loci of modified bases. These methods include both 1) affinity mapping of modified-base containing DNA fragments; and 2) techniques which identify DNA modifications at single-base resolution. In addition, the development of third-generation sequencing technology may allow the routine sequencing of modified bases in the future.

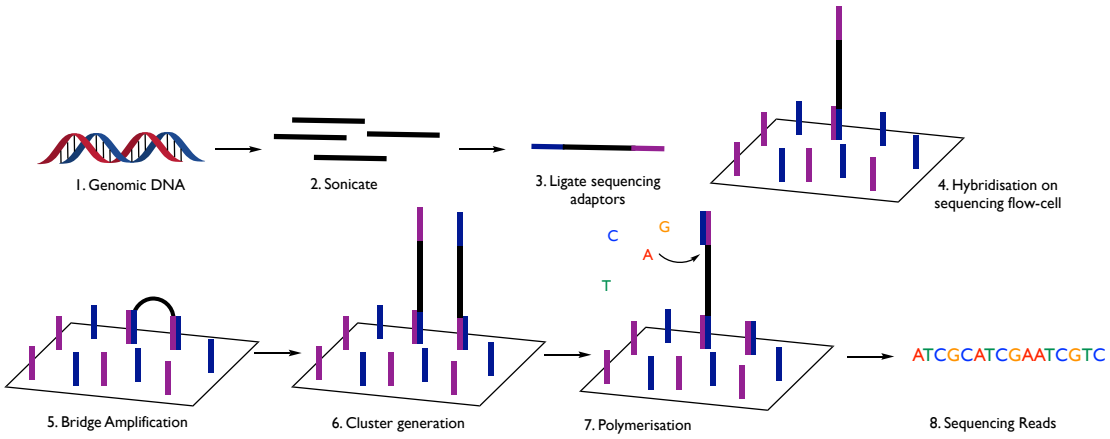


Figure 11: Schematic representation of the workflow of NGS sequencing technology.

1.5.2.ii. Affinity Mapping of Modified DNA

One way to determine the genomic loci of DNA modifications is to affinity-enrich fragments of DNA which contain modified bases prior to NGS (Figure 12). This can be achieved either via 1) selective chemical labelling with an affinity tag and subsequent enrichment (e.g. – with biotin and enrichment with magnetic streptavidin beads) or 2) affinity enrichment with antibodies that are specific for the DNA modification in question. Fragments are enriched via immobilisation, and non-modified DNA fragments can be removed in successive wash steps. After preparing enriched fragments for NGS, high throughput sequencing of these fragments leads to a build-up of 'reads' at particular genomic locations; this identifies where modifications occur in the genome (Figure 12). Affinity mapping typically generates 'low resolution' peaks (typically hundreds of base-pairs), but is a relatively low cost method which gives a great insight into where modified bases arise.

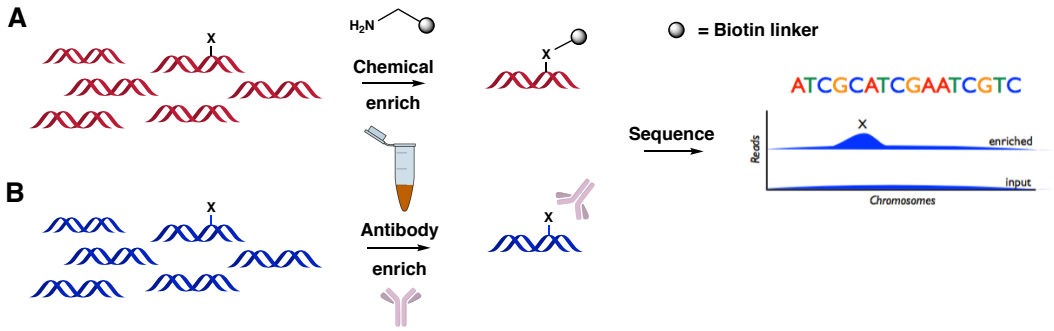


Figure 12: Affinity enrichment methods to determine the genomic loci of modified bases in DNA via A) chemical enrichment using an affinity tag (e.g. – biotin) or B) using a specific antibody for enrichment.

1.5.2.iii. Chemical Enrichment Methods

Both chemical and chemoenzymatic DNA enrichment techniques have been utilised to map a number of modified bases (Figure 13). 5-hmC "pulldown" or enrichment methods exploit the natural enzymatic glucosylation reaction of 5-hmC by the T4 phage β-glucosyltransferase. After glucosylation of DNA containing 5-hmC, Rao and co-workers used sodium periodate oxidation to cleave the vicinal diols on the sugar moiety to generate reactive dialdehydes,[116] which were subsequently tagged via oxime formation using a biotinylated oxyamine. This method was named GLIB (glucosylation, periodate oxidation, biotinylation)-seq (Figure 13). He and co-workers instead employed N_3-UDP-glucose as an alternative non-natural substrate for T4 β-GT.[117] The resulting azido-glucosylated 5-hmC can be tagged with an alkyne-linked biotin using 1,3-dipolar cycloaddition "click" chemistry (Figure 13)

Previous work within the Balasubramanian group[51,53] employed biotinylated oxyamines to generate genome-wide maps of 5-fC, exploiting its reactive aldehyde moiety to form an oxime (Figure 13). 8-oxoG has instead been mapped via selection oxidation, using the mild 1-electron oxidant potassium hexabromiridate. This is achievable due to the reduced redox potential of this base compared to G and the other canonical nucleosides; the resultant electrophilic intermediate can be trapped with a biotinylated amine nucleophile and utilised for enrichment of DNA fragments containing 8-oxoG (Figure 13).

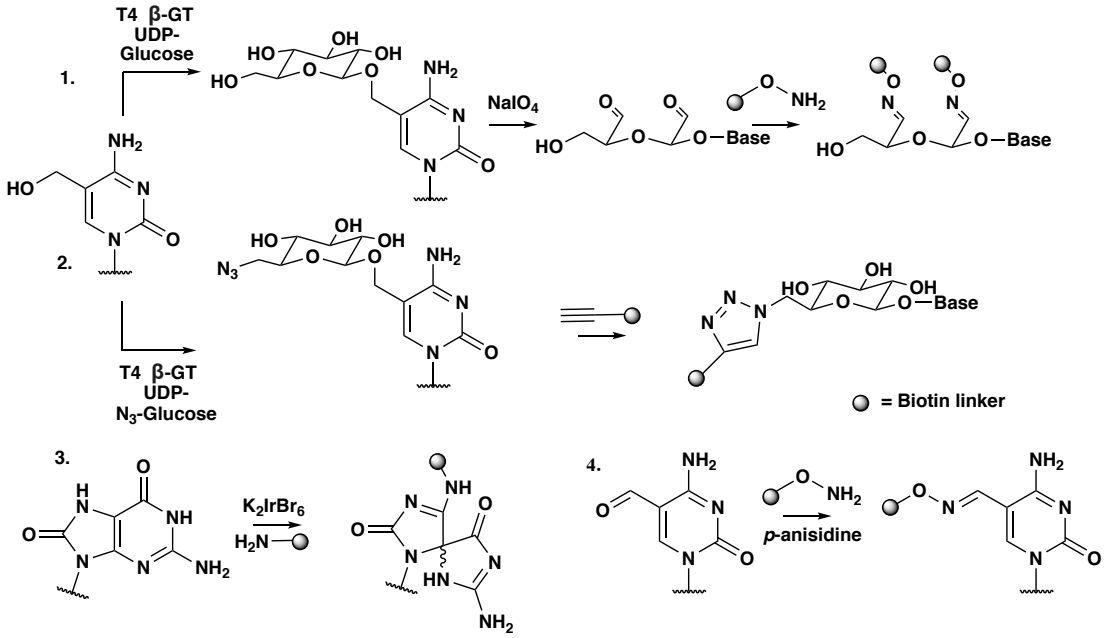


Figure 13: Chemical functionalisation with biotin used in existing methods for enrichment of modified bases. 1) GLIB-seq periodate oxidation of 5-glchmC followed by oxyamine functionalisation of the aldehydes, 2) Glucosylation of 5-hmC using UDP-N_3-glucose followed by Huisgen 1,3-dipolar click chemistry with a biotinylated alkyne tag, 3) Selective 1-electron oxidation of 8-oxoG followed by nucleophilic addition of a biotinylated amine, 4) Oxime formation with 5-fC using a biotinylated oxyamine.

1.5.2.iv. Antibody Enrichment Methods

Antibodies with a particular affinity to a modified base can be used to generate genome-wide maps using DNA immunoprecipitation (DIP). This has enabled mapping of 5-mC (MeDIP)[118], 5-hmC (hMeDIP-seq)[119], 5-fC (5fCDIPseq)[42], 5-caC (5caCDIPseq)[42]and Base J.[98] A potential drawback of DIP methods compared to chemical enrichment is a density-dependent bias,[120] where fragments that contain a higher abundance of modification are more likely to be captured. However, these techniques have had great utility, especially for bases that do not have a reactive chemical handle, such as 5-mC.

1.5.2.v. Single-base Resolution Methods – Restriction Endonucleases

Other sequencing methods generate a read-out for a modified base at single-base resolution. Restriction endonucleases are enzymes that cleave the phosphodiester DNA backbone at a particular sequence. A subset of these enzymes can differentially cut depending on the presence or absence of a modification site, and subsequent DNA sequencing can thus determine the genomic location of modified bases. For example, AbasI is an enzyme that preferentially cuts at sites of 5-glchmC, and can be used to map 5-hmC sites after treatment with T4 β-glucosyltransferase (Figure 14). This is a simple way of detecting modified bases with good resolution, however natural restriction enzymes that specifically cleave certain modified bases need to be firstly identified. Furthermore, sites may also be cleaved with varying efficiency.[121]

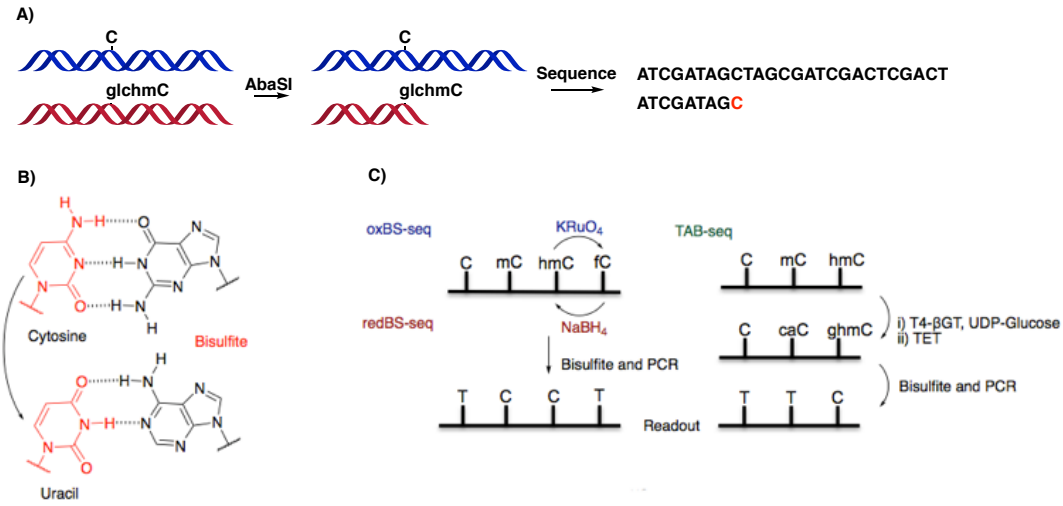


Figure 14: Single-base resolution sequencing methods. A) Restriction endonucleases can specifically cleave the phosphodiester backbone at sites of modified bases; NGS after endonuclease treatment can reveal where modifications arise. B) Bisulfite conversion of C to U allows discrimination of 5-mC (resistant to deamination) and C (deaminated to T) after NGS. C) ox-BS seq, red-BS-seq and TAB-seq can be used to distinguish between different C-modifications after NGS.

1.5.2.vi. Single-Base Resolution Methods - Chemical Transformation

Chemical transformation methods instead differentially alter the Watson-Crick base-pairing pattern of modified bases from the canonical bases; this enables a discriminative readout of modified bases at single-base resolution after NGS. The bisulfite reaction has been essential for distinguishing between C and its most abundant modification, 5-mC. Prior to bisulfite treatment, both C and 5-mC are read as 'C' via NGS, since they both base-pair with G. The bisulfite reaction catalyses the hydrolytic deamination of C to U (read as T), whilst 5-mC (read as C) remains intact (Figure 14).[122]

However, the bisulfite reaction fails to distinguish between the other oxidised C derivatives: 5-hmC, like 5-mC, is resistant to bisulfite deamination;[123] whilst 5-fC and 5-caC deaminate under bisulfite conditions. Further chemical transformations prior to bisulfite have now been utilized to generate definitive maps of each C-modification (Figure 14). Oxidative bisulfite (oxBS)-seq[124], uses potassium perruthenate ($KRuO_4$) to oxidise 5-hmC to 5-fC prior to bisulfite, which generates a definitive map of 5-mC. Subtraction of an oxBS dataset from a BS dataset identifies sites of 5-hmC. 5-fC is further distinguished by treatment with sodium borohydride (redBS-seq) to form 5-hmC, followed by comparison with BS and ox-BS datasets.[125] Many other alternative bisulfite-based techniques (e.g. Tet-assisted bisulfite (TAB)-seq – Figure 14)[126] have now been developed to distinguish these modifications based on similar principles. In addition, Xia and coworkers have recently developed a "bisulfite-free" method to detect 5-fC at single-base resolution; selective tagging of 5-fC with 1,3-indianone derivatives, and subsequent PCR leads to C-to-T transitions at 5-fC sites.[127]

1.5.2.vii. Third-generation Sequencing Methods

The development of third generation-sequencing methods has the potential to routinely decode modified bases without prior chemical transformation. Single molecule real-time sequencing (SMRT-seq) uses a processing polymerase to insert nucleotides opposite a template strand. The identity of each base is determined by its unique kinetic signature, thus, modified bases can be distinguished from the canonical ones in this manner.[128] Chemical labelling can be used in conjunction with SMRT-seq, since bulkier bases have a more unique signature which can aid detection.[129] Additional advantages of SMRT-seq include the ability to decode multiple modifications in one sequencing run, and the ability to sequence longer reads (up to 20000 bp). Unfortunately, current technology is still low throughput and is only feasible for the sequencing of small genomes, or in larger organisms after affinity enrichment. Furthermore, this sequencing technology is currently less widespread and accessible, and requires large amounts of optimization for each modified base. Regardless, SMRT-seq has already been used to map cytosine modifications in smaller organisms such as fungi[130], Base J in trypanosomes[131] and is likely to be an extremely important tool in the future. Similarly, nanopore-sequencing measures changes in ionic current while a nucleotide strand electrophoretically passes through a nanopore.[132,133] Modified bases have a different current to the canonical bases, and hence can be distinguished. However, the accuracy of this sequencing technology requires improvement before it is routinely used for this purpose in the future.[134]

1.5.2.viii. Application of Sequencing Methods in this Study

The development of chemical and biological tools have been essential to decode the genomic location and function of modified bases in eukaryotic DNA by sequencing. The current methods applicable to an NGS platform are hence summarised in Table 2.

Modified base	Chemical Enrichment techniques	Antibody Enrichment techniques	Single-base resolution techniques
5-mC	No technique published	5Me-DIP[135]	Bisulfite ox-BS[124]
5-hmC	GLIB-seq[136] hMeSeal-seq[117]	5hme-DIP[137] CMS-DIP[136] JBP1-DIP[138]	ox-BS[124] TAB-seq[126] SCL-exo[139] AbaSI-seq[140] PvuRts1[141]
5-fC	Aldehyde reactive probe[51,53] fCSeal-seq[52]	5fC-DIP[42]	fC-CET[127] red-BS[125] fCAB-seq[52] MAB-seq[142] CLEVER-seq[143]
5-caC	No technique published	5caC-DIP[42]	DIP-Cab-seq[144] MAB-seq[145]
5-hmU	No technique published	No technique published	No technique published
5-fU	No technique published	No technique published	No technique published
Base J	No technique published	Base J-DIP[98]	No technique published
8-oxoG	OG-seq[146]	8-oxoG DIP[56]	No technique published

Table 2: Current methods to sequence modified bases in DNA which are compatible with Illumina NGS technology

There are currently no methods to determine the genomic location of the T-modifications 5-hmU and 5-fU. Thus, chemical and biological tools will be developed and utilized in this thesis to enable genome-wide mapping of these modifications in both mammalian and trypanosomatids. This will provide further insight into the role of T-modifications in these biological systems.

1.5.3. Proteomics and Modified Base Protein Interactions

Other methods to probe the biological function of modified bases in DNA arises from the study of DNA-protein interactions. Modified bases possess different chemical moieties which protrude into the DNA major groove and as such, these moieties have the propensity to ascribe unique molecular binding interactions with specific proteins. Protein binders may include "writers" which generate the modification, "readers" which recognise and interact with the modification, and "erasers" which excise the modification from the genome.[147]

Proteomics, profiling proteins by mass spectrometry, is typically used to decipher which proteins are recruited to modified base containing fragments. Biotinylated "bait" modified DNA strands can be incubated in the presence of cellular protein extract, and interacting proteins can be identified by mass (Figure 15).[49]

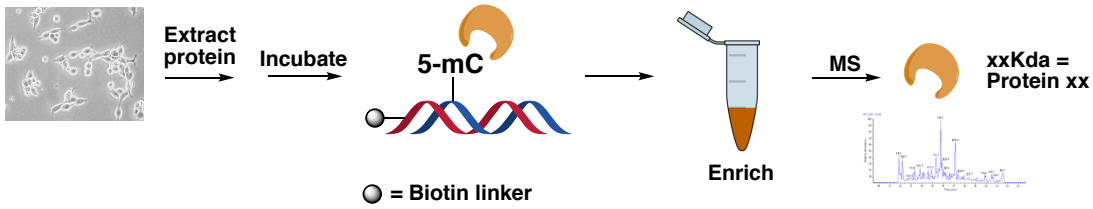


Figure 15: Proteomics can be used to decipher proteins that preferentially interact with modified bases.

Examples of such techniques include those which led to the discovery of MCBP2.[148] This protein is found to bind specifically to methylated CpG dinucleotide containing sequences, and further work led to the discovery of a whole host of other MCBPs. These proteins are found to be essential regulators of the transcriptional state of the epigenome, and facilitate epigenetic crosstalk by interactions with histone modifying proteins.[25] Proteomics studies have thus far been reported for all the C derivatives and 5-hmU in mammalian tissue, therefore 5-fU protein binders will be explored in this thesis.

1.5.4. Effect of Modified Bases on Nucleosome Structure

Studies can be performed to probe the effect of modified bases on higher order chromatin structure. Chromatin consists of strings of nucleosome subunits connected by linker DNA, where each subunit is composed of eight histone proteins (H2A, H2B, H3, H4 subunits) wrapped round 147bp of DNA.[149] Nucleosomes are the basic fundamental units of chromatin and control access to genetic information. The dynamics and occupancy of nucleosomes is therefore a huge determinant of gene regulation, and the

establishment of dynamic regulatory regions marked by nucleosomes is essential for the gain of cell identity. The presence of DNA modifications is shown to directly alter and affect both nucleosomal positioning and stability, leading to a downstream effect on gene expression.[150]

Due to the complexity of chromatin, researchers often study one nucleosomal unit to study chromatin interactions with DNA *in vitro.* DNA CpG methylation, for example, demonstrates more rigid wrapping around the histone octamer in this model, which may reflect 5-mC's association with gene silencing.[151] 5-hmC modified DNA demonstrated increased nucleosomal stability compared to unmodified C, however the presence of 5-hmC showed weakened interactions with the H2A-H2B subunit dimer.[152]

The relationship between DNA modifications and nucleosome occupancy can also be probed using Micrococcal Nuclease (MNase)-seq. In this method, DNA that is not associated with chromatin is enzymatically digested, whilst nucleosomal DNA remains intact for sequencing. This generates a genome-wide map of nucleosome structure, which can be directly correlated with the presence, or absence, of DNA modifications.[153] DNA with distinctive periodicity (10bp) of certain dinucleotides (CpG) is characteristic of positioned nucleosomes, and the methylation status at these sites has been shown to alter DNA affinity towards the histone octamer.[154] Methylated CpG dinucleotides are strongly correlated with nucleosome occupancy,[155] while 5-hmC loci and TET-binding sites are instead associated with labile MNase-sensitive nucleosomes, poised for eviction, in mESCs.[153]

The study of modified-base nucleosomal interactions can provide great insight into how modified bases can fundamentally alter nucleosomal organisation and downstream gene expression. However, there are currently limited studies on the effect of formylated modified bases, such as 5-fC and 5-fU, on higher chromatin structure. The role of these marks will be explored in this thesis, to provide a unique mechanistic insight into the association of these modifications with nucleosomal occupancy.

2. Global Measurements of Modified Bases by LC-MS/MS

2.1. Introduction of LC-MS/MS Methods and Workflow

Quantification of modified bases is a powerful tool to compare global levels of modifications in different organisms, tissues and phenotypes; liquid chromatography tandem mass spectrometry (LC-MS/MS) is the most sensitive and discriminating method for the accurate measurement of modified bases.

2.1.1 LC-MS/MS Workflow

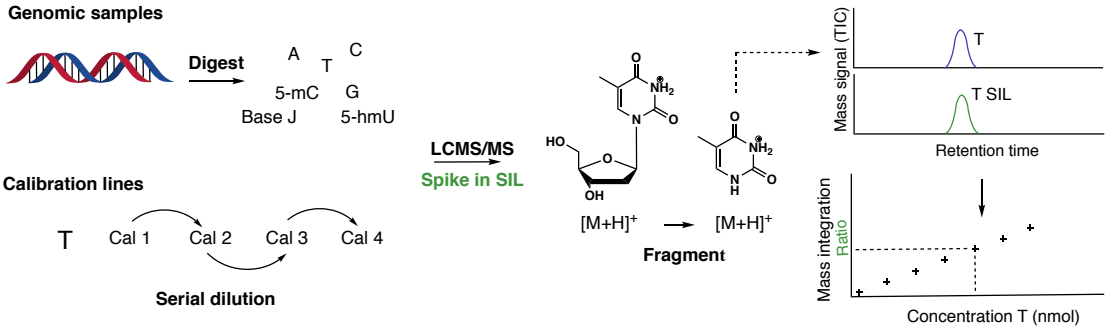


Figure 16: Schematic demonstrating the workflow of LC-MS/MS measurement, labels in green demonstrate internal calibration using an internal SIL standard.

Genomic DNA, extracted from tissue or cells, is initially enzymatically digested into its constituent mononucleosides.[40] Mononucleosides are then separated by chromatographic retention followed by mass analysis, where the parent ion of each modified base is targeted for further fragmentation. Accurate mass signals from the corresponding nucleoside fragments are extracted from the total ion count (TIC) and subsequently integrated for quantification; this integrated mass signal is used to determine the concentration of a nucleoside within a genomic sample, by comparison with a calibration line of nucleoside standards (Figure 16).

The accuracy of mass quantification can be improved via the use of an internal standard that is "spiked" into both calibration standards and biological samples. Quantification is instead determined via the mass integration area ratio of nucleoside/internal standard. This corrects for matrix effects which arise in complex biological samples,[156,157] where molecules that originate from the sample matrix co-elute with the biological target causing suppression (or enhancement) of the TIC. The gold-standard of mass quantification employs an internal standard that is an isotopically labelled (SIL) variant of the target compound.[113,157] A SIL has the same chromatographic and ionisation

properties as the non-labelled modified base, however it has a distinct mass signal that can be integrated separately.

After digestion, several modified bases can be targeted and quantified in parallel. Comparison of the modified base concentration with that of a canonical nucleoside (e.g. T) determines the percentage modified base present within a genomic sample. Levels are therefore reported as a proportion of one nucleoside (e.g. per T) or of the total number of nucleosides (e.g. per N, taking into consideration the percentage GC content for the organism in question). Modern mass spectrometry can accurately quantify over a wide range of concentrations and is sensitive enough to detect femtomole quantities of modified bases.[40,158] As such, LC-MS/MS methods have been integral in determining the global levels of DNA modifications in different cell types, phenotypes and tissues within the modified bases field.[48,159,160]

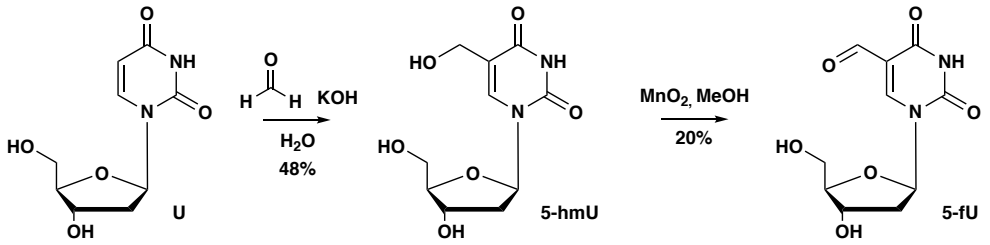
2.1.2. Global Measurement of T-modifications in DNA

The aim of this chapter was to develop methods to accurately quantify the T-modifications, 5-hmU, 5-fU and Base J, in genomic DNA. The ability to measure the global levels of these bases would highlight biological targets for further study, designed to complement the development of T-modification enrichment-sequencing methods (See Chapters 3 and 4). Furthermore, LC-MS/MS could be used to assess changes in T-modification global levels in response to biological perturbation, thereby providing greater insight into the role and metabolism of these marks. A main target was the accurate measurement of T-modifications in both trypanosomatids and mammalian systems. At the outset of this project, several groups had reported LC-MS/MS measurement of T-modifications in mammals[40,160,161,162]; yet, accurate quantification of these marks had yet to be reported for trypanosomal DNA. LC-MS/MS measurements for Base J and 5-hmU have since been reported in bloodstream-form (BSF) trypanosomes, at levels of 0.5% and 0.017% per T for Base J and 5-hmU respectively.[90]

2.2. Synthesis of 5-hmU and 5-fU Nucleoside Standards

In order to accurately quantify T-modifications in genomic samples using the internal calibration strategy, nucleoside standards (both with and without SIL) were required. 5-hmU mononucleoside was synthesised in one step by reaction of deoxyuridine (U) mononucleoside with formaldehyde in the presence of potassium hydroxide, utilising a published procedure (Scheme 1).[163] The 5-fU mononucleoside was subsequently synthesised by 5-hmU oxidation using manganese dioxide,[163] a 1-electron oxidant with

selective reactivity towards benzylic or allylic alcohols.[164] This enabled chemoselective oxidation of the allylic primary hydroxyl group on the nucleobase, leaving the primary hydroxyl group on the sugar intact.



Scheme 1: Reaction of U with formaldehyde under basic conditions yields 5-hmU mononucleoside. Subsequent radical oxidation of 5-hmU generates the 5-fU mononucleoside.

Analogous syntheses were employed to generate +3 5-hmU (hmU-SIL) and +2 5-fU (fU-SIL) standards utilising heavily labelled formaldehyde ($^{13}CD_2O$) (Figure 17). Deuterated solvents were used as a precaution to prevent any proton exchange and loss of isotopic integrity. Other mononucleoside standards (C, T, Base J, 5-mC and 5-hmC) were either commercially sourced or had been synthesised by other members of the Balasubramanian group.

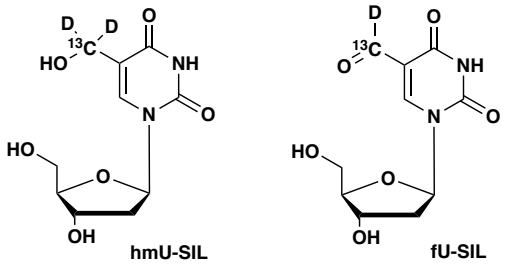


Figure 17: Synthesised isotopically labelled standards: hmU-SIL and fU-SIL

2.3. Background and Improvement of nano-HPLC Q-exactive set-up

A Dionex 3000 nano-HPLC coupled with a Q Exactive™ (QE) mass spectrometer was utilised for LC-MS/MS measurements. The Balasubramanian group had previously used this set-up for quantification of C-modifications, using a custom-packed 0.075 µm x 5 mm Hypercarb column, suitable for nano-flow.[45,46] Hypercarb is a porous graphitic carbon (PGC) column material, and has superb retention and separation of polar analytes such as nucleosides. However, the former LC set-up required custom-made columns which were prone to leakage and variability in chromatography.

Chromatographic reproducibility was therefore vastly improved with the use of a 0.18 µm Hypercarb KAPA capillary column. This column could be directly connected to the Dionex 3000 using two fingertight nanoviper connectors, which greatly reduced previous issues with dead-volume and leakage (Appendix – Chapter 2).

2.4. Validation of Nucleoside Standards, Calibration Curves and Detection Limits

The synthetic nucleoside and SIL standards were initially validated using the LC-MS/MS set-up, which confirmed the co-elution of SIL and non-labelled standards (Figure 18). It was also ensured that the SIL standards contained no detectable signal derived from the unlabelled nucleoside, to avoid contamination of genomic samples.

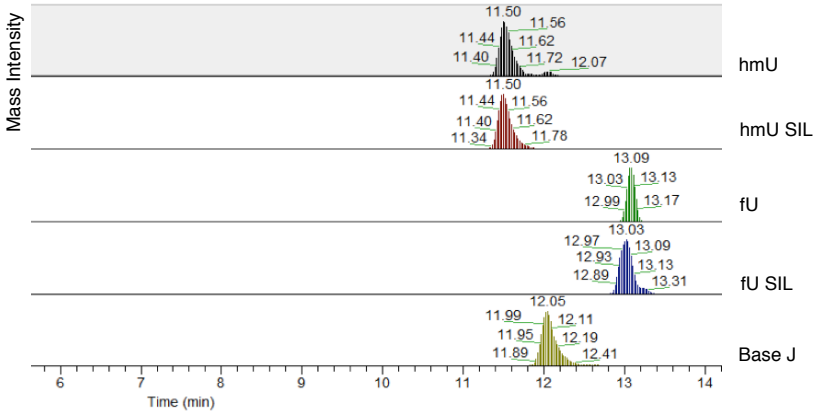


Figure 18: Validation of unlabelled and SIL nucleoside standards of 5-hmU, 5-hmU SIL, 5-fU, 5-fU SIL and Base J.

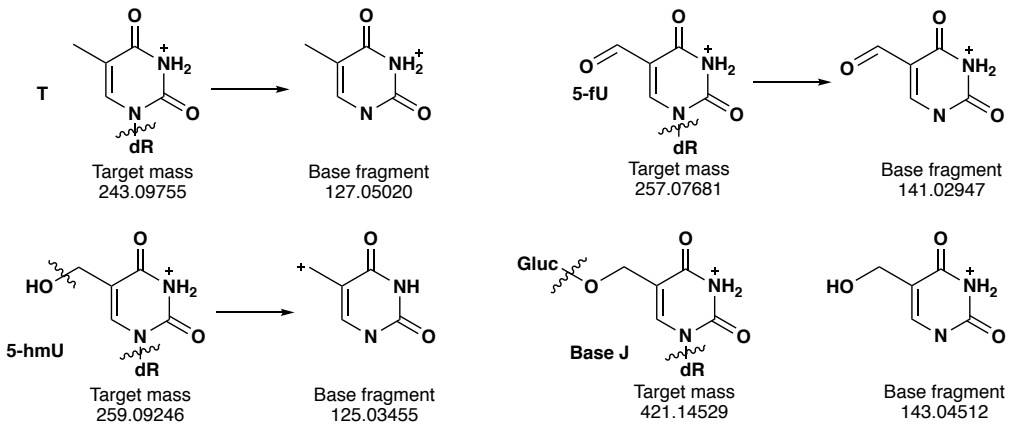


Figure 19: Schematic of fragmentation pattern of T-modifications and their most abundant fragment ion.

Next, the most ionisable fragment for each nucleoside was determined to be used for quantification (Figure 19). For T and 5-fU, the most ionisable fragment corresponded to the protonated base fragment after cleavage of the *N*-glycosidic bond (T 243 → 127, T SIL: 146 → 130; fU 257 → 141, 5-fU SIL: 259 → 143 respectively). For 5-hmU the most ionisable fragment corresponded to loss of the deoxyribose group and hydroxyl group

(5-hmU: 259 → 125, 5-hmU SIL: 262 → 128), whilst in Base J this corresponded to loss of deoxyribose and the glucose moiety (421 → 143). Accurate calibration lines were constructed for each of the above T-modifications and their relative detection limits were determined (Figure 20, Table 3). A Base J SIL was not required since the use of hmU-SIL as an internal standard led to excellent linearity (Figure 20). Notably, the T-modifications were observed to have much higher limits of detection compared with the analogous C-modifications (Table 3); this is likely due to the lower proton affinity possessed by T-modifications.[165]

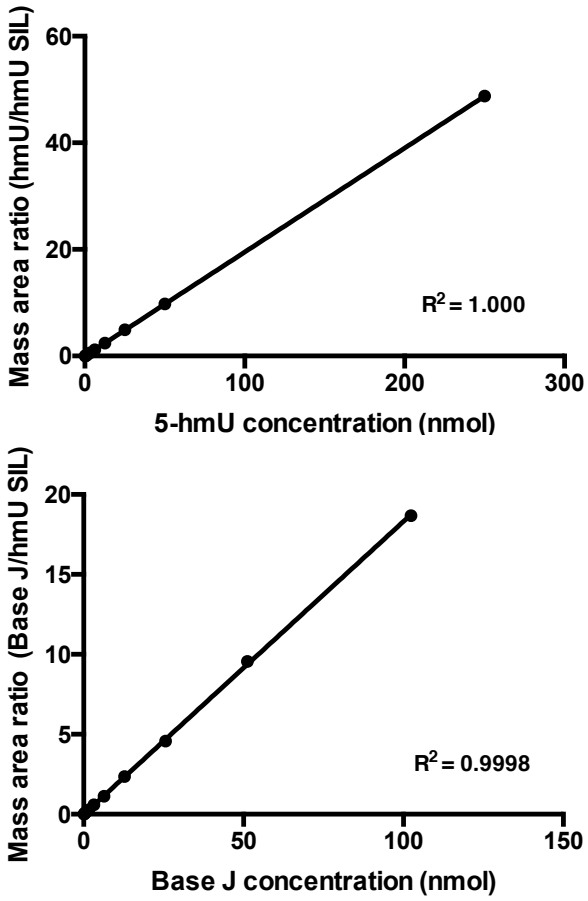


Figure 20: Calibration lines for quantification of Base J and 5-hmU.

Modification	Quantitative detection limit (fmol)
5-mC	0.9
5-hmC	0.2
5-fC	2
5-hmU	2
5-fU	30
Base J	15

Table 3: Table demonstrating relative quantitative detection limits of modified bases on the Q-Exactive spectrometer, determined from calibration standards.

2.5. Detection and Quantification of Modifications in Trypanosomatids

The optimised and validated method was used to accurately quantify Base J and 5-hmU in the trypanosomatids *Leishmania major* (*L.major*) and *Trypanosoma brucei (T. brucei)*. The measurements obtained were in accordance with the reported LC-MS/MS levels in bloodstream-form (BSF) *T. brucei*,[90] suggesting the method was robust. Levels of these marks in another trypanosomatid, *L. major,* were found to be similar (Figure 21).

The next aim was to detect and quantify T-modifications in procyclic form (PCF) *T. brucei*. There were several references to suggest that this life-form lacks Base J,[166] however, one report had detected 5-hmU in this life-form by ^{32}P radiolabelling.[108] Thus, the presence of 5-hmU in PCF trypanosomes was firstly confirmed, followed by accurate LC-MS/MS quantification. 5-hmU levels were found to be ~5-fold lower than in BSF trypanosomes, whilst Base J was undetected (Figure 21).

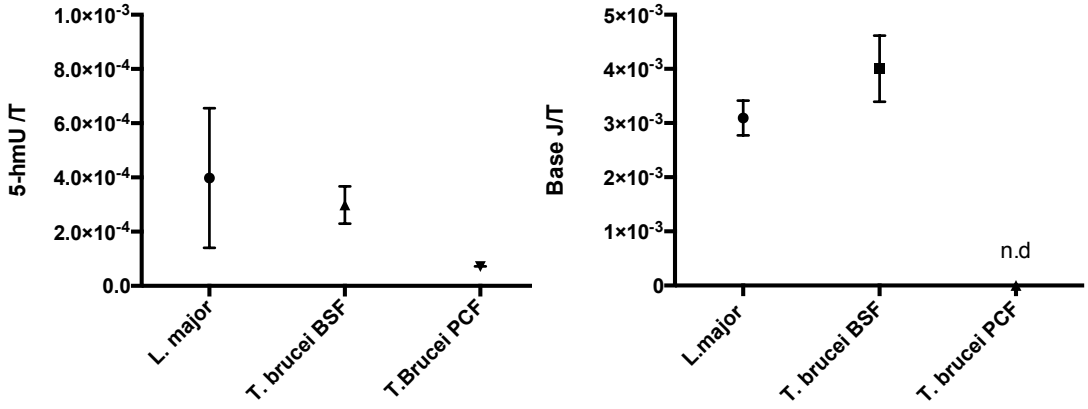


Figure 21: Accurate T-modification quantification in *L. major* and the BSF and PCF forms of *T. brucei* (Left: 5-hmU Right: Base J). n.d. = not detected. For *T. Brucei* Genomic DNA was extracted in the presence of antioxidant inhibitors, to minimise spurious oxidation during sample processing.[40] *L. major* samples were measured using the 0.075 mm x 5 mm custom-packed hypercarb nano-column. *T.Brucei* samples were measured using the 0.18 mm x 3 mm commercial capillary hypercarb column.

The presence of 5-hmU and associated lack of detectable Base J in PCF trypanosomes is of particular interest. The JBP enzymes 1 and 2, responsible for T oxidation in trypanosomes, are downregulated by ~10 and ~5 fold respectively in PCF trypanosomes.[167] Liu *et al.* also detected the presence of 5-hmU in JBP-null bloodstream-form trypanosomes, suggesting 5-hmU may be generated in a JBP-independent process.[90] Furthermore, J-GT, the glucosyltransferase responsible for converting 5-hmU to Base J, is still expressed in procyclic trypanosomes. Indeed, it has been reported that Base J is formed in PCF *T.Brucei* when cells are artificially fed with 5-hmU.[166,92] It was

therefore unclear why 5-hmU is observed, and not Base J, based on the current dogma; however, 5-hmU generation by spurious oxidative damage cannot be ruled out.

Furthermore, the presence of 5-fU and 5-mC modifications in trypanosomes was also investigated. Whilst 5-fU was undetectable in these samples, likely as a result of its high detection limit (Table 3), 5-mC was detected in both BSF and PCF trypanosomes. The presence of 5-mC in this organism is supported by Militello *et al.*, who identified a DNMT gene in *T. brucei,* expressed at similar levels in both-life forms.[168] Unfortunately, the levels of 5-mC were too low to be accurately quantified. Methods therefore needed to be developed for the quantification of low abundance modifications using the nano-HPLC coupled QE set-up.

2.6. Towards Quantification of Low-abundance Modifications in Genomic Samples

Modified base quantification is challenging when modifications are in low abundance or are poorly ionising. This is complicated by the high background signal of the canonical nucleosides, and other components of the digestion mixture, which leads to ion suppression. T-modifications in mammalian tissues are much less abundant than in trypanosomatids (Introduction 1.4.1), and T-modifications have much higher limits of detection compared to the analogous C-modifications (Table 3). The nano-flow HPLC instrumentation restricts the amount of DNA that can be digested and subsequently loaded onto the column, thus accurate quantification of low abundance modifications (e.g. 5-hmU and 5-fU in mammalian samples) was not possible using this set-up. Alternative strategies were therefore required to accurately quantify such modifications. Once developed, these methods would be applicable for the quantification of any low abundance modified base.

2.6.1. HPLC Enrichment

HPLC pre-enrichment of modified bases before LC-MS/MS injection was therefore employed as a method for low abundance base quantitation (Figure 22).[162] This allowed larger quantities of DNA to be digested before enrichment, thus improving the mass intensity signal of modified bases for more accurate quantification. An added benefit was that HPLC pre-enrichment enabled purification of modified bases from the higher abundance canonical bases within the digestion mixture, which helped to alleviate background signal and ion suppression.

Synthetic nucleoside standards were first used to determine the necessary timepoints for enrichment of each modified base using HPLC. Enrichment of a known amount of 5-hmU and 5-fU mononucleoside demonstrated full recovery, indicating the method is quantitative. The enrichment method was then applied to the digested genomic samples, before injection into the QE mass spectrometer. Quantification of a canonical nucleoside (e.g. T) in a diluted portion (e.g. 1 in 50) of digested samples enabled modified bases to be reported as a proportion of total nucleosides.

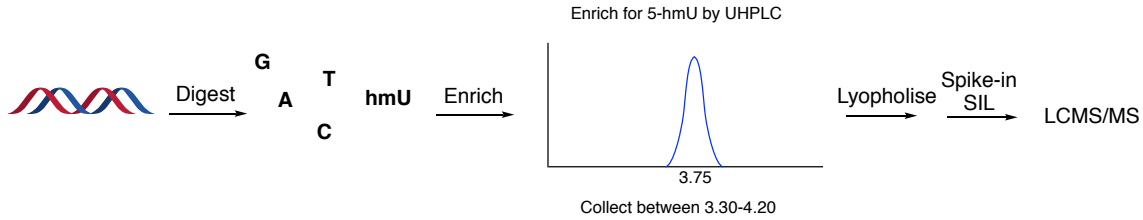


Figure 22: HPLC pre-enrichment of modified bases before LC-MS/MS injection

2.6.1.i C-modifications in Trypanosomatids

The HPLC pre-enrichment strategy was first used to accurately quantify the level of 5-mC in trypanosomatids (Figure 23 - Left). Similar levels of 5-mC were found in both BSF and PCF *T. brucei,* as expected by the similar levels of reported DNMT expression.[168] Interestingly, the presence of 5-hmC in this organism was also detected, which was supported by the findings of Valentine *et al.*[169] Since *T. brucei* lack TET enzymes, this indicates that JBPs may also have the potential to oxidise 5-mC → 5-hmC along with T → 5-hmU oxidation.

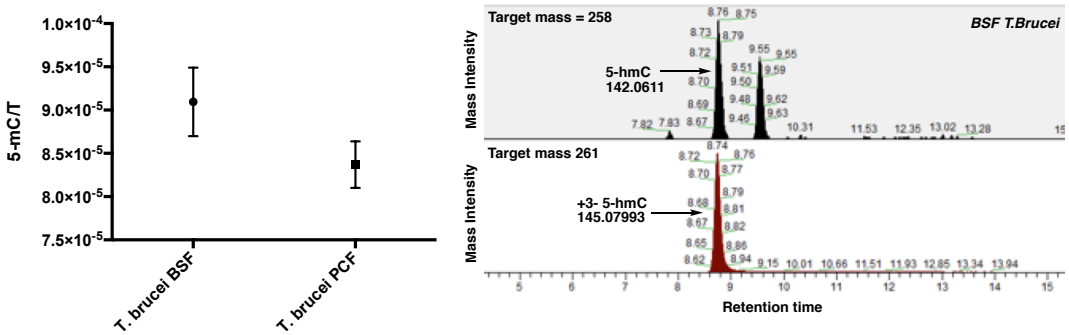


Figure 23: Left: Accurate quantification of 5-mC in both BSF and PCF *T. brucei*. Right: 5-hmC detection in *T.brucei* and confirmation using hmC-SIL.

2.6.1.ii 5-hmU and 5-fU Quantification in Biological Samples

The enrichment strategy was next applied to quantify the levels of 5-hmU and 5-fU in a range of organisms, including mammals (Figure 24 - Left). Large amounts (> 20μg) of DNA were digested for each replicate to ensure that the 5-hmU and 5-fU signals were at a concentration sufficiently above the limit of detection and background level. This proved to be more problematic for 5-fU, hence measurements were only recorded in a subset of the samples screened (Figure 24 - Right).

The data suggests no significant increase in 5-hmU and 5-fU levels in cancer cell-lines (MCF7) in comparison to somatic tissue (mESC and HEK293T cells). High 5-hmU levels in the blood have previously been suggested to be a biomarker of cancer[68,69], however in these circumstances, measurements reflected the level of free mononucleoside rather than the level of modified base present within genomic DNA. Furthermore, the data acquired suggests that 5-hmU levels are elevated in the model plant *A. Thaliana* compared to the other organisms screened. This may result from increased ROS susceptibility due to constant UV radiation exposure,[170] whilst there is also no known 5-hmU glycosylase in this organism.[171]

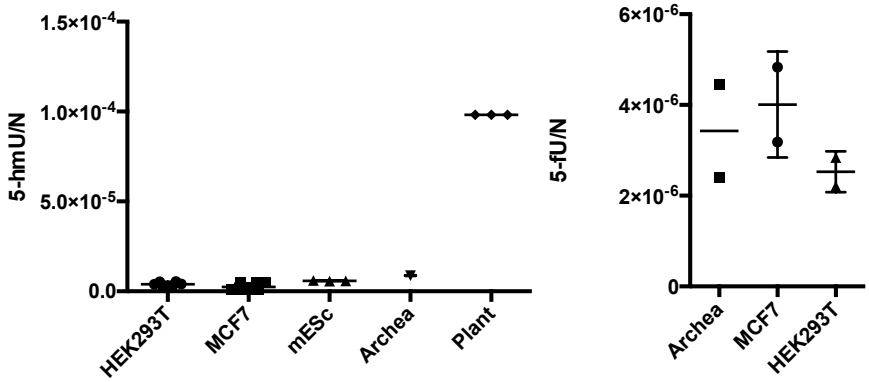


Figure 24: Left: 5-hmU measurements in different organisms, Right: 5-fU measurements in different organisms.

The observed 5-hmU level in HEK239T genomic DNA was on average ~5-fold higher than those reported by Carell and co-workers,[40] yet in accordance with levels reported by Liu *et al.*[162] The difference in levels is likely explained by variance in oxidative damage within biological systems/technical replicates, which can also account for the variability of 5-hmU levels reported in the literature for mammalian samples.[40,161,162]

2.7. Conclusion

A focus of this chapter was the development of methods to accurately quantify T-modifications in biological systems. In the process, methods to quantify low abundance modifications have been investigated and developed, and these methods can now be used for the global measurement of any low-abundance modified bases in DNA.

Accurate T-modification quantification is demonstrated in both trypanosomatids and mammalian samples, and LC-MS/MS measurements can now be used to support other studies described in this thesis (Chapter 4). Whilst LC-MS/MS enables the global measurement of T-modification levels, this method fails to provide any sequence context information. Thus, methods need to be explored to affinity-enrich DNA fragments containing T-modifications in combination with NGS, to determine the genomic loci of T-modifications in biological samples (Chapters 3 and 4).

Due to potential variance in T-modification oxidative damage in mammalian samples, several biological/technical replicates should ideally be measured when comparing global levels of T-modifications between different phenotypes or tissues. This requires a large amount of genomic material, thus a future goal would be to explore derivatisation chemistry to improve the detection limit of low abundance modifications in genomic samples. Reactive handles on modified bases, (e.g. the formyl group of 5-fU) can be tagged with a charged or highly ionisable moiety to facilitate their detection and quantification in smaller quantities of genomic DNA.[172] A derivatisation strategy for 5-fU was briefly explored (Appendix – Chapter 2), although further optimisation is required to enable accurate quantification at this stage.

3. Selective chemical labelling of T-modifications

3.1. Introduction

In order to further understand the role of T-modifications in both mammals and trypansomatids, the aim of this chapter was to develop a chemistry-based affinity method to enrich DNA fragments containing 5-hmU and 5-fU from genomic samples. This would require selective chemical tagging (e.g. biotinylation), followed by affinity purification or "pulldown" of tagged fragments using streptavidin. In combination with next-generation DNA sequencing (NGS), this method could be used to generate genome-wide maps of T-modifications. Ultimately, these maps can be correlated with other available data sets (e.g. gene expression, nucleosome structure, protein binding) to help understand the function of T-modifications in biological systems. A chemical enrichment method would be complementary to an antibody-based hmU-DIP method that was being developed in parallel within the Balasubramanian group.

Whilst many methods have been developed to map the analogous C-modifications, there were no reported methods for the affinity-enrichment of 5-hmU and 5-fU (Introduction – 1.5.2.viii). Therefore, this study aimed to establish a chemical or chemo-enzymatic tagging method for the T-modifications that was selective over other DNA modifications and the canonical bases. In order to avoid any uncontrolled DNA damage (e.g. depurination or backbone fragmentation), such a reaction needed to proceed in mild aqueous conditions (pH > *ca.* 5). Furthermore, the reaction ideally required quantitative or near-quantitative labelling for efficient T-modification pulldown.

Due to the presence of a reactive aldehyde moiety on 5-fU, it was envisaged that Schiff base, alkoxime or hydrazine formation could be utilised for labelling this base. (Figure 25)

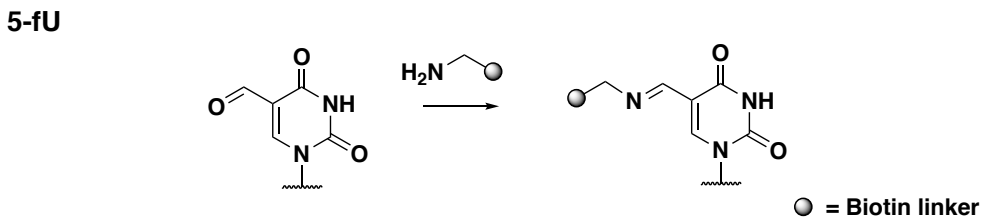


Figure 25: Strategy to tag 5-fU via formation of a Schiff base, hydrazone or oxime

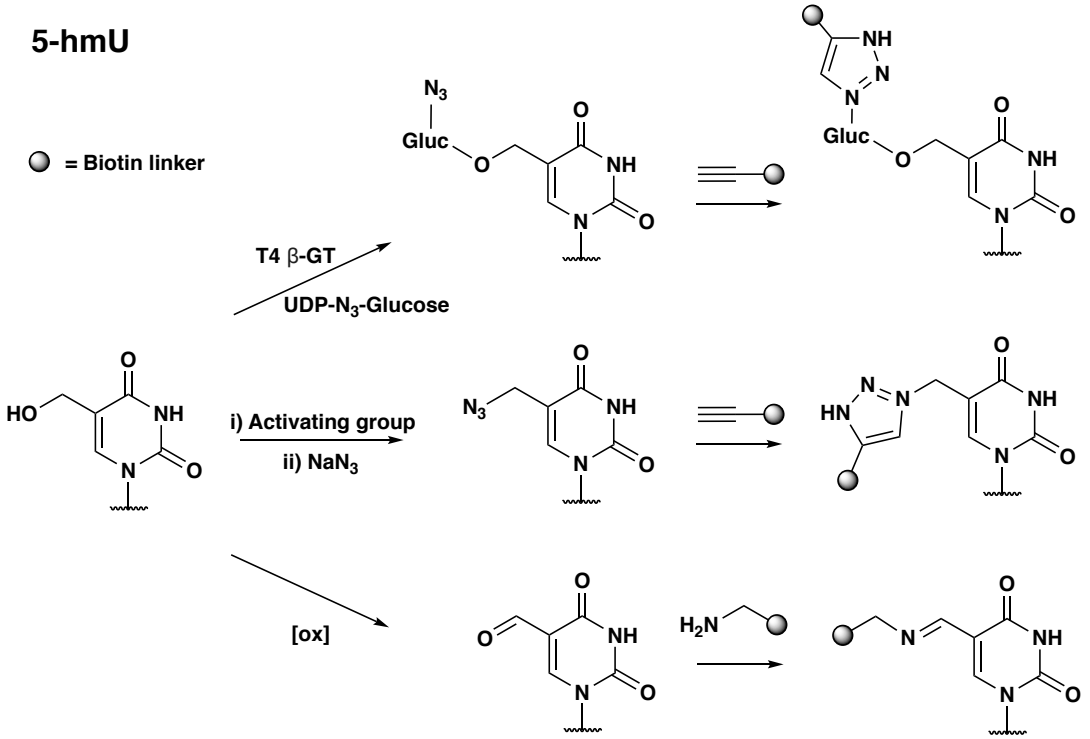


Figure 26: Strategies to tag 5-hmU via i) β-glucosylation and subsequent click chemistry, ii) azide substitution at the hydroxyl group followed by click chemistry or iii) oxidation and 5-fU tagging.

For 5-hmU, it was hypothesised that the base could be functionalised via three different strategies (Figure 26). Firstly, 5-hmU was likely to be a natural substrate for the T4-β-glucosyltransferases and could therefore be tagged chemoenzymatically, analogously to 5-hmC (Introduction – 1.5.2.iii).[116] Secondly, the hydroxyl group of 5-hmU could be directly functionalised by activating the hydroxyl group followed by nucleophilic substitution with sodium azide, followed by 1,3-dipolar cycloaddition "click" chemistry with a biotinylated alkyne.[173] Thirdly, 5-hmU could be synthetically oxidised to 5-fU, which could be subsequently tagged using the same strategy for 5-fU as discussed above. A major criterion for every method would be a way to selectively discriminate the T-modifications from canonical and other naturally occurring modified bases. One obvious challenge for this strategy was the very similar chemical reactivity of 5-hmU and 5-hmC, and also of 5-fU and 5-fC, all of which co-exist in genomic DNA. This was especially a concern for 5-hmC, which has a much higher overall abundance than 5-hmU in mammalian tissue (Figure 27).[40]

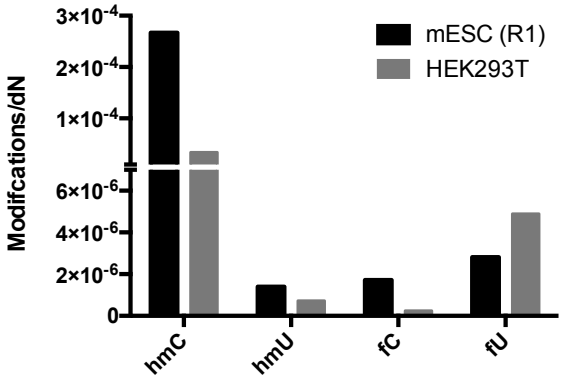
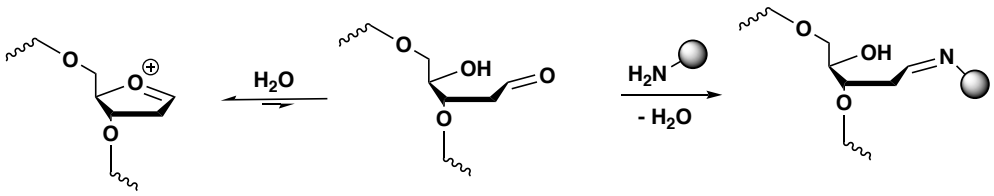


Figure 27: Relative modification levels of hmC, hmU, fC and fU in mESC (R1) and HEK293T cells using data reported by Carell and co-workers.[40]

Furthermore, selectivity is also required over apurinic/apyrmidinic (AP) sites; these sites are defined by the lack of a DNA base, and are another potential source of aldehydes present in genomic samples. Although AP sites thermodynamically prefer the ring-closed form of the sugar, they are in equilibrium with the ring-open aldehyde-containing form, which can be trapped in the presence of an aldehyde probe (Figure 28). AP sites are formed enzymatically as intermediates in the BER pathway (Introduction 1.3.2), but also arise due to DNA damage (depurination or depyrimidination). Reports estimate the level of AP sites to be 1-10 per 10^6 nucleosides in murine tissue, however these estimates do not consider other aldehydes present in DNA.[174] In addition, as well as natural DNA damage, AP sites can also be induced during DNA extraction or in response to harsh chemical treatment of DNA.[175] As a result, a chemical pulldown strategy must be extremely chemoselective for the T-modifications to ensure no background reactivity or subsequent pulldown of C-modifications or AP sites.



Apurinic/Apyrmidinic site

Figure 28: AP site equilibrium between the closed and open sugar form, which can be trapped as a result of Schiff base, hydrazone or oxime formation.

3.2. Chemoenzymatic Selective Glucosylation of 5-hmC

As already highlighted, 5-hmC can be tagged by exploiting a natural enzymatic glucosylation reaction which occurs in T4 phages (Introduction 1.5.2.iii). It was therefore investigated whether the same strategy could be applied towards 5-hmU.

To assess the 5-hmC/5-hmU substrate preference for T4 β-GT glucosylation, synthetic double stranded DNA (dsDNA) containing 5-hmC or 5-hmU (GhmCAT, GCAhmU and GhmCAhmU) was synthesised by polymerase chain reaction (PCR). To synthesise modified base-containing DNA, modified triphosphates can be used to replace the canonical base in the PCR reaction (e.g dhmCTP replaces dCTP),[124] where modified bases are inserted in all non-primer regions. Synthetic DNA strands bearing either 5-hmU or 5-hmC were incubated with UDP-glucose in the presence of T4 β-GT. After purification, the DNA was enzymatically digested into its composite nucleosides, and analysis was carried out by measuring HPLC consumption of 5-hmC and 5-hmU. While the 5-hmC signal was depleted after the glucosylation reaction, the 5-hmU signal remained intact (Figure 29), suggesting that the β-glucosylation is selective for 5-hmC over 5-hmU. Subsequent LC-MS/MS analysis (Chapter 2) revealed the conversion of 5-hmC to 5-glchmC was >99%, whilst only a small proportion of 5-hmU was consumed (4%).

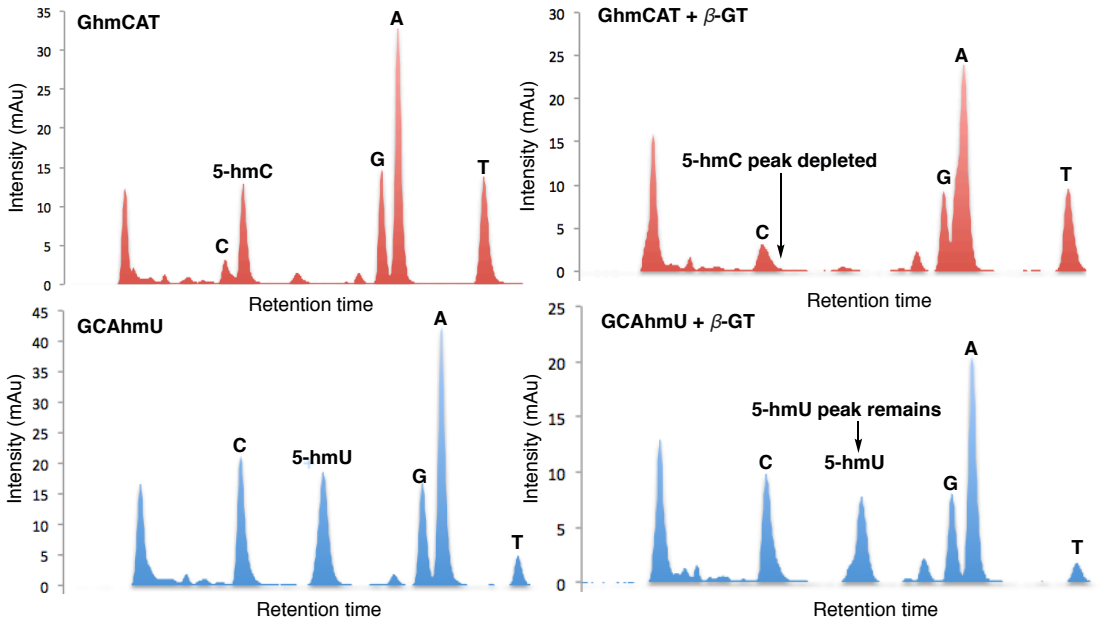


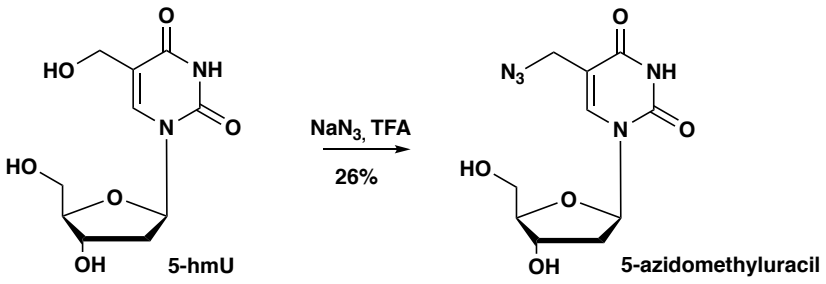
Figure 29: HPLC trace of digested mononucleotides of DNA. Top: Left - GhmCAT starting material, Right - GhmCAT after treatment with T4 β-GT (the 5-glchmC product is likely co-eluting with a canonical nucleoside, most likely A), Bottom: Left - GCAhmU starting material, Right - GCAhmU after treatment with T4 β-GT, the 5-hmU peak is not consumed.

These results have since been corroborated by Wang and co-workers[176] where efficient glucosylation of 5-hmU with T4-β-GT was shown to only occur when 5-hmU was mispaired with guanine. Since the goal was comprehensive tagging of 5-hmU regardless of sequence (hybridisation) context, this raised questions about the utility of 5-hmU tagging via the T4-β-GT strategy. However, 5-hmC glucosylation could be useful to block 5-hmC from further functionalisation, allowing selective reactivity of 5-hmU.

He and co-workers have since utilised this result to develop a proof-of-concept enrichment method for fragments containing hmU:G mispairs using an azido-glucose based enrichment strategy analogous to that used for 5-hmC.[177] This method could be used to look specifically for such mispairs generated by 5-hmC deamination. However, since most naturally occurring 5-hmU occurs in a T:A base-pair context,[40] this casts doubt over the usefulness of this overall approach.

3.3. Direct Hydroxyl Group Activation

The feasibility of direct functionalisation of the 5-hmU-hydroxyl group was next explored. A publication by Zhou and co-workers had highlighted a method to directly functionalise the hydroxyl group of 5-hmU selectively over 5-hmC at the mononucleoside level, using sodium azide in trifluoroacetic acid (TFA) (Scheme 2).[178] If an analogous reaction could be optimised for reaction with DNA oligomers, the resultant azide could then be coupled with a biotinylated alkyne using 1,3 dipolar cycloaddition click chemistry.



Scheme 2: Reaction of 5-hmU with sodium azide to give target molecule 5-azaU which was used as an LC-MS standard to assess the potential of this strategy.

As TFA is an unsuitable solvent for reactions involving DNA, due to the extensive depurination and depyrimidination that occurs under acidic conditions, a range of milder acidic agents were examined in the presence of sodium azide, including the use of the water-soluble Lewis acids scandium triflate and ytterbium triflate. However, no

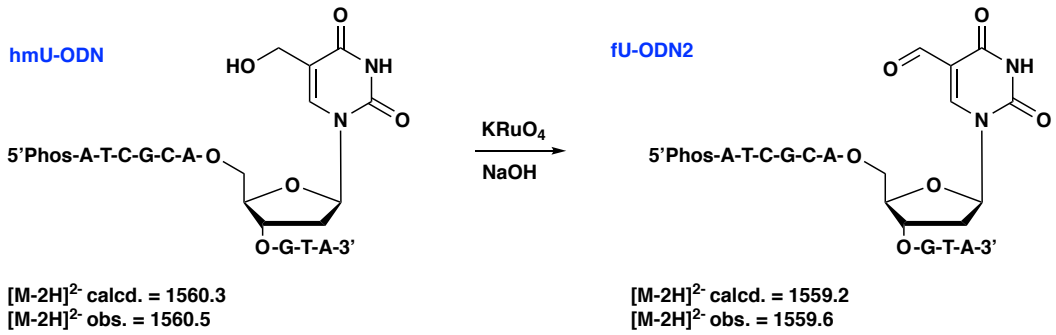
azide substitution was observed by LC-MS, highlighting the challenge of dehydrative activation of the hydroxymethyl group in aqueous conditions. An analogous strategy using the water-soluble coupling agents 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium chloride (DMT-MM) and *N*-(3-dimethylaminopropyl)-*N*'-ethylcarbodiimide hydrochloride (EDC) was examined, where it was postulated that the hydroxyl group could be activated to nucleophilic attack. DMT-MM has previously been used to selectively activate and functionalise the anomeric hydroxyl group of mono- and oligosaccharides in water for further glucosylation reactions.[179] This agent is also a water-soluble analogue of cyanuric chloride, which has been used to catalyse azide functionalisation of hydroxyl groups.[180]

Whilst LC-MS analysis suggested that 5-hmU mononucleotide reactions with both DMT-MM and EDC resulted in a product with the expected mass (see Chapter 3 - Appendix for further discussion), functionalisation did not occur at the desired location. Since direct activation of the 5-hmU hydroxymethyl group appeared to be challenging, the oxidation approach was pursued instead.

3.4. Oxidation and Aldehyde Tagging Strategy

3.4.1. 5-hmU Oxidation

Oxidation of 5-hmU to 5-fU would allow functionalisation of the resultant reactive aldehyde group. Potassium perruthenate in the presence of sodium hydroxide was used to oxidise a deoxynucleotide-oligomer (ODN) that contained a single 5-hmU (hmU-ODN), utilising conditions previously used for the oxidation of 5-hmC.[124] The desired oxidation to 5-fU was observed by LC-MS with the expected mass shift (Scheme 3).



Scheme 3: hmU-ODN oxidation to fU-ODN2 showed the correct mass shift by LC-MS analysis

Quantitative oxidation of 5-hmU to 5-fU was further corroborated by mononucleoside composition analysis by LC-MS/MS after nuclease digestion. Analysis of the mass signals for C, T, A, G, 5-hmU and 5-fU showed the complete disappearance of the 5-hmU peak and formation of a new 5-fU peak after potassium perruthenate oxidation (Figure 30).

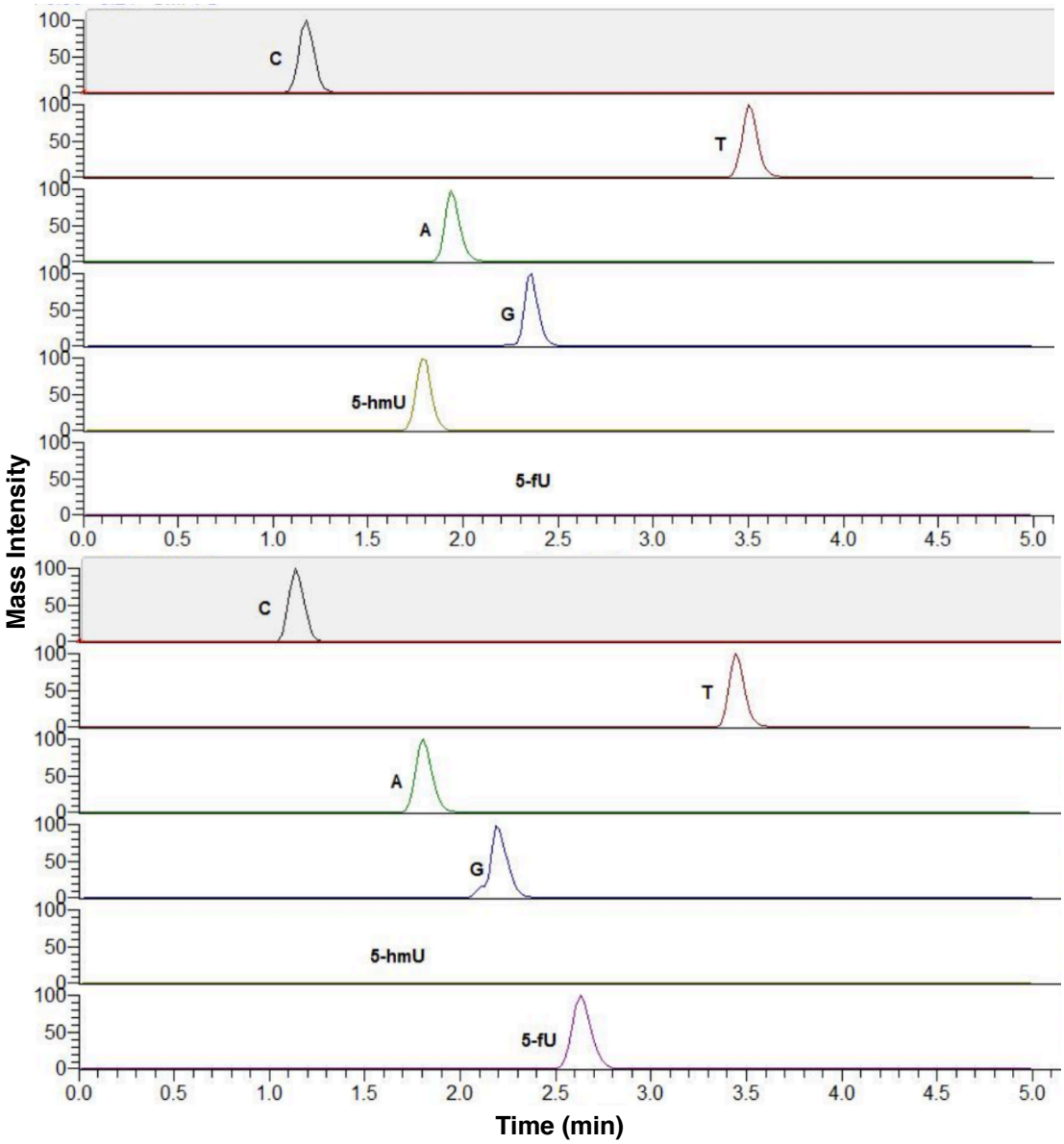


Figure 30: HPLC-MS extracted $[M+H]^+$ fragment ion count for C, T, A, G, 5-hmU and 5-fU deoxynucleosides after digestion of top) hmU-ODN and bottom) hmU-ODN after treatment with potassium perruthenate.

3.4.2. 5-fU Tagging with a Biotinylated Oxyamine (ARP)

Using the aldehyde group to tag 5-fU would provide a strategy for the chemical enrichment of both 5-fU, and 5-hmU after chemical oxidation. It was envisaged that 5-fU could be tagged with a biotinylated oxyamine or hydrazine to form an oxime or hydrazone respectively. This was first demonstrated using the commercially available oxyamine aldehyde reactive probe (ARP). Resultant oxime formation of ARP with 5-fU was quantitative, while no reaction was observed with a control ODN of the same general sequence which contained only the canonical nucleosides (Figure 31). The oxime product is stable with respect to reversion due to the alpha effect.

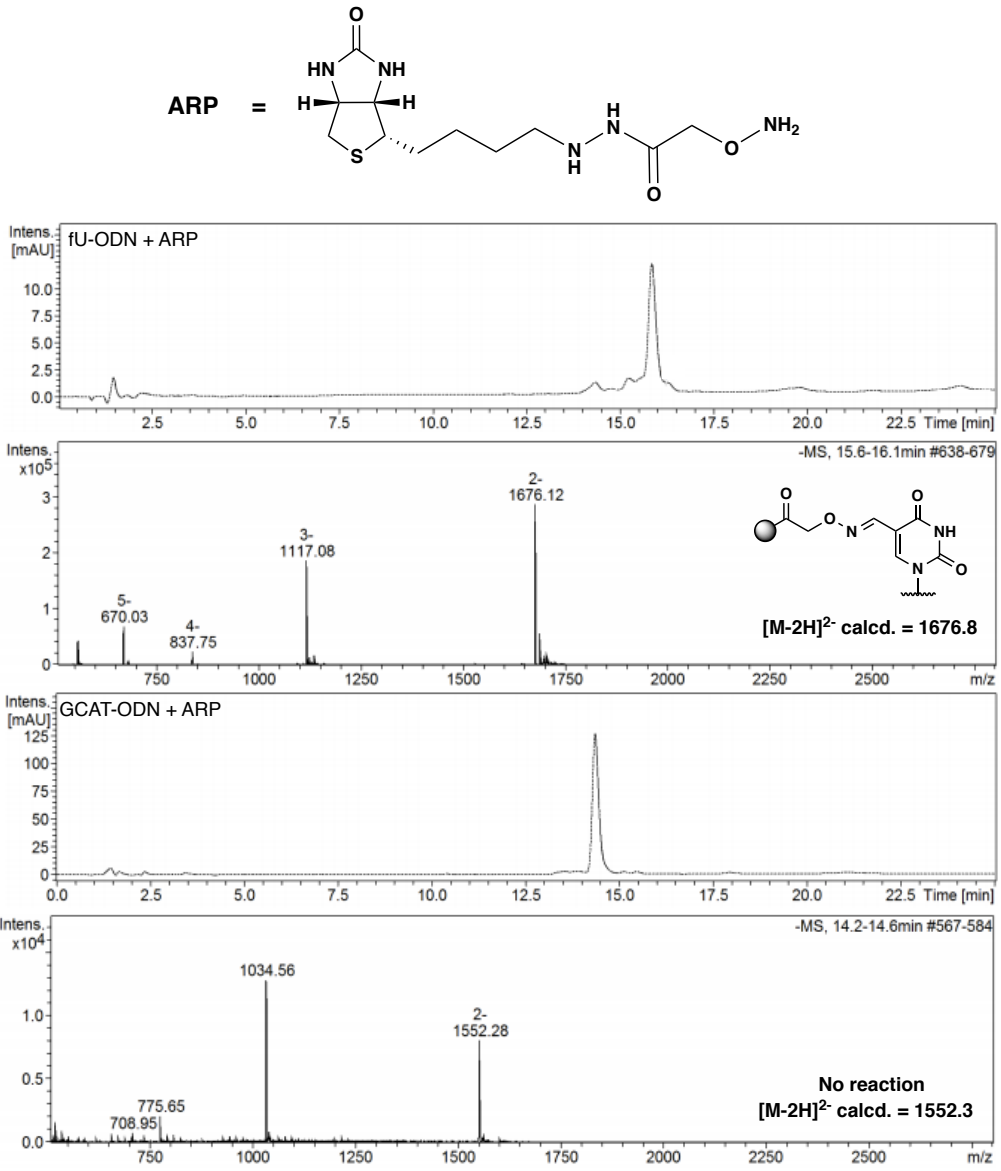


Figure 31: Top: Structure of ARP. Bottom: LC-MS trace demonstrates quantitative reaction of fU-ODN with ARP, demonstrating the scope of the 5-fU tagging strategy. No reaction was observed with a non-modified control ODN (GCAT-ODN)

However, ARP can also react with 5-fC, and this probe had previously been utilised for 5-fC genome-wide mapping.[51] Functionalisation of 5-fC using ARP required a long reaction time (24 hr), acidic reaction media, and the presence of a nucleophilic catalyst, *p*-anisidine. Under these conditions, ARP is unable to discern between the two formylated bases (Table 4 – Entry 1), preventing selective T-modification enrichment.

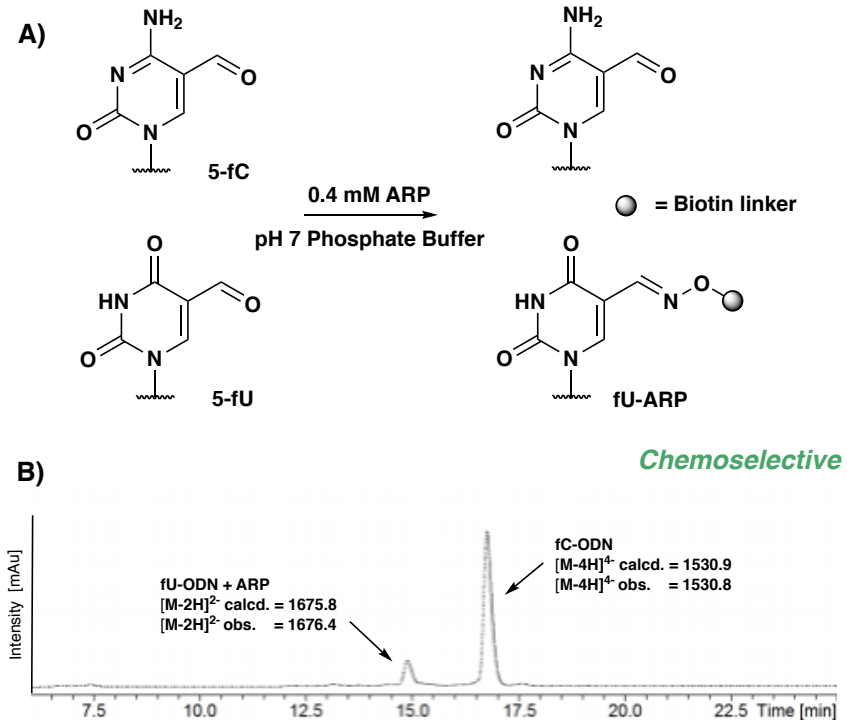


Figure 32: A - Oxime formation between 5-fU and a biotinylated oxyamine adduct (ARP) is chemoselective over 5-fC; this can be a strategy for selective T-modification enrichment. B – LC-MS trace demonstrating the selective reaction of fU-ODN with ARP in the presence of fC-ODN. Conditions (Table 4 – Entry 4).

Therefore, conditions were optimised to enable a chemoselective method for 5-fU tagging that discriminated between the T and C-modifications. Quantitative oxime formation occurred with 5-fU without the addition of the nucleophilic catalyst *p*-anisidine. Subsequent optimisation by modulating the pH sought to completely eliminate background reactivity of 5-fC. In pH 6 buffer, full conversion of 5-fU was observed after 4 hr and 5-fC functionalisation was undetected (Table 4 – Entry 4, Figure 32). The unexpected high reactivity and enhanced electrophilicity of 5-fU indicated the feasibility of 5-fU chemoselective functionalisation in the presence of 5-fC; furthermore, a 5-hmC glucosylation blocking step (Section 3.2) would not be required to achieve selective 5-hmU pulldown.

	Conditions (ARP)	fU-ODN (%)	fC ODN (%)
1	24 hr, pH 5, 100 mM *p*-anisidine,	100	94
2	24 hr, pH 5	100	26
3	24 hr, pH 6	100	3
4	4 hr, pH 6	100	n.d
5	24 hr, pH 7	64	n.d
6	24 hr, pH 8	21	n.d

Table 4: Reaction conversions with 0.4 mM ARP under various conditions. % conversion refers to integration of LC-MS product and starting material signal at 260nm. n.d. signifies that formation of product was not detected. Conversion becomes less efficient with increasing pH.

3.4.3. 5-fU Tagging with a Biotinylated Hydrazide (BH)

To screen for alternative biotinylated tagging reagents, the reactivity of fU-ODN with commercially available (+)-biotinamidohexanonic acid hydrazide (BH) was next explored. BH also demonstrated enhanced reactivity with fU-ODN over fC-ODN, as had been observed with ARP. Absolute selectivity for 5-fU was observed after 4 hr at ambient temperature with 10 mM BH (Table 5 - Entry 5, Figure 34), whereas analogous quantitative conversion of 5-fC required addition of *p*-anisidine and heating (Table 5 – Entry 1). The efficiency of this reaction under mild pH-neutral conditions (pH 7) was also suited for its applicability to DNA.

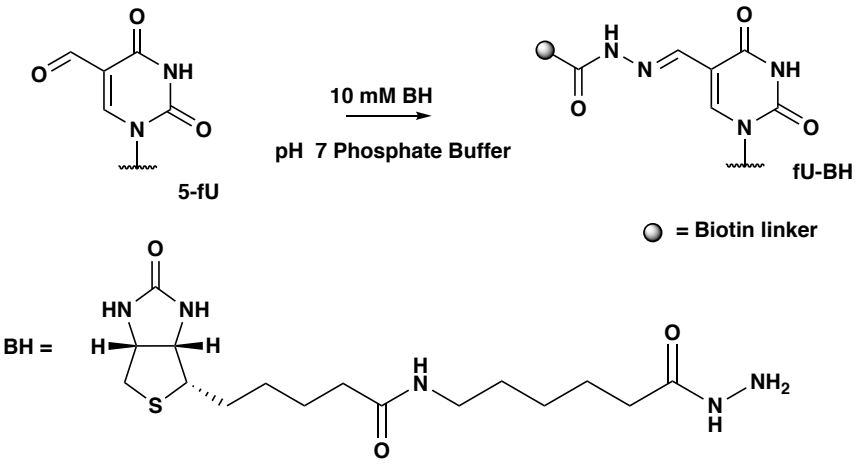


Figure 33: Top: Reaction scheme of 5-fU with biotinylated hydrazide (BH) to form a hydrazine, bottom: structure of BH.

	Conditions (BH)	fU-ODN (%)	fC-ODN (%)
1	20 mM BH, 24 hr, pH 5, 40 °C, 100 mM *p*-anisidine	100	100
2	10 mM BH, 24 hr, pH 6	100	13
3	10 mM BH, 4 h,r pH 6	100	2
4	10 mM BH, 24 hr, pH 7	100	1
5	10 mM BH, 4 hr, pH 7	100	n.d

Table 5: Reaction conversions with BH at various concentrations and conditions. % conversion refers to integration of LC-MS product and starting material signal at 260nm. n.d. signifies formation of product was not detected.

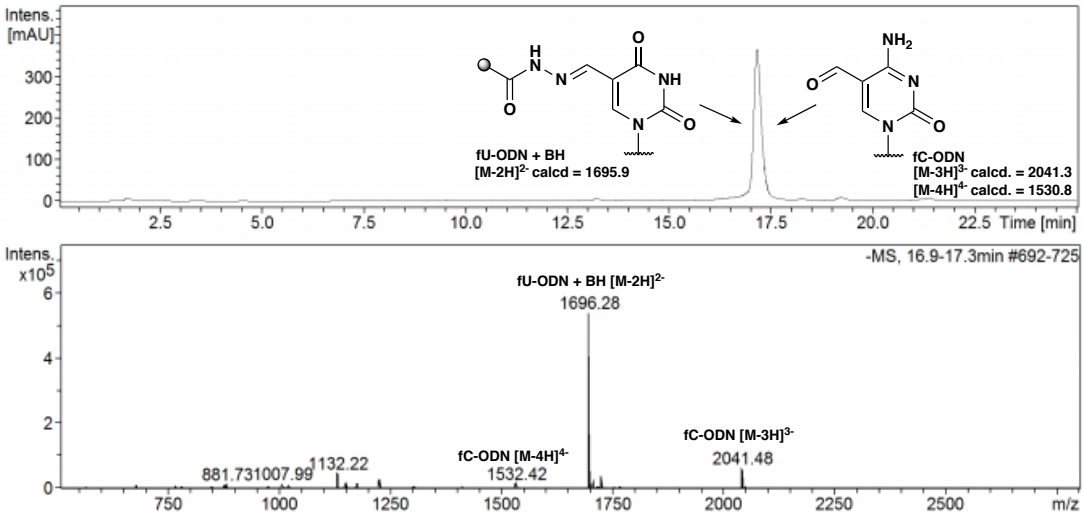
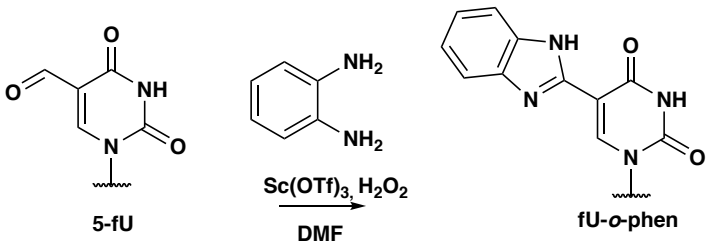


Figure 34: Selective reaction of fU-ODN with BH, where the fU-ODN BH adduct and fC-ODN peaks overlap.

3.4.4. 5-fU Tagging with *o*-phenylenediamine and Derivatives

Zhou and co-workers had reported that the 5-fU mononucleoside reacted with *o*-phenylenediamine in dimethyl formamide to form a stable benzimidazole product, after oxidation with hydrogen peroxide (Scheme 4).[163] It was therefore examined whether this reaction could also be exploited for the chemical tagging of 5-fU in DNA. One potential advantage of the benzimidazole product would be its stability to reversion under pulldown conditions.

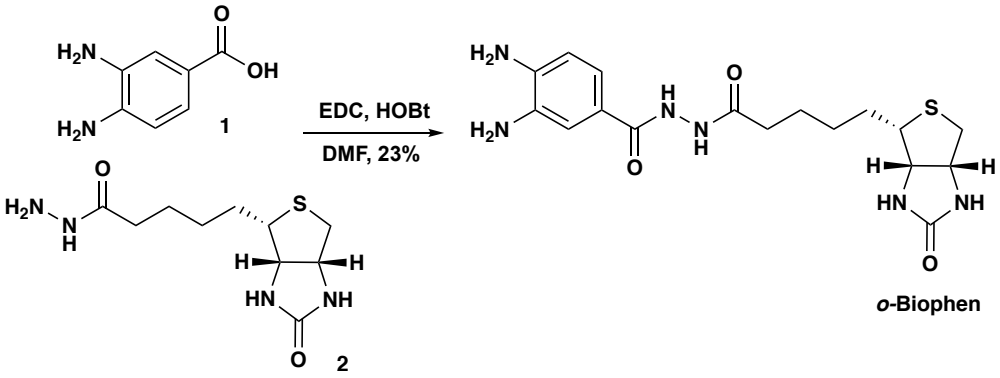


Scheme 4: Literature reaction of 5-fU with *o*-phenylenediamine

Thus, fU-ODN was treated with *o*-phenylenediamine to test whether analogous reactivity was achievable in aqueous conditions. The reaction proceeded quantitatively in pH 6 buffer, without the addition of any oxidant beyond the oxygen dissolved in water, to give the stable aromatic benzimidazole product. Under analogous conditions using 100mM *o*-phenylenediamine, fC-ODN also underwent adduct formation (Table 6 – Entry 1). However, the reaction with 5-fC was supressed at a neutral pH, and reduced concentration of reagent (5 mM), while quantitative reaction with fU-ODN was still achievable (Table 6 – Entry 4). This further demonstrated that T-modifications can be tagged chemoselectively due to the enhanced reactivity of 5-fU over 5-fC, although background reactivity was not completely eliminated with this moiety.

	Conditions (*o*-phenylenediamine*)*	(fU-ODN (%)	fC-ODN (%)
1	100 mM, 24 hr, pH 6	100	82
2	5 mM, 1 h,r pH 6	100	17
3	100 mM, 24 hr, pH 7	100	56
4	5 mM, 1 hr, pH 7	100	2
5	100 mM, 24 hr, pH 8	100	13
6	5 mM, 1 hr, pH 8	100	2

Table 6: Reaction conversions with *o*-phenylenediamine under various conditions. % conversion refers to integration of LC-MS product and starting material signal at 260nm. The % reaction of fC-ODN refers to extent of fC-ODN consumption since two peaks were observed corresponding to 5-fC functionalisation (Chapter 3 - Appendix).



Scheme 5: Synthesis of biotinylated *o*-phenylenediamine (*o*-Biophen) using EDC-activated peptide coupling chemistry.

To further examine the scope of differential reactivity in the pulldown strategy, a molecule was designed and synthesised that incorporated both *o*-phenylenediamine and a biotin linker. Biotinylated *o*-phenylenediamine (*o*-Biophen) was synthesised in one step from 4,5-diaminobenzoic acid (1) and biotin hydrazide (2), using EDC peptide coupling chemistry in the presence of hydroxybenzotriazole (HOBt), which acts as a nucleophilic catalyst[181] (Scheme 5). Subsequent reaction of the *o*-Biophen probe with the ODN models demonstrated that this reagent could also discriminate between the two formylated bases, which would allow selective 5-fU enrichment (Figure 35).

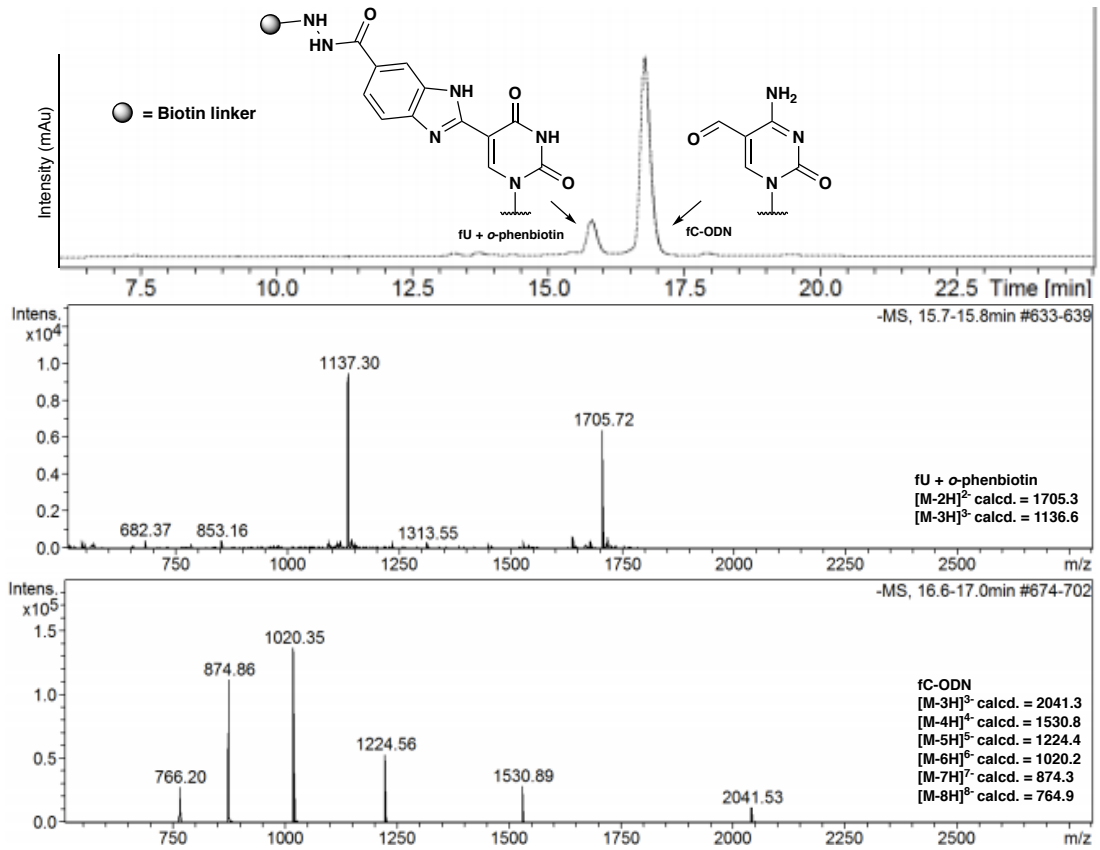


Figure 35: LC-MS trace demonstrating the selective reaction of fU-ODN with *o*-Biophen probe in the presence of fC-ODN.

3.5. Proof of Concept Chemical Enrichment Pulldown Studies

To validate the selectivity and feasibility of this approach for a pulldown sequencing method, the tagging strategy and extent of enrichment was assessed by quantitative-PCR (qPCR). 80mer dsDNA models containing 5-fU, 5-fC or non-modified T (fU-DNA, fC-DNA and GCAT-DNA) were designed to 1) incorporate two modified bases per strand, and 2) contain different primer regions so that each model could be distinguished separately by qPCR. Following a modified protocol from Rao and co-workers[116], DNA strands were subjected to the optimised chemoselective tagging conditions with each probe, before purification using size exclusion chromatography. The tagged DNA was incubated in the presence of magnetic streptavidin beads, before subsequent removal of the supernatant, which contained non-bound DNA, and stringent washing of the beads. The DNA was eluted from the beads by heating in formamide, which destroys the biotin/streptavidin interaction, and subsequently purified by size filtration.

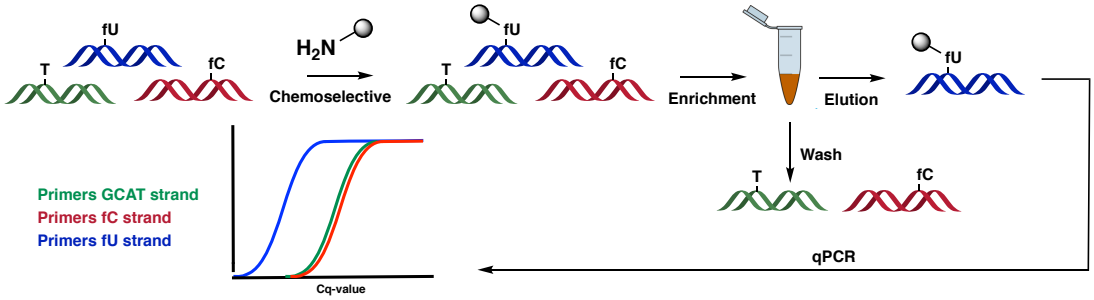


Figure 36: Workflow of chemical enrichment experiment followed by qPCR quantification.

The enriched DNA was quantified by qPCR in comparison with calibration lines for each model DNA strand (Figure 36). An approximately 150-fold enrichment of fU-DNA over fC-DNA was demonstrated with each probe after chemoselective tagging (Figure 37). The extent of 5-fU enrichment was similar over both fC-DNA and GCAT-DNA for the BH probe, indicating that captured fC-DNA was at the background level and hence unlikely to be caused by covalent reactivity.

The selectivity of 5-fC enrichment conditions[51], using ARP in the presence of *p*-anisidine, were also tested by qPCR. As expected, limited discrimination between 5-fU and 5-fC was observed using 5-fC tagging chemistry; this suggested that previously generated 5-fC genome-wide maps would have simultaneously enriched for the 5-fU modification.[51,53]

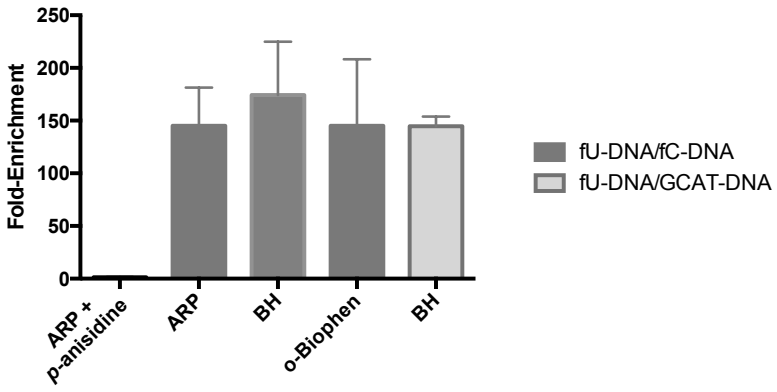


Figure 37: Extent of enrichment of fU-DNA over fC-DNA or GCAT-DNA under different conditions (Mean + SD).

3.6. *Ab initio* Quantum Mechanical Calculations on 5-fU and 5-fC Reactivity

The following ab initio work was completed in collaboration with Dr A. Sahakyan, Balasubramanian group.

To obtain a theoretical insight on what might facilitate the increased reactivity of 5-fU over 5-fC, *ab initio* quantum mechanical calculations were performed on simplified model systems. Both 5-fU and 5-fC were modelled with a methyl group at the *N*-glycosidic position (5-fU_m and 5-fC_m), while methylamine was taken to be the model reactant (Figure 38). When modelling transition states, the rate-determining step of the reaction was considered to be nucleophilic addition to the aldehyde to form the hemiaminal.

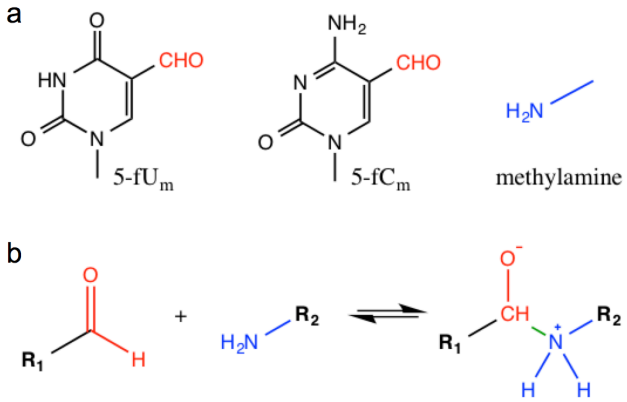


Figure 38: a) The simplified molecules for the computational study. b) The first stage of the addition reaction, expected to be the rate limiting step.[182]

3.6.1. Aldehyde Rotation Barriers in 5-fU_m and 5-fC_m.

Since 5-fU_m and 5-fC_m can exist in two rotameric states, *syn* or *anti*, the minimum energy structures (MES) of these two nucleotides and the associated rotation barrier between the two conformations was initially considered. It was found that while 5-fU_m prefers the *anti* conformation in its ground state (11.58 kcal/mol rotation barrier), 5-fC_m prefers the *syn* conformation (15.43 kcal/mol rotation barrier), likely due to a stabilising H-bond interaction between 4-NH_2 and the aldehyde oxygen (Figure 39). The rotation barriers are substantial enough that the molecules are likely to occupy their minimum energy structures at ambient temperature, which is supported by crystal structures of 5-fU- and 5-fC-containing small molecules.[183,184,185,186]

It was therefore feasible that the differences in conformational preference of 5-fU and 5-fC could contribute to the different reactivities observed for these formylated nucleobases. Since the biotinylated probes are more sterically bulky than the methyl group, it was conceivable that the products of the reaction would prefer the *anti*

configuration. Whilst the *anti* arrangement is favoured by 5-fU, the same arrangement for 5-fC would require a rotation barrier to be overcome. Moreover, upon hemiaminal formation, the H-bonding interaction responsible for the 5-fC *syn* conformation would be disrupted, a further energetic cost.

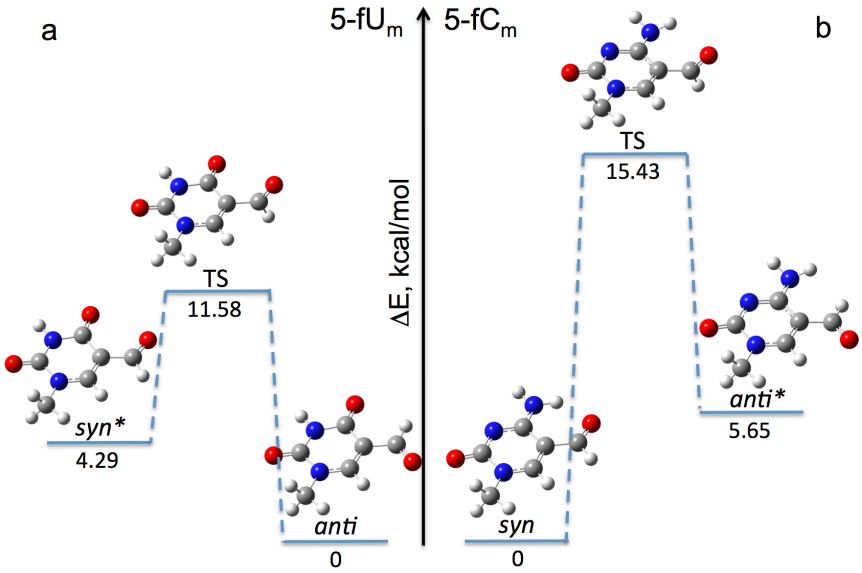


Figure 39: The energy minima (*syn* and *anti*) and the transition states (TS) along the aldehyde group rotation pathway in a) 5-fU$_m$ and b) 5-fC$_m$. The higher energy conformations are marked with asterisks.[182]

3.6.2. Partial Charges and LUMO Orbital Energies at the Aldehyde of 5-fU$_m$ and 5-fC$_m$.

The partial charges at the aldehyde carbon for 5-fU$_m$ and 5-fC$_m$ were next considered, to determine if electrostatics contributed to the difference in reactivity. Both Mulliken[187] (Mull) charges and Merz-Singh-Kollman[188,189] (MSK) were considered for 5-fU$_m$ and 5-fC$_m$ in both rotameric states. The computations, however, gave inconclusive results, since the calculated charge greatly varied depending on the model used (Figure 40). When considering the ground rotameric states, 5-fC appeared to have either a higher (MK) or similar (Mull) positive charge, which contradicted the experimental observation. This suggested that, unsurprisingly, the reaction was not driven purely by electrostatics.

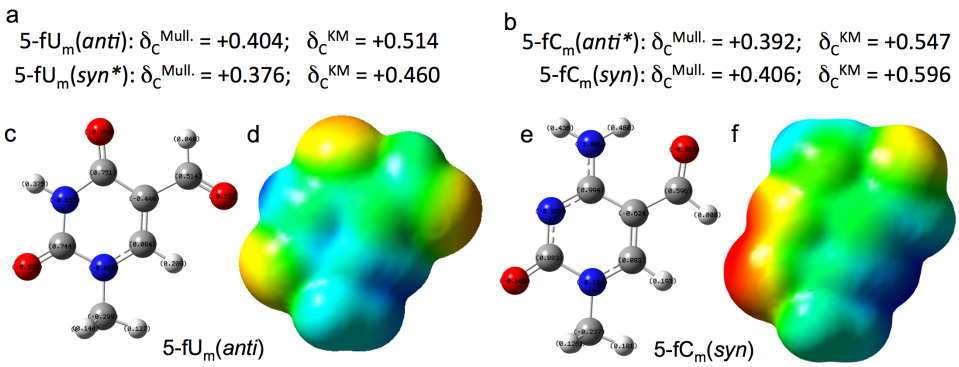


Figure 40: The Mulliken (Mull) and Merz-Singh-Kolmann (MSK) partial charges at the aldehyde carbon calculated for both *syn* and *anti* conformers of a) 5-fU$_m$ and b) 5-fC$_m$. The minimum energy conformers along with the MSK charges are shown in c) and e) for 5-fU$_m$ and 5-fC$_m$ respectively. Charge distribution (colour pallet from red to blue for -0.08 to +0.08 charge range) is demonstrated in d) and f).[182]

3.6.3. Natural Bond Orbital Analysis of 5-fU$_m$ and 5-fC$_m$.

Next, natural bond orbital (NBO) analysis[190,191] was used to determine whether 5-fU$_m$ and 5-fC$_m$ orbital energies might reveal a core electronic difference between the two formylated bases. The most marked difference was the calculated orbital energy of the C$_{ring}$-C$_{aldehyde}$ bonding orbital, which was found to be 18.37 kcal/mol more stable in 5-fC$_m$ compared to 5-fU$_m$ when considering their rotameric ground states. Such a difference suggested a higher degree of conjugation in 5-fC compared to 5-fU. Upon hemiaminal formation, the aldehyde carbon becomes more tetrahedral, leading to a loss of stabilizing conjugation; this effect would therefore be less disruptive for 5-fU, thereby contributing to its enhanced reactivity. In addition, when considering LUMO orbital energies, 5-fU$_m$ was found to have a lower orbital energy regardless of rotameric state (8.63 kcal/mol difference based on preferred 5-fU$_m$ vs 5-fC$_m$ conformation), indicating a much better orbital energy overlap between the LUMO of the aldehyde and the HOMO of the incoming nucleophile.

For both model reactions with 5-fU$_m$ and 5-fC$_m$, the location of a transition state for the nucleophilic addition was attempted. When considering the interactions between 5-fU$_m$, the incoming nucleophile and a water molecule, an energy minimum was identified (Figure 41), which was more stable than a system of non-interacting individual molecules (-13.79 kcal/mol). This intermediate structure demonstrates that a 6-membered hydrogen transfer ring is formed, where the 4-O of 5-fU forms a stabilising hydrogen bond with the incoming nucleophile. It was also observed that the aldehyde carbon loses planarity upon interaction with the nucleophile, which would aid C-N bond formation. It was, however, not possible to locate a similar energy minimum for 5-fC.

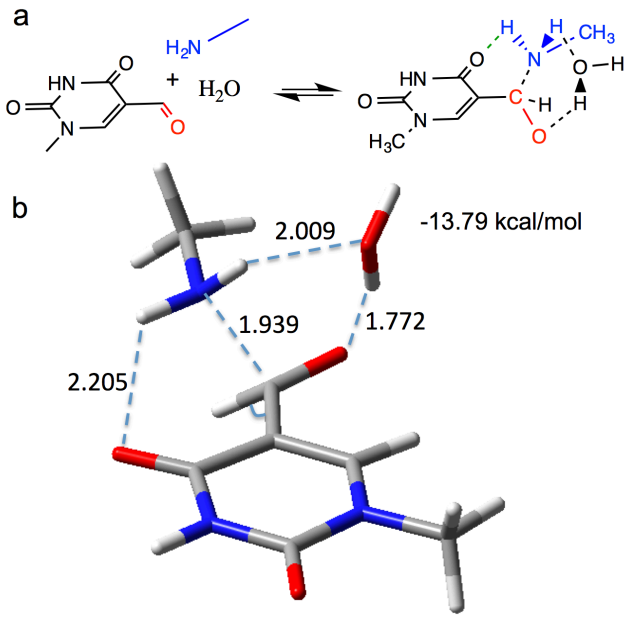


Figure 41: The intermediate state (ΔE=-13.79 kcal/mol) formed during hemiaminal formation with 5-fU$_m$ (a). The structure is stabilised via the extra hydrogen bond between the amino group and the 4-O of 5-fU$_m$. The aldehyde carbon has partially gained tetrahedricity (b) upon C-N bond formation, and slightly rotated to facilitate the formation of the above-mentioned hydrogen bond. All the outlined distances are measured in Å.[182]

In conclusion, the alternative preferred conformation of 5-fU and 5-fC may go some way to explaining such a difference in reactivity, however core-electronic differences corroborate the finding that 5-fU is a better electrophile, providing an explanation for its enhanced reactivity compared to 5-fC.

3.7. Apurinic/Apyrimidinic (AP) Site Selectivity

Although it had been demonstrated that 5-fU could be tagged selectively over 5-fC, it was possible that AP sites could also be trapped in the presence of an aldehyde reactive probe. ARP has in fact been utilised for AP site quantification,[192] hence it was necessary to determine the potential background of AP sites in the proposed T-modification enrichment-sequencing method. AP site cross-reactivity was firstly investigated under the optimised conditions for 5-fU tagging with BH, ARP and *o*-phenylenediamine probes. BH demonstrated the least cross-reactivity with AP-ODN, suggesting this probe may be advantageous to supress any signals from AP sites (Table 7).

	Probe	AP-ODN Reactivity (%)
1	0.4 mM ARP, pH 6, 4 hr	90
2	10 mM BH, pH 7, 4 hr	10
3	5mM, *o*-phenylenediamine, pH 7, 1 hr	28

Table 7: Percentage labelling of probes with AP-ODN where % conversion refers to integration of LC-MS product and starting material signal at 260nm using optimised conditions for 5-fU tagging.

However, AP sites have been reported to cause polymerase-stalling during primer extension or PCR of DNA fragments.[193] Since the amount of DNA recovered after chemical pulldown is likely to be very small due to the low abundance of 5-fU (and DNA containing AP sites), several cycles of PCR amplification are anticipated after chemical enrichment to generate sufficient DNA for sequencing (e.g. 14-16 cycles of PCR are required for 5-fC chemical enrichment).[51,53] Full polymerase extension is essential for NGS since both sequencing adaptors are required for amplification on the DNA sequencing flow-cell (Section 3.8.1). It was therefore likely that tagged AP sites would fail to amplify, regardless of tagging efficacy.

To confirm this finding, an 80bp oligomer containing two AP sites (AP-DNA), was synthesised from the enzymatic treatment of U-containing DNA (U-DNA) with Uracil DNA glycosylase (UNG) enzyme and incubation at 37 °C (Figure 42).[51]

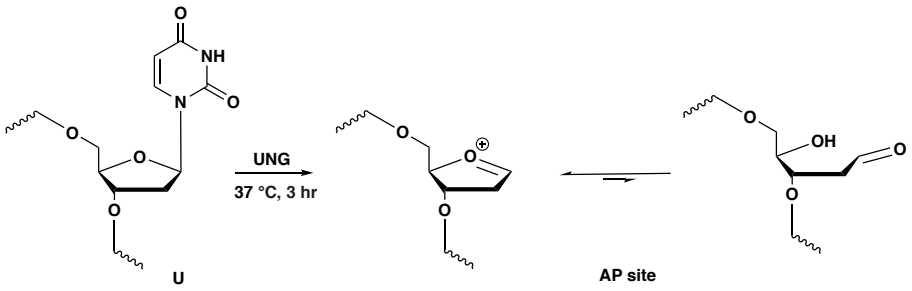


Figure 42: AP-DNA can be formed via excision of U from U-DNA using the UNG enzyme. The resulting AP site is in equilibrium between closed and open sugar form, the latter containing a reactive aldehyde.

A primer-extension experiment was next performed on AP-DNA and an equivalent 80-mer bearing two 5-fU modifications (fU-DNA), which had both been treated with the BH probe under optimised conditions (Table 5 – Entry 5). While full polymerase extension was observed in the latter case, stalling predominantly occurred at the first AP site in AP-DNA, and no full-length product was observed (Figure 43). This confirmed the expectation that AP sites are likely to be under-represented by sequencing even in the event of promiscuous reactivity.

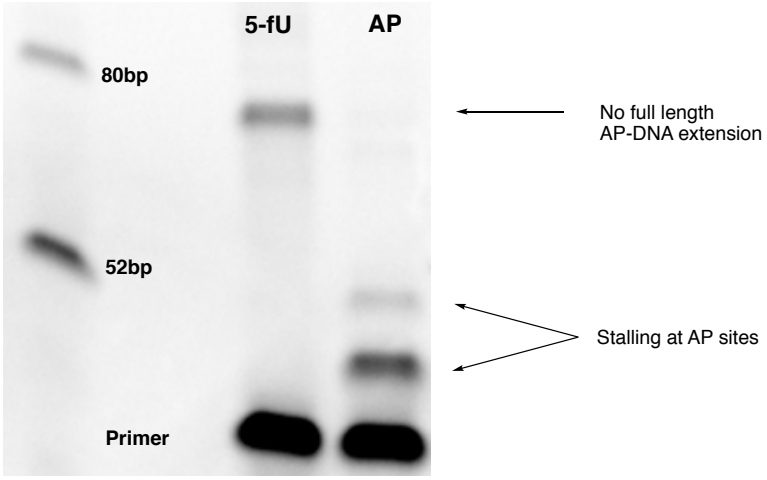


Figure 43: Polymerase extension of 80mer DNA containing either two 5-fU modifications (fU-DNA) or two AP sites (AP-DNA) which had been treated with the BH probe. No full-length primer extension of AP-DNA is observed, however two polymerase stalling sites corresponding to AP loci are seen.

Furthermore, DNA fragments containing AP sites were prone to fragmentation under the basic 5-hmU → 5-fU oxidation conditions[194] (Figure 44). Since oxidation occurs after the ligation of adaptors for sequencing (Section 3.8.3), this would further render DNA bearing AP-sites unsequenceable after fragmentation.

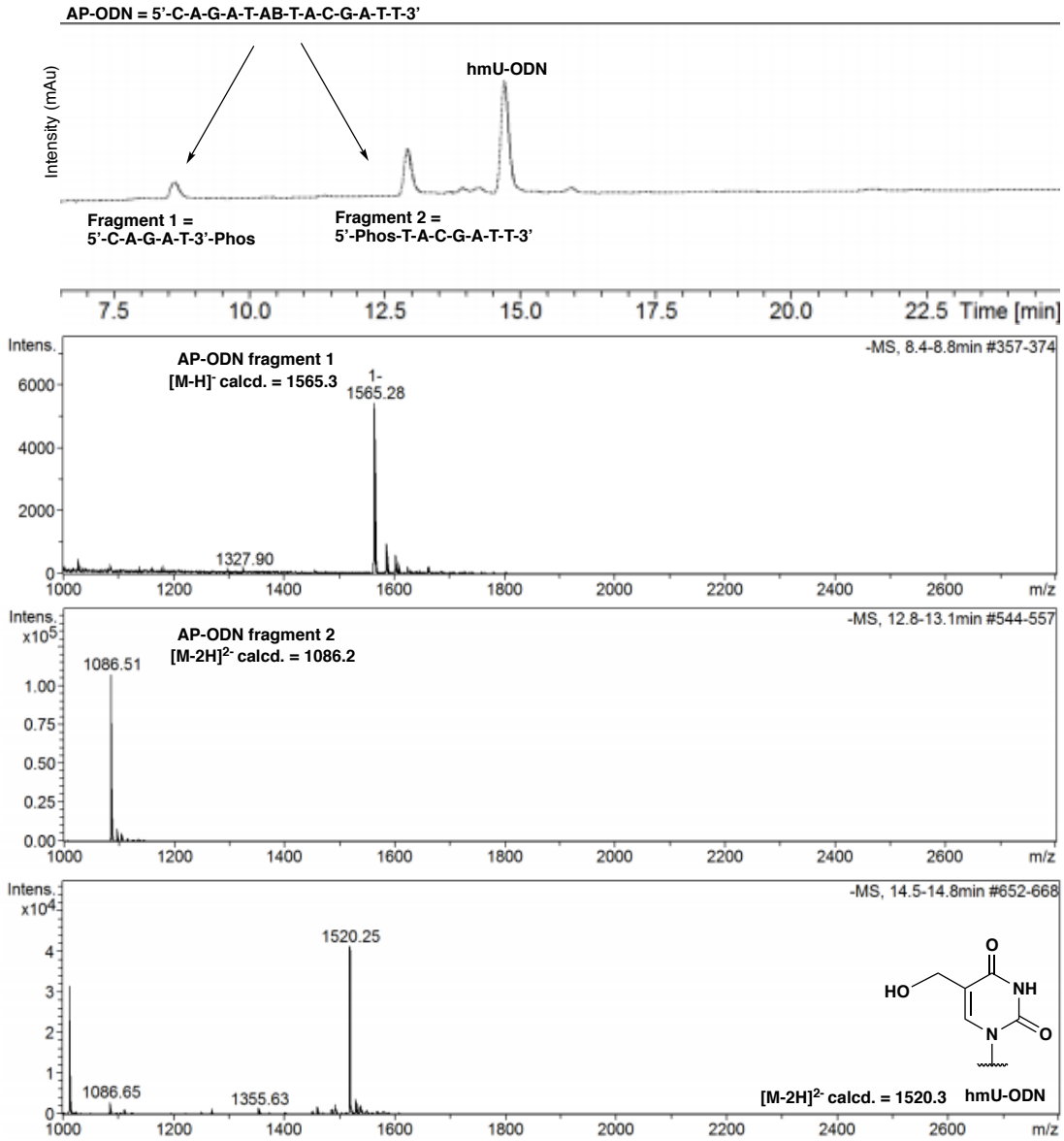


Figure 44: Under denaturation conditions for hmU-ODN oxidation (50 mM sodium hydroxide), AP-ODN fragmentation is observed.

3.8. Incorporation of Chemical Tagging into the NGS Library Preparation Workflow

3.8.1. Background - Library Preparation of Samples for NGS

Proof of principle selective tagging and enrichment had now been demonstrated for 5-hmU and 5-fU. The chemical tagging method next needed to be incorporated into an NGS library preparation workflow to enable T-modification-enrichment sequencing of genomic samples.

Library preparation for NGS (Figure 45) includes: 1) a fragmentation step, where DNA is first sonicated (broken down) into smaller fragments of around 100-1000 bp; 2) an end-repair step, where an enzymatic mixture of T4 polymerase, Klenow fragment and T4 polynucleotide kinase fills in 5' overhangs, resects 3' overhangs to generate blunt ends and phosphorylates the 5' end; 3) an A-tailing step, where a DNA polymerase (e.g. Klenow fragment) adds dATP to the 3' ends; 4) Adaptor Ligation, where sequencing adaptors containing complementary dT overhangs are ligated to DNA fragments by a DNA ligase (these adaptors are necessary for hybridisation onto the sequencing flow-cell, and for subsequent PCR enrichment) and 5) PCR amplification, where ligated fragments are amplified by PCR using primers complementary to the sequencing adaptors; this generates adequate amounts of DNA for sequencing.[195,196]

Chemical tagging and/or enrichment needed to be incorporated into this workflow, either after sonication or before the PCR step (Figure 45); furthermore, potential PCR biases associated with the biotinylated probes needed to be considered.

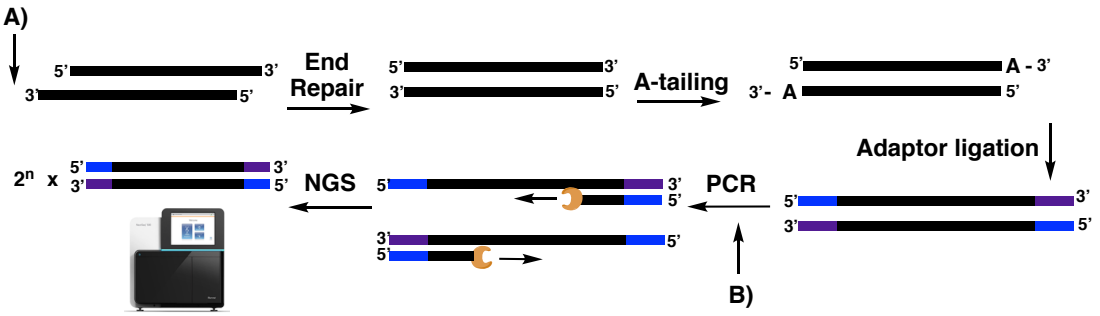


Figure 45: Workflow of library preparation for NGS. Chemical tagging and/or enrichment can be incorporated at either points A) or B).

3.8.2. Probe Reversibility and Minimisation of PCR Biases

Considering that T-modifications in genomic samples are low in abundance (Chapter 2), the PCR step would be essential after affinity enrichment to generate adequate amounts of DNA for sequencing. However, the presence of a large chemical moiety on DNA (i.e. biotin linker) can cause polymerase stalling, affecting the PCR amplification efficiency of chemically-tagged DNA fragments.[123,197] This would be of particular concern in densely modified regions, leading to a potential underrepresentation of genomic loci that are biologically relevant. It had already been demonstrated that the biotinylated oxyamine ARP caused significant polymerase at 5-fC sites after tagging.[197] This was alleviated when a cleavable biotinylated oxyamine ARP variant was employed, leaving behind a smaller oxime chemical residue (Figure 46).[197] The cleavable variant was therefore used for subsequent chemical 5-fC mapping.[53]

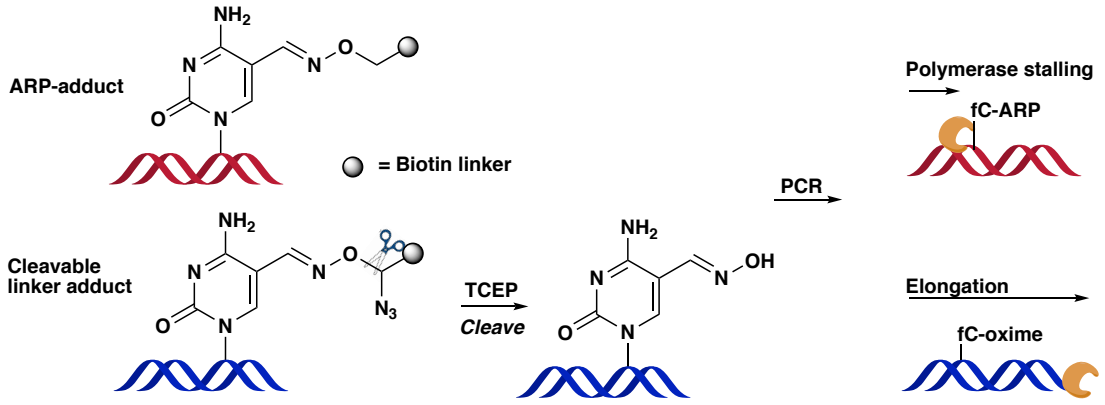
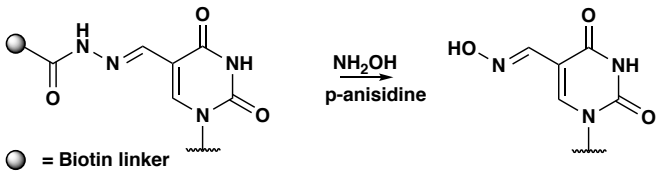


Figure 46: Chemical tagging of 5-fC with ARP led to polymerase stalling. A cleavable oxyamine linker using TCEP-mediated Staudinger chemistry results in a 5-fC-oxime adduct which demonstrated reduced PCR polymerase stalling compared to ARP.[197]

As an alternative to the cleavable linker strategy, it was found that the fU-BH adduct could be reversed. This occurred quantitatively via transimination in the presence of hydroxylamine and the nucleophilic catalyst *p*-anisidine,[198,199] to leave a smaller oxime moiety on 5-fU (Scheme 6). ARP adduct reversion was possible yet required harsher conditions, while the benzimidazole formed from *o*-phenylenediamine was stable to reversion (Appendix – Chapter 3).



Scheme 6: 5-fU-BH adduct could be reversed in the presence of *p*-anisdine + NH_2OH

The transimination reaction with the BH-adducts also enabled chemoselective elution of 5-fU-modified DNA from the beads, while the fC-BH-adduct remained intact. This exploited the enhanced electrophilicity of 5-fU and its related compounds, as had been modelled by quantum mechanical calculations. Selective elution was achievable using neutral conditions with moderate heating (40 °C) in the presence of *p*-anisidine (Figure 47). For the 5-fC-BH adduct, probe reversibility required more acidic conditions (pH 5) and longer reaction times in order to go to completion (Appendix – Chapter 3).

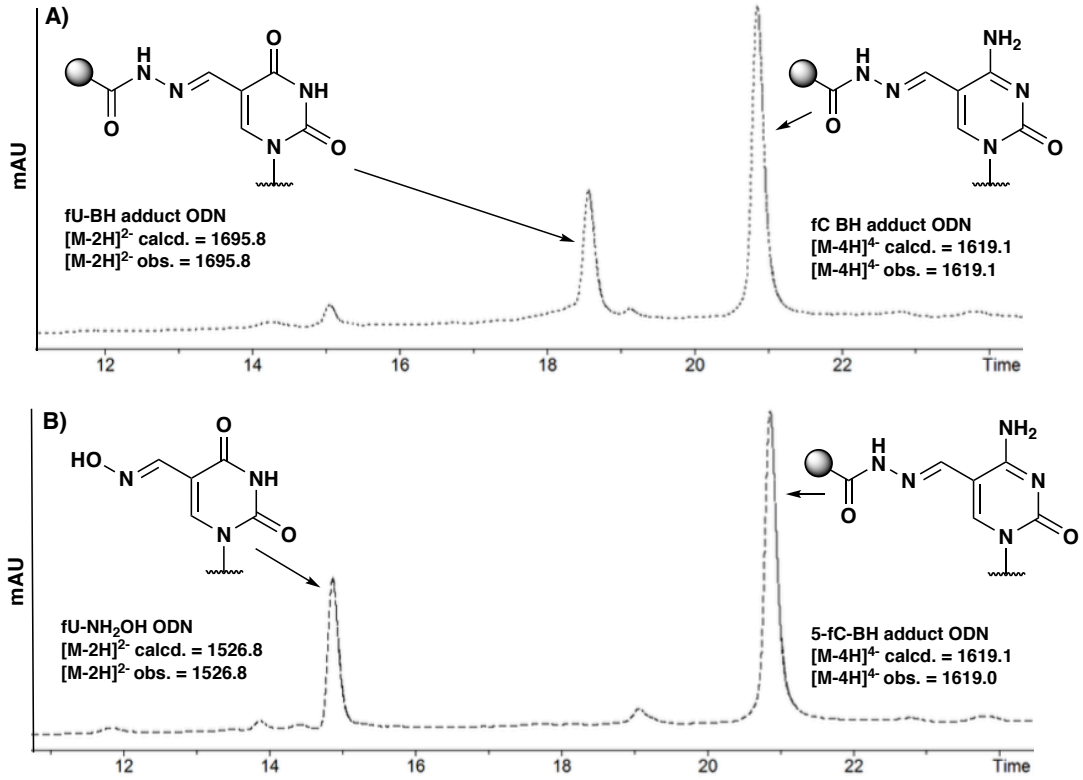


Figure 47: 5-fU-BH adduct could be reversed chemoselectively while 5-fC-BH adduct remained stable under chemical elution conditions. A) fU-ODN and fC-ODN BH adducts generated via quantitative tagging with BH in the presence of *p*-anisidine, B) Selective transamination observed for fU-BH adduct ODN to form fU-NH_2OH ODN in the presence of *p*-anisdine.

This strategy enabled a second-round of enrichment for 5-fU modified DNA fragments; any 5-fC adducts resulting from promiscuous reactivity in the tagging step remain attached to the streptavidin beads. An approximately 9-fold additional increase in selectivity for fU-DNA was observed using NH_2OH-mediated cleavage as determined by qPCR (Figure 48). This elution method further ensured selective T-modification enrichment, whilst also minimising PCR biases.

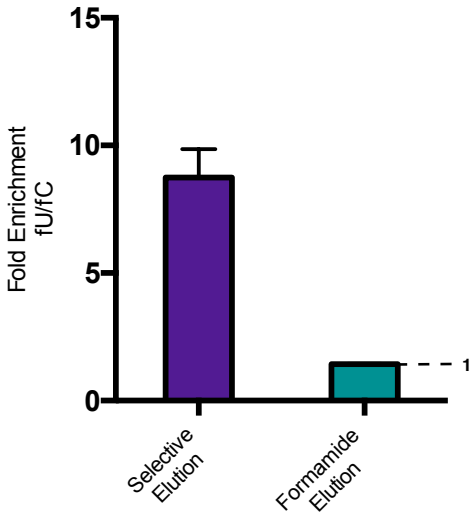


Figure 48: 5-fU/5-fC selective enrichment (mean with SD) of BH adducts using NH_2OH mediated selective chemical elution

3.8.3. Library Preparation for T-modification Using Model DNA

Chemoselective tagging and chemoselective elution had now been demonstrated for the T-modifications. In order to confirm T-modification chemical-enrichment by sequencing, these steps were incorporated into the library preparation procedure. Model ODNs (fU-DNA/hmU-DNA, fC-DNA2/hmC-DNA2, GCAT-DNA) were utilised to validate the procedure and to determine the extent of enrichment after library preparation and NGS.

For 5-fU enrichment sequencing, sonicated DNA fragments were firstly subjected to the optimised tagging conditions, followed by purification. The fragments were prepared for DNA sequencing via standard library preparation for Illumina sequencing using the NEB Ultra II library preparation kit; this was followed by streptavidin enrichment of biotinylated DNA fragments, selective NH_2OH-mediated elution and subsequent PCR.

For 5-hmU chemical enrichment sequencing, the workflow was slightly altered to accommodate the fact that chemical oxidation resulted in the formation of single-stranded DNA.[124] Due to the inefficiency of single-stranded ligation, ligation of double-stranded DNA fragments was performed prior to the chemical tagging reaction. Modified adaptors (5'-OMe and 3'-Phos) were used to avoid any unspecific oxidation of terminal 5'-OH or 3'-OH hydroxyl groups on DNA fragments. The DNA was denatured (50 mM sodium hydroxide, 37 °C, 30 min) to single-stranded DNA before addition of the oxidant. The oxidised DNA was subjected to chemoselective tagging conditions followed by chemical enrichment, selective NH_2OH-mediated elution and subsequent PCR. The oxidation step was extremely sensitive to the presence of alcohol or residual enzyme

from prior ligation steps. Thus, prior to chemical enrichment, purification steps to remove any trace impurities were essential.

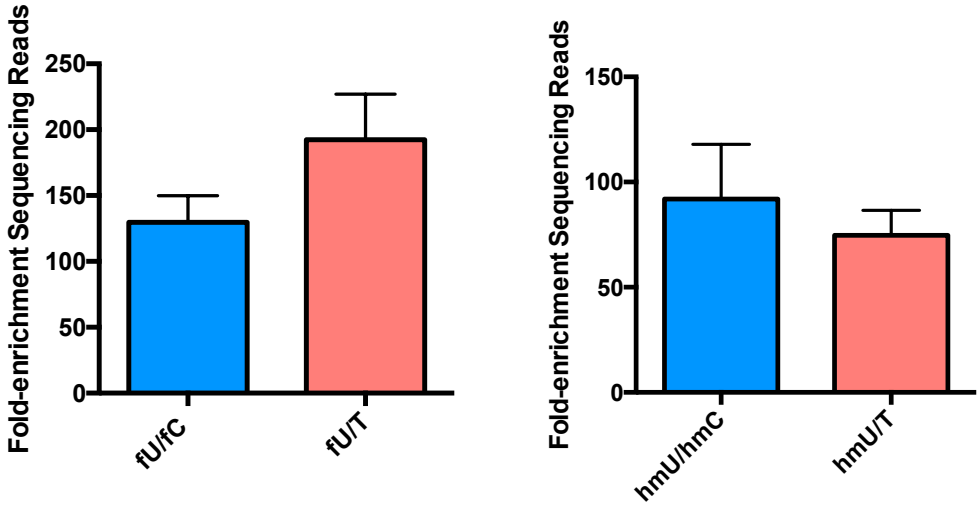
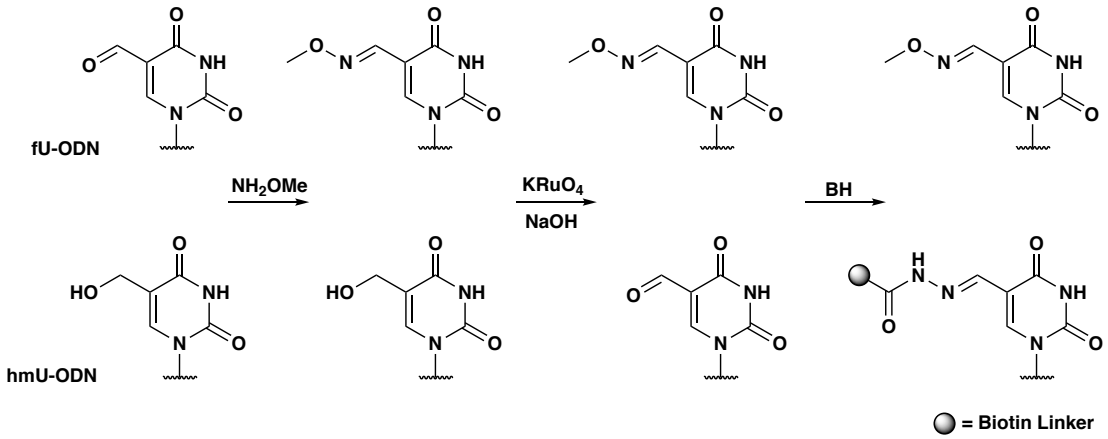


Figure 49: 5-fU and 5-hmU enrichment (mean with SD) demonstrated by fold enrichment of sequencing reads.

By comparing the number of sequencing reads of model ODNs after chemical-enrichment sequencing, significant enrichment of fU-DNA (> 100-fold) was observed over fC-DNA2 and GCAT-DNA models (Figure 49 - Left). This was possible by utilising both chemoselective 5-fU tagging and NH_2OH-mediated elution. 5-hmU enrichment was also observed via sequencing reads (Figure 49 - Right); enrichment efficacies in some replicates were more variable, likely due to varying efficiency of the oxidation step. NGS of the DNA models confirmed that the chemical enrichment method for 5-fU and 5-hmU was compatible when incorporated into the library preparation procedure. Thus, these methods could now be applied for T-modification enrichment sequencing in biological samples.

3.9. Chemical Discrimination Between 5-hmU and 5-fU

Finally, since 5-hmU is also tagged using the same chemistry as for 5-fU, a proof-of-principle method was demonstrated by LC-MS to chemically discriminate between the two T-modifications. It was shown that 5-fU can be chemically blocked prior to 5-hmU oxidation using *N*-methylhydroxylamine. 5-hmU can then be selectively tagged using the BH probe following oxidation (Scheme 7). Although chemical discrimination may be useful, it was envisaged that the presence of 5-hmU enriched regions could instead be determined bioinformatically by comparing chemical enrichment-sequencing maps in the presence or absence of an oxidation step.



Scheme 7: 5-hmU can be tagged selectively over 5-fU via prior blocking of 5-fU with *N*-methylhydroxylamine, oxidation and tagging with BH.

3.10. Conclusion

In conclusion, reactivity and proof-of-concept enrichment studies highlighted the feasibility of using a chemoselective method to tag 5-fU (and 5-hmU after chemical oxidation). This enables the enrichment of DNA fragments that contain T-modifications selectivity over the analogous C-modifications.[182] The increased reactivity of 5-fU, over 5-fC, was rationalised using *ab initio* quantum mechanical calculations. Selectivity of 5-fU versus AP sites was also assessed; it was demonstrated that DNA containing AP sites are unlikely to be efficiently amplified by DNA polymerases, hence, AP sites would be underrepresented in any chemical enrichment method where DNA sequencing provides the readout.

The chemical enrichment method was next incorporated into the experimental workflow for DNA sequencing. A selective chemical elution method using NH_2OH-transiminiation was shown to provide further T-modification discrimination, while minimising potential PCR biases. Using model DNA fragments, selective enrichment of both 5-hmU and 5-fU was demonstrated by NGS.

The chemical-enrichment method could now be used to generate T-modification maps in trypanosomatid and mammalian targets in order to probe their function; indeed, this method was utilised to generate the first 5-hmU affinity map in the trypanosomatid Leishmania in combination with a hmU-DIP method (Section – 4.2.1). An alternative proof of principle chemoenzymatic method to enrich 5-hmU has recently been published by Bullard *et al;* this involves enzymatic glucosylation of 5-hmU with UDP-glucose and J-GT, followed by Base J immunoprecipitation.[200] There are, however,

several limitations of this method compared to the strategy proposed in this chapter. Firstly, the J-GT enzyme is not widely available; secondly, the use of Base J immunoprecipitation prevents the discrimination of 5-hmU and Base J loci in trypanosomatid targets and thirdly, the method depends on J-GT not having any sequence specificity, an assumption which is questioned in Chapter 4.

Although the work in this chapter was designed with the chemical enrichment of T-modified bases in mind, analogous chemistry inspired by this work has subsequently been used for chemoselective fluorescence visualisation of 5-fU in DNA.[201] Furthermore, the chemistry developed could be applied in the future to tag T-modifications prior to third-generation sequencing methods; the presence of a large chemical tag will help to discriminate their kinetic signal from the canonical bases.[129]

Finally, the work in this chapter highlights that conditions currently utilised for 5-fC-enrichment sequencing fail to discriminate between 5-fC and 5-fU; this should therefore be acknowledged for any future 5-fC-based enrichment sequencing. Since 5-hmU sites can be distinguished from 5-fU sites by prior 5-fU blocking, this type of strategy (e.g. selectively blocking 5-fU tagging via irreversible benzimidazole formation with *o*-phenylendiamine) could be utilised to achieve selective 5-fC tagging.

4. Exploring the Role of T-modifications in Trypanosomatid and Mammalian Systems by Sequencing

4.1. Introduction

The aim of this chapter was to utilise NGS to explore the role of T-modifications in both trypanosomatids and mammalian targets. In trypanosomatids, 5-hmU is enzymatically generated via JBP-oxidation of thymine, which is further glycosylated by J-GT to form Base J (Introduction – 1.4.3). A goal of this chapter was to explore the origin of distinct 5-hmU loci in trypanosomes, independent from Base J loci, suggesting a unique role for 5-hmU aside from being a Base J intermediate. In mammals, 5-hmU and 5-fU are known products of ROS, whilst 5-hmU has also been suggested to be a product of enzymatic TET-mediated oxidation, and implicated in gene regulation.[40] Thus, a second aim of this chapter was to generate the first T-modification maps in the human genome, to provide further insight into the origin, biological role and consequence of oxidised T derivatives in mammalian systems.

4.2. T-modification Mapping in Trypanosomatids

4.2.1. Introduction - T-modification Mapping in *Leishmania*

Kawasaki *et al.* generated the first genome-wide map of 5-hmU in the trypanosomatid Leishmania.[202] 5-hmU maps were generated based on the chemical method developed in Chapter 3, and an antibody affinity approach using a goat polyclonal antibody specific for 5-hmU (hmU-DIP). Furthermore, since this organism also contains the hypermodified Base J, chemical enrichment of Base J was performed using a modified GLIB-seq approach (Figure 50). This method utilises sodium periodate oxidation of the glucose moiety followed by aldehyde-tagging with the biotinylated hyrazide probe BH.

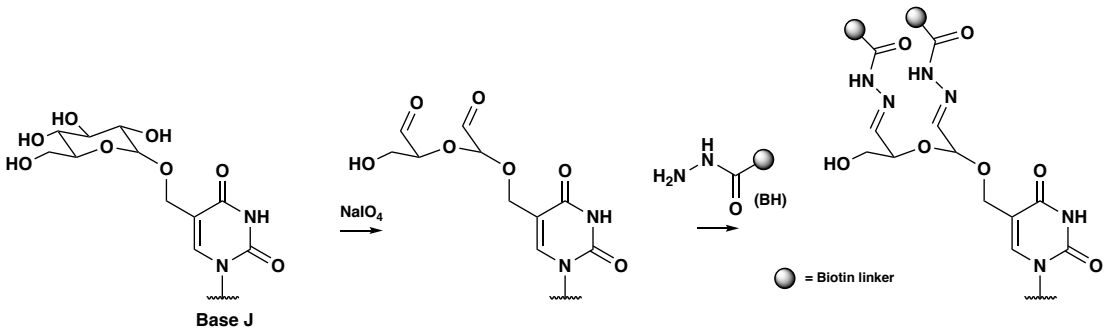


Figure 50: Chemical method utilised for Base J enrichment-sequencing; this involves periodate oxidation of Base J followed by BH tagging of the dialdehydes.

Genomic regions enriched with modifications were determined by a build-up of sequencing reads at certain loci.[203] Genomic regions enriched with 5-hmU were found to overlap with Base J loci (93%) (Figure 51 – A), consistent with this modification being an intermediate in Base J synthesis (Figure 51 - C). However, a notable outcome of this study was the detection of "5-hmU only" loci (Figure 51 – B); these loci were distinct genomic regions enriched for the 5-hmU modification, but not Base J. This indicated a unique function for the 5-hmU mark in addition to being a Base J intermediate.

Sabatini and co-workers have suggested that Base J loci is determined via sequence specific JBP1/JBP2 mediated 5-hmU formation. This is based on an understanding that J-GT efficiently converts 5-hmU to Base J with no sequence specificity.[95] The presence of distinct 5-hmU-loci in the Leishmania trypanosomatid challenges this pre-existing view, since it is not clear why 5-hmU in these regions are protected from further glucosylation (Figure 51 – C). Furthermore, there are numerous regions where Base J is found to be specifically enriched in the absence of 5-hmU. This suggests that 5-hmU in certain sequence contexts is instead more quickly or preferentially glycosylated by J-GT (Figure 51 – C). These observations raised questions about the existence and function of 5-hmU within distinct 5-hmU enriched regions in trypanosomatids, and the sequence specificity of Base J formation.

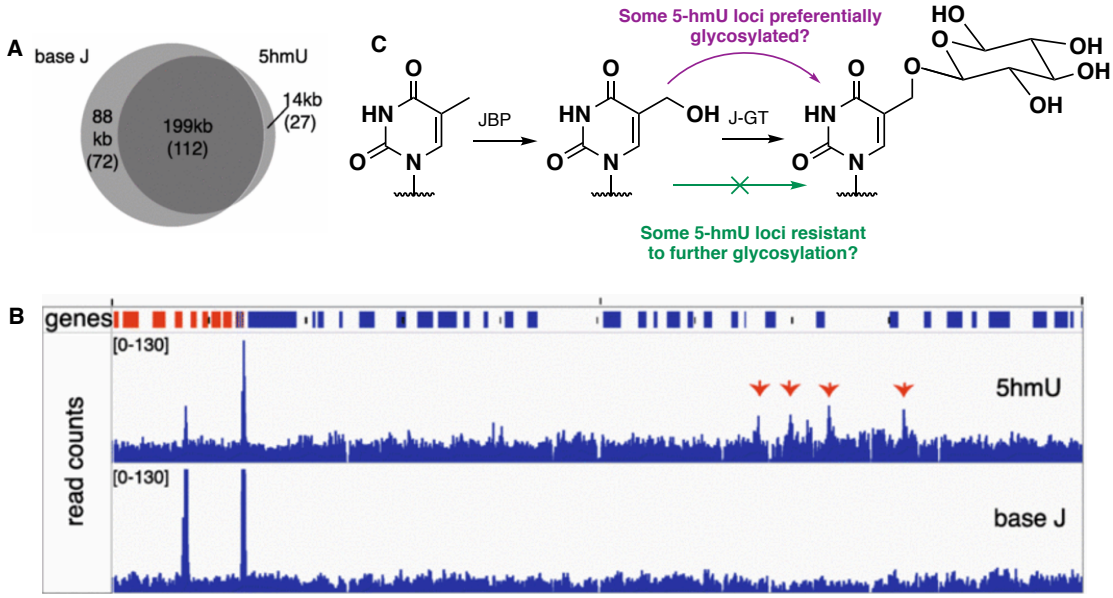


Figure 51: A) Overlap of 5-hmU-enriched loci with Base J-enriched genomic loci in the Leishmania genome as determined by chemical mapping, B) Gene-viewer demonstrating the presence of hmU-only enriched regions (demonstrated by red arrows) determined by 5-hmU chemical mapping (top panel), which are not enriched by Base J chemical-mapping (bottom panel) C) The biosynthetic pathway of Base J formation in Leishmania and other trypanosomatids. The presence of distinct Base J and distinct 5-hmU loci may indicate preferential glucosylation of certain 5-hmU marks. A and B parts of figure are from reference [202]

4.2.2. T-modification Sequencing in *T.brucei*

This was a collaborative project where T-modification sequencing was carried out by myself or Dr F. Kawasaki where indicated; bioinformatics analysis was performed by Dr F. Kawasaki, Dr S. Martinez-Cuesta or Dr D. Beraldi; trypanosome culture was either performed by Dr Janaina Freitas, Carrington group (Differentiation experiment) or myself (5-hmU spike-in experiment); Dr F. Kawasaki assisted with sample preparation for LC-MS/MS. All LC-MS/MS measurements were carried out by myself.

To further probe the role of 5-hmU in trypanosomatids, a collaboration was initiated with the Carrington group (Department of Biochemistry, University of Cambridge) to probe these marks in *T.brucei.* In this organism, 5-hmU is present in both bloodstream form (BSF) and procyclic form (PCF), whilst Base J is only detectable in BSF (Section 2.5). Base J and 5-hmU enrichment-sequencing was firstly performed in both life-stages of *T.brucei* (Sequencing by Dr F. Kawasaki). Distinct 5-hmU loci were also present in BSF *T.brucei* (Figure 52 – A) and were more abundant compared to Leishmania. Many of these regions were commonly enriched with 5-hmU enriched regions observed in PCF (Figure 52 - B), which lacks Base J. This provided further evidence that some 5-hmU marks were resistant to further glucosylation, and suggested that 5-hmU could have a distinct epigenetic role beyond being a Base J intermediate. Studies were therefore designed to assess the sequence specificity, formation and dynamics of J-GT mediated β-glucosylation in PCF *T.Brucei*, to address the existence of distinct 5-hmU loci in these organisms.

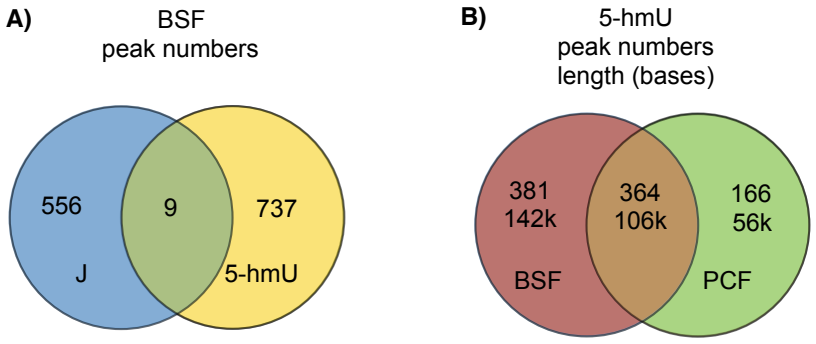


Figure 52: A) Overlap of Base J and 5-hmU distinct loci in BSF trypanosomes, B) Overlap of 5-hmU loci in BSF and PCF life-stages.

4.2.3. Exploring the Specificity of J-GT Enzyme

To further understand J-GT sequence specificity, β-glucosylation of artificially incorporated 5-hmU in PCF *T.brucei* was assessed. Borst and co-workers had previously demonstrated that supplementation of the 5-hmU mononucleoside into PCF culture medium led to formation of Base J.[109] The 5-hmU mononucleoside is assimilated into the nucleobase pool and incorporated into the triphosphate pathway leading to generation of 5-dhmUTP. This leads to the stochastic incorporation of 5-hmU at T-sites throughout

the genome, which can be further glycosylated to form Base J. By replicating this experiment and chemically-mapping Base J formation, the sequence specificity of J-GT could be assessed by determining areas of the genome enriched or depleted of Base J (Figure 53).

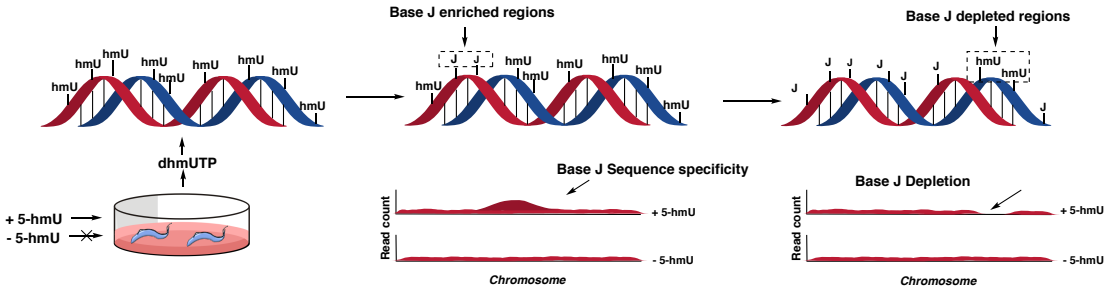


Figure 53: Workflow of hmU-spike in experiment in PCF trypanosomes. 5-hmU incorporation and subsequent Base J chemical-mapping will enable assessment of J-GT sequence-specificity

PCF trypanosomes were therefore cultured in media supplemented with the 5-hmU mononucleoside (1 mM), and left to proliferate without dilution, while control cultures were grown in the absence of mononucleoside. Subsequent DNA extraction and accurate measurement of T-modifications via LC-MS/MS confirmed the incorporation of 5-hmU into artificially-fed PCF DNA. Global 5-hmU levels were ~13-fold higher in 5-hmU-fed samples than in control cultures which had not been supplemented, while global levels were ~2.75-fold higher than the natural abundance of 5-hmU in BSF (Figure 54- Left). Base J formation in the artificially fed trypanosomes was also confirmed by LC-MS/MS (Figure 54- Right).

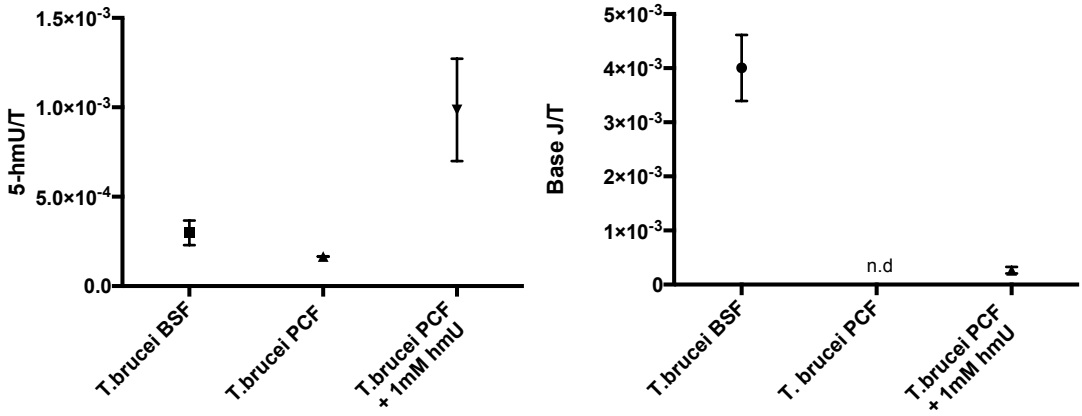


Figure 54: T-modification levels (Left- 5-hmU, Right – Base J) in PCF trypanosomes that had been cultured in the presence of 5-hmU, compared with natural levels in BSF and PCF trypanosomes. n.d = not detected.

5-hmU levels appeared to remain relatively constant irrespective of culture-time in the presence of 5-hmU mononucleoside (Figure 55 - Left). In contrast, Base J levels increased with proliferation time, likely explained by a time-lag associated with enzymatic glucosylation (Figure 55 - Right). Base J levels reached ~6% of those observed in BSF, despite the relative 5-hmU levels being higher. Since not all 5-hmU was subject to glucosylation, this further indicated that J-GT activity may have a sequence-context dependence.

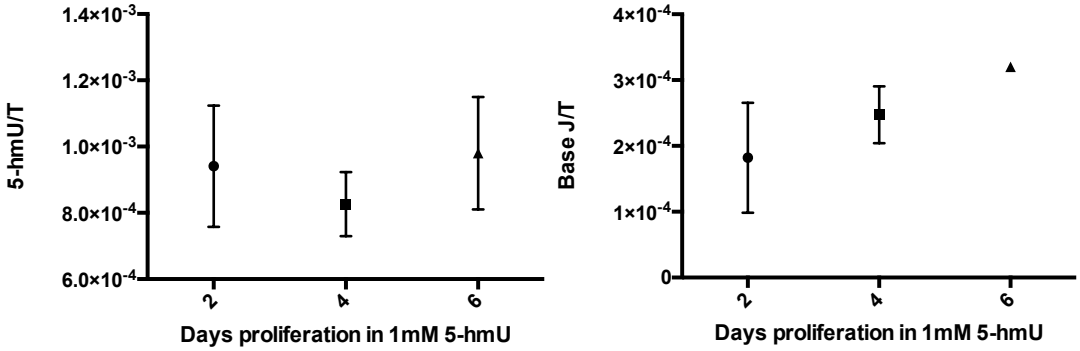


Figure 55: T-modification levels (Left- hmU, Right – Base J) in PCF trypanosomes cultured in the presence of 5-hmU mononucleoside and relative time of culture/proliferation.

To determine the loci of Base J in these samples, and hence J-GT specificity, Base J was mapped via chemical-enrichment sequencing. Since 5-hmU is randomly incorporated into the genome when added to the culture media, areas of Base J enrichment relative to the rest of the genome are indicative of regions where Base J preferentially forms (Figure 53). Furthermore, areas of the genome protected or resistant to β-glucosylation (e.g. the 5-hmU mark fails to be converted to Base J) would have a depleted signal relative to the rest of the genome (Figure 53). The extent of Base J enrichment was normalised against control samples, which had been cultured in the absence of 5-hmU mononucleoside.

Analysis of Base J mapping found that Base J was enriched (e.g. indicative of preferential Base J formation) in genomic regions where Base J is naturally observed in BSF trypanosomes. This indicated a sequence preference for J-GT glucosylation in these regions (Figure 56 – A), although it cannot be ruled out that other factors (e.g. chromatin) may play a role. The extent of Base J enrichment relative to surrounding sites was diminished with increasing proliferating time, as surrounding 5-hmU sites became glucosylated. Notably, after a longer proliferation time, the Base J signal was

strongly depleted in regions of the genome associated with distinct 5-hmU loci (Figure 56 – B). This corroborated the hypothesis that 5-hmU modifications in certain genomic regions are protected from further glucosylation, indicating that Base J formation is sequence specific.

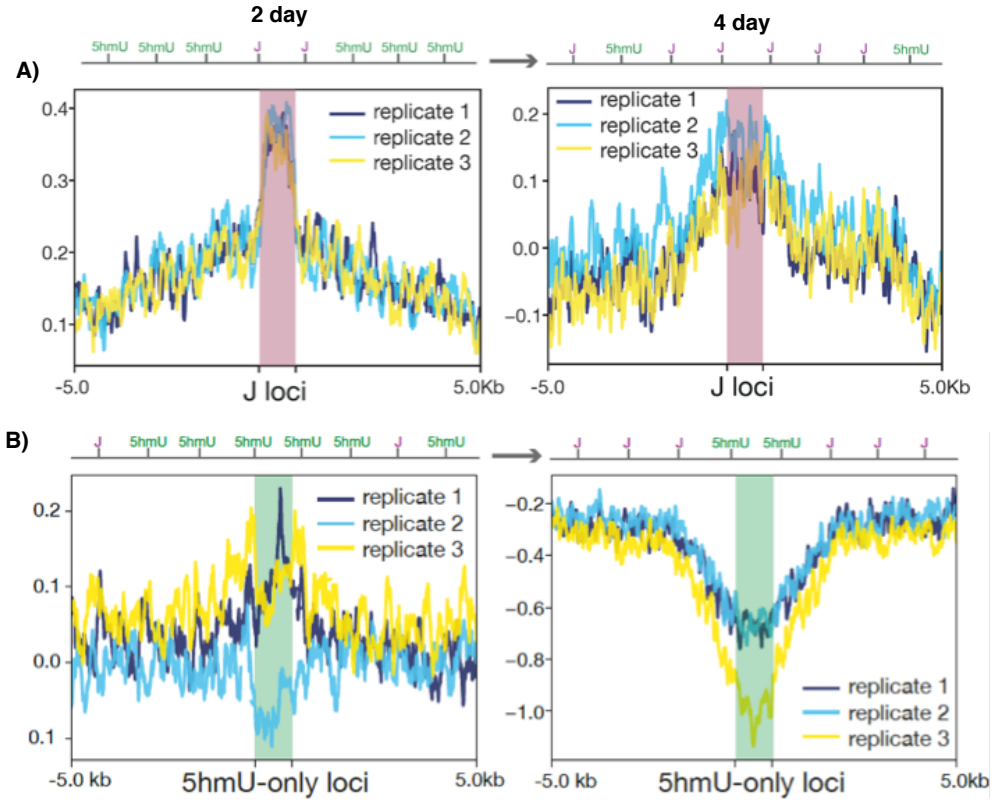


Figure 56: A) Base J chemical enrichment across Base J-enriched regions (pink panel) determined by Base J mapping in BSF, B) Base J chemical enrichment across 5-hmU only regions (green panel) as determined by hmU-chemical enrichment sequencing.

4.2.4. LC-MS/MS Differentiation Study

Currently, a lack of Base J signal in PCF trypanosomes has been attributed to a lack of Base J maintenance due to reduced JBP expression; thus, Base J levels are passively removed by dilution via proliferation.[101] However, an alternative hypothesis could be that Base J is actively removed (e.g. enzymatically excised in a replication-independent manner) from the genome. This could potentially explain the absence of Base J in PCF trypanosomes and the observance of distinct hmU-only loci in both life-forms.

Since Base J levels vary drastically between two life forms (Base J is only detected in BSF, section 2.5), Base J dynamics were assessed during the differentiation process (BSF → PCF) via LC-MS/MS. The differentiation process in BSF trypanosomes was initiated *in vitro* via 1) cold-shock stimulation, and 2) chemical treatment with citrate/*cis*-aconitate. Since BSF trypanosomes reside in the homoeothermic environment (37 °C) of the mammalian host, a cold-shock is thought to signal a change of environment indicative of the tsetse fly host, which initiates differentiation.[204] In the latter case, BSF trypanosomes cannot respire in the presence of citrate/*cis*-aconitate, thus addition of these chemicals leads to differentiation and production of procyclic-specific enzymes capable of breaking down these metabolites.[205] Some time after differentiation initiation, cells begin to proliferate; thus, Base J levels assessed by LC-MS/MS should be compared to theoretical levels expected by dilution proliferation.

Interestingly, Base J levels were estimated to be generally lower than that expected due to dilution by proliferation, however this effect was quite slight (Figure 57). Most notable is a loss of Base J after 24 hr after *cis*-aconitate/citrate addition, despite cells not proliferating until > 24 hr. This provided some indication that Base J may be actively removed from the genome. Such a phenomenon could be one potential explanation for the observation of 5-hmU, and absence of Base J in PCFs. However, further work needs to be performed to conclusively demonstrate this process; more accurate measurements of cell proliferation (e.g. via isotopic labelling) would be highly beneficial.

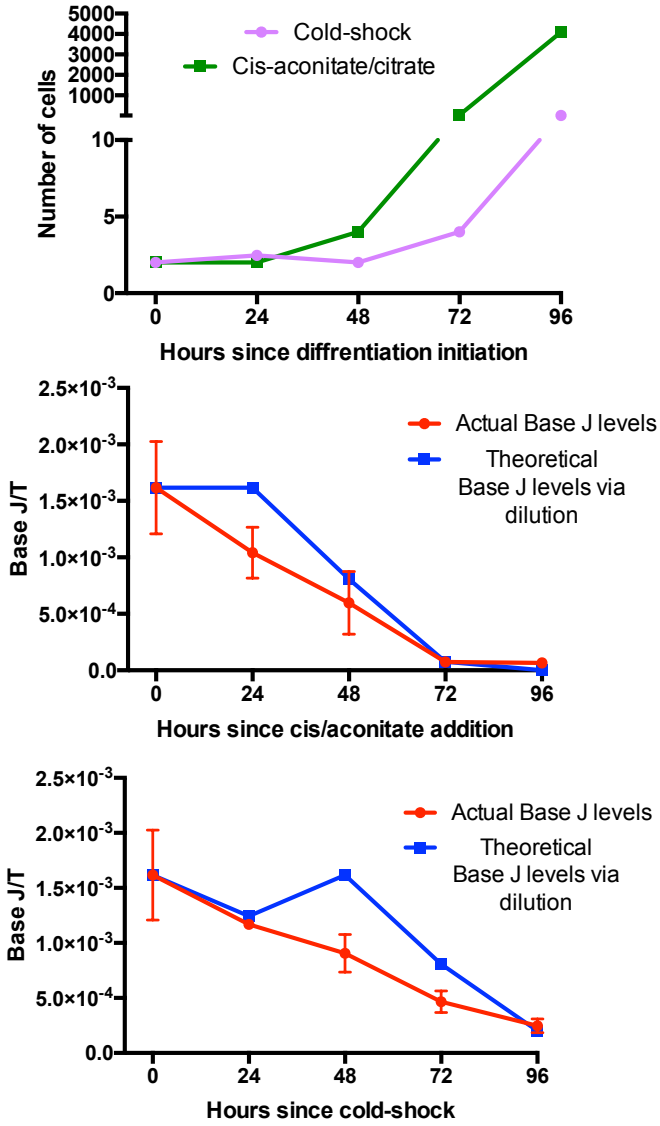


Figure 57: Top: Number of cells and timepoints after differentiation initiated by both chemical (green) and cold-shock (lilac) methods. Middle (*cis*-aconitate/citrate) and bottom (cold-shock): LC-MS/MS levels of Base J (red) and those expected due to Base J dilution by proliferation as determined by number of cells (blue). All plotted against hours after differentiation initiation.

The identification of 5-hmU-only loci, distinct from Base J, indicates a unique role for 5-hmU in this organism aside from being a Base J intermediate. These areas remain depleted with Base J when 5-hmU is stochastically incorporated into the trypanosomatid genome, providing evidence that Base J formation occurs in a non-random manner. This may highlight that the J-GT enzyme possesses sequence specificity, or could reflect active Base J removal from the genome at certain sites. These studies therefore provide further insight into the dynamics and biosynthesis of T-modifications in *T.Brucei.*

4.3. Sequencing T-modifications (5-hmU/5-fU) in Mammalian Tissue

Bioinformatic analysis was performed by Dr Sergio Martinez-Cuesta

4.3.1. Introduction and Design of Experiment

T-modification enrichment sequencing methods had been validated in the model organism *Leishmania*; thus, these methods could now be utilised to generate the first maps of T-modifications in the human genome. This would be important to further investigate the biological role of these marks in mammals.

SMUG1 knockdown HEK293T cells were chosen as the first mammalian biological target for T-modification enrichment sequencing. The design of this experiment was analogous to 5-fC affinity-mapping studies performed by the Balasubramanian lab[51,53] (chemical enrichment) and Zhang lab (antibody enrichment)[42]. 5-fC enrichment-sequencing in TDG depleted samples was found to generate more regions with significant enrichment compared to the wild-type model. Since SMUG1 is the main 5-hmU and 5-fU DNA glycosylase, knockdown of this protein should reduce the extent of T-modification excision and hence increase the levels and lifetime of these DNA modifications in the genome. This would lead to accumulation of T-modifications in regions where they naturally form; thus, subsequent mapping of these regions would help to infer their biological relevance (Figure 58). HEK293T cells were chosen as this cell-line is easily amenable to RNA interference. Furthermore, the T-modification maps generated could be subjected to association studies with pre-existing datasets available in this cell-line (e.g. DNaseI, TET-binding sites) for further functional investigation.[206]

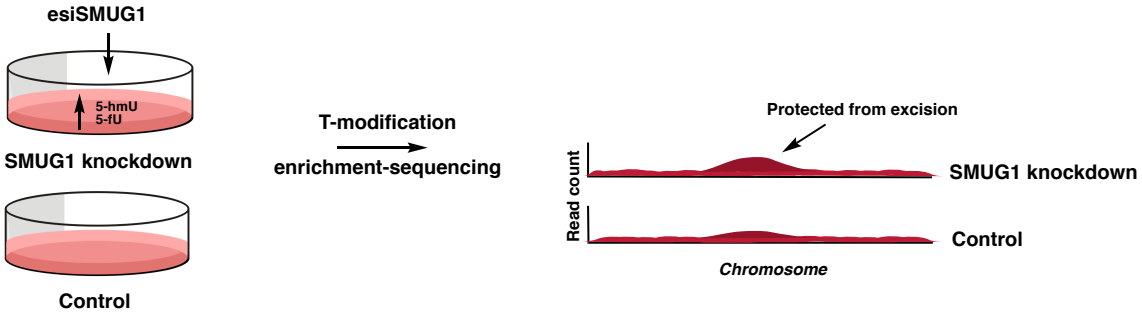


Figure 58: Design of experiment using SMUG1 knockdown for T-modification enrichment sequencing

5-hmU and 5-fU levels had been shown to increase by 40% and 70% respectively in cells where SMUG1 had been knocked down with 60% transfection efficiency.[40] Furthermore, a recent study suggested that SMUG1 knockout in mouse brain led to ~26-fold increase in 5-hmU levels.[207]

4.3.2. HEK293T esiRNA SMUG1 Knockdown

HEK293T cells were transfected with SMUG1 esiRNA for 96 hrs. esiRNA is a heterogeneous mixture of different siRNA sequences that target the same mRNA transcript, thus, this approach has generally higher transfection efficacies and exhibits less off-target effects compared to a single sequence target.[208,209] It was reasoned that longer transfection times would: 1) increase transfection efficiency and 2) allow T-modifications a longer time to naturally accumulate without repair. As a control, cells were cultured in the presence of an esiRNA targeting Renilla Luciferase (RLuc), a gene not present in the human genome.

SMUG1 knockdown and esiRNA transfection efficiency was firstly monitored by reverse-transcription-PCR (RT-PCR) of SMUG1 transcripts normalised to a housekeeping control gene (e.g. actin). This allowed the relative quantification of mRNA knockdown, normalising for amount of RNA input. Following this, RNA was extracted from the harvested cells and reverse transcribed to cDNA before qPCR. Relative quantification was assessed using four combinations of primer pairs (SMUG1 vs. actin) in SMUG1 knockdown samples compared to the RLuc esiRNA transfected negative control. An average 94% transfection efficiency was observed via qPCR after 96 hr (Figure 59).

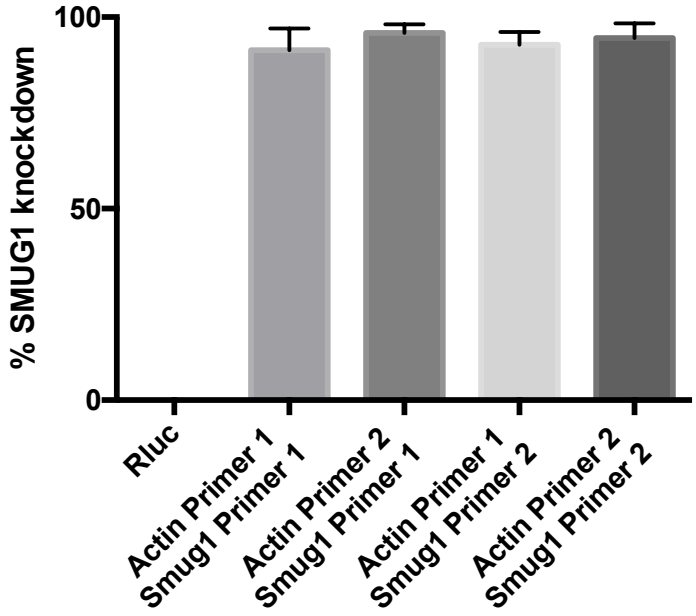


Figure 59: Knockdown of SMUG1 as determined by mRNA Smug1 transcript levels versus actin via RT-PCR (Mean + SD).

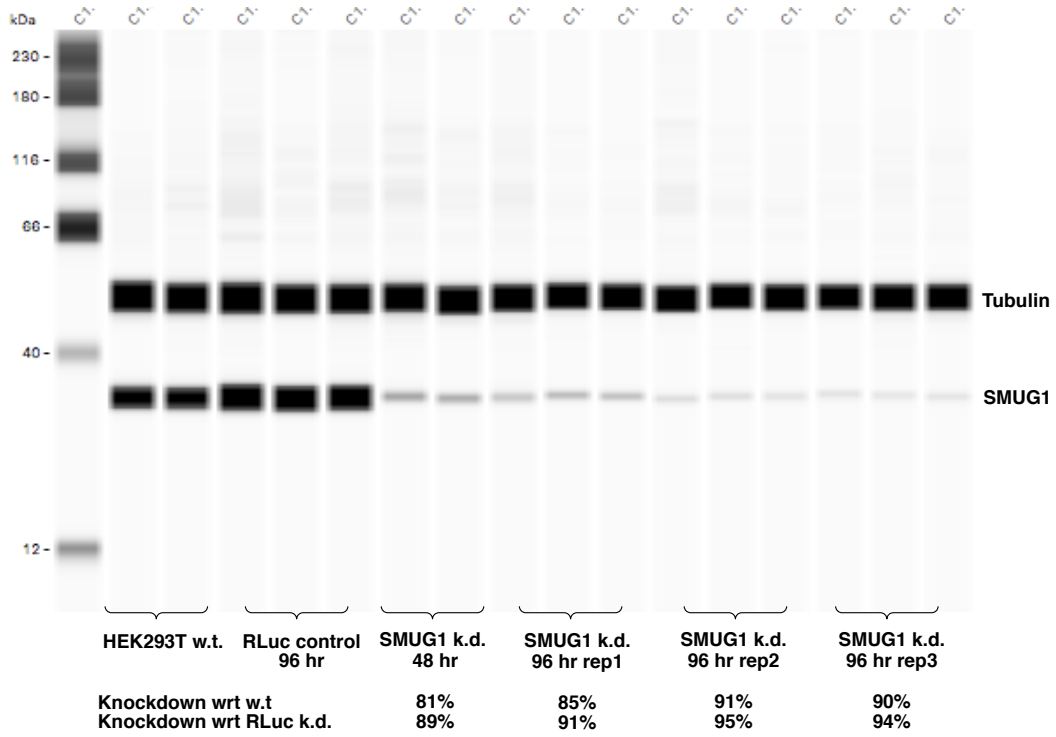


Figure 60: Western blot demonstrated knockdown of SMUG1 protein after esiRNA transfection relative to wild-type HEK293T and RLuc transfected negative control. k.d. = knockdown, w.t. = wildtype

The SMUG1 esiRNA transfection was also assessed using a Western blot plot to determine the percentage knockdown of the SMUG1 protein. Using rabbit monoclonal antibodies against SMUG1 and tubulin, disappearance of a 35 kDa band corresponding to the SMUG1 protein was observed in knockdown protein extracts (Figure 60). The average percentage knockdown of SMUG1 protein was ~88% across all replicates when normalised to wild-type and ~94% compared to the RLuc transfected negative control.

Next, T-modification levels in transfected cells were assessed by LC-MS/MS. As discussed in Chapter 2, variation between biological and technical digestion replicates is observed; however, the overall average level of 5-hmU was found to be higher in SMUG1 knockdown samples compared to that of wild-type (~2-fold). Accurate measurements of 5-fU weren't possible at this stage, impeded by its higher limit of detection

4.3.3. Mammalian hmU-modification Enrichment Sequencing via hmU-DIP

5-hmU enriched regions in both SMUG1 knockdown and RLuc samples were firstly mapped using the hmU-DIP method, which utilises a commercial goat polyclonal antibody that has been raised against the 5-hmU mononucleoside.[202] After ligation of sequencing adaptors, DNA was denatured by heating before incubation with the antibody at 4 °C overnight. The mixture was then immunoprecipitated using magnetic

beads that bind to Protein G, while non-bound DNA was removed in successive wash steps. Bound-DNA fragments were subsequently eluted from the antibody by heating in the presence of Proteinase K.

5-hmU enriched regions were determined by an increased number of sequencing reads as detected by the MACS peak caller[203] (p-value < 10^{-5}) compared to input (non-enriched DNA). Within technical replicates (n = 2), 29989 consensus peaks (~50% overlap) were found between SMUG1 knockdown libraries. To exclude the possibility that peaks arise due to the inherent reactivity of the 5-hmU antibody, peaks in RLuc transfected controls were also called. The consensus peaks in the SMUG1 knockdown libraries were approximately 10-fold higher than consensus peaks observed in the RLuc transfected negative controls (3225). The elevated levels likely reflect the increase in 5-hmU due to depletion of SMUG1-mediated 5-hmU excision upon knockdown. In addition, SMUG1 peaks were further validated by control hmU-DIP libraries that had been enzymatically treated with SMUG1 (e.g. to remove 5-hmU) prior to affinity-enrichment; this led to the disappearance of peaks (Appendix – Chapter 4).

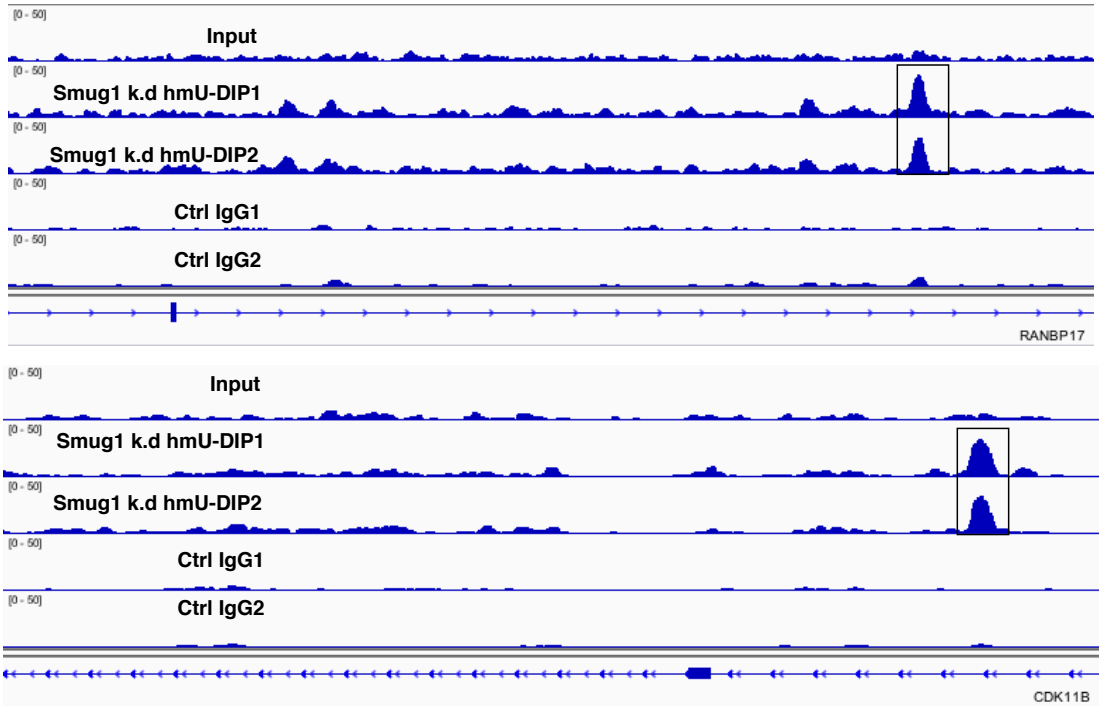


Figure 61: Example of 5-hmU enriched regions in SMUG1 knockdown samples assessed by hmU-DIP sequencing.

To exclude method-derived artefacts in a stringent manner, control-DIP sequencing was performed using a non-specific goat polyclonal IgG antibody. This is a recommended control for antibody-affinity sequencing methods,[210] and accounts for PCR efficiency biases which may arise in low-diversity sequencing libraries.[211] Thus, any genomic regions that also demonstrated a large number of sequencing reads with the control IgG antibody were eliminated from the SMUG1 knockdown 5-hmU dataset; this gave 11937 high confidence peaks which were used for downstream analysis. The non-random genomic distribution of 5-hmU indicated that this mark may have a functional role (Figure 61).

4.3.3.i Motif Analysis, Origin in Cells and Association with Chromatin

5-hmU in mammalian cells is proposed to derive from either ROS or TET oxidation of T.[40] To try and establish the origin of 5-hmU in HEK293T cells, motif-sequence analysis of 5-hmU-enriched regions was firstly determined using the Multiple EM for Motif Elicitation (MEME) and Discriminative Regular Expression Motif Elicitation (DREME) tool.[212]

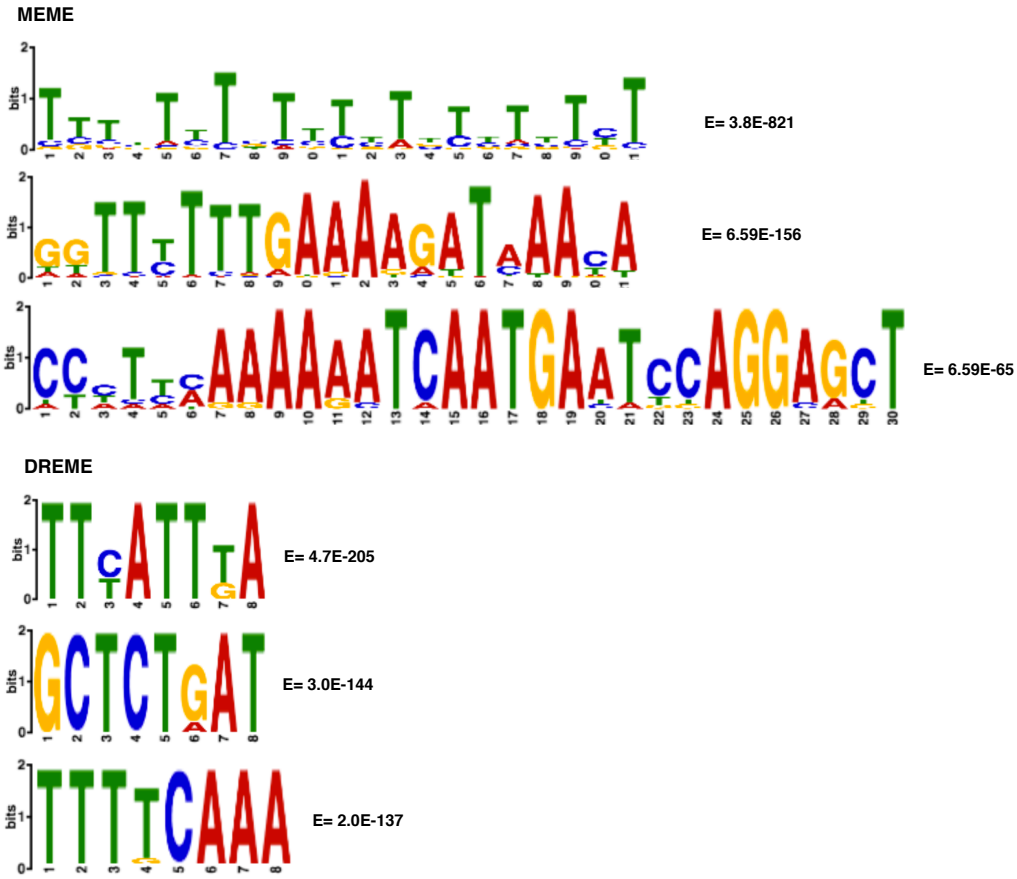


Figure 62: Common motifs of hmU-enriched regions as determined by MEME (Top) AND DREME (Bottom) analysis. E-value is a measure of statistical significance that a certain motif is enriched within a dataset compared to a random dataset of the same size and length.

Enriched motifs within 5-hmU-loci were found to be generally T-rich in nature, containing distinct motifs (Figure 62). Observed motifs did not resemble known TET binding sites, which are usually enriched for CpG dinucleotides.[213,214]

To further probe the origin of 5-hmU in genomic DNA, 5-hmU peaks were overlapped with TET2 and TET3 chromatin immunoprecipitation (ChIP) datasets, generated in HEK293T cells (ENCODE GEO # GSM897576 and GSM897577).[213] Only 0.3% and 0.2% peaks overlapped with TET2 and TET3 binding sites respectively; this suggested that the majority of 5-hmU loci (as determined by hmU-DIP) arises from a TET-independent process in HEK293T cells. This may, however, reflect the low expression of TET enzymes in this cell-line.[40]

5-hmU is therefore likely to be formed in ROS-dependent processes in HEK293T cells, and thus may have a similar biological role to 8-oxoG (Introduction 1.3.5). Thus, 5-hmU loci were correlated with chromatin accessibility, as accessible genomic loci may be more susceptible to oxidative DNA damage.[215] 5-hmU enriched loci were compared to regions of open chromatin, as determined by a DNaseI-seq dataset in HEK293T cells (ENCODE GEO # GSM1008573); DNaseI digests fragments of nucleosome-depleted DNA, which are then sequenced by NGS.[216] Only 1.3% of hmU-loci were found to overlap with open chromatin regions, and 5-hmU loci were instead depleted from open chromatin regions compared to a random distribution of genomic loci (Figure 63). This implied that the 5-hmU mark may be associated with heterochromatin and silenced genes, however this could also reflect preferential repair of 5-hmU in open chromatin regions.[217,215]

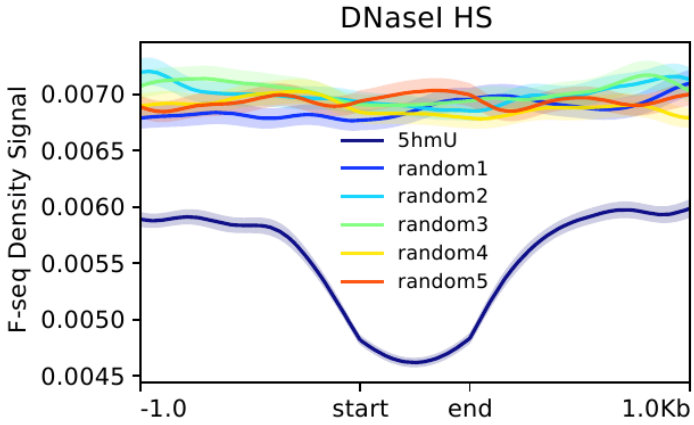


Figure 63: Within 5-hmU enriched regions, DNaseI density signal is depleted compared to randomly selected genomic regions (x5) of the same size distribution as the 5-hmU genomic region. F-seq density signal refers to regions of enriched signal from DNaseI experiments based on F-seq peak-calling.[218]

4.3.3.ii Genomic Location and Association with Gene Expression

To investigate any role of 5-hmU and its association with genes, the distribution of 5-hmU loci among functional features was determined using the genomic-annotation tester (GAT) tool.[219] 5-hmU was found to be mainly located in intergenic (65%) regions or gene deserts, while a large proportion of 5-hmU enriched regions also occurred in intragenic regions (33%, Figure 64). Notably, 5-hmU loci were generally depleted in other genomic regions compared to the normal genomic distribution (Figure 64). This included 1) a depletion in promoters, where oxidised cytosine derivatives are typically enriched[53,220] and 2) a strong depletion in exons, a phenomenon also reported for 8-oxoG.[56,146] This could indicate preferential repair or suppression in the creation of these marks in coding regions, which would eliminate potential mutagenesis caused by their existence or repair.

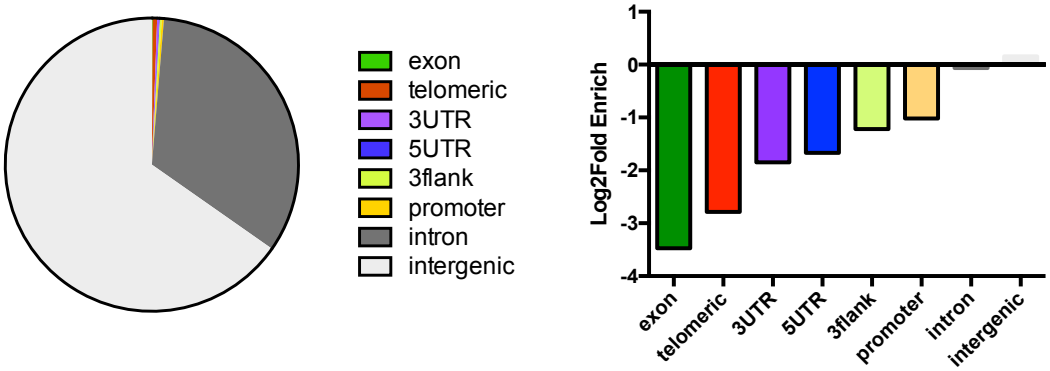


Figure 64: Left - Genomic distribution of 5-hmU-enriched regions determined by GAT, Right - Log_2Fold enrichment of 5-hmU enriched regions relative to the typical genomic distribution, determined by peak width.

To further investigate the role of 5-hmU associated genes, gene ontology was performed using DAVID (the database for annotation, visualisation and integrated discovery) functional analysis[221,222] on the subset of 5-hmU enriched regions in the vicinity (±1 kB) of a gene; both GO_biological process (Table 8), and KEGG_pathway (Table 9) ontology tools were used. A noticeable link was observed between 5-hmU enriched-loci and genes involved in cellular signalling, and a number of these functions were specifically associated with synaptic transmission and the central nervous system.

GO_Biological Process	Number of genes	q-value
Regulation of GTPase activity	25	9.90E-04
Protein phosphorylation	89	1.60E-02
Extracellular matrix organisation	47	1.70E-02
Cell adhesion	90	2.00E-02
Peptidyl-tyrosine phosphorylation	38	3.70E-02
Microtuble cytoskeleton organisation	22	5.30E-02
Ephrin receptor signalling pathway	25	5.50E-02
Synaptic transmission, glutamatergic	11	5.70E-02
Positive regulation of GTPase activity	100	1.00E-01

Table 8: Functional enrichment as defined by GO_biological process terms, number of genes and associated q-value.

KEGG_pathway	Number of genes	q-value
Morphine Addiction	30	8.3E-05
Retrograde endocannabinoid signalling	30	4.6E-04
ECM-receptor interaction	27	4.4E-04
Focal adhesion	48	3.7E-04
cAMP signalling pathway	45	9.8E-04
Axon guidance	33	1.2E-03
Calcium signalling pathway	41	1.4E-03
Glutamergic synapse	30	1.5E-03
PI3K-Akt signalling pathway	67	1.6E-03
Oxytocin signalling pathway	36	4.3E-03
Cholinergic synapse	28	4.4E-03
Circadian entrainment	25	4.9E-03

Table 9: Functional enrichment as defined by KEGG pathway, number of genes and associated q-value.

The 5-hmU mark, and subsequent SMUG1 excision, may be important for influencing gene regulation, potentially in response to environmental factors, such as ROS. Thus, to further investigate the relationship between 5-hmU and gene expression, differential mRNA levels between SMUG1 knockdown and control cells were assessed via RNA-seq. To prepare samples for RNA-seq, total RNA was first extracted and mRNA was selected for using oligo-(dT) magnetic beads which bind to poly(A) regions. The RNA was fragmented and reverse transcribed, followed by second-strand DNA synthesis, both using random priming. The resultant cDNA was prepared for NGS via standard library preparation (Section - 3.7.1).

There were a number of differentially expressed genes (both upregulated and downregulated) which contained 5-hmU in the SMUG1 knockdown sample, but not in the RLuc negatively transfected control (Figure 64). This included GFI1B (4.6-fold upregulated), a transcriptional repressor important for blood-cell development,[223] and DAB1 (2.2-fold downregulated), associated with signal transduction in the central nervous system,[224] among others (Appendix – Chapter 4). This implied that 5-hmU (e.g. by transcription factor or recruitment of chromatin remodelling proteins),[40] or BER-excision (shown to both initiate or repress transcription),[57,225] could be responsible for altered expression at specific genes.

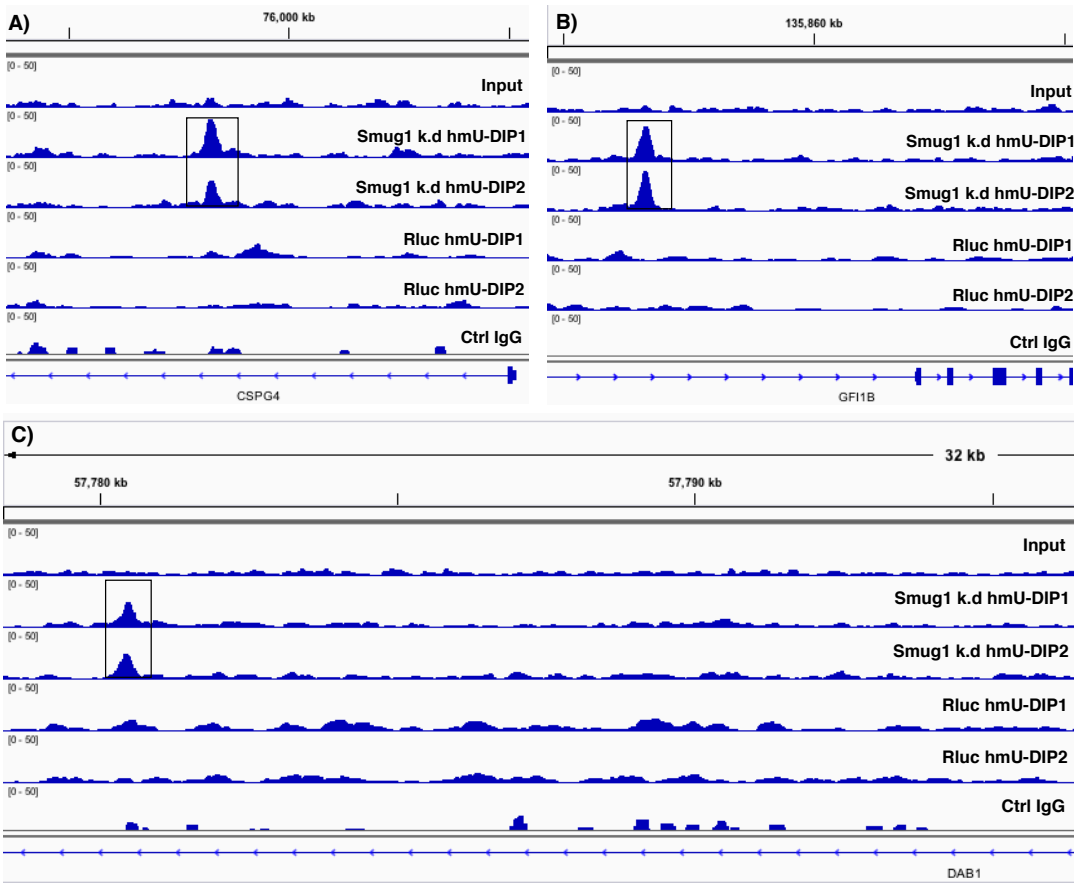


Figure 64: The presence of 5-hmU peaks in SMUG1 knockdown samples, but not RLuc samples in differentially expressed genes. A) CSPG4 (1.5-fold upregulated in SMUG1 knockdown, q = 9.5E-03) associates with malignant tumor progression[226], B) GFI1B (4.6-fold upregulated in SMUG1 knockdown, q = 1.6E-02) C) DAB1 (2.2-fold downregulated in SMUG1 knockdown, q = 6.6E-03)

However, within genes proximal to 5-hmU, there was no significant change in global transcription upon SMUG1 knockdown (Figure 65).

Figure 65: No significant change in RNA read count occurs upon SMUG1 knockdown within genes proximal to 5-hmU (1670 genes contained hmU peaks in both SMUG1 hmU-DIP replicates and in neither of the RLuc controls (1670).

Furthermore, there was no clear correlation between differentially expressed genes and those that contained 5-hmU. In total, 564 genes were differentially expressed (q-value < 0.05). Of 244 downregulated and 321 upregulated genes, only 14% and 11% contained 5-hmU respectively. This indicated that SMUG1-mediated 5-hmU excision does not globally alter gene expression.

4.4. 5-fU and 5-hmU Chemical-enrichment Sequencing

T-modification sequencing was also attempted in the SMUG1 knockdown HEK293T samples using chemical enrichment-methods (Section 3.7.3). For 5-fU chemical sequencing, efficient enrichment was confirmed by model ODN spike-in controls, however, a limited number of enriched regions (< 200 peaks) were determined using the MACS[203] peak-caller ($p < 1E-05$). Lowering the significance threshold led to an increased number of peaks; however, a limited number of consensus regions were observed between replicates (Table 10). This may indicate that the chemical method suffers from technical biases, potentially against clusters of modifications, or high background noise, which could inhibit 5-fU detection.

	Peaks Rep 1	Peaks Rep 2	Consensus peaks
p < 1E-05	155	26	5
p < 1E-04	919	127	12

Table 10: Number of peaks identified by MACS in 5-fU chemical enrichment libraries and consensus peaks between technical replicates.

Furthermore, 5-hmU chemical enrichment sequencing with the SMUG1 knockdown sample also failed to identify regions of 5-hmU loci (34 peaks, $p < 1E-05$) at this stage. As such, further optimisation is required to enable detection of T-modifications in mammalian samples using this method. It is currently not clear why there is discrepancy with the hmU-DIP method, although the chemical method is also found to detect fewer peaks than hmU-DIP in trypanosomatid genomes. The chemical method suffers from a relatively large background signal; thus, the method may not be sensitive enough to detect T-modifications in mammals where they are less abundant.

4.5. Conclusion

To investigate potential functions of T-modifications in both trypanosomatids and mammalian cells, genome-wide localisation profiles of these marks, as well as their global levels, were investigated using NGS and LCMS/MS.

In trypanosomatids, 5-hmU was observed in distinct loci from its downstream product, Base J, suggesting 5-hmU itself may have a unique function in these organisms. By tracing Base J formation by chemical-sequencing, 5-hmU appeared to be maintained in certain loci without being converted to Base J. This challenges the current understanding that 5-hmU is the substrate of J-GT regardless of sequence contexts. Further work should address the biological functions of 5-hmU in such loci by 1) perturbing relevant enzymes involved with T-modification biosynthesis (e.g. JBP1, JBP2, J-GT), in combination with any changes in T-modification profile and gene expression and 2) identifying reader proteins of 5-hmU and Base J in trypanosomatid systems. Proteomic studies may also identify candidate enzymes which can dynamically regulate 5-hmU and Base J (e.g. via active removal of Base J.)

In mammalian systems, where T-modifications are efficiently removed by BER, esiRNA SMUG1 knockdown was performed to investigate the consequence of T-modification excision deficiency. A significant increase of 5-hmU loci were observed in SMUG1 knockdown HEK293T cells compared to controls. These loci were found not to correlate with reported TET-binding sites, indicating that the majority of 5-hmU loci originate from TET-independent processes in HEK293T cells. 5-hmU is therefore instead likely to arise from ROS; 5-hmU loci were found to be mainly located in intergenic and intronic regions, and were depleted in coding regions, similar to another major ROS product 8-oxoG.

Genes that were proximal to 5-hmU loci were enriched with genes associated with cellular signalling and the central nervous system by gene ontology analysis. A number of candidate genes proximal to 5-hmU were differentially expressed upon SMUG1 knockdown, however, 5-hmU was found to have no significant global effect on differential mRNA levels.

Future work should focus on continuing to elucidate a potential biological role or consequence for this modification in the mammalian genome, although it cannot be ruled out that 5-hmU solely occurs due to random oxidative damage with no functional role. Biological perturbation studies (e.g. increased exposure to ROS) and their corresponding effect on hmU-loci and gene expression will help to establish their functional role or consequence. This is of particular importance for genes proximal to 5-hmU that are differentially expressed upon SMUG1 knockdown. Other biological targets, e.g. SMUG1 knockout mice, may yield a more significant result, since levels of 5-hmU change by ~26-fold upon SMUG1 depletion. In addition, similar hmU-DIP experiments could also be performed in a mESC model; mESCs have higher natural TET expression;[40] thus, TET-mediated 5-hmU formation may be more relevant in this cell-line.

For further analysis, efforts should focus on the optimisation of T-modification chemical enrichment methods (e.g. reduce background signal), which currently fail to identify substantial regions of T-modification enrichment. This would be an important tool to co-validate 5-hmU regions identified by the hmU-DIP method.

5. Identification of 5-fU protein binders by proteomics

5.1. Introduction

DNA modifications bear unique chemical moieties that protrude into the major groove, and thus have the propensity to alter protein recognition and recruitment to DNA. DNA binding proteins include those that can 1) create DNA modifications ("writer"); 2) recognise DNA modifications ("reader") and 3) remove or excise DNA modifications ("eraser").[147] Writer and eraser proteins control DNA modification dynamics, while the presence of these marks can alter the way in which protein readers, such as transcription factors, can bind or interact with DNA. This ultimately can have a downstream effect on the regulation of genes.[25,49]

Affinity-based enrichment of proteins that tightly bind to DNA modifications (i.e. pulldown), in combination with mass spectrometry-based proteomics, is a powerful tool in functional genomics (Figure 66). Biotinylated DNA baits, containing modified bases of interest, are incubated with protein extract, followed by enrichment of bound proteins via the streptavidin-biotin interaction.[25,40,49,148] Subsequent proteomics experiments utilise the unique sequence and LC-MS/MS fragmentation pattern of peptides to identify proteins in an enriched pool. Proteins are enzymatically digested into smaller polypeptide units and identified by comparison with comprehensive polypeptide mass databases for the organism in question. In a data-dependent acquisition (DDA) approach, a full MS spectrum acquired alongside MS^2 fragmentation enables a large number of proteins to be identified simultaneously.[227,228] Integration of the mass signal can be subsequently used for protein quantification.

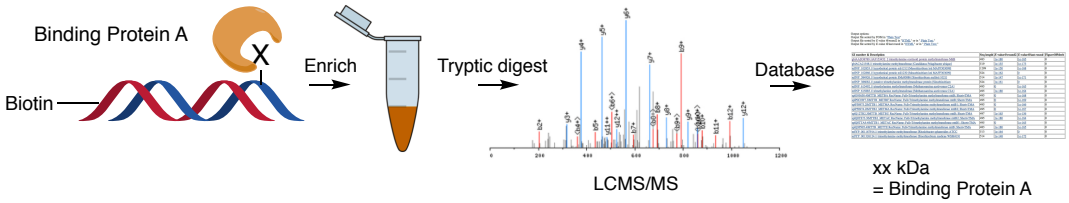


Figure 66: Workflow of protein pulldown and identification of proteins via mass spectrometry based proteomics.

Proteomics studies have identified a number of proteins that preferentially interact with the C-modifications in mammals,[49,50] and also with the T-modification 5-hmU.[40] 5-hmU protein pulldown identified a number of transcription factors and chromatin remodelling proteins that are specifically enriched, giving further credence to 5-hmU

having a regulatory role. However, no such study had been carried out to determine global protein interactions with 5-fU, a potentially mutagenic mark implicated in disease. Thus, 5-fU protein interactions needed to be explored, in order to reveal further information about its potential role, function or biological consequence.

5.2. 5-fU Pulldown and Proteomics

HEK293T nuclear protein extract was provided by Dr Sabrina Huber, Balasubramanian group.

Biotinylated 5-fU modified dsDNA (fU-DNA1) was synthesized by PCR, along with a non-modified control (GCAT-DNA1), to be used as protein baits. The DNA sequence used for pulldown corresponded to a region of the p53 tumor suppressor gene, prone to mutation in cancer tissue.[229] A modified proteomics pulldown procedure was utilised based on that which had been previously reported for 5-fC pulldown;[49] biotinylated DNA was pre-incubated with streptavidin beads, before incubation with nuclear protein extract at 4 °C. The supernatant was removed and the magnetic beads were washed to remove non-interacting proteins. Pulled-down proteins were eluted from the beads, ran on a Bis-Tris Nupage gel and visualised with Coomassie blue stain. A clear enrichment of proteins in the presence of biotinylated DNA was observed compared to the no-DNA control; this suggested that non-specific binding to streptavidin beads was minimal (Figure 67).

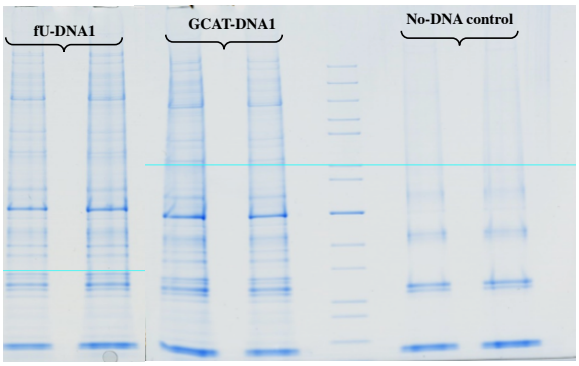


Figure 67: Biotinylated DNA pulldown of proteins, compared (HEK293T nuclear extract) with no-DNA/"beads-only" control. 4-20% SDS-PAGE gel ran in MOPS buffer, stained with Coomassie blue.

The gel-lanes were cut, subjected to tryptic digest and submitted for proteomic analysis. Using Scaffold software,[230] pulled-down proteins were compared semi-quantitatively via t-test, normalising for total protein spectral counts;[230] a method utilised for many relative quantification proteomics studies.[231,232,233,234] This permitted identification of proteins that were either enriched or depleted in the presence of 5-fU modified DNA compared to the non-modified DNA control (Figure 68). For two replicates, a p-value

significance threshold of > 0.1, a protein threshold of 95%, a peptide threshold of 50%, and a minimum identification of two peptides for each protein was required. Of the 1283 proteins identified, 41 were found to be consistently enriched in the 5-fU DNA-bound pool (> 0.4 log_2-fold-change 5-fU/control, Table 11), whilst 53 proteins were found to be consistently depleted in the 5-fU DNA-bound pool (> 0.4 log_2-fold-change control/5-fU, Table 13)

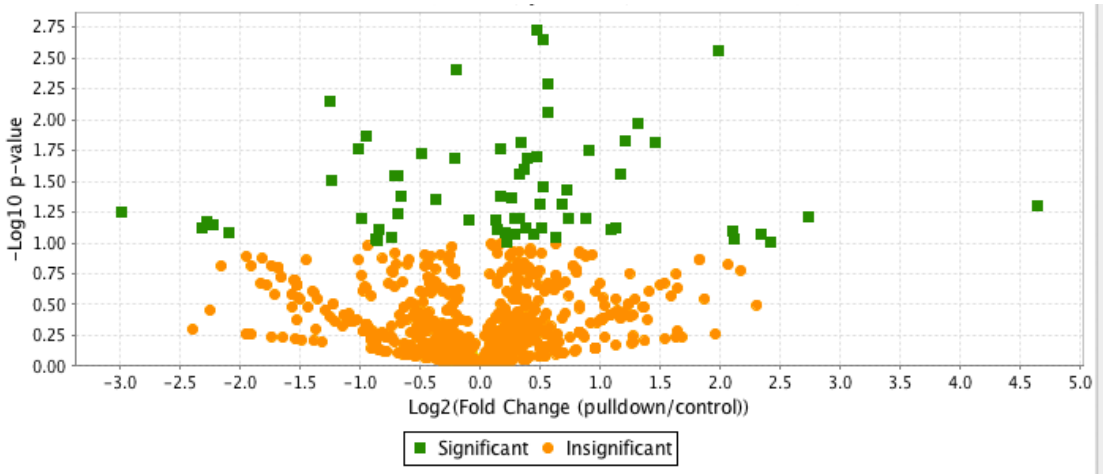


Figure 68: Volcano plot showing fold-change enrichment and their significance after t-test using Scaffold software.

Protein	Log2Fold Enrichment fUDNA1/GCATDNA1	p-value
RNA-binding protein 4 (RBM4)	4.43	8.39E-02
Vigilin (HDLBP)	4.19	8.52E-02
DNA primase small subunit (PRIM1)	4.16	2.10E-03
TAR DNA-binding protein 43 (TARDBP)	3.09	1.99E-02
Transcriptional repressor protein YY1 (YY1)	2.56	5.06E-02
DNA Primase large subunit (PRIM2)	2.50	7.76E-02
Mediator of RNA polymerase II transcription subunit 6 (MED6)	2.26	3.43E-02
28S ribosomal protein S23, mitochondrial (MRPS23)	2.26	2.45E-03
Meiotic nuclear division protein 1 homolog (MND1)	2.26	4.98E-03
Centrosome-associated protein 350 (CEP350)	2.24	8.59E-03
Procollagen galactosyltransferase 1 (COLGALT1)	2.17	9.14E-02
Protein FAM50A (FAM50A)	2.11	9.14E-02
Heterogeneous nuclear ribonucleoprotein M (HNRPM)	1.99	4.92E-02
Integrator complex subunit 3 (INTS3)	1.99	3.74E-02
5'-3' exoribonuclease 2 (XRN2)	1.67	6.33E-02
Glutathione peroxidase (Gpx)	1.67	6.38E-02
Staphylococcal nuclease domain-containing protein 1 (SND1)	1.46	1.79E-02
Cell growth-inhibiting protein 34	1.37	7.89E-02
General transcription factor IIH subunit 1 (GTF2H1)	1.37	7.64E-02
ATP-dependent RNA helicase (DDX1)	1.32	2.78E-02
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	1.21	1.53E-02
tRNA-splicing ligase RtcB homolog (RTCB)	1.17	1.10E-02
UPF0568 protein(C14orf166)	1.14	9.35E-02
Treacle protein (tcof1)	1.09	9.35E-02
Histone deacetylase complex subunit (SAP130)	0.91	1.54E-02
Ubiquitin-associated protein 2-like (UBAP2L)	0.88	6.26E-02
Uracil-DNA glycosylase (UNG)	0.74	6.26E-02
Pre-mRNA cleavage complex 2 protein Pcf 11 (PCF11)	0.72	7.29E-02
Bifunctional methylenetetrahydrofolate dehydrogenase/cyclohydrolase (MTHFD2)	0.68	2.75E-03
Procollagen-lysine,2-oxoglutarate 5-dioxygenase 3 (PLOD3)	0.63	9.05E-02
LRPAP1 (Alpha-2-macroglobulin receptor-associated protein)	0.63	7.86E-02
Signal recognition particle receptor subunit beta (SRPRB)	0.57	3.86E-02
Paired amphipathic helix protein Sin3a (SIN3A)	0.56	1.43E-02
Emerin (EMD)	0.53	1.43E-02
Epidermal growth factor receptor pathway substrate 8 (EPS8)	0.52	1.43E-02
Nitric oxide synthase interacting protein (NOSIP)	0.51	5.55E-02
RNA binding motif protein 26 (RBM26)	0.51	1.60E-02
Procollagen-lysine,2-oxoglutarate 5-dioxygenase 1 (PLOD1)	0.48	6.44E-02
DnaJ homolog subfamily C member 8 (DNJC8)	0.47	9.61E-04
Flap endonuclease 1 (FEN1)	0.44	5.12E-02
Uveal autoantigen with coiled-coil domains and ankyrin repeats (UACA)	0.44	9.93E-03

Table 11: Specifically enriched (> 0.4 log_2-fold-change fUDNA1/ GCAT1-DNA control) 5-fU DNA-binding proteins with their Log_2Fold change and their associated p-value.

5.3. Functional Analysis of 5-fU enriched Proteins

To determine the role of proteins enriched by 5-fU, functional annotation analysis of 5-fU binding proteins was performed using DAVID using the UP_keywords function.[221,222,235]

Function	Number	Proteins	p-value
Acetylation	28	XRN, DDX1, PCF11, PBM2, SIN3A, SAP130, C14orf166, EMD, FAM50A, FEN1, GTF2H1, GAPDH, HDLBP, INTS3, MED6, MDN1, MTHFD2, MRPS23, PRIM1, SND1, TCOF1, UBAP2L, UNG, UACA	2.80E-10
Nucleus	25	XRN1, DDX1, PCF11, RTCB, RBM4, SIN3A, SAP130, TARDBP, YY1, CEP350, C14ORF166, EMD, FAM50A, FEN1, GTF2H1, GAPDH, HDLBP, INTS3, MED6, MND1, NOSIP, SND1, TCOF1, UNG, UACA	2.40E-07
Phosphoprotein	26	XRN2, DDX1, LRPAP1, PCF11, RTCB, RBM26, SIN3A, SRPBP, SAP130, SIN3A, TARDBP, YY1, CEP350, EMD, EPS8, FEN1, GTF2H1, GAPDH, HDLBP, INTS3, NOSIP, PRIM2, SND1, TCOF1, UBAP2L, UNG	3.40E-04
DNA repair	5	YY1, FEN1, GTF2H1, INTS3, UNG	1.70E-03
Transcription	12	XRN2, DDX1, SIN3A, SAP130, TARDBP, YY1, C14orf166, GTF2H1, MED6, PRIM1, PRIM2, SND1	2.00E-03
Ubi conjugation	10	PCF11, RBM26, SIN3A, SAP130, TARDBP, YY1, EPS8, GAPDH, TCOF1, UACA	2.20E-03
mRNA processing	5	XRN2, DDX1, PCF11, RBM4, TARDBP	2.60E-03
DNA damage	5	YY1, FEN1, GTF2H1, INTS3, UNG	3.20E-03
Primosome	2	PRIM1, PRIM2	3.50E-03
Exonuclease	3	XRN2, DDX1, FEN1	3.60E-03
RNA-binding	5	DDX1, RBM26, RBM4, TARDBP, C14orf166, HDLBP	5.70E-03
Methylation	7	XRN2, PCF11, SAP130, TARDBP, FEN1, GAPDH, UBAP2L	7.4E-03
DNA replication	4	FEN1, PRIM1, PRIM2	1.20E-02
Isopeptide bond	3	PCF11, RBM26, SAP130, TARDBP, YY1, TCOF1, UACA	1.30E-02
Transcriptional regulation	10	XRN2, DDX1, SIN3A, SAP130, TARDBP, YY1, C14orf166, GTF2H1, MED6, SND1	1.70E-02
Nuclease	3	FEN1, XRN2, DDX1	2.30E-02
Coiled coil	2	LRPAP1, PCF11, RBM26, SIN3A, CEP350, FAM50A, FEN1, HDLBP, MDN1, TCOF1, UACA	3.20E-02
Vitamin C	2	PLOD1, PLOD3	3.40E-02
DNA directed RNA polymerease	2	PRIM1, PRIM2	6.60E-02
Oxidoreductase	4	GAPDH, MTHFD2, PLOD1, PLOD3	8.10E-02
Repressor	4	SIN3A, SAP130, YY1, TARDBP	8.40E-02

Table 12: 5-fU enriched proteins corresponding to different biological functions and their associated p-value of enrichment in annotation categories, analysed using DAVID functional analysis.

Furthermore, functional protein association networks were identified *via* the STRING algorithm[236]. This determined protein clusters within the 5-fU protein interactome (Figure 69); 5-fU enriched proteins were primarily associated in two main clusters.

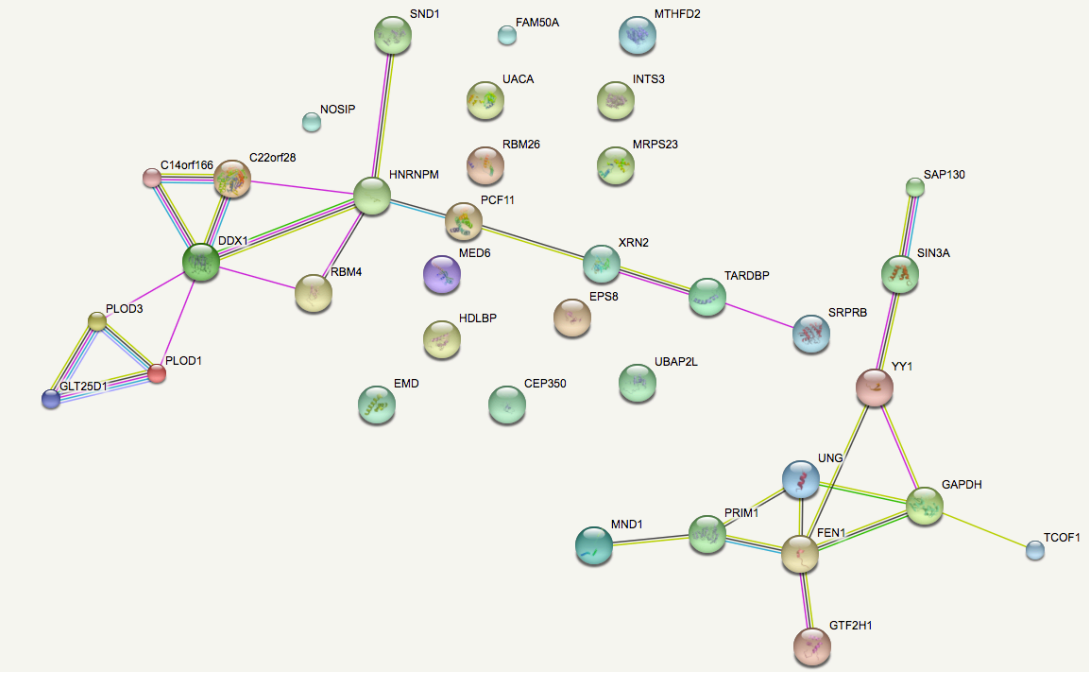


Figure 69: Associated protein networks of 5-fU enriched proteins from STRING.[236]

Consistent with 5-fU being a product of oxidative DNA damage, multiple proteins were revealed to be associated with DNA damage and DNA repair. The YY1 transcription factor is reported to be recruited to sites of laser-induced DNA damage,[237] while FEN is an important protein involved in removing overhangs after long-patch BER.[31] In addition to those highlighted by functional analysis, GAPDH is reported to bind to AP sites in DNA, and is involved in the recruitment and activation of APE1 which cleaves AP sites after BER.[238] Many of the proteins enriched in the 5-fU DNA-bound pools were associated with stress response pathways. TCOF1 has been implicated in the oxidative stress response, where it is shown to have antioxidant ROS cytoprotective action,[239,240] whilst GPx reduces ROS by enzymatic reduction of hydrogen peroxide to water.[241] One of the largest networks involves the DNA-binder DDX1, which is rapidly redistributed in cells that are exposed to ionising radiation, and important for stress survival.[242,243] Furthermore, commonly enriched proteins were found between this 5-fU proteomics study and a study where cells had been subjected to oxidative stress *via* cellular steatosis (HDLBP, MTHFD2, RTCB, SND1).[244] These findings indicate that 5-fU may serve as a molecular marker for oxidative stress in cells.

Notably, a number of 5-fU enriched proteins also appear to be linked with transcription and transcriptional regulation, including several transcription factors and mediators of RNA polymerases [245,246,247]. SAP130 has literature precedent for binding preferentially to UV-damaged DNA.[248] This protein associates with SIN3a, which functions as a transcriptional co-repressor in association with histone deacetylases (HDAC) and has roles in chromatin remodelling.[249] TARDBP also functions as a transcriptional repressor, repressing *acrv1* gene expression during spermatogenesis in mouse.[250] As a result, this highlighted that 5-fU may also have an influence on gene expression.

5.4. 5-fU Suppressed Proteins

A subset of proteins were instead preferentially depleted in fU DNA-bound pools (Table 13). Impeded binding of several transcription factors, including FOXC1 and ELF2, was observed, the former being important for cardiovascular development.[251] Furthermore, DNA-binding protein DNA-methyltransferase related protein 1 (DMAP1) was suppressed. This protein is implicated in epigenetic regulation and associates with both DNMT1 and HDAC, leading to transcriptionally inactivated genes after replication.[252] Several chromatin remodellers were also suppressed: CHAF1B, integral for nucleosome assembly[253]; ASH2 histone methyltransferase subunit, responsible for methylation of H3K4[254]; and SAGA complex associated factor 29 (SGF29).[255] The reduced binding of these proteins suggests that the presence of 5-fU may impair the normal epigenetic processing of DNA.

Protein	Log2Fold Enrichment GCATDNA1/fUDNA1	p-value
Transcriptional repressor p66-beta (GATAD2B)	3.95	5.52E-02
Phenylalanyl-tRNA synthetase alpha chain (FARSA)	3.87	5.45E-04
Poly [ADP-ribose] polymerase 1 (PARP-1)	3.74	1.44E-02
Poly(A) RNA polymerase (MTPAP)	3.27	7.57E-02
SAGA Complex Associated Factor 29 (SGF29)	3.25	2.82E-06
Methyltransferase-like protein 13 (METTL13)	3.22	8.27E-02
DNA methyltransferase 1-associated protein 1 (DMAP1)	3.14	7.73E-03
ADP Ribosylation Factor GTPase Activating Protein 2 (ARFGAP2)	2.80	9.13E-04
Dedicator Of Cytokinesis 1 (DOCK1)	2.69	5.16E-02
Engulfment And Cell Motility 2 (ELMO2)	2.69	8.41E-02
Leucine-rich repeat and WD repeat-containing protein 1 (LRWD1)	2.69	8.41E-02
Cysteine-rich protein 2-binding protein (KAT14)	2.69	8.41E-02
Ribosomal Protein L23 (RPL23)	2.65	2.54E-02
Nucleus Accumbens Associated 1 (NACC1)	2.65	2.54E-02
Metastasis associated 1 family member 2 (MTA2)	2.53	4.20E-02
Tho complex 1 (THOC1)	2.53	4.20E-02

Chromatin assembly factor 1 subunit B (CHAF1B)	2.53	4.20E-02
BUD31 homolog (BUD31)	2.53	4.20E-02
Migration-inducing gene 14	2.48	9.57E-02
Pyruvate dehydrogenase (acetyl-transferring)] kinase isozyme 3 (PDK3)	2.48	9.57E-02
Forkhead box protein C1 (FOXC1)	2.35	6.52E-03
NADH dehydrogenase [ubiquinone] flavoprotein 1 (NDUFV1)	2.35	6.52E-03
60S ribosomal protein L22 (RPL22)	2.35	6.52E-03
Protein phosphatase 2C 56 (ABI1)	2.14	1.01E-02
Replication factor C subunit 5 (RFC5)	2.14	1.01E-02
PCNA-associated factor (PAF)	2.14	1.01E-02
Heat shock 70 kDa protein 14 (HSPA14)	2.14	1.01E-02
Large neutral amino acids transporter small subunit 1 (SLC7A5)	2.14	1.01E-02
Nuclear receptor coactivator 6 (NCOA6)	2.14	1.01E-02
Talin-1 (TLN1)	2.14	1.01E-02
Protein lin-9 homolog (LIN9)	2.14	1.01E-02
Rho GTPase-activating protein 5 (ARHGAP5)	2.02	6.22E-02
Microtubule-associated protein RP/EB family member 2 (MAPRE2)	2.02	5.24E-02
ETS-related transcription factor Elf-2 (ELF2)	1.93	5.29E-02
Nucleolar and spindle-associated protein 1 (NUSAP1)	1.93	5.29E-02
Interferon regulatory factor 2-binding protein 2 (IRF2BP2)	1.89	6.75E-02
Density-regulated protein (DENR)	1.25	7.12E-03
NADH dehydrogenase [ubiquinone] flavoprotein 3 (NDUF3V)	1.24	3.14E-02
Epididymis tissue sperm binding protein Li 14m	1.02	1.66E-02
Coatomer subunit delta (ARCN1)	0.98	6.34E-02
REST corepressor 1 (RCOR1)	0.95	1.32E-02
Delta-1-pyrroline-5-carboxylate synthase (ALDH18A1)	0.87	9.42E-02
ATP synthase subunit alpha (ATP5A1)	0.86	9.60E-02
CREB/ATF bZIP transcription factor (CREBZF)	0.84	7.83E-02
AFG3-like protein 2 (AFG3L2)	0.73	9.03E-02
NADH-ubiquinone oxidoreductase 75 kDa subunit (NDUFS1)	0.71	2.84E-02
ELM2 and SANT domain-containing protein 1 (ELMSAN1)	0.69	2.87E-02
Set1/Ash2 histone methyltransferase complex subunit ASH2 (ASH2L)	0.68	5.82E-02
Pleiotropic regulator 1 (PLRG1)	0.66	4.08E-02
Neural Wiskott-Aldrich syndrome protein (WASL)	0.48	1.79E-02

Table 13: Specifically enriched (> 0.4 log_2-fold-change GCAT1-DNA control/fU-DNA1) 5-fU DNA-binding proteins with their log_2fold change and associated p-value.

5.5. UNG as a 5-fU Binder

In this study, fU-DNA baits led to enrichment of the DNA glycosylase, uracil deglycosylase (UNG). This protein is typically known for the selective excision and repair of U in DNA;[256] yet, a previous study demonstrated a potential correlation between the pKa of the N(3)H uracil derivative and UNG binding affinity using a UNG structural mimic.[257] The reduced pKa of 5-fU (due to its electron withdrawing formyl group) was rationalised to have tighter UNG binding over unmodified DNA.

Since UNG possesses glycosylase capacity, the capability of UNG to excise 5-fU was assessed. fU-ODN, a 10mer containing one 5-fU modification, was incubated with *E. coli* UNG, which shares the same catalytic domain (200 amino acid residues) as the human UNG isoform.[256] A ~12% conversion to AP-site DNA was observed, indicating UNG has a slight capacity to excise 5-fU (Figure 70), although large quantities of enzyme were required (50 U).

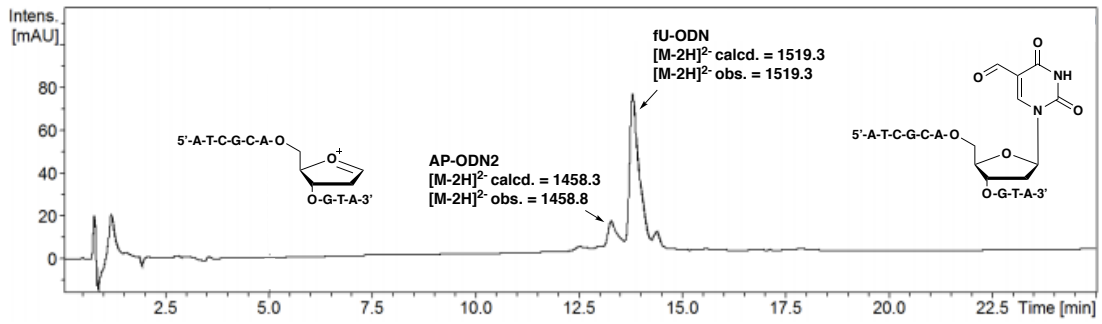


Figure 70: UNG incubation with fU-ODN led to formation of AP-site containing ODN (AP-ODN2). % conversion was calculated by integration of peak at 260 nm between starting fU-ODN and AP-ODN2

The mammalian 5-fU glycosylase SMUG1, considered as the main 5-fU glycosylase,[62] was not identified in this 5-fU proteomics study. However, the level of SMUG1 abundance is low in HEK293 cells[258] (~ 5 fold lower than UNG2) and a large drawback of classical global proteomics is the lack of sensitivity for low abundance proteins in the presence of those that are much more abundant.[259] Since the rate of excision is found to be much slower and less efficient than that observed with hSMUG1 protein *in vitro*, hSMUG1 is still likely to be the dominant 5-fU glycosylase in humans.

UNG's mode of action involves scanning DNA and forming a pseudo-base pair with uracil/thymine-derived bases as they partially emerge from the DNA double helix.[260] This interaction leads to the DNA base being flipped out, and whilst U can be incorporated into the active site and subsequently excised, the methyl group of thymine

sterically clashes with a tyrosine residue (Y66) in the active site, allowing discrimination between these two bases.[261] Thus, UNG has limited tolerance for uracil bases with substituents at the 5' position; indeed, UNG has also been shown to have no excision activity against 5-chloro, 5-bromo or 5-iodo-uracil,[261,262,263] yet, shows limited excision activity towards the smaller 5-fluorouracil.[264] Observation of the UNG protein co-crystallised with U (Figure 71) suggests that the planar formyl group of 5-fU may facilitate its accommodation into the UNG active site, enabling its limited excision.

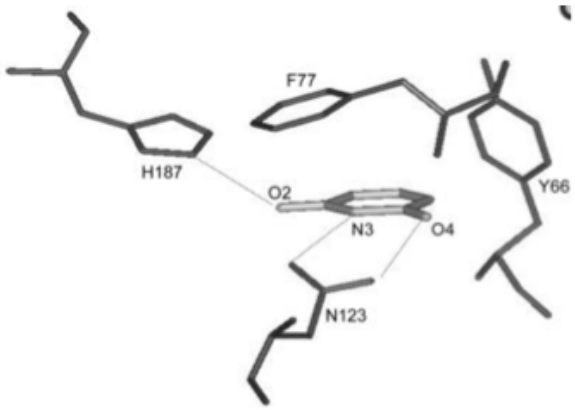


Figure 71: Crystal structure of uracil co-crystallised with UNG. Steric clash with Y66 leads to selectivity for U incorporation over T, where the central molecule represents the U base.[261]

5.6. Conclusion

Mass spectrometry-based proteomics has identified a number of key proteins that are consistently enriched in the presence of 5-fU compared to a non-modified control. A large number of these proteins appear to be involved in the DNA damage response, consistent with the current understanding that 5-fU is caused by ROS-activated oxidation of T. Furthermore, a number of proteins are also involved in redox regulation and antioxidant activity, suggesting that 5-fU could potentially act as a marker for oxidative stress.

A subset of proteins linked with gene regulation were also enriched or suppressed in the presence of the 5-fU modification, suggesting the mark could have an effect on gene expression. Furthermore, UNG is identified as a 5-fU binder, and subsequent follow-up study shows UNG has a slight capacity to excise or 'erase' 5-fU from DNA.

Follow-up work should explore 5-fU protein pulldown using alternative sequence baits to determine if 5-fU-protein recruitment is sequence-specific. In addition, proteomic baits using less heavily-modified 5-fU may be more biologically relevant, reflecting the low abundance of 5-fU in mammalian DNA (Section 2.6.1.ii). To improve future experimental design, the addition of competitor non-modified DNA would allow more stringent selection of 5-fU preferential binders. Furthermore, the use of stable isotope labelling with amino acids in cell culture (SILAC) nuclear extracts is likely to increase the reliability of quantitative measurements, particularly for low-abundance proteins.

In this current study, it is not clear which proteins are direct binders to 5-fU, or which are enriched as part of a wider complex. ELISA studies or EMSA-assays with native proteins would definitively identify direct 5-fU binders, and the strength of this interaction could be subsequently assessed.

6. Probing interactions between formylated DNA bases and histone proteins

6.1. Introduction

Nucleosomes are the fundamental units of chromatin and control access to genetic information. Nucleosomal organisation is influenced by a number of factors, including DNA sequence and chromatin remodelling proteins,[265,266] and is a key determinant of gene expression. Regulatory regions depleted of nucleosomes, and those marked with particular histone modifications (e.g. lysine acetylation and methylation), are correlated with transcriptionally active genes.

Certain dinucleotides are shown to favour interactions with the histone core (e.g. CG) and hence facilitate the bending and rotational positioning of DNA within the nucleosome, which occurs with ~10bp periodicity.[267] The presence of 5-mC, within these nucleotides, is shown to alter nucleosomal stability.[154] In addition, there have been numerous reports assessing how 5-mC and 5-hmC alter nucleosomal structure and positioning (Introduction, 1.5.4). However, there has been less exploration into the effect of formylated nucleobases on chromatin architecture.

Proteomics studies have shown that formylated bases can influence DNA-protein binding (Chapter 5); 5-fC, in particular, is found to specifically recruit a number of chromatin remodelling proteins.[49] Furthermore, genome-wide maps of 5-fC show this mark is correlated with tissue-specific active enhancers, lysine acetylation and lysine monomethylation,[53] raising questions about its role in chromatin biology. Both 5-fU and 5-fC are shown to alter the physical properties of DNA,[150,268] and hence have the propensity to influence protein recognition and biological function. A novel DNA structure in the presence of multiple 5-fC modifications, termed F-DNA, was solved by X-ray crystallography[269]; although the relevance of this structure has recently been challenged by Brown and co-workers.[270] In relation to nucleosome structure, an *in vitro* study by Ngo *et al.,* using FRET and optical tweezers, demonstrated that 5-fC increased the flexibility of DNA, which in turn enhanced nucleosomal stability compared to non-modified C.[150]

Thus, this chapter will describe studies aimed to determine the influence of formylated modified bases on nucleosomal structure. A particular focus was to explore whether

direct Schiff base formation could occur between formylated nucleobases and histone protein side-chains, which may contribute to nucleosomal stability and positioning.

6.1.1. Effect of Formylated DNA on Nucleosomal Occupancy *in vitro* and *in vivo*

In vivo and in vitro studies performed by Dr E.A. Raiber or Z.Li

Work within the Balasubramanian group had demonstrated that DNA containing both 5-fC (heavily-modified and ~1% modified) and 5-fU (heavily-modfiied) increased nucleosome occupancy *in vitro (unpublished work).* Nucleosome occupancy was assessed by determining the ratio of Cy5-labelled DNA between nucleosomal and unbound DNA fractions after nucleosome assembly (Figure 72 – A and B). A further study, using salt titration to dissociate the nucleosome, confirmed that 5-fC DNA (~1% modified) increased nucleosomal stability ($c_{½}$ = 531 ± 1.0 mM) compared to a non-modified C control ($c_{½}$ = 480 ± 1.1 mM) (Figure 72 - C).

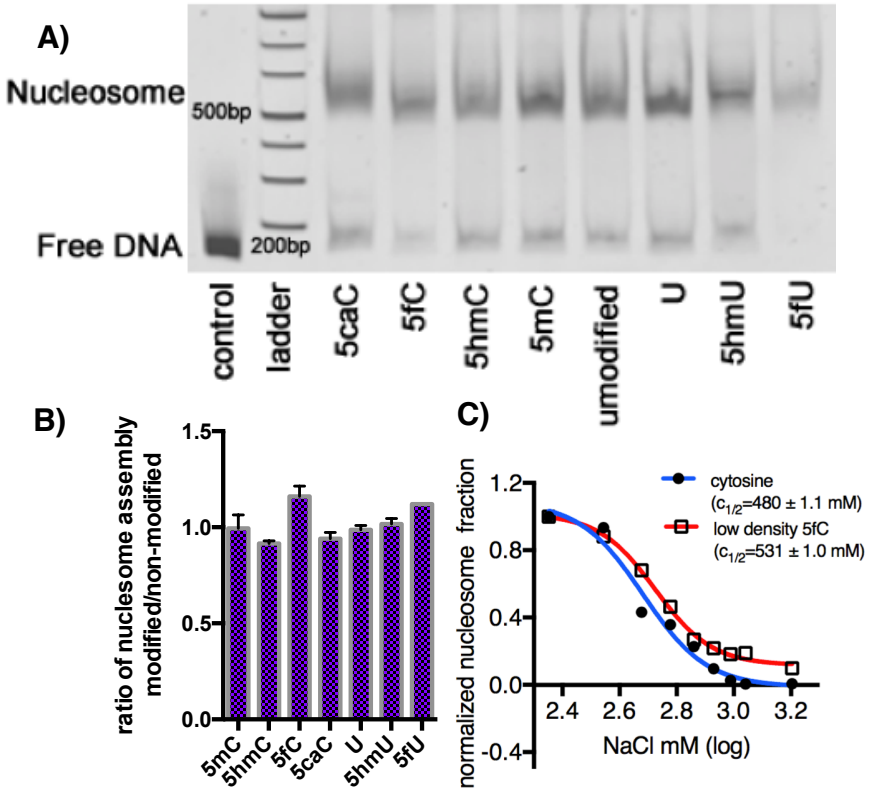


Figure 72: A) Gel demonstrating the nucleosome assembly of fluorescently-labelled modified DNA with histone proteins in the presence of chaperones. B) Relative nucleosomal occupancy determined by the relative ratio of DNA (by Cy5 fluorescent signal) in nucleosomal and free DNA fractions. 5-fC and 5-fU DNA markedly increase nucleosome occupancy compared to other modifications *in vitro*. C) Salt titration, where salt dissociates the nucleosome, was used to assess the stability of nucleosomes in the presence of 5-fC. 5-fC DNA increased the stability of the nucleosome with respect to a non-modified control. Low density refers to DNA that is ~1% modified.

A nucleosome reconstitution experiment, performed by E.A Raiber, further demonstrated that the presence of 5-fC influenced nucleosomal positioning *(unpublished work).* Mouse hindbrain DNA (E = 11.5 days), shown to be enriched in 5-fC,[53] was reconstituted with nucleosomes and treated with MNase to generate a map of nucleosomal occupancy. As a control, DNA was treated with sodium borohydride before nucleosome reconstitution;[125] this reagent reduces 5-fC to 5-hmC, and hence the effect of 5-fC on nucleosomal positioning could be assessed independent of sequence context. Nucleosomes were found to be enriched at 5-fC sites, but not in the reduced (5-hmC) control (Figure 73 – A). Next, to determine genome-wide nucleosomal organisation *in vivo*, MNase-seq (peformed by E.A. Raiber) was performed directly on mouse hindbrain tissue, where nucleosome positioning was correlated with genome-wide maps of 5-fC.[53] Consistent with the *in vitro* studies, higher average nucleosome occupancy at 5-fC sites was observed compared to other positioned nucleosomes. Furthermore, 5-fC containing CpGis demonstrated higher nucleosomal occupancy compared to non-fC containing CpGis (Figure 73 – B). This was of relevance as CpGis are usually associated with depletion of nucleosomes *in vivo*.[271,272]

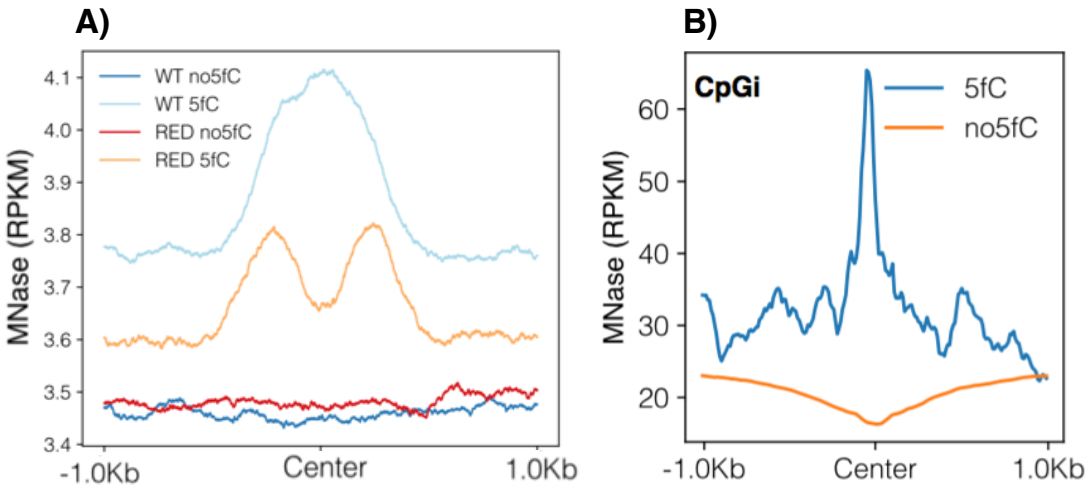


Figure 73: a) A nucleosome reconstitution experiment with mouse hindbrain DNA and subsequent MNase treatment and NGS demonstrated higher nucleosome occupancy (determined by MNase RPKM) at 5-fC sites (determined in TDG knockout hindbrain), and not in a $NaBH_4$ reduced control (5-hmC). b) MNase treatment of mouse hindbrain tissue revealed higher nucleosome occupancy (determined by MNase RPKM) within 5-fC-enriched regions at CpGi loci, compared to CpGi which do not contain 5-fC.

My role was to extend this study to more fully understand the molecular interactions between formylated bases and nucleosomes; this would provide a mechanism underpinning the enhanced nucleosomal occupancy at 5-fC-enriched loci observed *in vivo* and the increased nucleosomal formation observed with both 5-fC and 5-fU *in vitro*.

6.1.2. Basis for Probing Schiff Base Formation between Formylated DNA Nucleobases and Histone Proteins.

Since histone proteins are lysine-rich, including their N-terminal tails, it was reasoned that reversible Schiff-base interactions beween lysines and formylated bases (5-fC and 5-fU) could alter the positioning and stability of nucleosomal structure in DNA (Figure 74). It had already been demonstrated that the formyl groups in both 5-fU and 5-fC could readily form Schiff bases with various nitrogen-nucleophiles present in excess (Chapter 3). Furthermore, 5-fU had been shown to form a Schiff base with 5-aminocytosine as a non-natural base-pair,[273] suggesting the formyl group can interact with amines in close proximity. Therefore, interactions between formylated bases and lysine residues were explored to assess the potential relevance of Schiff base formation within the nucleosome.

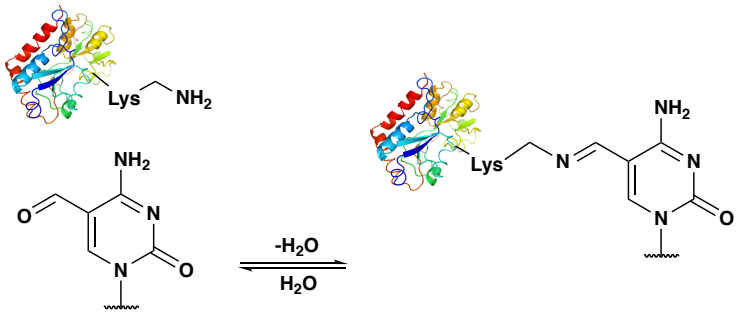


Figure 74: Proposed Schiff base formation between formylated bases (demonstrated here by 5-fC) and lysine side-chains of histone proteins may contribute to increased formylated-DNA nucleosomal occupancy and stability.

6.2. Schiff Base Formation between Formyl Groups and Lysine

Schiff-bases can be covalently trapped by sodium cyanoborohydride ($NaBH_3CN$) reduction; this reagent is selective for imine reduction over aldehyde reduction under neutral conditions.[274] Imine reduction by sodium cyanoborohydride has been routinely applied in chemical biology to probe key lysine residues involved in enzymatic reactions.[275] In addition, crosslinking between aldehyde-bearing DNA and proteins, including for AP sites and DNA damage products, had been demonstrated previously suggesting proof of principle for this work.[276,277,278,279] Selective reduction could thus be used as a tool to probe potential associations between 5-fU/5-fC and proximal lysine residues.

6.2.1 Reactivity of Lysine with 5-fU and 5-fC-bearing ODNs

Reactivity of lysine with 5-fC and 5-fU was assessed using ODNs bearing one formylated base (fU-ODN or fC-ODN). In the presence of lysine (10 mM) and sodium cyanoborohydride (25 mM), trapped Schiff base adducts with fU-ODN were observed via LC-MS (Figure 75).

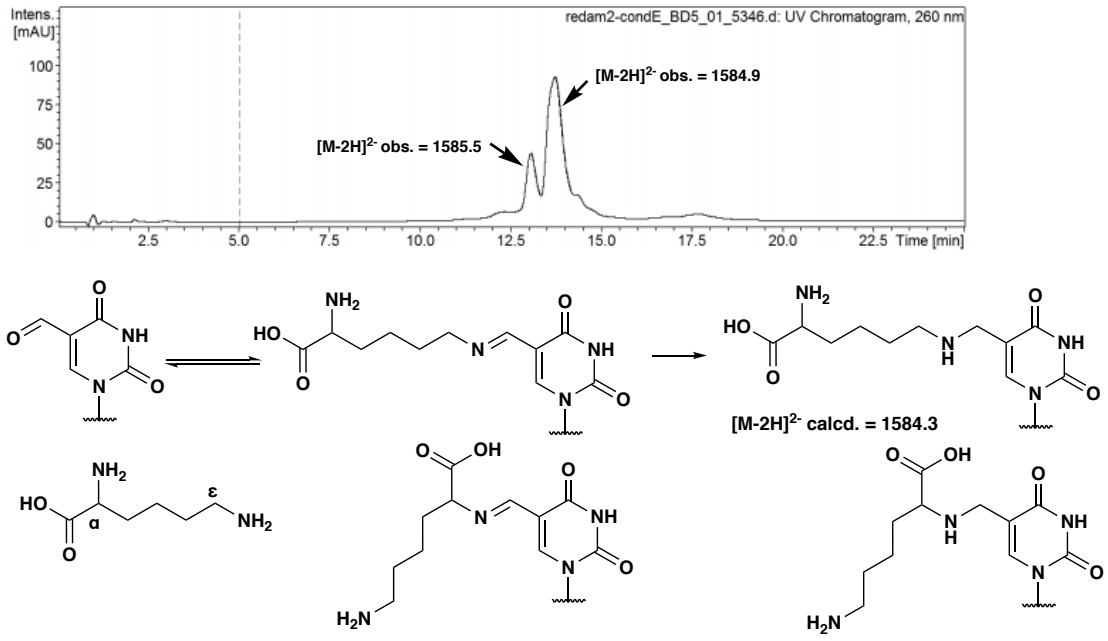


Figure 75: LC-MS trace showing sodium cyanoborohydride reduction of the Schiff base adduct formed between lysine and fU-ODN. Two products were observed demonstrating Schiff base formation with either the α- or ε- amine of lysine (bottom). Reactivity of fU-ODN was confirmed with both 5-aminovaleric acid and glycine as models for both α- or ε- amines (Chapter 6 Appendix).

Using the same crosslinking conditions as those used for fU-ODN, no crosslink was observed for 5-fC. This demonstrates the reduced electrophilicity of 5-fC and its adducts compared with 5-fU, as discussed previously in this thesis (Chapter 3). Further optimisation revealed that lysine-5-fC crosslinking was possible in the presence of a much more concentrated solution of lysine (500 mM) and with heating to 37 °C (Figure 76 - top). At this concentration of lysine, Schiff base formation without reduction was also observable (Figure 76 – bottom), however, subsequent purification of the oligomer from lysine led to re-equilibriation and recovery of fC-ODN, demonstrating the reversibility of Schiff base formation.

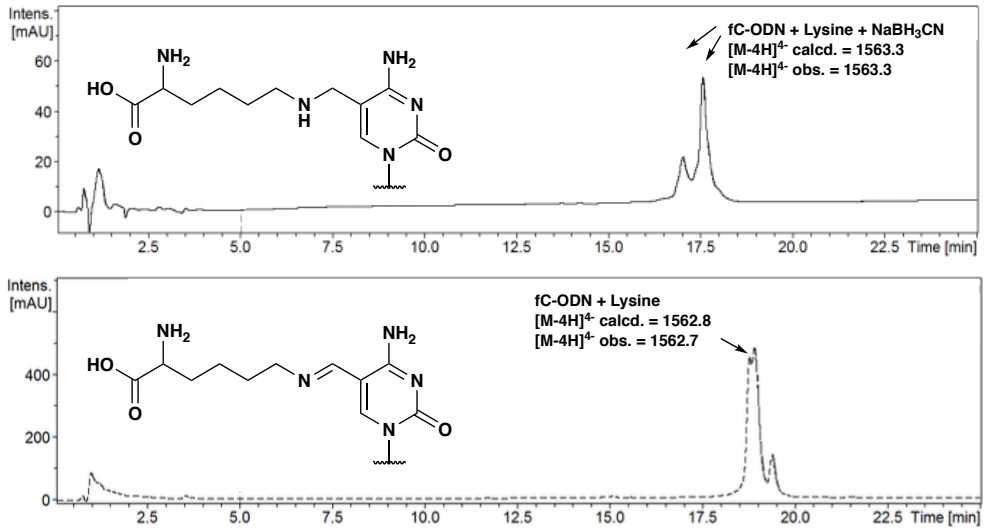


Figure 76: Top: fC-ODN crosslinked with lysine after incubation with sodium cyanoborohydride, Bottom: fC-ODN Schiff base interaction with lysine in the absence of reducing agent. Two products were observed demonstrating Schiff base formation with either the α- or ε- amine of lysine. Reactivity of fC-ODN was confirmed with both 5-aminovaleric acid and glycine as models for both α- or ε- amines.

6.3. Crosslinking with Model Proteins

Having demonstrated the feasibility of crosslinking DNA bearing 5-fU or 5-fC with lysine, the applicability of crosslinking DNA to proteins was next investigated; this had already been demonstrated for aldehyde-bearing DNA containing the damaged base 7-deaza-7-(2,3-dihydroxypropyl)-guanine.[279] Thus, the model proteins Bovine Serum Albumin (BSA, 59 lysines) and Ribonuclease A (Ribo A, 8 lysines) were incubated with fU, fC or non-modified DNA (fU-DNA1, fC-DNA1, GCAT-DNA1) in the presence of reducing agent. After incubation, analysis by gel electrophoresis demonstrated shifted gel-bands for fU-DNA1 corresponding to DNA-protein crosslinks with both model proteins, BSA (MW = 66.5 kDa) and Ribo A (MW = 13.7 kDa), whilst no shifted band was observed for the non-formylated GCAT-DNA1 control (Figure 77). Crosslinking of fU-ODN with Ribo A was also demonstrated by mass spectroscopy (Figure 78).

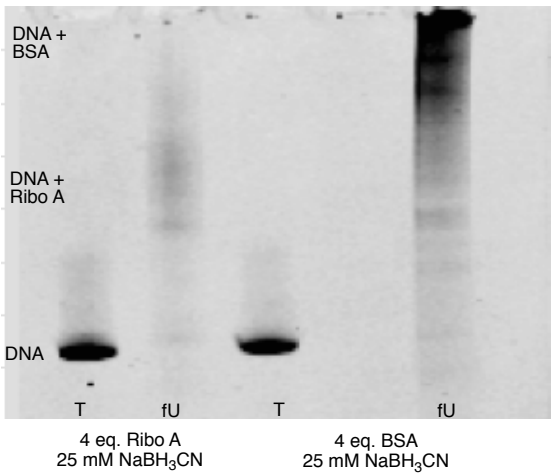


Figure 77: Incubation of fU-DNA1 and model proteins in the presence of sodium cyanoborohydride led to fU-DNA1 shifted gel-bands demonstrate crosslinking with Ribo A and BSA model proteins. 12% SDS gel ran in MES buffer.

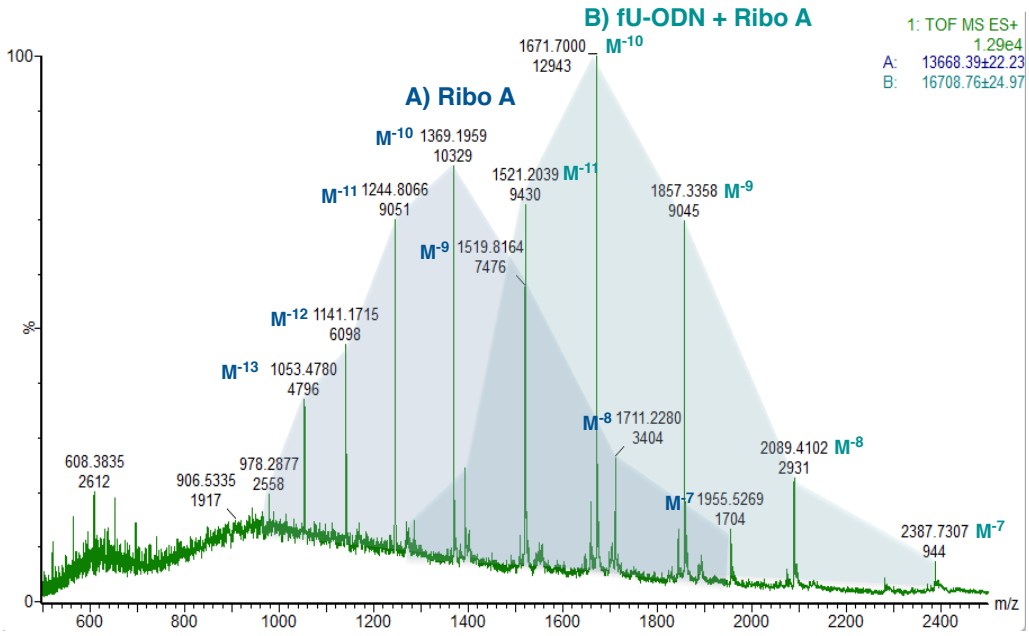


Figure 78: Mass spectrum demonstrating covalent crosslink (MW = 16.71 kDa) between fU-ODN (MW = 3.04 kDa) and Ribo A (13.67 kDa).

In contrast, no 5-fC crosslink was observed with the model proteins, either by gel electrophoresis or mass spectroscopy; this was even with the increased concentration of reducing agent necessary for crosslinking on the ODN model. A lack of reactivity with fC-DNA1 demonstrated that reduction of 5-fC Schiff-base adducts may only be possible with interacting proteins in close proximity, as in the nucleosome. Thus, Schiff base formation was instead probed using an *in vitro* nucleosome model for further exploration.

6.4. Crosslinking in the Nucleosome

Nucleosome assembly was performed by Z. Li or Dr E.A. Raiber.

Nucleosomes were assembled either via chaperone or salt dilution using the Widom DNA sequence, selected for its high affinity towards histone proteins.[280] Formylated bases (5-fU and 5-fC) replaced non-canonical bases (T and C) in non-primer regions of the Widom sequence (leading to 55 and 70 formylated bases respectively). Nucleosomes were incubated with sodium cyanoborohydride (100 mM) at 37 °C for 18 hr. The reducing agent was removed using size-exclusion chromatography and the resulting mixture was ran on a denaturing protein gel and imaged for subsequent analysis.

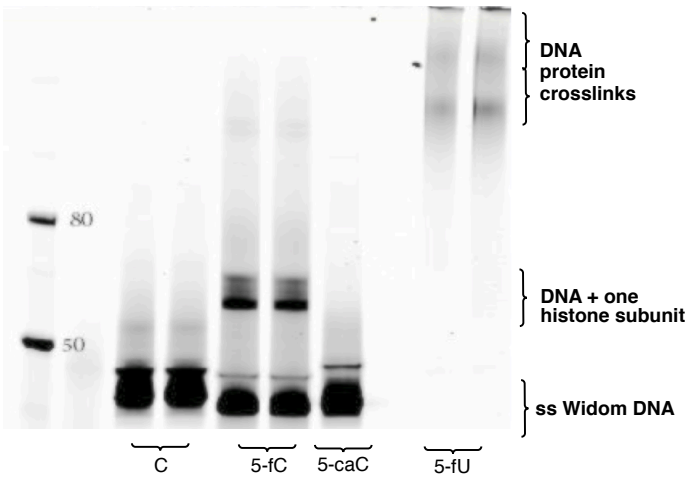


Figure 79: 12% SDS gel ran in MES buffer after incubation of modified-DNA containing nucleosomes in the presence of 100mM sodium cyanoborohydride for 18 hr. Crosslinking of 5-fC is observed corresponding to DNA + one histone subunit. 5-fU shows shifted gel-bands > 100 kDa. No shifted gel-bands were observed for non-modified or 5-caC containing DNA controls.

A shifted gel-band was observed for labelled 5-fC-containing DNA corresponding to the size of the DNA (44 kDa) plus one histone subunit (15-20 kDa), indicating Schiff base formation within the nucleosome (Figure 79). No analogous band was seen for GCAT-Widom DNA or 5-caC-Widom DNA controls, which are incapable of forming Schiff bases with lysine. It was reasoned that reductive trapping with 5-fC was possible within the nucleosome due to their close proximity with histone protein lysine side-chains.

In contrast, 5-fU crosslinking led to the formation of higher molecular weight bands (Figure 79). Since the Widom DNA sequence contains more than one 5-fU per strand, the results suggested that 5-fU was either crosslinking to more than one histone subunit simultaneously, or was alternatively/in addition crosslinking to other high abundance proteins in the nucleosome assembly mixture such as BSA (as demonstrated by model

potein studies) or the Nucleosome Assembly Protein 1 (NAP1) chaperone. This demonstrated the higher reactivity of 5-fU adducts with sodium cyanoborohydride. Efforts to subject the 5-fU nucleosome to milder reduction conditions still showed shifted gel-bands corresponding to > 80 kDa under these conditions, indicating the promiscuity of 5-fU reactivity (Appendix – Chapter 6).

6.5. Probing Crosslinking via Proteomics Experiments

6.5.1. Analysis of Shifted Gel-bands via Proteomics

To provide conclusive evidence of crosslinking between formylated DNA and histone proteins within the nucleosome, proteomic mass-spectrometric analysis was utilised to confirm the presence of histone proteins within DNA-protein crosslinked bands. Bands were extracted and subjected to tryptic digest before proteomics analysis. For crosslinked 5-fC DNA, polypeptides corresponding to histone protein subunits were identified by mass spectrometry, providing conclusive evidence for crosslinking in the nucleosomal model. H2B and H4 subunits were seen in all 5-fC crosslinked samples and the H3 subunit was observed in 2/3 replicates (Table 14). This indicated that Schiff base formation occurs at multiple positions within the nucleosome. An analogous molecular-weight band corresponding to GCAT-Widom (which had been exposed to crosslinking conditions) was submitted as a control, where no histone proteins were detected.

Subunit	fC-Rep1	fC-Rep2	fC-Rep3	C	fU
H2A	✗	✗	✗	✗	✗
H2B	✓	✓	✓	✗	✗
H3	✓	✗	✓	✗	✗
H4	✓	✓	✓	✗	✗

Table 14: Peptides of histone subunits identified by mass-spectrometry based proteomics.

In addition, no histone subunits were identified by proteomics in the 5-fU DNA nucleosome shifted gel-band; this indicated that 5-fU DNA is instead crosslinking with other components of the nucleosome assembly mixture (e.g. Nucleosome Assembly Protein 1). However, it is possible that excessive DNA crosslinking to histone peptide fragments may have inhibited their identification.

6.5.2. Identifying Crosslinking Sites via Proteomics

It was envisaged that proteomics could be utilised to identify the exact sites of lysine crosslinking via a histone peptide mass-shift (Figure 80). These sites would be indicative of stabilising interactions with formylated-Widom DNA.

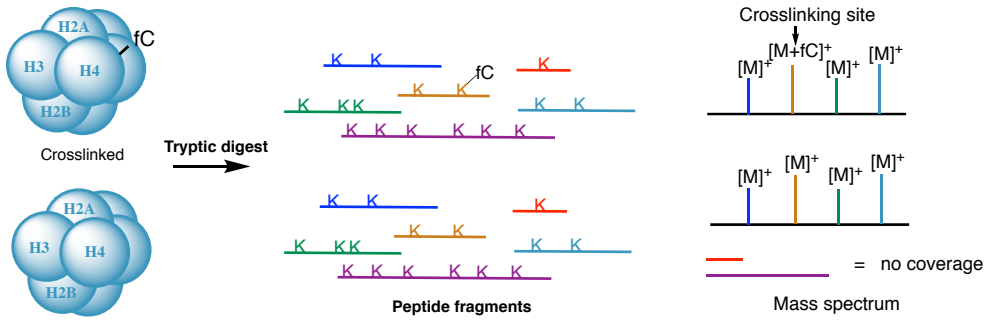


Figure 80: Schematic demonstrating tryptic digest of histone to different-sized peptide fragments. Site of crosslinking can be determined by a mass-shift of 5-fC. Some peptide units after proteolytic digest can be too small or large to be observed via mass spectrometry.

In order to determine sites of crosslinking via mass-shift, efficient peptide coverage of crosslinking sites (e.g. lysine) would be necessary to explore the extent and occurrence of Schiff base formation (Figure 80). To improve peptide coverage, peptides should be cut into chains 6-20 amino acids in length.[227] Fragments < 6 are too small to be unique for a particular protein in database searches, whereas fragments > 20 suffer from poor ionisation. Whilst 100% coverage is unlikely to ever be attainable, different proteolytic enzymes were screened; these enzymes cut at different amino acid residue restriction sites, and hence are likely to give a different distribution of peptide chain lengths and sequences. Trypsin, the most typically used protein digestion enzyme for proteomic study, cuts at the C-terminus between lysine and arginine, apart from in the instances where these amino acids follow a proline residue.[281] In contrast, chymotrypsin (high specificity) cuts at C-terminus tryptophan, tyrosine and phenylalanine, Asp-N cuts at N-terminus aspartic acid and glutamine, while Arg-C cuts at C-terminus arginine.[282,283]

	Trypsin %	Trypsin Lysine %	Chymotrypsin %	Chymotrypsin Lysine %	Arg-C %	Arg-C Lysine %	Asp-N %	Asp-N Lysine %
H2A	27	14	22	14	17	21	7	0
H2B	51	36	22	9	56	36	50	36
H3	33	31	26	8	62	69	45	31
H4	52	36	22	9	56	27	34	18

Table 15: Overall coverage and percentage lysine coverage for each histone subunit after treatment with different proteolytic enzymes.

Unfortunately, the overall percentage lysine coverage was poor irrespective of proteolytic enzyme used, where the best scenario reached 69% for the H3 subunit using Arg-C (Table 15). This approach was therefore not suitable to assess 5-fC-lysine histone interactions in an unbiased manner, as crucial interactions may not be detected in regions devoid of peptide coverage. As an alternative strategy, crosslinking sites could be determined by a polymerase stop-assay and subsequent NGS. Sites of polymerase stalling, caused by the presence of a bulky histone subunit crosslink, would thus be indicative of sites where Schiff base formation occurs.

6.6. Polymerase Stalling Assay and NGS to identify 5-fC crosslinking sites

Bioinformatic analysis of sequencing data was performed by Dr Sergio Martinez Cuesta, Molecular modelling of the nucleosome was performed by Dr Guillem Portella.

Polymerase stop assays have previously been used to determine sites of guanine alkylation damage[284] and DNA-protein conjugates.[285] These techniques can be coupled with DNA sequencing to show the exact site of DNA-polymerase stalling, an approach that has recently been used to determine the genomic location of DNA-cisplatin adducts.[286] This experiment, instead of the proteomics approach, would provide information about every potential 5-fC crosslinking site in an unbiased manner. Furthermore, any significant stalling sites can be computationally modelled to identify potentially stabilising interactions with proximal lysine residues within the nucleosome.[287]

A workflow for a polymerase extension experiment was designed in order to determine the sites of significant 5-fC-lysine interactions via crosslinking (Figure 81); free 5-fC-containing DNA in the absence of histone proteins was used as a control to account for natural polymerase-stalling events. 5-fC-containing nucleosomes were first exposed to the crosslinking conditions (100 mM sodium cyanoborohydride, 37 °C, 18 hr) before being purified from the reducing agent using size exclusion chromatography. The DNA was then denatured by heating and subjected to single primer extension, using forward and reverse primers of the Widom DNA sequence. Following extension, the mixture was treated with Proteinase K, to digest the crosslinked histone subunit. The mixture was purified and subsquently treated with the 5' → 3' $RecJ_f$ exonuclease to remove the resultant ssDNA overhang to enable efficient ligation of sequencing adaptors. After purification, DNA was subjected to the standard library preparation procedure for NGS (Section 3.7.1), and submitted for DNA sequencing.

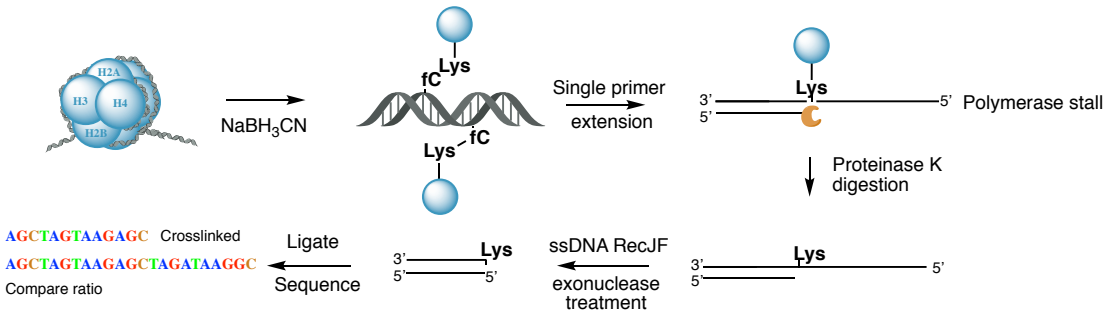


Figure 81: Design and workflow of polymerase stall assay after 5-fC nucleosome crosslinking.

To determine the extent of crosslinking, the ratio of truncated DNA sequences for the cross-linked nucleosome and the control sequence was calculated, normalising for the size of each library. Less full-extension product was observed in the sodium cyanoborohydride treated 5-fC nucleosome condition compared to the control in both forward (Rev – 6-fold) and reverse (Fw – 1.2-fold) primer extension conditions. This indicated greater polymerase stalling due to DNA-histone protein crosslinking, in line with expectations.

Next, the extent of stalling was assessed at each individual site on the Widom sequence by ratio of truncated sequences. A moving average (±1) of fold-change stalling was plotted against each position of the forward and reverse Widom template (Figure 82). A clear pattern of stalling emerged, with distinct periodicity. All stalling maxima had a 5-fC in the vicinity, either directly at the site of stalling, or at the nucleotides surrounding the crosslinking site. Greater stalling sites were indicative of greater Schiff base formation in the nucleosome.

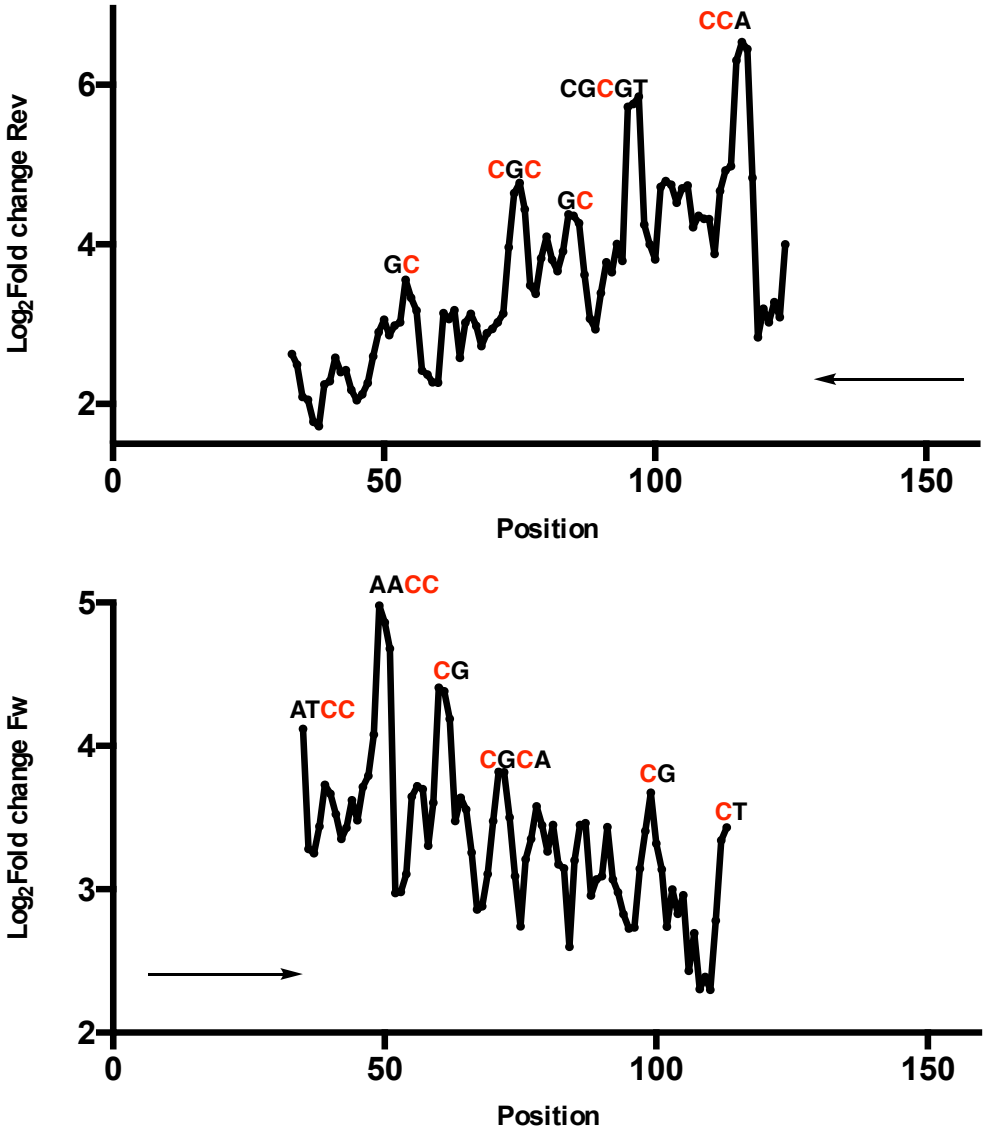


Figure 82: Log_2Fold-change in polymerase stalling between crosslinked nucleosomes and the free DNA control. The likely 5-fC crosslinked site(s) are highlighted in red. Arrow demonstrates the direction of polymerase.

To determine the significance of the periodic stalling sites, the polymerase data set was overlapped with a molecular dynamics simulation of the nucleosome. Notably, stalling sites were found to be periodically and almost symmetrically distributed either side of the nucleosome dyad, demonstrating ~10-12 bp periodicity. Significant stalling sites were more likely when 5-fC was facing the major groove towards the histone core (Figures 83 and 84). This correlated with molecular modelling data that determined likely nucleotide-lysine contact frequency.

In combination with modelling, the most likely 5-fC crosslinking sites were identified by identifying the most proximal lysine (Table 16). This indicated that crosslinking occurs with mainly H4, H3 and H2B histone subunits, in accordance with the initial proteomics data.

This study implies that specific and potentially stabilising interactions between 5-fC and lysine can exist in the nucleosome; these interactions provide a potential molecular mechanism for 5-fC-directed nucleosomal positioning and the increased nucleosomal occupancy associated with this mark.

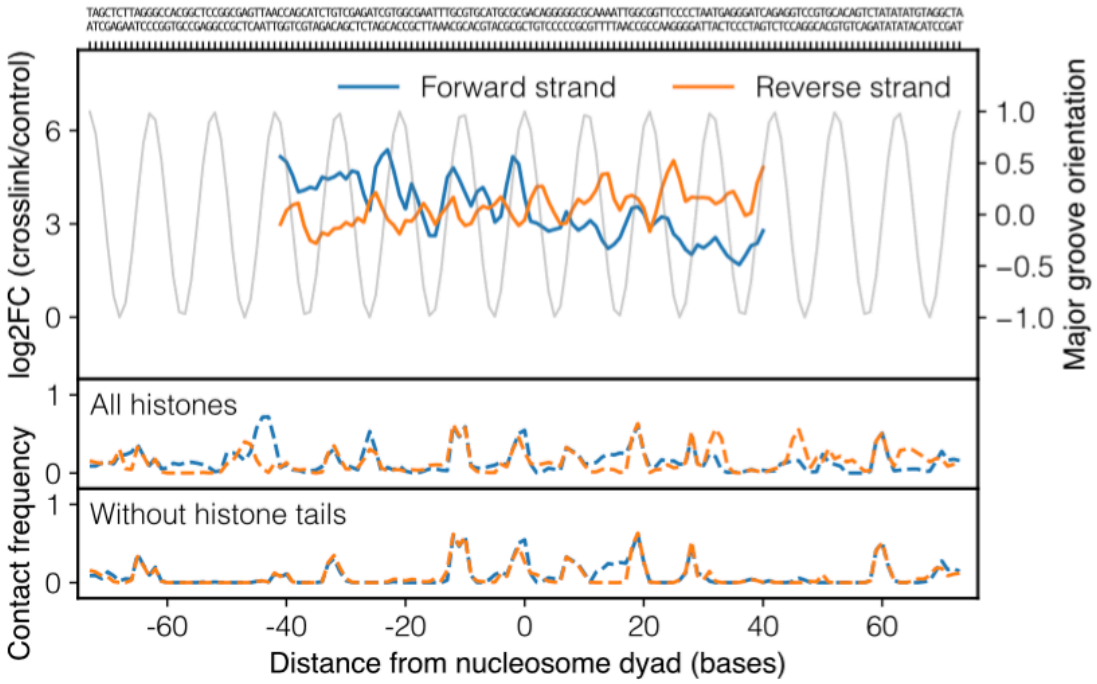


Figure 83: Upper panel: Polymerase stop data demonstrates the log_2fold change between the number of truncated reads at a given position compared to the number of reads in an untreated 5-fC sample, both for forward (blue) and reverse strand (orange). The grey line indicates the orientation of the major groove with respect to the histone core, 1 indicates core-facing and -1 indicates where 5-fC points away from the histone core.

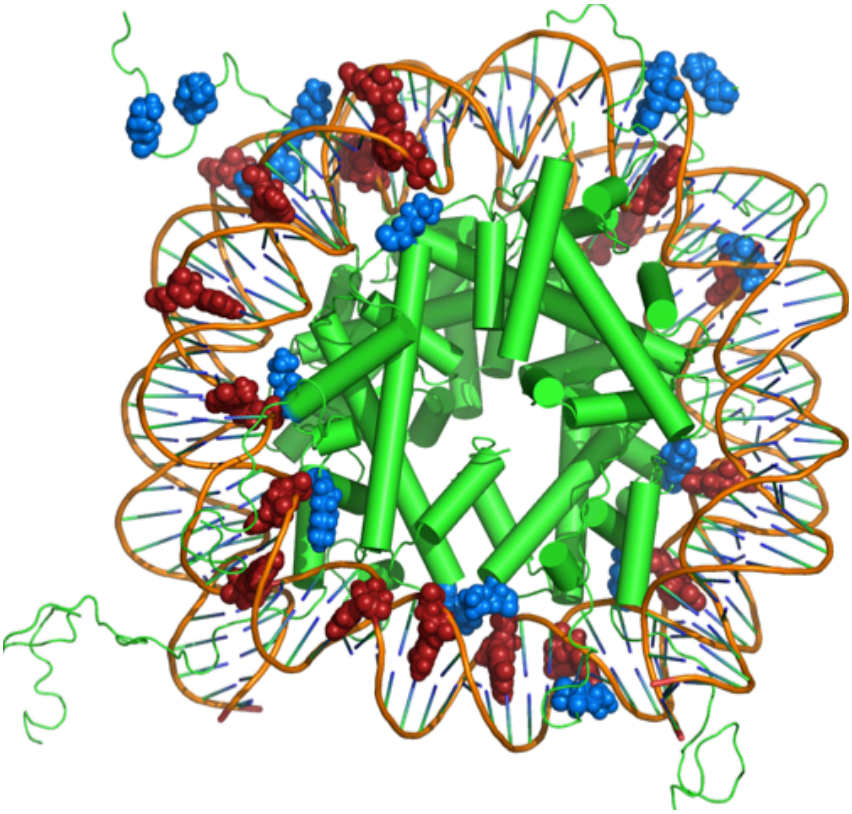


Figure 84: Model of nucleosome demonstrating 5-fC positions that indicated significant stalling (red spheres) and proximal lysine residues (blue spheres).

fC residue	Lys residue	Probability	Histone subunit
Fw (38)	2	0.01	H2B_1 Tail
Fw (52)	16	0.04	H4_2 Tail
Fw (64)	31	0.11	H4_2
Fw (76)	115	0.30	H3_2
Fw (92)	16	0.04	H4_1
Fw (112)	8	0.04	H2B_2_Tail
Rev (33)	36	0.02	H2A_1 Tail
Rev (50)	12	0.15	H2B_2 Tail
Rev (63)	23	0.12	H3_1
Rev (73)	115	0.11	H3_1
Rev (94)	16	0.04	H4_2

Table 16: Most likely Schiff-base sites between 5-fC and histone protein side-chains modelled using polymerase stop data and proximal lysine distances via molecular simulation. Table gives details of DNA Widom position; proximal lysine histone subunit and whether lysine is part of the core histone subunit or histone tail.

6.7. Conclusion

The identification of significant 5-fC interactions with histone proteins to form Schiff bases provides new mechanistic insight into the enhanced assembly and stability of 5-fC-DNA nucleosomes *in vitro*. This data provides support for the hypothesis that specific Schiff base formation between 5-fC and histone residues occurs preferentially at certain positions within the nucleosome, hence indicating that Schiff base formation could contribute to nucleosomal positioning and stability.

Furthermore, Schiff base formation between 5-fC and histone proteins provides a mechanistic hypothesis for the association of 5-fC and nucleosomes *in vivo*. It is now speculated that specific reversible Schiff base formation may determine the positioning and anchoring of nucleosomes at the 5-fC mark. This has wider implications for the role of this modified base with regard to chromatin architecture and gene expression.

At the time of writing, the ability of 5-fC to form Schiff base interactions with histone proteins was corroborated by two other studies. Li *et al.* investigated the effect of crosslinking using DNA models bearing a single 5-fC site at four different positions in the sequence.[288] Whilst a different DNA sequence was used, (hence sequence-specific interactions cannot be directly compared), greater crosslinking was also observed when the 5-fC site pointed towards the histone core, in agreement with the results in this chapter. In addition, Tretyakova and co-workers demonstrated the presence of 5-fC lysine adducts after sodium cyanoborohydride reduction in HEK293T cells; ~1/100 5-fC sites were estimated by LCMS/MS to form a Schiff base *in vivo*.[289]

Further work could validate significant stalling sites observed in the polymerase stalling assay by using Widom sequence DNA with single 5-fC sites at key positions. Structural analysis of the 5-fC-Widom nucleosome (e.g. X-ray crystallography), would provide conclusive evidence for 5-fC-lysine Schiff base formation in this model. Most importantly, it should now be determined whether lysine-5-fC interactions are functionally relevant in biology. A genome-wide polymerase-stalling assay, or primer extension at specific regions of interest, could be utilised to assess the genomic location of 5-fC-lysine Schiff bases. Alternatively, a modified MNase-seq experiment could be performed after crosslinking and nucleosomal dissociation, where an enrichment in MNase signal would determine where crosslinking sites arise.

5-fU also demonstrated the ability to form a Schiff base with lysine, and DNA-protein crosslinking was validated with model proteins. However, increased reactivity of the 5-fU lysine adduct, which lead to excessive crosslinking, made 5-fU nucleosomal interactions difficult to probe. To further investigate 5-fU and its association with nucleosomes 1) Widom-DNA containing single 5-fU sites at key positions could be used or 2) reduction conditions could be further optimised. Analogous polymerase stalling experiments, in the first instance, can then be performed using *in vitro* nucleosome models.

7. Materials and Methods

7.1. General

General chemistry remarks All solvents and reagents were purified by standard techniques reported in Armarego, W. L. F., Chai, C. L. L., Purification of Laboratory Chemicals, 5th edition, Elsevier, 2003; or used as supplied from commercial sources (Sigma Aldrich® unless stated otherwise). Thin layer chromatography (TLC) was performed on Merck Kieselgel 60 F254 plates, and spots were visualized under UV light. LC-MS was performed on an Amazon ESI-MS (Bruker) connected to Ultimate 3000 LC (Dionex) (Agilent Technologies, Santa Clara, CA). Flash chromatography was carried out using CombiFlash Rf (Teledyne Isco) with puriFlash columns (Interchim). NMR spectra were acquired on a Bruker® DRX-500 instrument using deuterated solvents as indicated and at ambient probe temperature (300 K). Notation for the NMR spectral splitting patterns includes: singlet (s), doublet (d), doublet of doublets (dd), triplet (t), doublet of triplets (dt) quartet (q), quinent (qn) multiplet/overlapping peaks (m). Signals are quoted as δ values in ppm, coupling constants (*J*) are quoted in Hz and approximated to the nearest 0.1. Data analysis for the NMR spectra was performed using MestReNova® software. HRMS were recorded on a Waters LCT Premier (ESI) spectrometer.

Temperature controlled reactions and qPCR Temperature controlled reactions and PCRs were performed either in a T100 Thermal cycler (Bio-Rad) or peqSTAR 96X Universal Gradient (Peqlab). qPCRs were carried out using a CFX96 Touch Real-Time PCR system (Bio-Rad) and data was analysed using CFX Software manager (BioRad).

Polymerase Chain Reaction for synthesis of modified DNA and non-modified DNA PCRs were made up with 4 x dNTP (1 mM), forward primer (1 μM), reverse primer (1 μM), template (0.01 μM), 10 × DreamTaq Buffer (2.5 μL) and DreamTaq Polymerase (Thermo Scientific, 0.25 μL) to give a final volume of 25 μL. Primers and templates were obtained from Sigma Aldrich or Invitrogen. dGTP, dCTP, dATP and dTTP were obtained from ThermoFischer, while modified dNTP (dfUTP, dfCTP, dhmCTP, dhmUTP) were obtained from Trilink Biotechnologies. The mixture was subjected to the following thermal cycle: 95 °C for 3 min, 40 cycles of (95 °C for 30 s, 60 °C for 60 s, 72 °C for 90 s), 72 °C for 5 min. The PCR products were purified using a GeneJET PCR Purification Kit (Thermo Scientific) according to the manufacturer's instructions. Formation of PCR products were confirmed by Tapestation 2200 (Agilent Technologies) using D1000

screen-tape. Confirmation of incorporated bases was confirmed by digestion of synthetic DNA and injection into a Q-Exactive MS spectrometer (Thermo Fischer).

Sonication of DNA Genomic DNA was sheared by sonication with the Covaris M220 Focused-ultrasonicator to an average fragment length of 200 or 300 bp. Fragmentation was confirmed by Agilent 2200 Tapestation using D1000 screentape.

Digestion of Synthetic DNA DNA (500ng - 1 μg) in the presence of Degradase Plus (1 μL, Zymo Research) and 10 × Degradase Plus reaction buffer (4 μL, Zymo Research) in a final volume of 40 μL was digested into its constituent nucleosides via incubation at 37 °C for 4 hr. The nucleoside mixture was subsequently purified by filtration using Amicon Ultra-0.5 mL Centrifugal Filters 10K (Millipore).

DNA and RNA extraction DNA was extracted from cell-lines using DNeasy blood and tissue kit (Qiagen). ATL, AL, AW1 and AW2 lysis and wash buffers were all supplemented with desferal (200 μM) and butylated hydroxytoluene (200 μM) as recommended to reduce spurious DNA oxidation during extraction.[40] RNA was extracted using RNEasy mini kit (Qiagen) and Qiashredder columns (Qiagen); RLT buffer was additionally supplemented with DTT (40 mM).

Quantification and screening of DNA/RNA Quantification of nucleic acids was performed using Nanodrop 1000 or Nanodrop one (Thermo Fischer). The size of DNA fragments was assessed via Tapestation using D1000 or High sensitivity DNA screentapes (Agilent).

Gel electrophoresis Gel electrophoresis was carried out using an X-cell surelock mini-cell electrophoresis system (Life Technologies), powered with Pharmacia power-supplies. DNA either carried Fluorescein, Cy5 or Cy3 labels for visualization, or was first stained with SYBR gold (Thermo Fischer) before imaging. Gels were imaged either using a G:Box (Syngene) or Typhoon.

DNA library preparation for NGS and library quantification DNA libraries were prepared using NEBNext® Ultra™ DNA Library Prep Kit for Illumina using Illumina TruSeq adaptors (2.5 μL) or custom adaptors (2.5 μL, 10 mM) using the NEBNext Ultra II End Prep Enzyme mix followed by the addition of NEB Next Ultra II Ligation Master Mix in the presence of NEBNext Ligation Enhancer. Genomic samples were purified and size-

selected using Ampure XP beads (Beckmann Coulter). Libraries were amplified using NEBNext Ultra II Q5 Master Mix unless otherwise stated. DNA libraries were quantified using KAPA Library Quantification kit (KAPA Biosystems) before NGS.

RNA library preparation for NGS (RNA-seq) and library quantification mRNA was isolated from Total RNA (1 μg) using NEBNext Poly (A) mRNA magnetic isolation module (NEB) and then prepared for NGS using NEBNext Ultra Directional RNA Library prep Kit for Illumina according to the manufacturer's instructions, instead using TruSeq adaptors (2.5 μL) and without USER digestion. Samples were purified and size-selected using Ampure XP beads (Beckmann Coulter). Resultant libraries were quantified using KAPA Library Quantification kit (KAPA Biosystems) before NGS.

General biotin-streptavidin affinity-enrichment-procedure A reported DNA enrichment protocol was used with some modifications.[116] MagneSphere streptavidin magnetic beads (50 μg, Promega) or MyOne C1 Dynabeads Streptavidin C1 (50 μg, Thermo Fischer) were washed with 1 × binding buffer (5 mM Tris pH 7.5, 0.5 mM EDTA, 1M NaCl, 0.05% Tween 20) (3 x 500 μL) and resuspended in 50 μL 2 × binding buffer (10 mM Tris pH 7.5, 1 mM EDTA, 2 M NaCl, 0.1% Tween 20). DNA and Salmon sperm DNA (10 μg, Invitrogen) were mixed and made up to a final volume of 50 μL, before addition to the magnetic beads and incubation for 15 min at RT. Beads were then washed with 1 × binding buffer (6 × 500 μL) before either formamide-heating or NH_2OH-mediated chemical elution.

NGS sequencing DNA sequencing was carried out internally and performed on either MiSeq or NextSeq500 instruments (Illumina Inc.) depending on experimental requirements. All consumables were purchased from Illumina Inc.

Oligomer LC-MS analysis LC-MS was performed using an XTerra MS C18 column (2.5 μM, 2.1 x 50 mm), using solvents A (100 mM 1,1,1,3,3,3- hexafluoro-2-propanol, 10 mM NEt_3) and B (MeOH), at a flow-rate of 0.2 mL/min, with a gradient 5%–30% B increasing at 1% per min. Reaction conversion was calculated by integration of UV signals of the starting material and product(s) at 260 nm. The identity of products was assessed by ESI-MS with negative polarity in ultra-scan mode. Data was acquired between 1000-2800 m/z.

Source of fU-ODN, fC-ODN and hmU-ODN fU-ODN (Table 19) was synthesized using a protected 5-formyldeoxyuridine phosphoramidite.[268] The identity of the product was confirmed by LC-MS analysis. fC-ODN (Table 19) was obtained from Eurogentec and subjected to further HPLC purification using a Agilent Technologies 1200 series HPLC to remove impurities. A Pursuit C18 column (5 μM , 150 x 10.0 mm, Agilent) was used, (solvent A = 50 mM NH_4OAc, solvent B = MeCN, flow-rate 4 mL/min, 3% B for 5 min, and a gradient of 3-10% B for 25 min). hmU-ODN and hmU-ODN2 (See Table 19) were purchased from ATD Bio, U-ODN was purchased from Sigma Aldrich and AP-ODN was synthesized from U-ODN as discussed below.

Synthesis of modified DNA by polymerase chain reaction (PCR).

fU-DNA was synthesised using template 1 and forward primer 1 and reverse primer 1 (Appendix, Table 18) in the presence of dATP, dCTP, dGTP and dfUTP.

hmU-DNA was synthesised using template 1 and forward primer 1 and reverse primer 1 (Appendix, Table 18) in the presence of dATP, dCTP, dGTP and dhmUTP.

fC-DNA was synthesised using template 2, forward primer 2 and reverse primer 2 (Appendix, Table 18) in the presence of dATP, dfCTP, dGTP, and dTTP.

hmC-DNA2 was synthesised using template 5, forward primer 5 and reverse primer 5 (Appendix, Table 18) in the presence of dTTP, dhmCTP, dATP, dGTP.

GCAT-DNA was synthesised using template 3, forward primer 3 and reverse primer 3 (Appendix, Table 18) in the presence of dATP, dCTP, dGTP and dTTP.

U-DNA was synthesized using template 1 and forward primer 1 and reverse primer 1 (Appendix, Table 18) in the presence of dATP, dCTP, dGTP and dUTP.

Widom-DNA was synthesised using Widom template and Widom forward primer and reverse primer (Appendix, Table 18) in the presence of dATP, dGTP, dTTP/dhmUTP/dfUTP and dCTP/dhmCTP/dfCTP/dcaCTP

fU-DNA1, fC-DNA1, GCAT-DNA1 for protein pulldown and model protein crosslinking studies was synthesised using template 4, forward primer 4 and reverse primer 4 (Appendix, Table 18) in the presence of dfUTP/dTTP, dfCTP/dCTP, dATP, dGTP.

fC-DNA2 was synthesised using template 5, forward primer 5 and reverse primer 5 (Appendix, Table 18) in the presence of dTTP, dfCTP, dATP, dGTP.

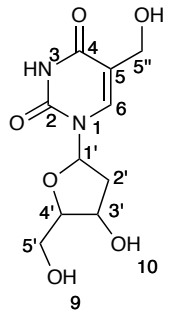
Proteomics-analysis (*performed and written by Cambridge Proteomics Centre)* 1D gel bands were cut, destained, reduced (DTT) and alkylated (iodoacetamide) and subjected to enzymatic digestion with trypsin, chymotrypsin, Arg-C or Asp-N overnight at 37 °C. After digestion, the supernatant was pipetted into a sample vial and loaded onto an

autosampler for automated LC-MS/MS analysis. All LC-MS/MS experiments were performed using a Dionex Ultimate 3000 RSLC nanoUPLC (Thermo Fisher Scientific Inc, Waltham, MA, USA) system and a Q Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific Inc, Waltham, MA, USA). Separation of peptides was performed by reverse-phase chromatography at a flow rate of 300 nL/min and a Thermo Scientific reverse-phase nano Easy-spray column (Thermo Scientific PepMap C18, 2mm particle size, 100 Å pore size, 75 mm i.d. x 50cm length). Peptides were loaded onto a pre-column (Thermo Scientific PepMap 100 C18, 5mm particle size, 100 Å pore size, 300 mm i.d. x 5mm length) from the Ultimate 3000 autosampler with 0.1% formic acid for 3 min at a flow rate of 10 mL/min. After this period, the column valve was switched to allow elution of peptides from the pre-column onto the analytical column. Solvent A was water + 0.1% formic acid and solvent B was 80% acetonitrile, 20% water + 0.1% formic acid. The linear gradient employed was 2-40% B in 30 min.

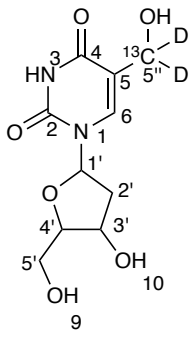
The LC eluant was sprayed into the mass spectrometer by means of an Easy-Spray source (Thermo Fisher Scientific Inc.). All m/z values of eluting ions were measured in an Orbitrap mass analyzer, set at a resolution of 70000 and was scanned between m/z 380-1500. Data dependent scans (Top 20) were employed to automatically isolate and generate fragment ions by higher energy collisional dissociation (HCD, NCE:25%) in the HCD collision cell and measurement of the resulting fragment ions was performed in the Orbitrap analyser, set at a resolution of 17500. Singly charged ions and ions with unassigned charge states were excluded from being selected for MS/MS and a dynamic exclusion window of 20 s was employed.

Post-run, the data was processed using Protein Discoverer (version 2.1, ThermoFisher). Briefly, all MS/MS data were converted to mgf files and the files were then submitted to the Mascot search algorithm (Matrix Science, London UK) and searched against the Uniprot human database (151984 sequences; 47833598 residues) and common contaminant sequences (115 sequences, 38274 residues). Variable modifications of oxidation (M), deamidation (NQ) and carbamidomethyl were applied. The peptide and fragment mass tolerances were set to 5 ppm and 0.1 Da, respectively. A significance threshold value of $p < 0.05$ and a peptide cut-off score of 20 were also applied.

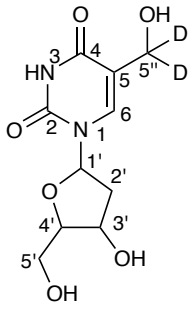
7.2. Chapter 2 Materials and Methods



5-hmU To a solution of 2-deoxyuridine (13.15 g, 57.6 mmol) in formaldehyde (15.19 mL, 37% by weight in H_2O) under an atmosphere of Ar was added 1M KOH (100 mL, 100 mmol). The reaction was sealed and heated for 72 hours at 60 °C. The reaction mixture was concentrated *in vacuo* and was then purified by flash column chromatography (5:1 CH_2Cl_2:MeOH) and recrystallized from MeOH to afford *5-hydroxymethyldeoxyuridine (5-hmU)* as white crystals (7.13 g, 48%); δH (500MHz, d_4-methanol) 7.98 (s, 1H, H6), 6.32 (dd, *J* = 7.3, 6.2, 1H, H1'), 4.42 (1H, dt, *J* = 6.5, 3.5, H3'), 4.34 (2H, s, H5''), 3.94 (1H, app q, *J* = 3.5, H4') 3.80 (1H, dd, *J* = 12.0, 3.5, H5'), 3.74 (1H, dd, *J* = 12.0, 3.5, H5') 2.28 (2H, m, H2'); δC (126MHz, MeOD) 161.1 (C4), 152.2 (C2), 139.4 (C6), 115.3 (C5), 88.9 (C4'), 86.5 (C1'), 72.3 (C3'), 62.9 (C5'), 58.0 (C5''), 41.3 (C2'). Data was in accordance with that reported in the literature.[290]

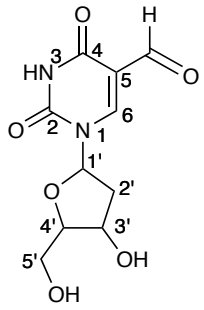


hmU-SIL To a solution of 2-deoxyuridine (173 mg, 0.75 mmol) in ^{13}C-d_2-formaldehyde (1.5 mL, 20% by weight in D_2O) under an argon atmosphere was added NEt_3 (1.35 mL, 9.70 mmol). The reaction was sealed and heated for 72 h at 60 °C. The reaction mixture was concentrated and then purified by flash column chromatography (CH_2Cl_2:MeOH = 6:1 to 5:1, v/v) to afford *5-[$^{13}CD_2$] hydroxymethyl-2'-deoxyuridine (hmU-SIL)* as a white solid. (72 mg, 37%); ^{1}H NMR (500 MHz, d_4-methanol) δ 7.96 (1H, d, J_{C-H} = 4.1, H6), 6.29 (1H, dd, *J* = 7.3, 6.2, H1'), 4.42 (1H, dt, *J* = 6.5, 3.4, H3'), 3.94 (1H, app q, *J* = 3.4, H4') 3.78 (1H, dd, *J* = 12.0, 3.4, H5'), 3.72 (1H, dd, *J* = 12.0, 3.4, H5') 2.24 (2H, m, H2'). ^{13}C NMR (126 MHz, MeOD) δ 161.5 (C4) 152.2 (C2), 139.5 (d, J_{C-C} = 4.6, C6), 115.1 (d, J_{C-C} = 53.0, C5), 88.9 (C4'), 85.5 (C1'), 72.3 (C3'), 62.9 (C5'), 57.4 (qn, J_{C-D} = 21.9, C5''), 41.3 (C2'); HRMS $C_9{}^{13}CH_{13}D_2N_2O_6$ $[M+H]^+$ calcd. 262.1089, found 262.1096.

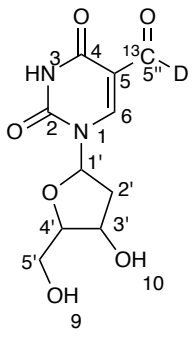


hmU-SIL2 To a solution of deoxyuridine (300 mg, 1.23 mmol) in formaldehyde-d_2 (2.24 mL, 20% by weight in D_2O) under an atmosphere of Ar was added NEt_3 (1.93 ml, 13.9 mmol). The reaction was sealed and heated for 72 hours at 60 °C. The reaction mixture was concentrated *in vacuo* before purification using a Combiflash (gradient 19:1 EtOAc: MeOH → 9:1 EtOAc: MeOH) followed by further purification by preparative HPLC (199:1 H2O:MeCN) to afford *d_2-5-*

hydroxymethyluridine (hmU-SIL2) as a white solid (8mg, 2%); δH (500MHz, *d*$_4$-methanol) 7.96 (1H, d, *J* = 1.5, H6), 6.29 (1H, dd, *J* = 7.3, 6.2, H1'), 4.85 (2H, d, *J* = 1.7, H5''), 4.39 (1H, dt, *J* = 6.5, 3.4, H3'), 4.06 (1H, t, J = 12.8, H3'), 4.05 (1H, dt, *J* = 4.1, H4'), 3.92 (1H, app q, *J* = 3.4, H4'), 3.78 (1H, dd, *J* = 12.0, 3.4, H5'), 3.72 (1H, dd, *J* = 12.0, 3.4), 2.25 (2H, m, H2'); δC (126MHz, D_2O) 165.2 (C4), 152.2 (C2), 139.5 (C6), 115.1 (C5), 88.9 (C1'), 86.5 (C4'), 72.3 (C2'), 62.9 (C3'), 60.5 (C5'), 57.5 (m, C5''), 41.2 (C2'); HRMS $[M+H]^+$ expected 277.1005, found 277.1009.



5-fU To a solution of 5-hydroxydeoxyuridine (65 mg, 0.25 mmol) in MeOH (1 mL) was added MnO_2 (109 mg, 1.25 mmol) and the reaction mixture was stirred at 50 °C for 18 hr. The catalyst was removed by filtering through Celite, which was washed with MeOH, and the filtrate was concentrated under reduced pressure. The crude product was purified by column chromatography (9:1 CH_2Cl_2:MeOH → 5:1 CH_2Cl_2:MeOH) to afford *5-formyldeoxyuridine (5-fU)* as a pale solid (13 mg, 20%); ^{1}H NMR (500 MHz, D_2O) 9.55 (1H, s, H5''), 8.68 (1H, s, H6), 6.15 (1H, t, *J* = 5.9, H1) 4.37 (1H, dt, *J* = 6.5, 4.7, H3'), 4.01 (1H, td, *J* = 4.6, 3.3, H4'), 3.79 (1H, dd, *J* = 12.6, 3.3, H5'), 3.68 (1H, dd, *J* = 12.6, 4.6, H5'), 2.38 (2H, m H2'); ^{13}C NMR (126 MHz, D_2O) 189.0 (C5''), 162.9 (C5), 151.2 (C4), 150.4 (C2) 111.2 (C5), 87.2 (C1'), 86.9 (C4'), 69.8 (C3'), 60.6 (C5'), 39.6 (C2'). Data was in accordance with that reported in the literature.[291]



fU-SIL To a solution of hmU-SIL (40 mg, 0.15 mmol) in MeOH (1 mL) was added MnO_2 (67 mg, 1.25 mmol) and the reaction mixture was stirred at 50 °C for 18 hr. The catalyst was removed by filtering through celite, washed with MeOH, and the filtrate was concentrated. The crude product was purified by flash column chromatography (CH_2Cl_2:MeOH = 9:1 to 5:1, v/v) to afford *5-[^{13}CD] formyl-2'-deoxyuridine (fU-SIL)* as a white solid (3 mg, 8%); ^{1}H NMR (500 MHz, D_2O) δ 8.68 (1H, d, J_{C-H} = 4.8, H6), 6.14 (1H, app t, J = 6.5, H1'), 4.37 (1H, dt, J = 6.5, 4.7, H3'), 4.01 (1H, dt, J = 4.7, 3.3, H4'), 3.79 (1H, dd, J = 12.6, 3.3, H5'), 3.68 (1H, dd, J = 12.6, 4.7, H5'), 2.43 (1H, m, H2'), 2.33 (1H, m, H2'). ^{13}C NMR (126 MHz, D_2O) δ 188.7 (t, J_{C-D} = 27.3, C5''), 163.0 (C6), 151.2 (d, J_{C-C} = 6.4, C4), 150.4 (C2), 111.2 (d, J_{C-C} = 61.4, C5), 87.2 (C1'), 86.9 (C4'), 69.8 (C3'), 60.6 (C5'), 39.6 (C2'); HRMS $C_9{}^{13}CH_{11}DN_2NaO_6$ $[M+Na]^+$ calcd. 281.0689, found 281.0692.

Source of Genomic DNA samples. HEK293T cell pellets were generated by myself or by Dr S. Huber (Balasubramanian group). MCF7 and mESC (derived from 129/C6 blastocyst) cell pellets were provided by Dr S. Mao, (Balasubramanian group). Bloodstream and procyclic form *T.brucei* cell pellets were provided by Prof. M. Carrington or Dr J. Freitas (Carrington group, Department of Biochemistry, University of Cambridge). L.major DNA was sourced from ATCC, *A. Thaliana* DNA was acquired from Prof. D. Baulcombe (Department of Plant Sciences, University of Cambridge), Archea DNA (*Sulfolobus acidocaldarius MW001)* was provided by Dr F. Werner (University College of London). DNA was extracted from tissue and cell-lines using the procedure outlined by Carell et al, using a Qiagen extraction kit in the presence of antioxidants BHT and desferral.[40] This procedure was designed to minimise artificial 5-hmU and 5-fU thymine auto-oxidation during DNA extraction. This was possible for trypanosome samples and mammalian cell-lines, however DNA from Leishmania, Archea and Arabadopsis were acquired after DNA extraction.

DNA digestion Trypanosomatid samples containing hypermodified Base J, were first sonicated to ~200 bp before addition of Degradase Plus (0.5 µL, Zymo Research), MNase (0.1 µL, NEB), UltraPure Benzoase (0.2 µL, Sigma Aldrich), Antarctic Phosphatase (0.4 µL, NEB) in the presence of 10 x Degradase Buffer and left for 12 hr. Mammalian samples were digested in the presence of Degradase Plus (1 µL per 2 µg DNA) in the presence of 10 x Degradase buffer and left for 4 hr. Digested mononucleosides were purified using Amicon Ultra-0.5 mL Centrifugal Filters 10K (Millipore).

Q-Exactive LC-MS/MS Set-up LC-MS/MS analysis was performed on a Q-Exactive mass spectrometer (Thermo Fischer) coupled with an UltimMate 3000 RSLC nano-HPLC (Dionex) and either a self-packed hypercarb column (50mm x 75 µm, 3 µm particle size) or a commercially sourced hypercarb kappaguard column (30 mm x 0.18 mm, 5 µm particle size, Thermo Fischer) connected with 2 x nanoviper connectors (75 µm x 150mm). Samples were injected onto the column via the loading pump in 95:5 0.1% formic acid H_2O:MeCN with a flow-rate of 2 µL/min. A valve switch to the NC pump followed after either 5 or 7 min; the flow-rate was set to 1.5 µL min^{-1} and ran with a gradient of 95:5 → 0:100 0.1% formic acid H_2O:MeCN with a run-time of 19 min. Parent ions were fragmented in positive ion mode with 10% normalised collision energy using parallel-reaction monitoring (PRM). MS^2 resolution was 35,000 with an AGC target of $2e^5$, a maximum injection time of 100 ms and an isolation window of 1.0 m/z. Extracted ion chromatograms (±5ppm) were used for detection and quantification, and

quantification was performed using XCalibur QuanBrowser software (Thermo Fischer) via internal calibration using SIL standards where possible; hmU-SIL was used as an internal standard for Base J internal calibration and T-SIL was used for Girard's T calibration.

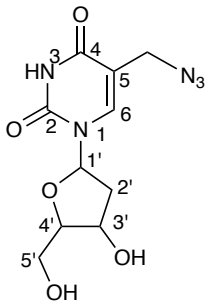
Calibration lines A dilution series of T or C was prepared in the range of 0.045 – 9000 nM. SIL standards were added to every calibration point, to give a final concentration of 25 nM T-SIL or C-SIL. For modified bases, dilution series of hmC, mC, Base J, hmU, fU were prepared in the range of 0.00009 nM – 450 nM. SIL standards (hmC-SIL, hmU-SIL, hmU-SIL2, fU-SIL or mC-SIL) were added to every calibration point to give a final concentration of either 5 nM or 2.5 nM. Concentration of known nucleoside standards were plotted against the mass integration area ratio of nucleoside/internal standard using QuanBroswer software (Thermo Fischer) to generate calibration lines for subsequent quantification.

LC-MS/MS Quantification To each digested genomic sample was added an equivalent concentration of SIL standards to that of the calibration line. The mass integration area ratio of nucleoside/internal standard was compared to the linear fit equation of calibration lines using QuanBrowser software to determine nucleoside concentration (Thermo Fischer).

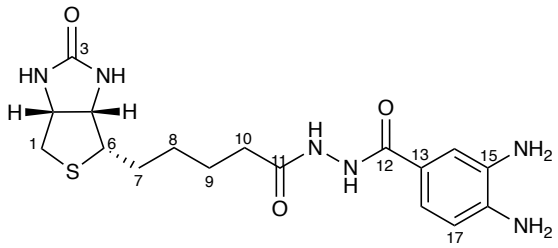
HPLC enrichment HPLC pre-enrichment of genomic samples was performed on Ultimate 3000 Dionex HPLC system (Thermo Fischer) equipped with a Waters HSS-T3 column (2.1 x 100 mm, 1.8 µm particle size). Flow rate was 0.35 mL/min and ran with a gradient of 98:2 to 0:100 (H_2O:MeCN 0.1% formic acid). Mononucleoside fractions were collected at particular timepoints (Appendix, Chapter 2 - Table 21), previously determined by the UV signals of nucleoside standards in model studies (Appendix, Chapter 2 - Figure 88). Collected fractions were lyophilised overnight and resolubilised in water for LC-MS/MS analysis. A proportion of the digested sample was kept for quantification of T or C without pre-enrichment.

fU-ODN with Girard's Reagent T Girard's T (50 mM) was added to fU-ODN (1-2 µL, 100 uM) in pH 7 phosphate buffer (40 mM) and left at RT for 3 hr. The sample was purified by oligo clean and concentrator kit (Zymo Research) and analysed by oligomer LC-MS analysis.

7.3. Chapter 3 Materials and Methods



5-azidomethyluracil Using a procedure outlined by Xu et al[178], 5-hmU (69 mg, 0.27 mmol) and NaN_3 (65 mg, 1 mmol) were dissolved in TFA and the reaction mixture was left to stir at rt for 18 hr. The reaction mixture was neutralised with aq. $NaHCO_3$ solution before concentration under reduced pressure. The reaction mixture was purified using a Combiflash gradient 19:1 → 9:1 CH_2Cl_2:MeOH) to afford 5-azidomethyluracil as a white solid (20 mg, 26%); NMR data was in accordance with the literature reports.[178] δH (500MHz, d_6-DMSO) 11.52 (1H, s, H3), 8.01 (1H, s, H6), 6.12 (1H, d, *J* = 6.7, H1'), 5.25 (2H, d, *J* = 4.3, C5'), 5.03 (1H, t, *J* = 5.2, H3'), 4.21 (1H, m, H3'), 4.04 (2H, d, *J* = 2.5, H5''), 3.76 (1H, q, J = 3.7, H4'), 2.01 (2H, m, H2'); δC (126MHz, d_6-DMSO) 162.9 (C4), 150.3 (C2), 140.0 (C6), 108.3 (C5), 87.5 (C4'), 84.3 (C1'), 70.3 (C3'), 61.3 (C5'), 47.0 (C5''), 41.0 (C2').



***o*-phenbiotin**. To a solution of 3,4-diaminobenzoic acid (70 mg, 0.46 mmol) in DMF (20 mL) was added 1-hydroxybenzotriazole (84 mg, 0.55 mmol), *N*-(3-dimethylaminopropyl)-*N*'-ethylcarbodiimide hydrochloride (96 mg, 0.50 mmol), NEt_3 (0.12 ml, 0.86 mmol) and (+)-biotin hydrazide (107 mg, 0.41 mmol). The reaction was stirred at 45 °C for 18 hr, and the solvent was then removed *in vacuo.* The reaction mixture was subjected to flash chromatography (CH_2Cl_2:MeOH=19:1 to 3:2, v/v) and size exclusion chromatography using Sephadex LH-20 to afford biotinylated phenylendediamine linker *o*-phenbiotin as a white solid. (35 mg, 20%); ^{1}H NMR (500 MHz, methanol-*d*4) δ 7.23 (d, *J* = 2.1, 1H, *H*14), 7.19 (dd, *J* = 8.2, 2.1, 1H, *H*18), 6.67 (1H, d, *J* = 8.2, *H*17), 4.49 (ddd, *J* = 7.9, 5.0, 0.9, 1H, *H*2), 4.32 (dd, *J* = 7.9, 4.5, 1H, *H*5), 3.22 (ddd, *J* = 8.4, 6.2, 4.5, 1H, *H*6), 2.93 (dd, *J* = 12.7, 5.0, 1H, *H*1), 2.70 (d, *J* = 12.7, 1H, *H*1), 2.32 (m, 2H, *H*10), 1.75 (m, 2H, *H*9), 1.67–1.45 (m, 4H, *H*7, *H*8); ^{13}C NMR (100 MHz, methanol-*d*4) δ 173.9, 168.2, 164.8, 140.0, 133.3, 121.2, 119.5, 115.2, 114.1, 61.8, 60.3, 55.5, 39.7, 33.2, 28.2, 28.0, 25.0; HRMS $C_{17}H_{25}N_6O_3S^+$ $[M+H]^+$ calculated 393.1709, found 393.1701.

7.3.2. β-glucosylation with UDP-Glucose Widom sequence DNA containing 1) 5-hmC, 2) 5-hmU or 3) 5-hmC and 5-hmU (500 ng) was subjected to β-glucosylation conditions using the Quest-hmC DNA enrichment kit. (Zymo Research). DNA was made up to 38μL, before the addition of 10 x hmC-GT buffer (5 μL), UDP-Glucose (5 μL, 1 mM) and 5-hmC β-GT enzyme (2 μL), and the mixture was incubated at 37 °C for 4 hr. The products were purified using Genejet PCR Purification kit (Thermo Fischer). DNA (starting material and β-glucosylated) was digested using the digestion of synthetic DNA protocol and purified by 10K Amicon centrifugal filters (Millipore). Nucleosides were separated using a Hewlett Packard Series 1100 HPLC and Eclipse XDB:C18 column heated to 45 °C, dimensions 3.5 uM, 3.0 x 150 mm, with a flowrate of 1 mL/min. Eluting buffers were A) 500 mM NH_4OAc pH = 5, B) MeCN, C) H_2O. Buffer A was kept at 1% throughout each run, with gradients 0 min – 1% B, 8 min – 4% B, 10 min – 95% B. Analysis revealed a consumption of hmC peak after glycosylation conditions, while the hmU peak remained. Subsequent LC-MS/MS Q-Exactive quantification of GhmCAhmU-DNA was used to determine hmC and hmU consumption. This was measured by comparing mass integration area ratios of 5-hmU and 5-hmC, normalized to C, in starting material DNA, and after glycosylation conditions. 99.9% of 5-hmC was consumed compared to 4.3% of 5-hmU after glycosylation.

Analysis of mononucleoside reactions by LC-MS. LC-MS analysis was performed on a Bruker Dionex Ultimate 3000 system, using a Kinetex 100A column dimensions 50 x 2.1 mM, using solvents A) 0.1% TFA in Water, B) 0.1% TFA in MeCN, flow-rate 1mL/min, with a gradient, 0.5% B 0-2 min, 0.5-100% B 2-3 min. Extracted ion count and subsequent integration of chromatograms gave % conversion of reaction.

Acid Catalysis No observation of target molecule 5-azaU was observed by LC-MS when NaN_3 (0.15 mmol) was added to 5-hmU in the presence of acid in aqueous conditions. To a solution of 5-hmU (4.0 mg, 0.01 mmol) in water (1 mL) was added either $Yb(OTf)_3$ (19.2 mg, 0.02 mmol), $Sc(OTf)_3$ (15.2 mg, 0.02 mmol) or TFA (20 μL, 2% by volume) and left for 18 hr.

Reactions with DMT-MM To a solution of 5-hmU (1.0 mg, 0.01 mmol) was added DMT-MM (5.0 mg, 0.1 mMol) and either i) NEt_3 (0.1 mmol) ii) 2,6-Lutidine (0.1 mmol), iii) 2,6-ditertbutylpyridine (0.1 mmol), iv) N-methylmorpholine (0.1 mmol) or v) no base and the reactions were stirred at 18 hr at a) RT, b) 50 °C. Reactions were analysed using general LC-MS analysis.

Reaction with DMT-MM and NaN_3 To see if the DMT-MM adduct could be further functionalized by sodium azide, to a solution of 5-hmU (15 mg, 0.06 mmol) was added DMT-MM (222 mg, 0.8 mmol) and NaN_3 (58 mg, 0.89 mmol) in H_2O. The reaction was analysed using general LC-MS analysis. No target product formation 5-azaU was observed by LC-MS.

Reaction with EDC To a solution of 5-hmU (13 mg, 0.05 mmol) in water, was added EDC (15 mg, 0.25 mmol) and NEt_3 (34 µL, 0.25 mmol) and the solution was left to stir at RT for 18 hr. The reaction was analysed using LC-MS analysis, demonstrating adduct formation (56%). Attempted preparative HPLC purification (Buffer A: 0.1% TFA in H_2O, Buffer B: 0.1% TFA in MeCN, gradient 5-100% B) lead to re-equibriliation of peak after separation. LC-MS [M+H] = 414.2.

ODN Reaction with DMT-MM To hmU-ODN, Phos-hmU-ODN or Phos-GCAT-ODN in water was added DMT-MM (75 µg) and NEt_3 (1 µL) to make a final reaction volume of 25 µL. The mixtures were incubated for 16 hr at 50 °C before purification using mni quick spin columns (Roche) which had been pre-washed with water (2 x 300 µL) oligomer LC-MS analysis.

Chemical oxidation of hmU-ODN. Using a protocol used for 5-hmC oxidation,[124] hmU-ODN (2 µL, 100 uM) (Appendix – Table 19) was incubated with NaOH (1.25 µL, 1M) and $KRuO_4$ (1 µL, 15 mM in 0.05 M NaOH) (Alfa Aesar) on ice for 1 hr. The reaction was purified by mini quick spin oligo column (Roche), which was pre-washed with water (2 × 300 µL). The ODN reaction was traced using general oligomer LC-MS analysis, and nucleobase composition of digested DNA was analyzed by a Q-exactive (Thermo Fischer) quadrupole-orbitrap hybrid tandem MS spectrometer in positive ion mode. Extracted ion chromatograms of base fragments were used corresponding to C, T, A, G, 5-hmU and 5-fU respectively. Gaussian smoothing (7 points) was applied. Analogous oxidation conditions can be achieved by using 1 µL of the oxidant contained in the TrueMethyl kit (Cambridge Epigentix) which has been diluted 10-fold in water.

Chemical tagging reactions with ARP. fU-ODN, fC-ODN, AP-ODN, GCAT-ODN (0.5 - 2 µL, 100 uM) (Appendix – Table 19) were incubated with ARP (0.4 mM) at RT, in the presence of absence of *p*-anisidine (100 mM) in sodium phosphate buffer or NH_4OAc (40 mM) at a range of different pH and incubation times (Table 4 and 7). Reactions on ODNs

were purified by mini quick spin oligo columns (Roche), which were pre-washed with water (2 × 300 µL). Reactions were traced using general oligomer LC-MS analysis. ARP (Biotinylcarbazoylmethyl)Hydroxylamine, Cayman Chemicals) was made up as a stock solution in DMSO (100 mM) and *p*-anisidine was made up as a stock solution in MeOH (1 M).

Chemical tagging reactions with BH. fU-ODN, fC-ODN, AP-ODN, GCAT-ODN (0.5 - 2 µL, 100 uM) (Appendix Table 19) were incubated with BH (10 or 20 mM), the presence of absence of *p*-anisidine (100 mM) in sodium phosphate buffer (40 mM) at a range of different pH, temperature and incubation times (Table 5 and 7). Reactions on ODNs were purified by mini quick spin oligo columns (Roche), which were pre-washed with water (2 × 300 µL). Reactions were traced using general oligomer LC-MS analysis. BH ((+)-Biotinamidohexanoic acid hydrazide, Sigma Aldrich) was made up as a stock solution in DMSO (100 mM) and *p*-anisidine was made up as a stock solution in MeOH (1 M).

Chemical tagging reactions with *o*-phenylenediamine. fU-ODN, fC-ODN, AP-ODN, GCAT-ODN (0.5 - 2 µL, 100 uM) (Appendix – Table 19) were incubated with *o*-phenylenediamine (100 or 5 mM) at RT, in sodium phosphate buffer (40 mM) at a range of different pH and incubation times (Table 6 and 7). Reactions on ODNs were purified by mini quick spin oligo columns (Roche), which were pre-washed with water (2 × 300 µL). Reactions were traced using general oligomer LC-MS analysis. *o*-phenylenediamine (Sigma Aldrich) was made up as a stock solution in MeOH (100 mM).

Chemical tagging reactions with *o*-biophen. fU-ODN, fC-ODN, AP-ODN, GCAT-ODN (0.5 - 2 µL, 100 uM, Appendix – Table 19) were incubated with *o*-biophen (5 mM) in pH 7 sodium phosphate buffer (40 mM) at RT for 4 hr. The ODN reaction was purified by mini quick spin oligo columns (Roche), which were pre-washed with water (2 × 300 µL). Reactions were traced using general oligomer LC-MS analysis. *o*-Biophen was made up as a stock solution in DMSO (100 mM).

Selective chemical labeling of hmU-ODN over fU-ODN. To a mixture of fU-ODN (3 µL, 100 µM) and hmU-ODN (3 µL, 100 µM) was added *N*-methylhydroxylamine hydrochloride (5 µL, 50 mM) and pH 6 Sodium Phosphate Buffer (40 mM) in a final reaction volume of 50 µL. The reaction was left for 3 hr at RT, before being purified using 2 x Bio-Spin P-6 Gel Columns, SSC Buffer (Bio-Rad), which had been pre-washed

with water (6 × 500 µL). The resultant solution was incubated with NaOH (2.5 µL, 1M) and $KRuO_4$ (2 µL, 15 mM in 0.05 M NaOH) (Alfa Aesar) on ice for 1 hr. The mixture was purified by mini quick spin oligo column (Roche), which was pre-washed with water (2 × 300 µL), before the addition of BH (10 µL, 100 mM) and pH 7 Sodium Phosphate Buffer (40 mM). Purification was completed via mini quick spin oligo column (Roche), which had been pre-washed with water (2 × 300 µL). Analysis was performed by general oligomer LC-MS analysis.

Synthesis of AP-site containing DNA (AP-ODN and AP-DNA) U-ODN or U-DNA (1 µg, Appendix Table 19) in the presence of UNG (1 µL, 5U, NEB) and 10 × UNG reaction buffer (3 µL, NEB) in a final volume of 30 µL was incubated at 37 °C for 3 hr to generate AP-ODN or AP-DNA respectively. AP-ODN was purified by mini quick spin oligo column (Roche). AP-DNA was purified by GeneJET PCR Purification Kit (Thermo Scientific) and confirmed by oligomer LC-MS analysis.

***Ab initio* Quantum Mechanical Calculations** (*Performed by Dr A. Sahakyan*) Closed-shell restricted Hartree-Fock (RHF) calculations were done with the Møller-Plesset correlation energy correction truncated at second-order (MP2)[292,293] and with the double-zeta cc-pVDZ Dunning's correlation consistent basis set.[294,295] All the calculations were done using the Gaussian 03 suite of programs.[296]

Energy minima were found through a fully relaxed geometry optimisation of two (syn and anti) rotameric structures constructed for both 5-fU_m and 5-fC_m. The transition states were located via synchronous transit-guided quasi-Newton search[297] in between the two minima. All the found stationary points were verified to be either energy minima or transition states (first order saddle points) via an additional vibrational frequency calculation to find out the number (or the absence) of the imaginary force constants.

AP-DNA and fU-DNA Primer extension A protocol was used based on that reported by McInroy *et al*.[197] AP-DNA or fU-DNA was firstly reacted with BH (10 mM) for 4 hr in pH 7 phosphate buffer (40 mM). The DNA was purified using mini quick spin columns (Roche). 200 ng of the reacted product was subjected to polymerase extension in the presence of dNTP (0.2 mM), 5'-Fluorescein-labelled Rev primer 1 (1 mM), 10 x DreamTaq Buffer (5 µL) and DreamTaq Polymerase (1 µL). After an initial 3 min denaturation at 95 °C, 10-cycles of denaturation (95 °C, 30s), annealing (52 °C,

30s) and extension (72 °C, 30s) were performed. Polymerase extension was monitored on a 10% TBE-Urea gel.

AP-ODN fragmentation AP-DNA and hmU-ODN2 were incubated in the presence of NaOH (50 mM) for 1 hr at 40 °C, before purification using mini quick spin columns (Roche). ODNs were analysed by oligomer LC-MS analysis.

qPCR enrichment studies and biotinylated adduct reversal

Reaction procedure for chemical enrichment studies. DNA (500 ng) was subjected to reaction conditions a, b, c, d or e. The reactions were purified using mini quick spin oligo columns (Roche) pre-washed with water (2 × 300 μL). The resulting purified mixture was diluted 100-fold to give an approximate concentration of 100 pg/μL per ODN. Reactions were done in duplicate or triplicate:

a) fU-DNA and fC-DNA (500 ng) was incubated with NH_4OAc buffer pH = 5 (40 mM), ARP (2 M) and *p*-anisidine (100 mM), to make a final reaction volume of 50 μL, at RT for 24 hr;

b) fU-DNA and fC-DNA (500 ng) was incubated with sodium phosphate buffer pH = 6 (40 mM) and ARP (0.4 mM) to make a final reaction volume of 50 μL, at RT for 4 hr;

c) fU-DNA and fC-DNA (500 ng) was incubated with sodium phosphate buffer pH = 7 (40 mM) and BH (10 mM) to make a final reaction volume of 50 μL, at RT for 4 hr;

d) fU-DNA and fC-DNA (500 ng) was incubated with sodium phosphate buffer pH = 7 (40 mM) and *o*-phenbiotin (5 mM) to make a final reaction volume of 50 μL, at RT for 4 hr;

e) fU-DNA and GCAT-DNA (500 ng) was incubated with sodium phosphate buffer pH = 7 (40 mM) and BH (10 mM) to make a final reaction volume of 50 μL, at RT for 4 hr.

Input DNA (10 μL, 1 ng) in the presence of Salmon sperm (10 μg) was incubated and enriched using the general affinity-enrichment procedure using Magnesphere streptavidin magnetic beads (50 μg, Promega) and eluted via formamide elution.

Formamide elution After streptavidin enrichment, magnetic beads were resuspended in 100 μL elution buffer (95% formamide, 10 mM EDTA) and were heated to 95 °C for 5 min. The eluent was removed from the beads and placed on ice. The step was then repeated using 50 μL elution buffer to remove residual DNA from the magnetic beads. The eluent was diluted with water (350 μL), and purified by filtration using Amicon Ultra-0.5 mL Centrifugal Filters 10K (Millipore), following a wash by water (450 μL) and centrifugation for 15 min. The Amicon filters were washed with water and centrifuged

for a further 15 min (2 × 450 μL). DNA was then recovered from the Amicon filter (25 μL) and enrichment was assessed by qPCR.

qPCR analysis for chemical enrichment studies. qPCRs were performed using a CFX96 Real-Time System (BioRad), and data was processed using the CFX Software manager (BioRad). Enriched DNA (1 μL) was added to a mixture of Brilliant III Ultra-Fast SYBR Green qPCR Master Mix (5 μL) (Agilent Technologies), forward primer 1, 2 or 3 (1 μM), reverse primer 1, 2 or 3 (1 μM) (Appendix, Table 18) and diluted with water to give a final volume of 10 μL. The mixture was subject to qPCR according to the protocol outlined by the manufacturer. DNA concentration was quantified by comparison with calibration lines of known concentration of input ODNs (Appendix, Chapter 2 – Figure 92).

Synthesis of BH-adduct starting materials for reversal study fU-ODN (2 μL) and/or fC-ODN (2 μL) were incubated with BH (10 mM) in the presence of *p*-anisdine (100 mM) in pH 6 phosphate buffer (40 mM) at 40 °C for 18 hr, before purification via mini quick spin oligo columns (Roche). fU-ODN-BH and fC-ODN-BH adduct formation was confirmed by general oligomer LC-MS analysis. The purified mixture was diluted 100 fold to give an approximate concentration of 100 pg/μL per ODN to be used for qPCR enrichment studies.

5-fU-ODN-adduct reversal fU-BH, fU-*o*-phen and fU-ARP ODN adducts were incubated in the presence of *p*-anisidine (100 mM) and the presence or absence of hydroxylamine (0.05% v/v) in pH 7 or pH 6 phosphate buffer (40 mM) or pH 5 NH_4OAc buffer for 2 hr. Reactions were purified by mini quick spin oligo columns (Roche) and analysed by oligomer LC-MS analysis.

5-fC-BH reversal fC-ODN-BH was incubated in the presence of *p*-anisdine (100 mM), and hydroxylamine (0.05% v/v) in either pH 5 NH_4OAc buffer (40 mM) or pH 6 phosphate buffer (40 mM) and heated at 40 °C for 4 hr or 24 hr. Reactions were purified by mini quick spin oligo columns (Roche) and analysed by oligomer LC-MS analysis.

Polymerase stop assay for probes: A primer extension experiment was performed based on conditions reported by McInroy et al.[197] fU-DNA was firstly reacted with i) BH (10 mM) or ii) NH_2OH (0.5% v/v) in pH 7 phosphate buffer (40 mM) for 4 hr or iii) ARP (0.4 mM), pH 6 phosphate buffer (40 mM). The DNA was purified using mini quick spin

columns (Roche). 200ng of the reacted product was subjected to polymerase extension in the presence of dNTP (0.2 mM), 5'-Fluorescein-labelled Rev primer 1 (1 mM), 10 x DreamTaq Buffer (5 µL) and DreamTaq Polymease (1 µL). After an initial 3 min denaturation at 95 °C, 10-cycles of denaturation (95 °C, 30s), annealing (52 °C, 30s) and extension (72 °C, 30s) were performed. The products were visualised on a 10% TBE-Urea gel.

Library preparation of DNA models

Chemical 5-fU enrichment-sequencing library preparation Fragmented DNA samples (1 µg) and spike-in ODNS (fU-DNA, fC-DNA2, GCAT-DNA, 100pg), (+)-biotinamidohexanoic acid hydrazide BH (10 mM) and pH 7 sodium phosphate buffer (40 mM) was incubated at RT for 4 hr followed by work-up with DNA Clean and Concentrator 5 (Zymo Research) or Micro Biospin-P6 columns in SDS buffer (BioRad). DNA fragments were subjected to the general NGS library preparation procedure, followed by general affinity-enrichment procedure using Magnesphere beads (Promega). DNA fragments were eluted off beads by incubating with (0.05% v/v NH_2OH), pH 7 sodium phosphate buffer (40 mM) and *p*-anisidine (100 mM), at 40 °C for 2 hr. DNA fragments were removed from the beads and purified by GeneJet PCR purification kit (ThermoFischer) or DNA Clean and Concentrator-5 (Zymo Research). PCR of enriched fragments was achieved using NEBNext Ultra II Q5 Master Mix (NEB).

Chemical 5-hmU enrichment-sequencing library preparation Fragmented DNA samples (1 µg) and spike-in ODNs (hmU-DNA, hmC-DNA2, GCAT-DNA 300 pg) were firstly ligated using general NGS library preparation procedure in the presence of custom adaptors 1) 5′-MeO-GAATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ (Eurogentec) and
2) GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCT TCTGCTTG-*O*-phosphate-3′ (Sigma Aldrich) which were annealed prior to use. After library preparation, Ampure beads were washed twice with 80% MeCN, and the eluted fragments were purified twice with Micro BioSpin-P6 columns in SDS buffer (BioRad). To 22.75 µL ligated DNA solution was added NaOH (1 µL, 1M) and the mixture was heated at 40 °C for 30 min, before the addition of ten-fold diluted oxidant solution (1 µL) provided in the TrueMethyl kit (Cambridge Epigenetics); the solution was incubated at 40 °C for a futher 30 min. Samples were purified using Micro-Biospin P-6 columns in SDS buffer (BioRad). To the eluted solution was added (+)-biotinamidohexanoic acid hydrazide BH (10 mM) and pH 7 sodium phosphate buffer (40 mM), and the mixture

was incubated at RT for 4 hr followed by work-up with DNA Clean and Concentrator 5 (Zymo Research) or Biospin-P6 columns in SDS buffer (BioRad). DNA fragments were subjected to the general affinity-enrichment procedure. DNA fragments were eluted off beads by incubating with (0.05% v/v NH_2OH), pH 7 sodium phosphate buffer (40 mM) and *p*-anisidine (100 mM), and heating at 40 °C for 2 hr. DNA fragments were purified by GeneJet PCR purification kit (ThermoFischer) or DNA Clean and Concentrator 5 (Zymo Research). PCR of enriched fragments was achieved using NEBNext Ultra II Q5 Master Mix (NEB).

7.4. Chapter 4 Materials and Methods

Artificial 5-hmU feeding experiment in *T.Brucei* Procyclic trypanosomes were cultured in DTM media (20 mL) in the presence or absence of 5-hmU mononucleoside (1 mM) for at least 48 hr. Cells were left to proliferate without dilution at 27 °C in an atmosphere of 5% CO_2 for 2, 4 or 6 days.

Differentiation by cis-aconitate/citrate in *T.Brucei* (*performed by Dr J. Freitas, Department of Biochemistry, Carrington lab, University of Cambridge*) Cis-aconitate/citrate differentiation was initiated using a procedure reported by Ziegelbauer *et al.*[298] Bloodstream *T.brucei* cells in mid-log phase were resuspended in diffrentiating trypanosome medium (DTM) supplemented with cis-aconitate (3 mM) and sodium citrate (3 mM). Cells were left for 96 hr at 27 °C in an atmosphere of 5% CO_2. Cell number was also determined at each timepoint to monitor cell proliferation.

Differentiation by cold-shock in *T.Brucei* (performed by Dr J. Freitas, Carrington lab, Department of Biochemistry, University of Cambridge) Cold-shock differentiation was performed using a procedure reported by Engstler *et al.*[204] Bloodstream *T.brucei* cells in mid-log phase were suspended in DTM and incubated at 20 °C for 16 hr. After 16 hr, cis-aconitate (6 mM) was added to the culture medium and cells were left to proliferate at 27 °C for 96 hr in an atmosphere of 5% CO_2. Cells were harvested, at 24 hr time points, and cell number was determined to as a measure of cell proliferation.

DTM Media Composition DTM media was used as described in Reference [[205]]. 1 x DTM for media: 10 x DTM (10 % v/v), H_2O (90% v/v), NaCl (6.8 g/L), HEPES (7.5 g/L), $NaHCO_3$ (2.2g/L), glutamic acid (240mg/L), glutamine (1.34 g/L), glycerol (720 mg/L), proline (640 mg/L), mercapoethanol (0.0014 % v/v), haemin (0.3% v/v), heat inactivated feline bovine serum (15 % v/v), 100 x penicillin-streptomycin (1% v/v), adjusted to pH = 7.2 using 5M NaOH where 10x DTM: KCl (4 g/L), CaCl (2 g/L), $H_2NaO_4P{\cdot}H_2O$ (1.4 g/L), $MgSO4{\cdot}7H_2O$ (2 g/L), sodium pyruvate (1.1 g/L), phenol red (100 mg/L), alanine (90 mg/L), arginine (1.26 g/L), asparagine (150 mg/L), aspartic acid (140 mg/L), cysteine (240 mg/L), glutamic acid (150 mg/L), glutamine (3 g/L), glycine (80 mg/L), histidine-$HCl{\cdot}H_2O$ (420 mg/L), isoleucine (520 mg/L), leucine (520 mg/L), lysine-HCl (730 mg/L), methionine (150 mg/L), phenylalanine (320 mg/L), proline (60 mg/L), serine (110 mg/L), threonine (480 mg/L), tryptophan (100 mg/L), tyrosine (360 mg/L), valine (460 mg/L), hypoxanthine (140 mg/L in NaOH 1% v/v))

LC-MS/MS measurements Global quantification of Base J and 5-hmU were performed as described in Materials & Methods – Chapter 2.

HEK293T cell culture HEK293T cells (ATCC) were cultured in DMEM culture media (Thermo Fischer) supplemented with FBS (10%) and penicillin-streptomycin (100U, Thermo Fischer). Cells were washed with PBS and detached using 0.05% Trypsin (Thermo Fischer).

esiRNA Transfection Cells were transfected either transfected in either: a) 6-well plate or b) T75 flask using either RLuc or SMUG1 MISSION esiRNA (Sigma Aldrich). esiRNA in OptiMEM (Thermo Fischer) (a) 1 μg in 125 μL , b) 6 μg in 500 μL) was added to Lipofectamine 3000 (Thermo Fischer) in OptiMEM (a) 7.5 μL in 125 μL, b) 45 μL) in 500 μL) and incubated for 5 min before being added supplemented to the DMEM culture media (a) 1.75 mL, b) 9mL). Cells were grown in the presence of transfection agents for 48 hr or 96 hr.

Protein extraction and quantification Cell pellets were lysed in ice-cold RIPA buffer (Thermo Fischer) and vortexed and incubated for 15 min on ice. The mixture was centrifuged and supernatant containing protein extract was retained. Protein lysate was quantified using Pierce BCA Protein Assay Kit, where proteins were compared to BCA standards by visualization at 562 nm on a SPECTROstar^{Nano} (BMG Labtech).

RT-qPCR Reverse transcription was achieved using High Capacity Reverse Transcription Kit (Thermo Fischer) with (800 ng – 1 μg) input RNA using random primers, following the manufacturer's instructions. Resultant cDNA (1 μL) was added to a mixture of Brilliant III Ultra-Fast SYBR Green qPCR Master Mix (5 μL) (Agilent Technologies), forward and reverse primers for actin or smug1 (1 μM) (Appendix, Chapter 4 - Table 42) and diluted with water to give a final volume of 10 μL. SMUG1 mNA expression was determined by relative quantification to actin mRNA expression.

Western blot Western blot was performed using Wes (Protein Simple) utilising the 12-230 kDa Wes Seperation Module and anti-rabbit detection module. 0.73 μg of protein extract was loaded per sample, where Rabbit monoclonal Anti-SMUG1 antibody (abcam - ab192240, 1 in 50 dilution) and β-Tubilin antibody (Cell signalling, #2146, 1 in 50 dilution) were mixed for detection and quantification of SMUG1 and tubulin

respectively. Relative protein expression was quantitatively determined using Compass for Simple Western software.

5-hmU and 5-fU chemical enrichment-sequencing Fragmented DNA samples (1 μg) and spike-in ODNs (fU-ODN/hmU-ODN, fC-ODN/hmC-ODN, GCAT-ODN; 100 pg) were prepared for NGS as described as described in Chapter 3 Materials and Methods. For 5-fU chemical-enrichment sequencing, 14 cycles of PCR generated adequate DNA for sequencing. For 5-hmU chemical-enrichment sequencing, 18 cycles of PCR generated adequate DNA for sequencing.

Chemical base J enrichment-sequencing[202] Fragmented DNA samples (250 ng-500 ng), $NaIO_4$ (50 mM) and sodium acetate buffer (50 mM, pH 5.5) were incubated at 40 °C for 1 hr followed by work-up with DNA Clean and Concentrator 5 (Zymo Research). The resultant DNA solution (25 μL), pH 6 phosphate buffer (40 mM), (+)-biotinamidohexanoic acid hydrazide BH (10 mM) and *p*-anisidine (2 mM) were mixed and incubated at 40 °C for 12 hr. The samples were subjected to the general NGS library preparation procedure, and Base J-containing DNA fragments were affinity-enriched using Dynabeads MyOne Streptavidin C1 (ThermoFischer). PCR of enriched fragments was achieved on beads using KAPA HiFi Uracil + Polymerase (KAPA Biosystems); 10 cycles rendered adequate DNA for sequencing. Control libraries were prepared following all the above steps in the absence of the $NaIO_4$ oxidation step.

5-hmU DIP sequencing[202] Fragmented DNA samples (1 μg) and spike-in ODNs (hmU-ODN, hmC-ODN2, GCAT-ODN; 500pg) were firstly subjected to general NGS library preparation procedure. The ligated sample was purified via Ampure beads (Beckmann Coulter). Salmon sperm was added (1 μL, 10mg/mL) to the ligated mixture and DNA was denatured via heating to 95 °C for 10 min before snap-cooling on ice. To the solution was added 2 x binding buffer (0.2% Tween 20 in PBS, 50 μL), anti-hydroxymethyluridine antibody ab19735 (10 μL, abcam) and rabbit anti-goat IgG H&L ab6697 (5 μL). The solution was incubated at 4 °C for 16 hr before incubation with Dynabeads Protein G (100 μL, Life Technologies) in binding buffer (0.1% v/v tween 20 in PBS) for 2 hr. The supernatant was removed and the beads were washed with binding buffer (x5). The magnetic beads were then suspended in elution buffer (100 μL, 50mM Tris-HCl, pH 8, 10mM EDTA, 0.5% SDS, 40 μg Proteinase K) and heated at 50 °C for 2 hr at 1300 rpm. The supernatant was removed from the magnetic beads and purified using DNA Clean and Concentrator 5 (Zymo Research). The eluted fragments were subjected

to PCR using NEBNext Ultra II Q5 Master Mix (NEB); 14 cycles rendered adequate DNA for sequencing. For control libraries, control Goat IgG, polyclonal isotype control (ab37373) was utilised in the place of anti-hydroxymethyluridine antibody.

NGS Sequencing Base J enrichment sequencing was ran on a MiSeq instrument in single-end mode, with a read-length of 150. 5-fU chemical-enrichment sequencing, 5-hmU chemical-enrichment sequencing and 5-hmU-DIP enrichment-sequencing and RNA-seq libraries were ran on a NextSeq instrument in paired-end, using a High output kit with a readlengths of 75.

SMUG1 treatment for "no-hmU" DIP-control To fragmented genomic DNA (1 μg) and spike-in contols (hmU-DNA, hmC-DNA2, GCAT-DNA, 500 pg) was added hSMUG1 (3 μL, 15U, NEB) and NEB buffer 1 (3 μL) to make a total reaction volume of 30 μL, and the mixture was incubated at 37 °C for 18 hr before purification using DNA Clean and Concentrator 5 (Zymo Research). The resultant mixture was then subject to the hmU-DIP sequencing protocol.

Sequencing data visualisation Sequencing data was visualised and screenshots were taken from IGV version 2.3.69.

Analysis of Base J chemical-sequencing data *(performed by Dr F. Kawasaki, Balasubramanian group)* Sequencing reads were trimmed to remove adaptors using cutadapt version 1.11. Reads were aligned to the trypanosome genome using bwa mem version 0.7.10.2. Primary alignments with mapping of >10 were retained. Filtering and manipulations were performed using samtoools version 1.1.3. The difference between samples fed with 5-hmU mononucleoside and control samples were quantified and analysed using Deeptools.

Analysis of T-modification enrichment data (*performed by Dr S. Martinez-Cuesta, Balasubramanian group*). The quality of raw FASTQ sequencing libraries were firstly checked using FastQC. Adaptors were trimmed and filtered, base quality was performed using cutadapt (-m 15, -q 20). Reads were aligned to the human genome hg19 using bwamem, duplicate reads were removed using sambamba. Data from different sequencing lanes were merged, alignments were filtered and indexed using samtools. Unmapped, duplicate reads, and mapping with low quality mapping of >10 wee

removed. Reads aligning to blacklisted regions were filtered (https;//sites.google.com/site/anshulkundaje/
projects/blacklists). Spike-in sequences were added to the hg file, and enrichment of spike-ins was obtained via samtools. Bam files were converting to tdf files for visualisation using intergrative genomics viewer (IGV) using the igvtools function.
Peak calling against the input library was determined using MACS2. The intersection of peaks was determined using bedtools, along with the depth of coverage of alignment files within called peaks. Nucleotide sequences within peaks were obtained using bedtools and getfasta. Motif analysis was performed using the online version of MEME-ChiP. Genomic regions of called peaks were determined using GAT using "-ignore-segment-tracks" and "—num-samples=10000', human genome annotations were obtained using the CGAT code applied to GRCh37. Unix tools (awk, cat, sort and uniq) were used for downstream file manipulation. Intersections of hmU-loci with different genomic datasets, and intersection compared to random shuffling of regions was performed using bedtools. Visualiation of DNaseI datasets were obtained using deeptools. Conversion of hg18 to hg19, to determine intersections with TET-binding peaks, was performed using liftOver.

Analysis of RNA-seq data (*performed by Dr S. Martinez-Cuesta, Balasubramanian group*) The quality of FASTQ sequencing files was determined via FASTQC. Illumina adaptors were trimmed, and base quality was called as for the analysis of T-modification enrichment data above. Reads were aligned to the human reference genome hg19 using tophat2. Data from different sequencing lanes was merged, alignments were filtered and indexed using samtools, low quality alignments (< 10) were removed. Read count of exons in hg19 was performed using htseq-count using options "-s no: and "-m intersection-strict", indicating that data is not from a strand-specific assay. Only reads from exons were kept. Differential gene expression between SMUG1 and RLuc libraries, along with statistical analysis, was performed using the R programming language, edgeR and GenomicFeatures. R packages data.table and reshape2 were used to manipulate matrices and tables and ggplot2 was used for data visualisation.

7.5. Chapter 5 Materials and methods

Nuclear protein extract (*performed by Dr S. Huber, Balasubramanian group*) Cells were harvested in PBS and resuspended in Hypotonic buffer (500 μL, 20 mM Tris-HCl pH 7.4, 10 mM NaCl, 3mM $MgCl_2$) in the presence of NP40 detergent (25 μL, 10%) before

centrifugation to separate the cytoplasmic extract (supernatant) and nuclear fraction (pellet). The nuclear pellet was resuspended in Cell Extraction buffer (10 mM Tris-HCl pH 7.4, 2mM Na_3VO_4, 100 mm NaCl, 1% Triton X-100, 1 mM EDTA, 10% glycerol, 1 mM EGTA, 0.1% SDS, 1mM NaF, 0.5% deoxycholate, 20 mM $Na_4P_2O_7$). Proteins were quantified using the Pierce BCA Protein Assay Kit, where proteins were compared to BCA standards by visualization at 562 nm on a SPECTROstar^{Nano} (BMG Labtech).

Protein pulldown Protein pulldown was carried out as done by Reik and co-workers, with some modifications.[49] Thermo Pierce streptavidin beads (10 μL) were firstly washed in buffer A (1 x PBS, 0.1% Triton X-100, x3) and incubated with 1 μg of biotinylated DNA fU-DNA1 or GCAT-DNA1 at RT for 1 hr. The beads were washed in buffer A, and were resuspended in buffer B (0.2 mM EDTA, 20% Glycerol, 20mM HEPES-KOH pH 7.9, 0.1M KCl, 1mM DTT, 1mM PMSF, 0.1% Triton X-100) and incubated with 100 μg HEK293T nuclear extract for 1 hr. The beads were subsequently washed with buffer B (x 6), once with PBS, and then heated in 1 x LDS loading buffer supplemented with 4 mM DTT at 95 °C for 5 min. The eluted proteins were then removed from the beads and ran on a 4-20% SDS-PAGE gel in MOPS buffer. The gel was stained with Coomassie blue, each lane was cut into 8-pieces and submitted for proteomics analysis.

Proteomic downstream analysis To determine enriched proteins for fU-DNA1 and GCAT-DNA1, a t-test ($p < 0.1$) was performed using Scaffold software (Version 4.4.8) to determine log_2fold change between modified and non-modified DNA total mass spectral counts. Low scoring matches were removed and no multiple test corrections were performed. A minimum of 2 peptides was required for identification. Using 95% protein and 50% peptide probability thresholds, as determined by algorithms in the Scaffold software, the false discovery rate was estimated to be 0.0%. Common contaminants such as keratin were discounted.

UNG excision fU-ODN or U-ODN (500 ng) (Appendix, Table 19) was incubated in the presence of UNG buffer (3 μL) in the presence of UDG (50 U, NEB) and was incubated at 37 °C for 18 hr. Analysis of 5-fU excision was performed via general LC-MS oligomer analysis.

hSMUG1 excision fU-ODN (500 ng) was incubated in the presence of UNG buffer (3 μL) in the presence of SMUG1 (25 U, NEB) and was incubated at 37 °C for 18 hr. Analysis of 5-fU excision was performed via general LC-MS oligomer analysis.

7.6. Chapter 6 Materials and Methods

fU-ODN crosslinking conditions fU-ODN (500 ng) was incubated in the presence of a) lysine, b) glycine, c) 5-aminovaleric acid, d) guanidium hydrochloride (10 mM) and $NaBH_3CN$ (25 mM) in PBS in a total volume of 25 μL. The reaction mixture was purified by mini quick spin column (Roche) or oligo clean and concentrator kit (Zymo Research). Reaction analysis was performed via general LC-MS oligomer analysis.

fC-ODN crosslinking conditions fC-ODN (500 ng) was incubated in the presence of a) lysine, b) glycine, c) 5-aminovaleric acid, d) guanidium hydrochloride (500mM) in the presence or absence of $NaBH_3CN$ (100 mM) in PBS in a total volume of 25 μL. The reaction mixture was purified by mini quick spin column oligo clean and concentrator kit (Zymo Research). Reaction analysis was performed via general LC-MS oligomer analysis.

Model protein crosslinking fU-DNA1, fC-DNA1 or T-DNA1 (1 μg) labelled with Fluorescein was incubated with of BSA (5.88 μg) or Ribonuclease A (1.21 μg) in the presence of $NaBH_3CN$ (25 mM) at 4 °C for 18 hr, or $NaBH_3CN$ (100 mM) at 37 °C for 18 hr. The mixture was purified by P-6 Micro Bio-spin columns (Bio-Rad) before gel electrophoresis on a denaturing 12% Bis-Tris Nu-page gels, which were ran in MES buffer.

Crosslink with model proteins for mass spectrometry fU-ODN or fC-ODN2 (400 ng) was incubated with Ribonuclease A (1.8 μg) in the presence of $NaBH_3CN$ (25 mM) at 4 °C for 18 hr, or $NaBH_3CN$ (100 mM) at 37 °C for 18 hr. DNA-protein crosslink was analysed by electrospray mass spectroscopy. The sample was ran on a ACQUITY UPLC® Protein BEH C4 Column (Waters) before injection into a Xevo-G2-S Q-Tof, using solvents A (0.1% formic acid in water) and B (0.1% formic acid in 95:5 MeCN:water), at a flow-rate of 0.2 mL/min, with a gradient 5%–95% B, increasing at 1% per min. Mass was acquired between 300-5000 m/z.

Nucleosome assembly (*performed by Z. Li or Dr E.A. Raiber, Balasubramanian group*)
Nucleosomes were assembled either using chaperones (Chromatin Assembly Kit, Active Motif) for the crosslinking and proteomics studies, or via Salt Dilution (EpiMark Nuclesome Assembly Kit, NEB) for the polymerase stop assay, according to manufacturer's instructions.

Crosslinking formylated Widom nucleosome Nucleosomes composed from Widom DNA (5-fC, C, 5-fU and 5-caC) labelled with Cy3 and Cy5 labels, and previously stained with GelRed, were incubated at 37 °C for 18 hr in the presence of $NaBH_3CN$ (100 mM). The mixture was purified using P-6 Micro-spin columns (Bio-Rad) and crosslinking formation was monitored via gel electrophoresis, using a denaturing 12% Bis-Tris Nu-page gel ran in MES buffer.

Optimisation of 5-fU nucleosome crosslinking 5-fU nucleosome composed from Widom DNA labelled with Cy3 and Cy5 were incubated at either 37 °C or 4 °C for 3 hr or 18 hr in the presence of varying concentrations of $NaBH_3CN$ (100 mM, 10 mM, 1 mM). After reduction, the mixture was purified using P-6 Micro-spin columns (Bio-Rad) before analysis by gel electrophoresis on a denaturing 12% Bis-Tris Nu-page gel ran in MES buffer.

Proteomics analysis 5-fU or 5-fU crosslinked nucleosome shifted gel-bands, and C band corresponding to the same molecular weight were cut and submitted for proteomics analysis. Protein threshold was set to 95% and a peptide threshold of 50%, with a mimimum of one peptide fragment for discovery. H4 and H2B were identified in the case of 5-fC in n = 3 crosslinked replicates, H3 was found in 2/3 replicates. No histone proteins were identified for the C-control and 5-fU band. Protein coverage was determined using Scaffold software. % Lysine coverage was determined by number of lysine residues in peptide fragments identified by proteomics/the total number of lysines within the histone subunit.

RecJF optimisation 5'-Phos-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT (10 mM) and 3'-TCTGCACACGAGAAGGCTAG-5'-Phos (10 mM) were annealed in 10mM Tris-HCl, 50mM NaCl and 1mM EDTA buffer by heating to 95 °C followed by cooling to 4 °C. 400 ng of annealed DNA was treated with $RecJ_F$ (3 μL, 30U or 9 μL 90U, NEB) in the presence of NEB 2 buffer (3 μL) in a total volume of 30 μL. The resultant mixture was ran on a 10% TBE-Urea denaturing gel.

Polymerase stalling experiment by NGS sequencing Nucleosomes composed of 5-fC Widom DNA were incubated at 37 °C for 18 hr in the presence 100mM $NaBH_3CN$. The mixture was purified by Micro Biospin P6 Tris columns (BioRad). Free 5-fC widom DNA was used as a control. Polymerase extension was achieved in the presence of dNTP (200 μM), fw or rev primer (0.5 μM), 1 x polymerase buffer and DreamTaq (1 μL, 5U) and was

heated at 95 °C for 30 s, 60 °C for 60 s and 72° C for 3 min. After extension, the mixture was treated with Proteinase K (40 µg) in Proteinase K buffer at 37 °C (750 mM Gu-HCl, 5% Tween 20, 30mM EDTA, 30mM Tris-HCl), before being purified by Oligo clean and concentrator Kit (Zymo Research). The samples were then treated with RecJF (3 µL, 90U) for 12 hr at 37 °C to remove excess primer and overhangs. The samples were repurified by Oligo clean and concentrator Kit (Zymo Research) before being ligated for DNA sequencing using NEB NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB) in the presence of standard Illumina adaptors (2.5 µL), and subject to 6 cycles of PCR using NEB Ultra II Q5 Master mix. Libraries were sequenced on a MiSeq instrument in single-end with 150 cycles.

Polymerase stalling by NGS data analysis *(performed by Dr S Martinez-Cuesta, Balasubramanian group).* BCL files were converted to FASTQ using bcl2fast1. Trimming and quality check of sequencing reads was performed as for T-modification enrichment sequencing libraries. Trimmed reads from the libraries were aligned to forward and reverse strands of the Widom sequence using bowtie2. Alignments were filtered and sorted using samtools. Counting of truncated sequences was performed using samtools. Unix tools and python were used to transform output files to tables. Data was analysed using the R programming software and statistical analysis was performed using edgeR and data.table. A moving average (±1) of fold-change was plotted against the template position of the Widom sequence for visualisation of stalling pattern.

Molecular Dynamics simulations of nucleosomal particles (*performed and written by Dr G. Portella, Balasubramanian group)* Molecular dynamics (MD) simulations were completed using Gromacs-4.5 software[299] and periodic boundary conditions and the particle mesh Ewald method[300] for long-range electrostatics. The short-range repulsive and attractive dispersion interactions were modelled using a Lennard-Jones potential with a cut-off of 1.0 nm. The Settle algorithm[301] was used to constrain bond lengths and angles of water molecules, and P-Lincs[302] was used for all other bond lengths, in combination with virtual interaction-sites[299,303] to remove the hydrogen vibrations and therefore use a time step of 4 fs. The temperature was kept constant as described in Bussi et al.[304]. The pressure was kept constant and it was controlled by coupling the simulation box to a pressure bath of 1 atm[305]. The amber99SB*-ildn[306] force-fields was used to describe the histone tails, and the amber99+parmBSC0[307] force field was used for the nucleosomal DNA. The solvent was modeled using the TIP3P water model, the sodium and chlorine ion were modeled using Dang's parameters[308], and manganese

atoms parameters taken from the Amber force field database. Two dinucleotide model initial conformations were built following the protocol described in Collepardo et al.[309], by stacking two nucleosome particles (X-ray structure with PDB code 1KX5)[310] on top of each other. In one model the inter-nucleosomal distance, as measured by the vector connecting the center of mass of the two nucleosomal was set to 6 nm, and in the second model was set to 7 nm. The histone tail sequences were replaced by the human sequences. The di-nucleosome systems were embedded in a truncated octahedron box containing ~200,000 water molecules, leaving 2 nm between the nucleosome atoms and the edges of the box. This separation is large enough to accommodate a fully extended H3 tail, which is the longest one. Approximately 900 sodium ions and 600 chlorine ions were added to balance the nucleosome charge and give an ionic concentration of 150 mM NaCl (the exact values depend on the model). Each di-nucleosome system was energy minimized and simulated twice (using two different random seeds) for 1.15 microsecond. To facilitate the analysis the MD trajectories of the di-nucleotide systems were split into individual nucleosomes, imaged to remove periodic boundary crossings, and then concatenated into one long trajectory containing one nucleosome. To alleviate auto-correlation effects, we analysed frames with a 2ns frequency. After discarding first 100 ns of each trajectory, the resulting concatenated trajectory contained 4200 structures. For each nucleotide, we collect the set of lysine residues whose side chain atoms are found within a cut-off distance of 1.2nm with respect to the nucleotides in any of the analysed frames. From these set of distances, we compute for each pair of reference position – lysine side-chain a time/ensemble averaged contact metric by means of a continuous switching function. $1/\left(1+e^{\left(b*(x-1.5*d0)\right)}\right)$ was used as such switching function, where b=10 and d0=0.5 and x represents the minimum distance between any atom in the lysine side chain with the reference nucleotide. The parameters were empirically chosen such that distances below 0.5 nm result in a value of ~1, and anything above 0.5 nm decays to 0 (at ~1 nm is almost zero). As proxy for the orientation of duplex DNA strands with respect to the nucleosome core, the phase angle φ the angle between a v_{bp} vector centered at the base pair centre of mass and pointing towards the minor groove with the vector connecting the centre of mass of the base pair and the centre of mass of the nucleosome. The vector v_{bp} was defined as the sum of the two vectors connecting the center of mass with the N9(Y)/N1(R) for a given base pair. The resulting curve was further refined by fitting a sinusoidal curve, maximum and minimum values were rescaled to the [-1, 1] range for visualization purposes. In our definition, a large φ value associated to a given base pair reports locations where the major groove faces the histone core.

8. Appendix

8.1. General

ODN	MW	ESI-MS
fU-ODN	3039.5	$[M-2H]^{2-}$ = 1519.3
fC-ODN	6128.1	$[M-3H]^{3-}$ = 2041.4 $[M-4H]^{4-}$ = 1530.8
hmU-ODN	3121.5	$[M-2H]^{2-}$ = 1560.3
GCAT-ODN	3105.5	$[M-2H]^{2-}$ = 1552.3
hmU-ODN + $KRuO_4$	3119.5	$[M-2H]^{2-}$ = 1559.2
fU-ODN + ARP	3354.7	$[M-2H]^{2-}$ = 1676.8
fU-ODN + BH	3392.3	$[M-2H]^{2-}$ = 1695.8
fU-ODN + *o*-phenylenediamine	3127.6	$[M-2H]^{2-}$ = 1563.3
fU-ODN + *o*-phenbiotin	3413.5	$[M-2H]^{2-}$ = 1705.3
fU-ODN + NH_2OMe	3068.6	$[M-2H]^{2-}$ = 1533.8
fC-ODN + ARP	6438.2	$[M-3H]^{3-}$ = 2145.7 $[M-4H]^{4-}$ = 1609.0
fC-ODN + BH	6478.3	$[M-3H]^{3-}$ = 2159.1 $[M-4H]^{4-}$ = 1619.1
fC-ODN + *o*-phenylenediamine	Benzimidazole = 6213.1 Intermediate = 6215.2	$[M-3H]^{3-}$ = 1552.8 $[M-4H]^{4}$ = 1553.4
fC-ODN + *o*-phenbiotin	6497.2	$[M-3H]^{3-}$ = 2165.4 $[M-4H]^{4}$ = 1624.8
hmU-ODN + $KRuO_4$ + BH	3472.7	$[M-2H]^{2-}$ = 1735.8
U-ODN	3035.7	$[M-3H]^{3-}$ = 1311.2
AP-ODN	3840.7	$[M-3H]^{3-}$ = 1279.9
AP-ODN + $KRuO_4$	Fragment 1 = 1566.3 Fragment 2 = 2174.4	$[M-H]^{-}$ = 1565.3 $[M-2H]^{2--}$ = 1086.2
AP-ODN + ARP	4153.8	$[M-3H]^{3-}$ = 1383.9
AP-ODN + BH	4193.9	$[M-3H]^{3-}$ = 1397.3
AP-ODN + *o*-phenylenediamine	3929.7	$[M-3H]^{3-}$ = 1309.2
hmU–ODN	3039.5	$[M-2H]^{2-}$ = 1520.3
hmU-ODN + DMT-MM	+1= 3260. +2 = 3383.6	$[M-2H]^{2-}$ = 1629.8 $[M-2H]^{2-}$ = 1699.2
GCAT-ODN + DMT-MM	+1=3244.5 +2 = 3383.6	$[M-2H]^{2-}$ = 1621.8 $[M-2H]^{2-}$ = 1691.3
hmU-ODN2 + DMT-MM	3180.6	$[M-2H]^{2-}$ = 1589.8
fU-ODN + NH_2OH	3054.5	$[M-2H]^{2-}$ = 1526.8
fC-ODN + NH_2OH	6140.1	$[M-4H]^{4}$ = 1534.5
fU-ODN + SMUG1/UNG	2997.5	$[M-2H]^{2-}$ = 1498.3
fU-ODN + Lysine + $NaBH_3CN$	3169.6	$[M-2H]^{2-}$ = 1584.3
fC-ODN + Lysine + $NaBH_3CN$	6255.2	$[M-4H]^{4}$ = 1562.8
fC-ODN + Lysine	6253.2	$[M-4H]^{4}$ = 1563.3
fU-ODN + Girard's T	3153.6	$[M-2H]^{2-}$ = 1575.8
fU-ODN + Glycine + $NaBH_3CN$	3096.6	$[M-2H]^{2-}$ = 1548.8
fC-ODN + Glycine + $NaBH_3CN$	6184.1	$[M-4H]^{4}$ = 1545.5
fU-ODN + 5-Aminovaleric acid + $NaBH_3CN$	3152.6	$[M-2H]^{2-}$ = 1569.8
fC-ODN + 5-Aminovaleric acid + $NaBH_3CN$	6240.2	$[M-4H]^{4}$ = 1556.0

Table 17: Mass data for ODNs used in this thesis and their reaction products

	Sequence
Widom Template	5'-ATCGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGC ACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCCAGGCACGTGTC AGATATATACATCCGAT-3'
Template 1	5'-TTCTTGGCTGTGGCTCTGCGTCCTTGTCCTGCCCACTGCCTGACGGGCGGAGGCACAACAGAGA GCAACACCGCCGAGGA-3'
Template 2	5'-CTAAATCTACTAAATCCTCTAAATCTATTCTATACATGAATCTTAGTTAAAGGTAGTAGTAGT AGATATAAGATGATAGG-3'
Template 3	5'-GCTCGCTTTGTTGGTTTCCTTGTTCTCTGTGCCCACTGCCTGACGGGCGGAAAGCAGCGCGAGC AAGCGAGACAGGACAC-3'
Template 4	5'- GGGGACCCTGGGCAACCAGCCCTGTCGTCTCTCCAGCCCCAGCTGCTCACCATCGCTATCTGAG CAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTC-3'
Template 5	5'-CTAATCTCCAATCCATCCTAATCTCATACTTATTTCTGAACTTTATGATTCTCTAACTACTTAC TTCAACTACTACTCTT-3'
Widom Reverse Primer	5'- ATCGAGAATCCCGGTGCCGA-3' where in some instances the 5' end is modified with Cy5 or Cy3
Widom Forward Primer	5'- ATCGGATGTATATATCTGACACGTGCCTGGAGA-3' where in some instances the 5' end is modified with Fluoriscein, Cy5 or Cy3
Forward Primer 1	5'-TTCTTGGCTGTGGCTCTGCGTCCTTGTCCT-3'
Reverse Primer 1	5'-TCCTCGGCGGTGTTGCTCTCTGTTGTGCCT-3' where in some instances the 5' end is modified with Fluorescein
Forward Primer 2	5'-CTAAATCTACTAAATCCTCTAAATCTATTC-3'
Reverse Primer 2	5'-CCTATCATCTTATATCTACTACTACTACCT-3'
Forward Primer 3	5'-GCTCGCTTTGTTGGTTTCCTTGTTCTCTGT-3'
Reverse Primer 3	5'-GTGTCCTGTCTCGCTTGCTCGCGCTGCTTT-3'
Forward Primer 4	5'-Biotin-GGGGACCCTGGGCAACCAGC-3'
Reverse Primer 4	5'-GAGGTTGTGAGGCGCTGCCC-3' where in some instances the 5' end is modified with Fluorescein.
Forward Primer 5	5'-CTA ATC TCC AAT CCA TCC TAA TCT CA-3'
Reverse Primer 5	5'-CTCTAACTACTTACTTCAACTACTACTCTT-3'

Table 18: Sequence of templates and primers used for PCR synthesis of DNA in this thesis.

ODNs	Sequences
fU-ODN	5'-ATCGCA**fU**GTA-3'
fC-ODN	5'-TAATTATC**fC**GGACTCATAAG-3'
U-ODN	5'-CAGAT**U**TACGATT
AP-ODN	5'-CAGAT**AP**TACGATT
hmU-ODN	5'Phos-ATCGCA**hmU**GTA-3'
fU-ODN2	5'Phos-ATCGCA**fU**GTA-3'
hmU-ODN	5'-ATCGCA**hmU**GTA-3'
GCAT-ODN	5'Phos-ATCGCATGTA-3'
fC-ODN2	5'-ATCG**fC**GCGTA

Table 19: Sequences of ODNs used in this thesis (modifications are highlighted in bold).

8.2 Appendix Chapter 2

Girard's Reagent T Derivatisation of 5-fU and associated improvement in formyl-group detection - *The following work was done in collaboration with a Masters' student J. Cross, under my supervision.*

It had been demonstrated that 5-fU could be quantified using large amounts of extracted DNA following enrichment, however this required lots of genomic material due to the high detection limit of 5-fU, meaning accurate 5-fU measurement remained problematic. To try and alleviate this problem, Wang and co-workers had previously reported that the detection limit for 5-fU could be improved with the use of a derivatising agent, Girard's Reagent T.[311] This molecule bears a positively charged ammonium moiety, and hence lowers the detection limit in positive ion mode MS. Thus, this strategy was investigated to improve the detection of 5-fU in genomic samples, with the aim of being able to detect and quantify 5-fU in smaller quantities (e.g. 1 µg of DNA).

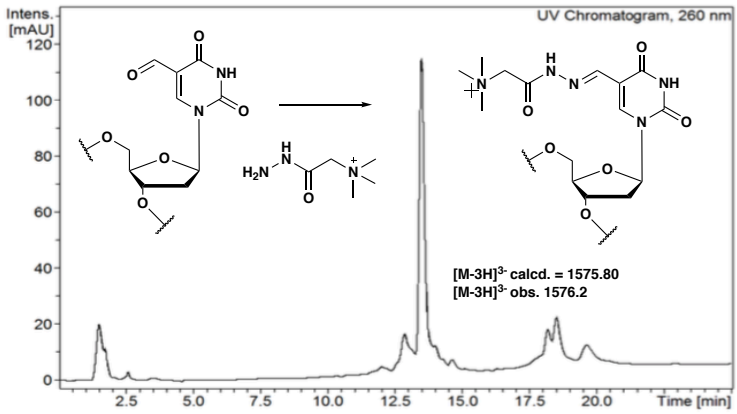


Figure 85: Derivatisation of 5-fU-ODN with Girard's Reagent T.

The reported derivatising strategy was modified so that 5-fU was tagged prior to DNA digestion rather than before LCMS/MS injection. This enabled the removal of any excess derivatising reagent via size exclusion, to avoid injection into the highly sensitive QE spectrometer. Derivatisation conditions were developed on a 10mer containing one 5-fU base (fU-ODN), using analogous 5-fU-tagging chemistry to that developed in Chapter 3 (Figure 85). These reaction conditions were subsequently applied to genomic material before DNA digestion. After adduct formation with Girard's T, 5-fU was detected in ~1 µg of genomic HEK293T sample, highlighting the scope of using derivatization chemistry for low abundance modifications. (Figure 86 - Left).

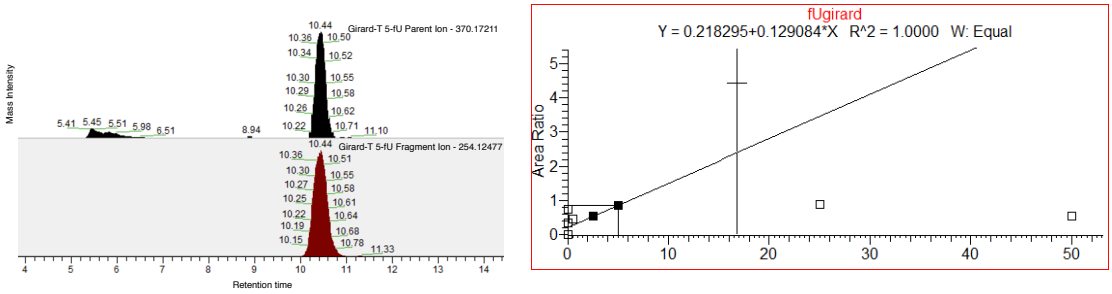


Figure 86: Left: Girard's-fU adduct detected in HEK293T sample, Right: Linear relationship between concentration and mass integration of 5-fU Girard's adduct standard was only attainable within a small concentration range.

Unfortunately, a linear relationship between concentration and mass integration was only observed within a small concentration range for the derivatised nucleoside (Figure 86 - Right). This highlights that the 5-fU derivatisation method cannot be used for accurate quantification at this stage. However, this method could still be a useful tool for relative quantification of 5-fU between samples. When genomic samples were diluted, a corresponding reduction in signal intensity was observed, providing validation for a relative quantification approach.

Accurate quantification could be improved with the use of an isotopically labeled variant of the derivatised base, to be used as a SIL internal standard. Furthermore, alternative derivatising probes should be screened, utilizing the chemistry developed in Chapter 3, which may show an improved linear relationship between concentration and mass signal. Derivatisation approaches would be highly beneficial to improve the detection limit of other low abundance modified bases; furthermore, derivatization could aid the discovery of new modified bases in genomic DNA.

Chapter 2 Supplementary tables and figures

	Parent Ion $[M+H]^+$	Fragment Ion $[M+H]^+$
T	243	127.0502
T-SIL (d_3-T)	246	130.06903
C	228	112.05054
C-SIL ($^{15}N_3$-C)	231	115.04164
mC	242	126.06619
mC-SIL (d_3-mC)	245	129.08502
hmC	258	142.0611
hmC-SIL (d_3-hmC)	261	145.07993
fC	256	140.04545
hmU	259	125.03455
hmU-SIL (^{13}C-d_2-hmU)	262	128.05046
hmU-SIL2 (^{13}C-d_1 - hmU)	261	127.04711
fU	257	141.02947
fU-SIL (^{13}C-d_1 - fU)	259	143.0391
Base J	421	143.04512
A	252	136.06177
G	268	152.05669
fU-Girard-T	370	370.17211 254.12477

Table 20: Parent and fragment ion masses used for detection and quantification via LC-MS/MS. All bases contain the deoxyribose sugar. T, C, mC, A and G standards were purchased from Sigma Aldrich; hmC, fC was sourced from Berry and Associates; C-SIL was purchased from Cambridge Isotope Laboratories; mC-SIL and hmC-SIL were purchased from Toronto Research Chemicals; T-SIL was purchased from Carbosynth; hmU, fU, hmU-SIL, hmU-SIL2 and fU-SIL were synthesised by myself, Base J was synthesised by Dr F. Kawasaki, Balasubramanian group; fU-Girard was synthesised by J. Cross, Balasubramanian group

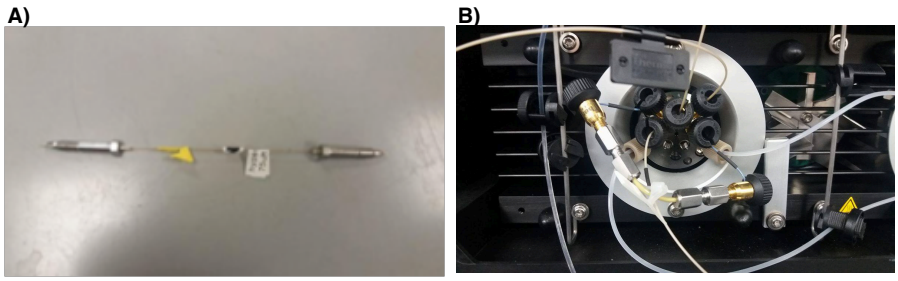


Figure 87: A) 0.075 mm nano-HPLC hypercarb column (custom-packed), B) 0.18 mm capillary KAPA hypercarb column (Thermo Fischer) connected to the Dionex 3000 with nanoviper connectors for more reliable and reproducible chromatography.

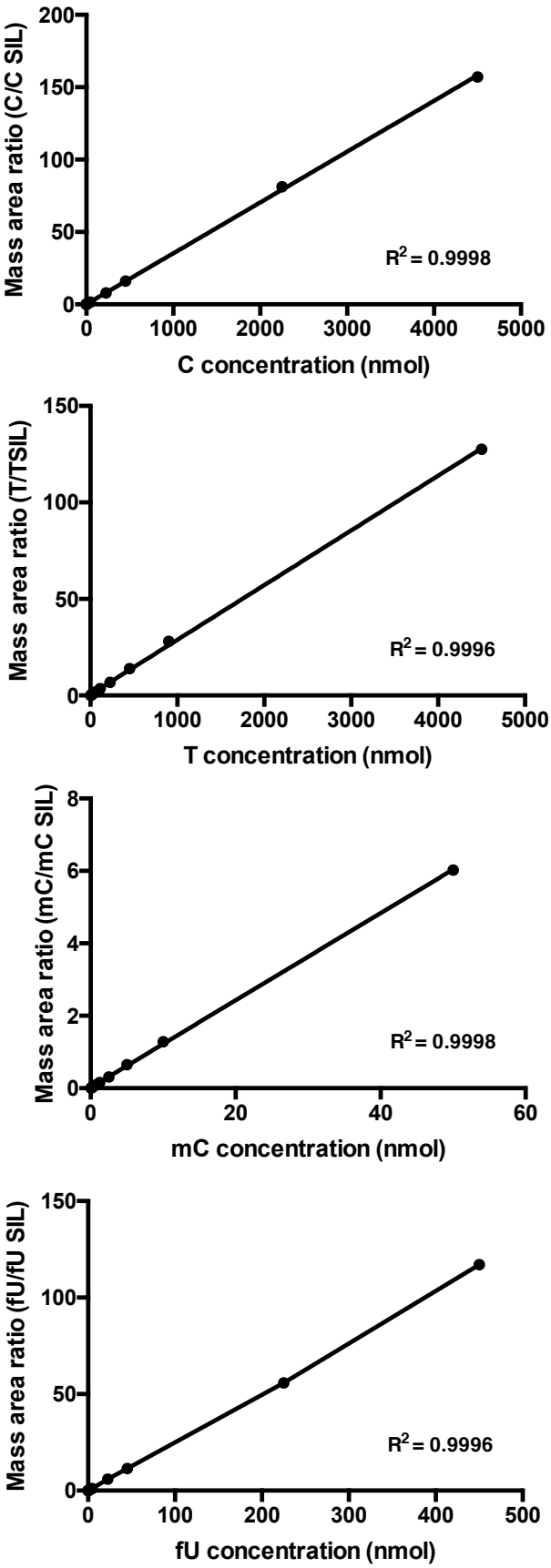


Figure 88: Example calibration lines for C, T, 5-mC and 5-fU LC-MS/MS quantification

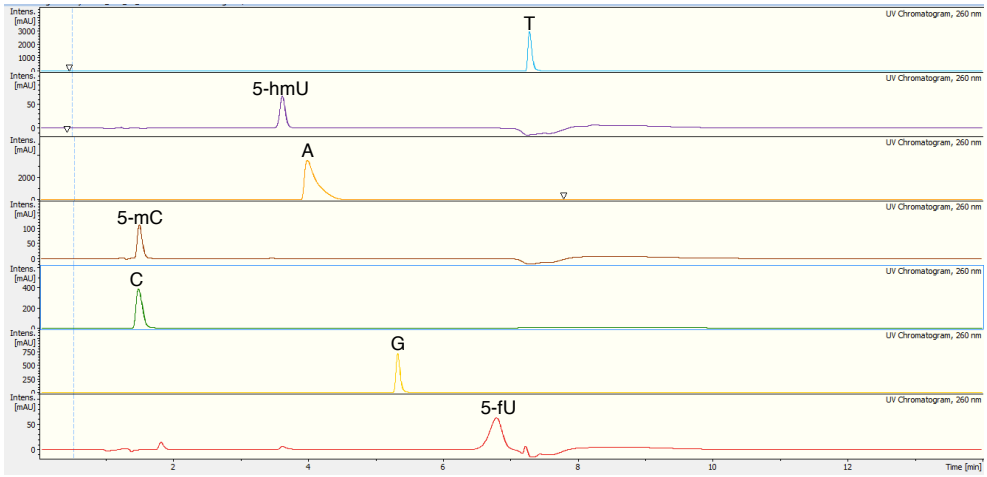


Figure 89: Retention times of canonical nucleosides and modified base standards via UHPLC, measured at 260 nm.

Nucleoside	Collected fraction (retention time)
mC	1.0-2.2
hmU	3.0-4.1
fU	6.0-7.2
hmC	1.0-2.2

Table 21: Timepoints of fraction collection using UHPLC pre-enrichment of DNA modified bases.

	hmU/T	Base J/T
Leishmania major	8.59E-04	2.86E-03
	1.73E-04	2.98E-03
	2.98E-04	2.69E-03
Bio 1	**4.43E-04**	**2.84E-03**
	3.07E-04	3.15E-03
	2.19E-04	3.30E-03
	5.31E-04	3.58E-03
Bio 2	**3.52E-04**	**3.35E-03**
Average	**3.98E-04**	**3.09E-03**
T.brucei BSF	3.65E-04	4.55E-03
Bio 1	3.48E-04	4.50E-03
	3.56E-04	**4.53E-03**
Bio 2	2.24E-04	3.78E-03
	2.57E-04	3.59E-03
	2.41E-04	**3.68E-03**
Average	**2.99E-04**	**4.11E-03**
T.brucei PCF		
Bio 1	7.11E-05	n.d
Bio 2	7.17E-05	n.d
Average	7.14E-05	

Table 22: Accurate 5-hmU and Base J measurements as a proportion of T in trypanosomatids. n.d = not detected.

	mC/N
T.brucei BSF	9.37E-05
	8.82E-05
Average	**9.09E-05**
T.brucei PCF	8.56E-05
	8.15E-05
Average	**8.37E-05**

Table 23: Accurate 5-mC measurements as a proportion of total nucleosides in T.Brucei

	hmU/per N	fU/per N
Archea	8.59E-06	2.41E-06
	8.89E-06	4.46E-06
Average	**8.73E-06**	**3.48E-06**
MCF7	1.94E-06	4.90E-06[1]
Bio 1 Tech 1	2.19E-06	4.77E-06[1]
	4.26E-06	2.98E-06[2]
Bio 1 Tech 2	4.25E-06	3.39E-06[2]
MCF7 Bio 1	**3.16E-06**	**4.01E-06**
	8.78E-07	
Bio 2 Tech 1	9.17E-07	
	1.03E-06	
Bio 2 Tech 2	1.03E-06	
MCF7 Bio 2	**9.64E-07**	
Average	**2.06E-06**	
mESC	5.66E-06	
	6.08E-06	
Average	**5.87E-06**	
A.thaliana	9.81E-05	
	9.82E-05	
Average	**9.82E-05**	
HEK293T	5.44E-06	
	5.10E-06	
	5.37E-05	
Bio 1 Tech 1 Average	**5.30E-06**	
	3.97E-06	
	4.10E-06	
Bio 1 Tech 2 Average	**4.04E-06**	
	1.36E-06	
	1.03E-06	
	1.18E-06	
Bio 1 Tech 3 Average	**1.19E-06**	
	3.76E-06	2.21E-06
	3.59E-06	2.85E-06
Bio 2 Average	**3.67E-06**	**2.53E-06**
HEK293T Average	**3.59E-06**	**2.53E-06**

Table 24: Accurate 5-hmU and 5-fU measurements as a proportion of total nucleosides in a variety of organisms/cell-lines

8.3. Appendix Chapter 3

Direction functionalization of hydroxyl-group

Reaction with EDC

Adduct formation was observed between 5-hmU and EDC by LC-MS, however investigation of the EDC adduct by crude ^{1}H NMR indicated no functionalisation at the 5' hydroxyl group. Only a small 5'' CH_2 shift was seen compared to that observed for the H6 proton (Figure 90). Other carbodiimides have been previously shown to exhibit adduct formation with T and U nucleosides, where functionalisation occurs at the endocyclic nitrogen.[312] It was therefore highly likely that it was the product forming shown in Figure 90-b, rather than the desired functionalization at the 5'' position.

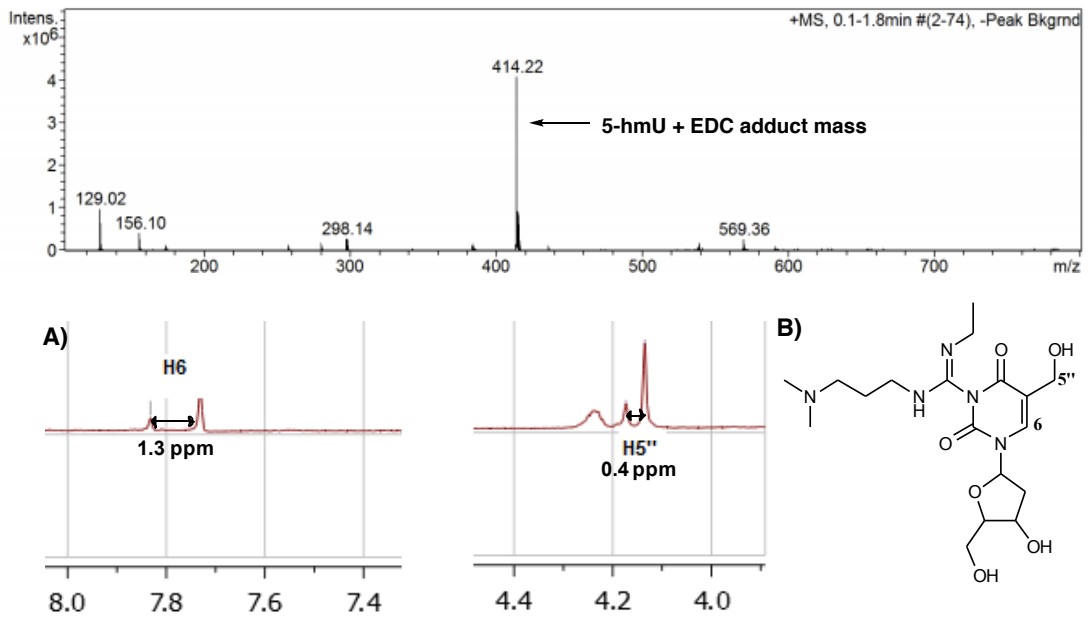


Figure 90: Top: LC-MS trace of hmU-EDC adduct mass. Bottom: a) ^{1}H NMR shows a greater chemical shift difference at H6 rather than H5''. b) Suggested structure of 5-hmU EDC adduct.

Reaction with DMT-MM

Mono-adduct formation of 5-hmU mononucleoside was also observable by DMT-MM by LC-MS analysis, however supplementation with sodium azide led to no formation of target molecule 5-azaU by LC-MS. Due to the excess of coupling reagent, purification was problematic. This reaction was therefore alternatively traced using a single-stranded oligomer which contained one 5-hmU (hmU-ODN). Using this model, a large excess of reagent could be added to the substrate, and facile purification was achieved via size exclusion chromatography. hmU-ODN and GCAT-ODN were treated with DMT-MM substrate, where quantitative conversion was observed in both cases (Figure 91). Since hmU-ODN and GCAT-ODN are both phosphorylated at the 5'-position, the reaction

was repeated with a non-phosphorylated variant of the oligomer hmU-ODN2, which harboured the same sequence. Reaction with unphosphorylated hmU-ODN (hmU-ODN2) revealed only ~1% desired adduct formation. This indicated that functionalization was mainly occurring on the phosphate ester with ODN models, and hence highlighted the unfeasibility of the direct hydroxyl activation approach.

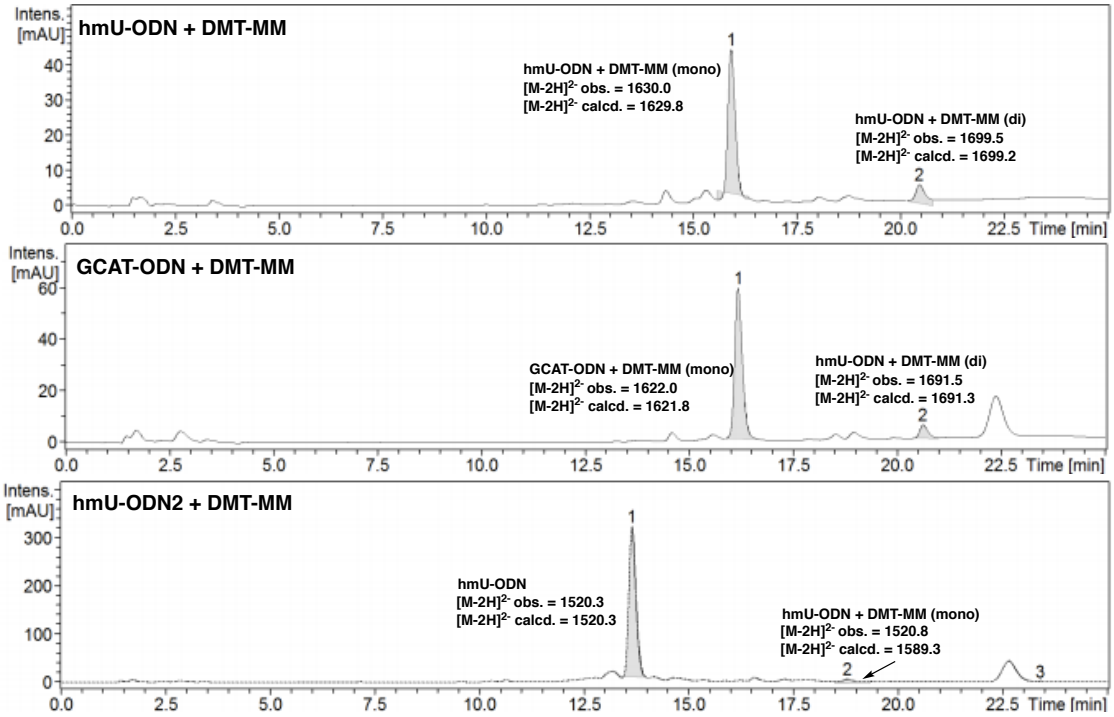


Figure 91: Top: mono and di-adduct formation of DMT-MM with 5'-phosphorylated hmU-ODN, Middle: mono and di-adduct formation of DMT-MM with 5'-phosphorylated GCAT-ODN, Bottom: <1% adduct formation of DMT-MM with hmU-ODN2.

Reaction of fC-ODN with *o*-phenylenediamine derivatives

fC-ODN and incubation with *o*-phenylenediamine saw the observance of two products (Figure 92). Further investigation by LC-MS revealed one product with mass corresponding to the reduced benzimidazole product, while the other corresponded to the non-reduced intermediate. Since both the reduced and non-reduced forms would enable 5-fC enrichment, both these reaction outcomes were minimised to enable chemoselective tagging of 5-fU.

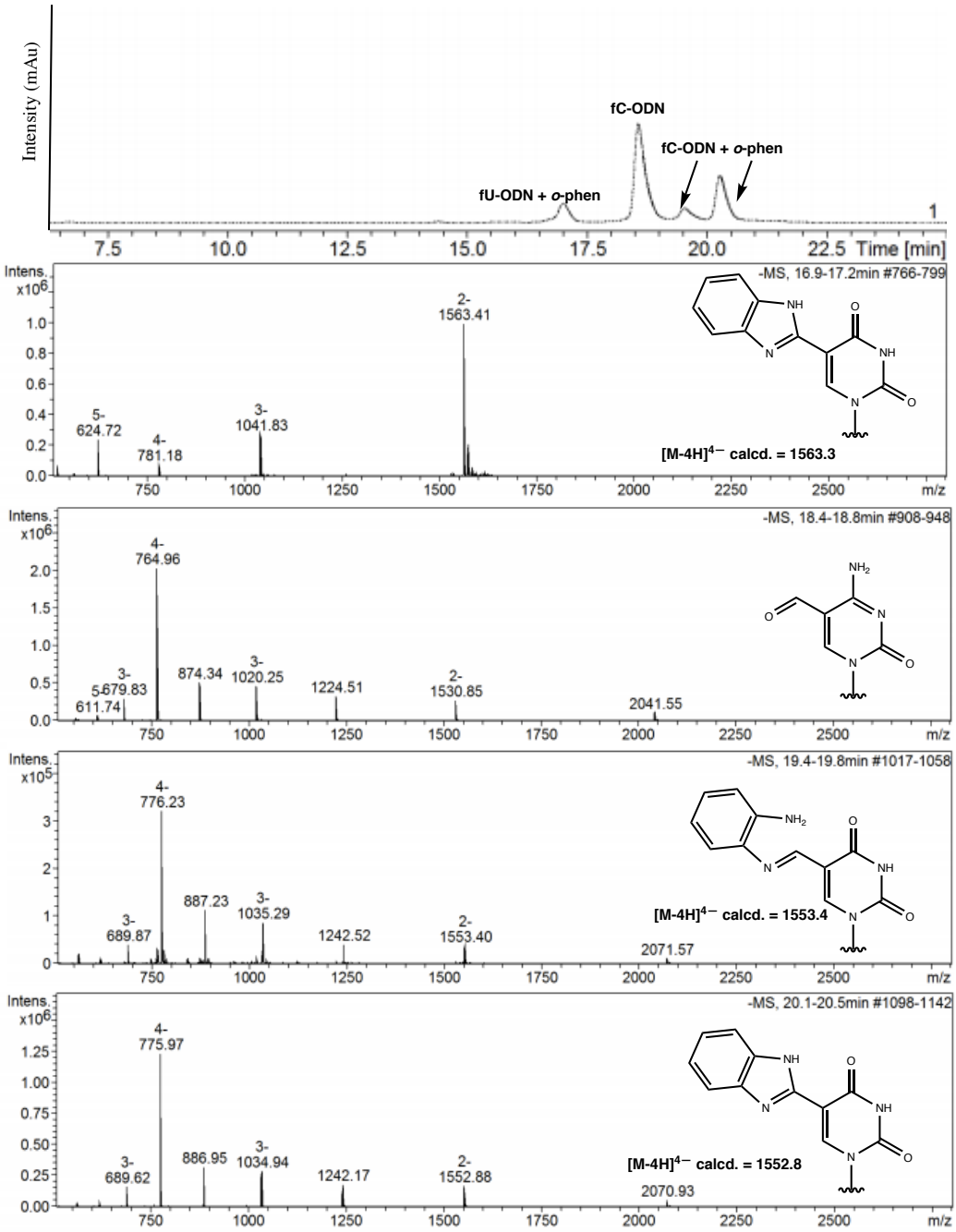


Figure 92: LC-MS trace and corresponding mass spectrum associated with each peak demonstrates a reduced 5-fC benzimidazole product and a non-reduced Schiff base adduct.

Biotinylated adduct reversibility

The fU-BH ODN adduct could be reverted back to fU-ODN in the presence of *p*-anisidine. Reversion of the biotinylated probe was found to be more effective by transamination in the presence of NH_2OH (Table 25). Reversion was also possible for the fU-ARP ODN adduct but this required a lower pH, whilst *o*-phenylenediamine was mainly stable to reversion (Appendix, Chapter 3 - Table 26). 5-fC-BH adduct reversion required harsher conditions compared to the 5-fU-BH adduct due its reduced electrophilicity.

Entry	Conditions	fC-ODN BH adduct Reversion (%)
1	pH 6, 2 hr	62
2	pH 7 2 hr	68
3	pH 7, 2 hr, 0.05% NH_2OH	100

Table 25: % reversion of fU-BH adduct to generate fU-ODN or fU-ODN-oxime determined by integration of the 260nM UV signal via LC-MS analysis.

Entry	Conditions (100 mM p-anisidine, 0.05% v/v NH_2OH)	Adduct Reversion (%)
1	ARP pH 6, 2 hr	83
2	ARP pH 5, 2 hr	30
3	*o*-phenylenediamine pH 6, 2 hr	14

Table 26: % reversion of fU-ARP and fU-*o*-phenylenediamine ODN adducts to form fU-oxime as determined by integration of the 260nm UV signal via LC-MS analysis.

Entry	Conditions (100 mM p-anisidine, 0.05% v/v NH_2OH)	fC-ODN BH adduct Reversion (%)
1	pH 6, 4 hr	62
2	pH 6, 4 hr	93
3	pH 5, 24 hr	100

Table 27: % reversion of fC-BH ODN adduct to form fC-oxime as determined by integration of the 260nm UV signal via LC-MS analysis. fC-BH ODN adduct reversion required harsher conditions compared to the 5-fU-BH adduct due its reduced electrophilicity.

	Enrichment/fC-DNA2 or hmC-DNA2			Average
fU-chem	114.7 111.7	122.7 118.8	158.3 152.1	129.7
hmU-chem	71.2 100.0	71.4 125.2		92.0

Table 28: Enrichment of T-modifications via sequencing reads of model ODN sequences over hmC-DNA2 and fC-DNA2

	Enrichment/GCAT-DNA			Average
fU-chem	163.1 161.2	207.0 161.2	225.4 236.2	192.4
hmU-chem	74.6 60.1	89.1 75.2		74.4

Table 29: Enrichment of T-modifications via sequencing reads of model ODN sequences over GCAT-DNA

Chapter 3 Supplementary figures and tables

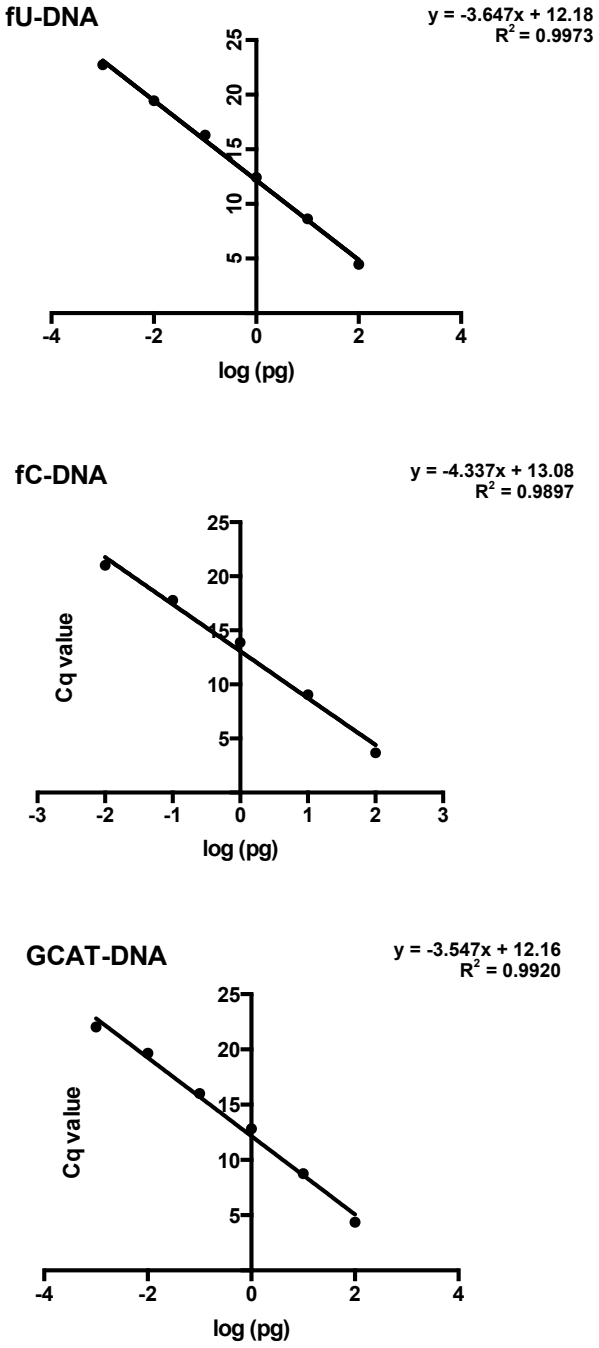


Figure 93: Example calibration lines for qPCR quantification of fU-DNA, fC-DNA and GCAT-DNA.

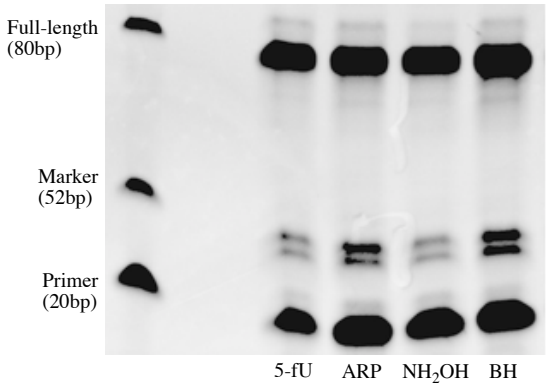


Figure 94: Polymerase stalling of a DNA model containing 2 modifications or adducts per strand. More significant polymerase stalling is observed for ARP and BH adducts compared with native 5-fU and the 5-fU oxime (NH_2OH) adducts formed after chemical elution.

Probe ARP + *p*-anisidine	**Selectivity fU-DNA/fC-DNA**	**Mean Selectivity**
Bio 1, Tech 1	1.8	1.5
Bio 1, Tech 2	1.1	
Bio 2, Tech 1	1.7	1.4
Bio 2, Tech 2	1.2	

Table 30: Selectivity determined by qPCR enrichment for Probe ARP + *p*-anisidine for fU-DNA and fC-DNA.

Probe ARP	**Selectivity fU-DNA/fC-DNA**	**Mean Selectivity**
Bio 1, Tech 1	197	184
Bio 1, Tech 2	236	
Bio 1, Tech 3	119	
Bio 2, Tech 1	143	112
Bio 2, Tech 2	54	
Bio 3, Tech 3	138	
Bio 3, Tech 1	164	139
Bio 3, Tech 2	114	

Table 31: Selectivity determined by qPCR enrichment for Probe ARP for fU-DNA and fC-DNA.

Probe BH	**Selectivity fU-DNA/fC-DNA**	**Mean Selectivity**
Bio 1, Tech 1	163	176
Bio 1, Tech 2	208	
Bio 1, Tech 3	156	
Bio 2, Tech 1	141	123
Bio 2, Tech 2	104	
Bio 3, Tech 1	236	224
Bio 3, Tech 2	211	

Table 32: Selectivity determined by qPCR enrichment for Probe BH for fU-DNA and fC-DNA

Probe *o*-phenbiotin	Selectivity fU-DNA/fC-DNA	Mean Selectivity
Bio 1, Tech 1	115	120
Bio 1, Tech 2	86	
Bio 1, Tech 3	160	
Bio 2, Tech 1	236	217
Bio 2, Tech 2	197	
Bio 3, Tech 1	94	98
Bio 3, Tech 2	102	

Table 33: Selectivity determined by qPCR enrichment for *o*-phenbiotin for fU-DNA and fC-DNA.

Probe BH	Selectivity fU-DNA/GCAT-DNA	Mean Selectivity
Bio 1, Tech 1	162	151
Bio 1, Tech 2	141	
Bio 1, Tech 3	151	
Bio 2, Tech 1	120	134
Bio 2, Tech 2	142	
Bio 2, Tech 3	139	
Bio 3, Tech 1	147	149
Bio 3, Tech 2	169	
Bio 3, Tech 3	131	

Table 34: Selectivity determined by qPCR enrichment for Probe 2 for fU-DNA and GCAT-DNA.

BH Chemical reversion	Selectivity fU-DNA/fC-DNA
Tech 1	7.9
Tech 2	10.0
Tech 3	8.4
Mean	8.8

Table 35: Selectivity determined by qPCR enrichment of fU-DNA over fC-DNA via NH_2OH-mediated chemoselective elution.

Model Molecule	Ground State	Orbital Energy (a. u.)
5-fC_m *anti*	no	-0.91041
5-fC_m *syn*	yes	-0.92741
5-fU_m *anti*	yes	-0.89813
5-fU_m *syn*	no	-0.90472

Table 36: Orbital energies (in atomic units) of the C_{ring}-$C_{aldehyde}$ bonding orbitals in the studied model molecules, as calculated via NBO analysis.

Model Molecule	Ground State	Orbital Energy (a. u.)
5-fC$_m$ *anti*	no	-1.36378
5-fC$_m$ *syn*	yes	-1.36138
5-fU$_m$ *anti*	yes	-1.35866
5-fU$_m$ *syn*	no	-1.36961

Table 37: Orbital energies (in atomic units) of the C=O bonding orbitals in the aldehyde moieties of the studied model molecules, as calculated via NBO analysis.

Model Molecule	Ground State	LUMO
5-fC$_m$ *anti*	no	0.07321
5-fC$_m$ *syn*	yes	0.07643
5-fU$_m$ *anti*	yes	0.06268
5-fU$_m$ *syn*	no	0.06850

Table 38: Orbital energies (in atomic units) of the LUMO of the studied model molecules, as calculated via NBO analysis.

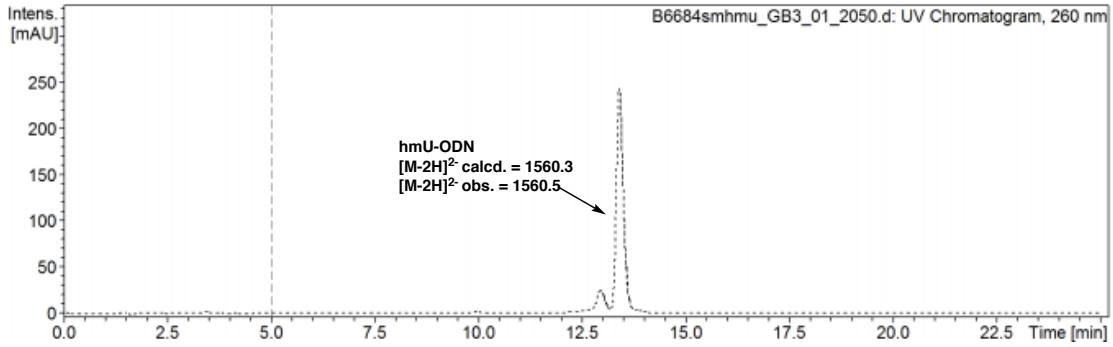


Figure 95: LC-MS trace of hmU-ODN, Chapter 3, Scheme 3.

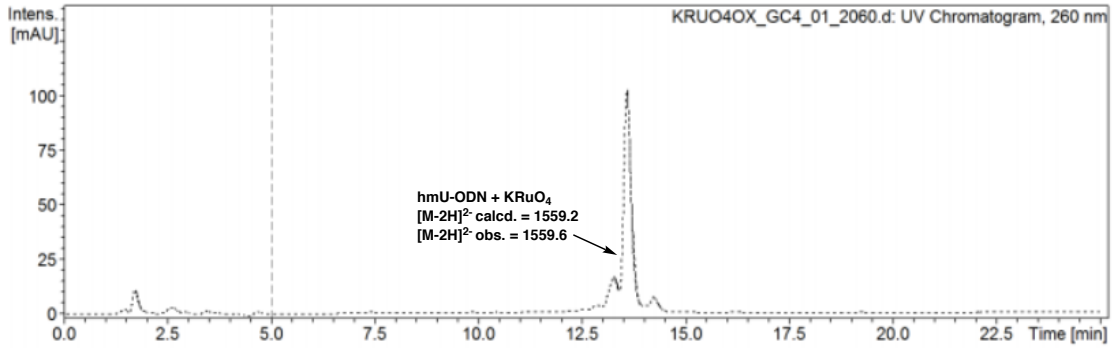


Figure 96: LC-MS trace of fU-ODN2 after hmU-ODN treatment with $KRuO_4$, Chapter 3, Scheme 3.

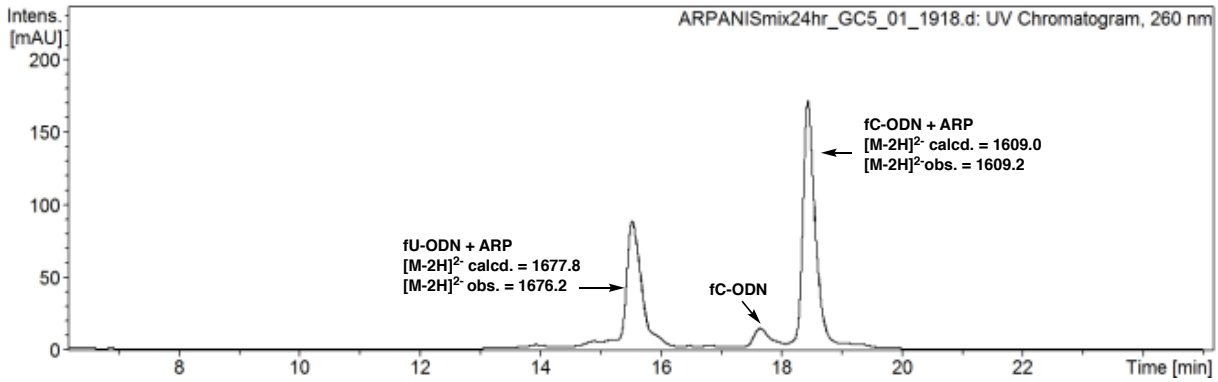


Figure 97: LC-MS trace Chapter 3, Table 4, Entry 1

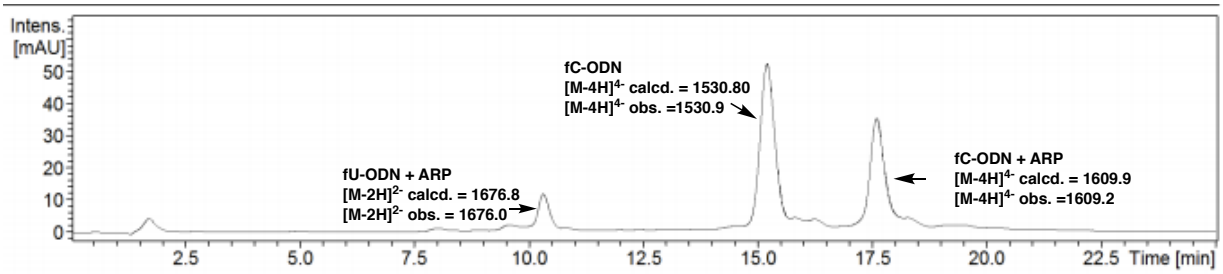


Figure 98: LC-MS trace Chapter 3, Table 4, Entry 2

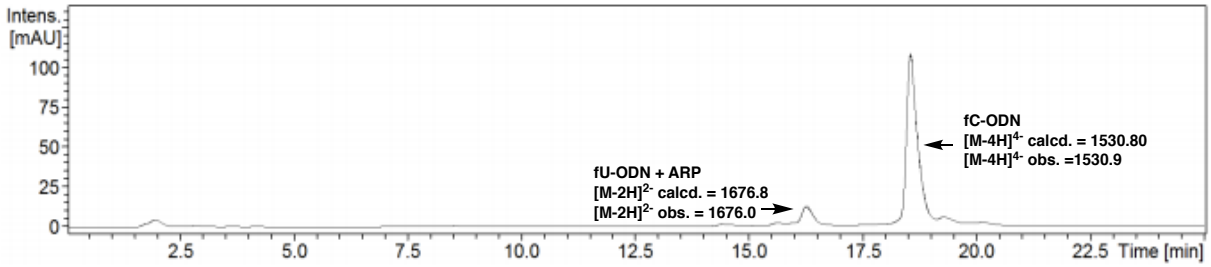


Figure 99: LC-MS trace Chapter 3, Table 4, Entry 3

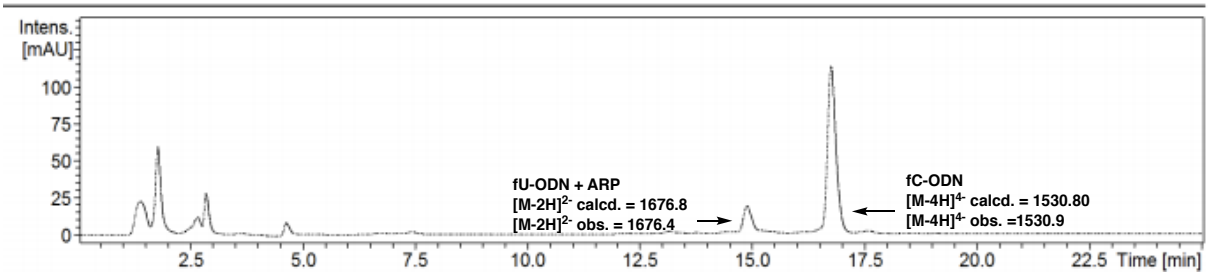


Figure 100: LC-MS trace Chapter 3, Table 4, Entry 4

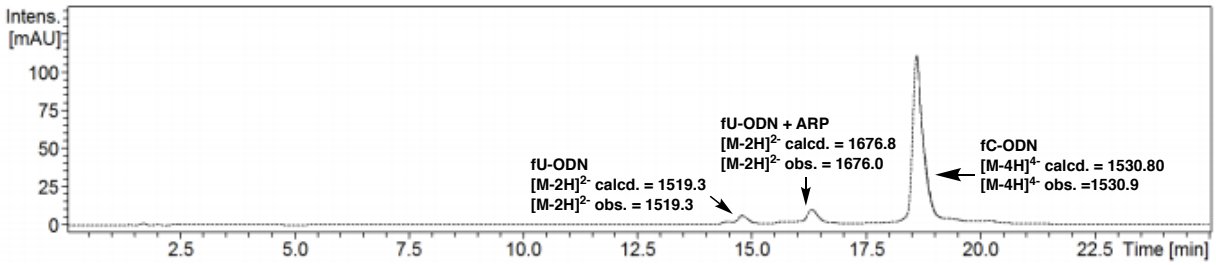


Figure 101: LC-MS trace Chapter 3, Table 4, Entry 5

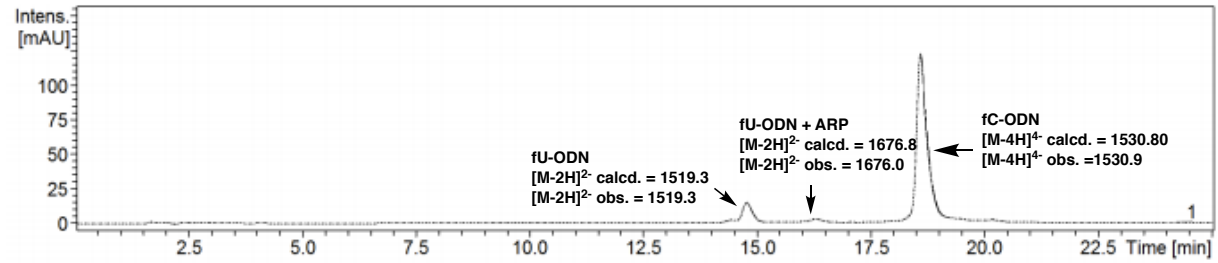


Figure 102: LC-MS trace Chapter 3, Table 4, Entry 6

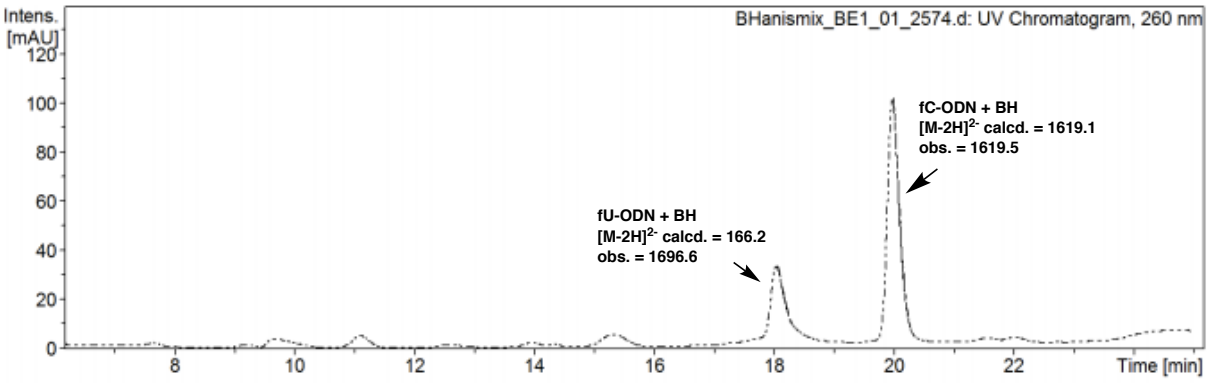


Figure 103: LC-MS trace Chapter 3, Table 5, Entry 1

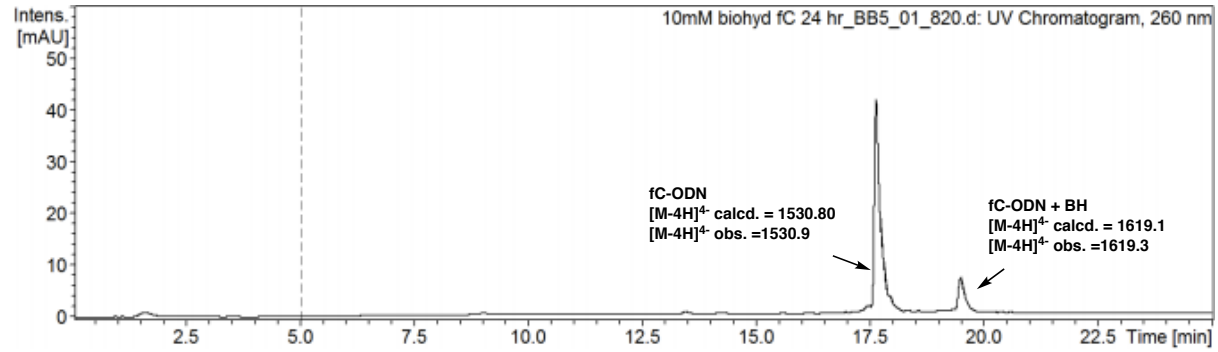


Figure 104: LC-MS trace Chapter 3, Table 5, Entry 2

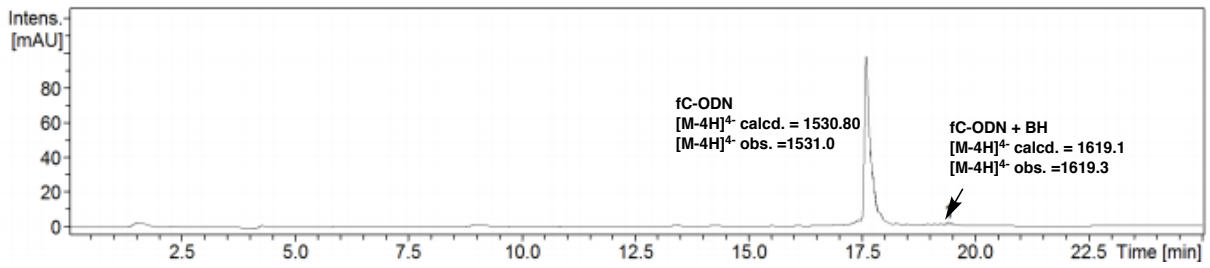


Figure 105: LC-MS trace Chapter 3, Table 5, Entry 3

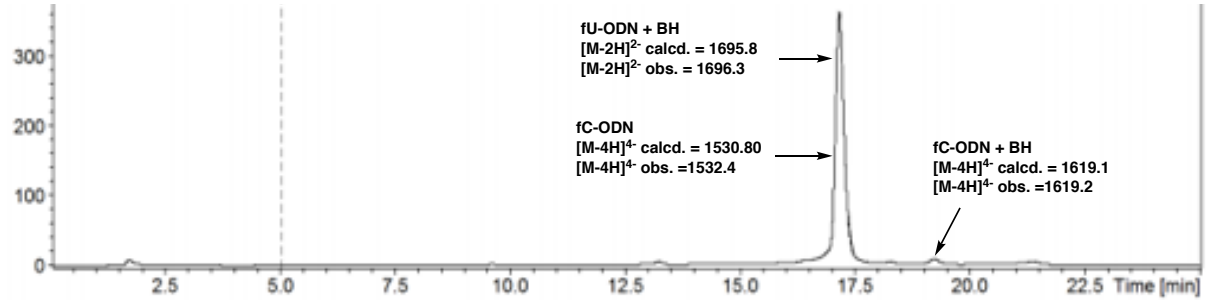


Figure 106: LC-MS trace Chapter 3, Table 5, Entry 4

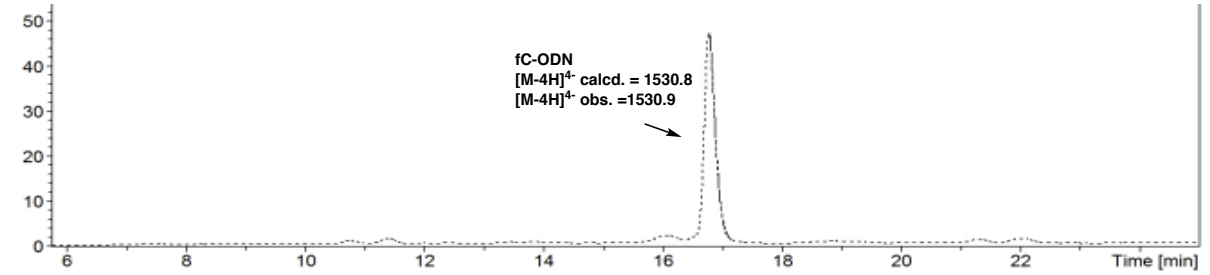


Figure 107: LC-MS trace Chapter 3, Table 5, Entry 5

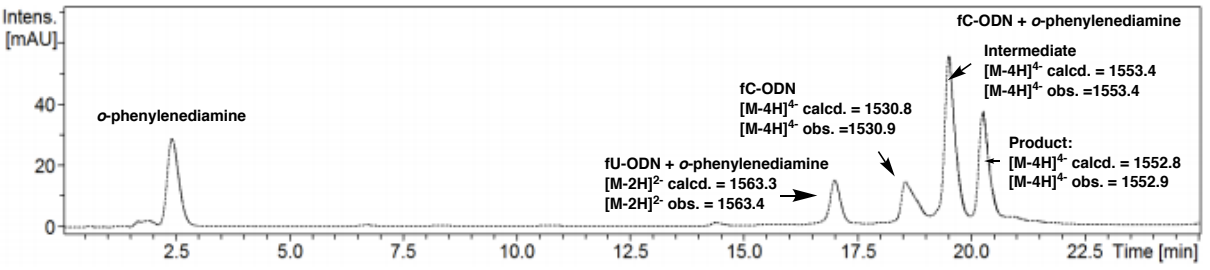


Figure 108: LC-MS trace Chapter 3, Table 6, Entry 1

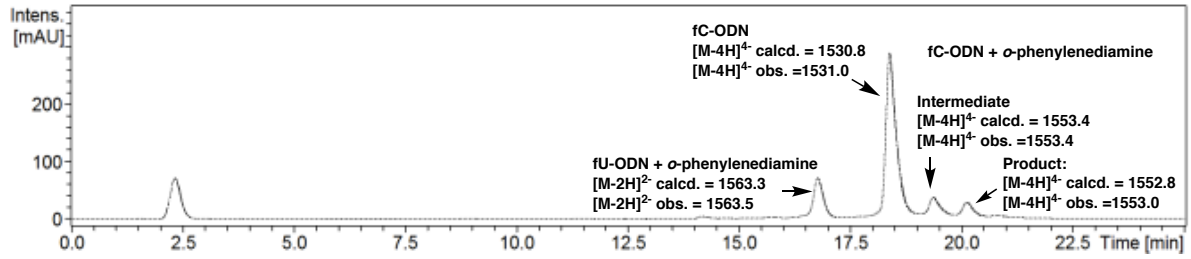


Figure 109: LC-MS trace Chapter 3, Table 6, Entry 2

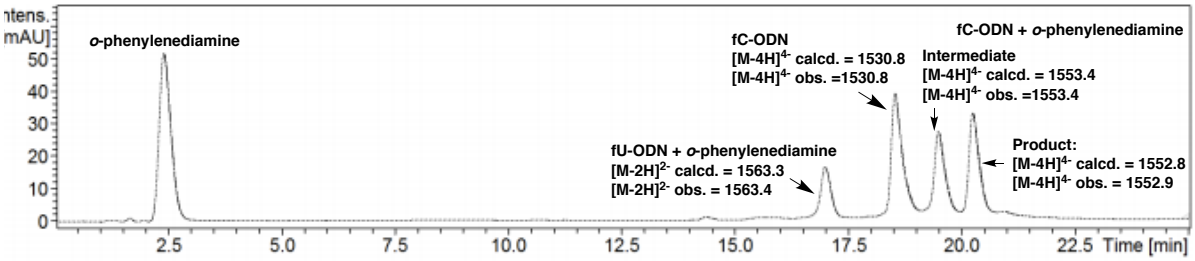


Figure 110: LC-MS trace Chapter 3, Table 6, Entry 3

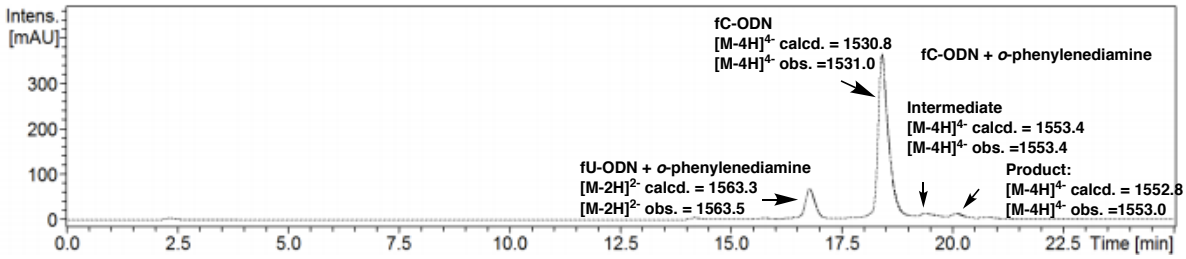


Figure 111: LC-MS trace Chapter 3, Table 6, Entry 4

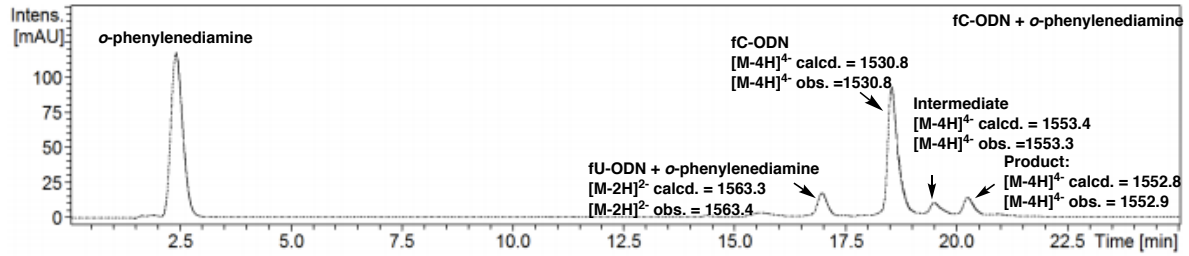


Figure 112: LC-MS trace Chapter 3, Table 6, Entry 5

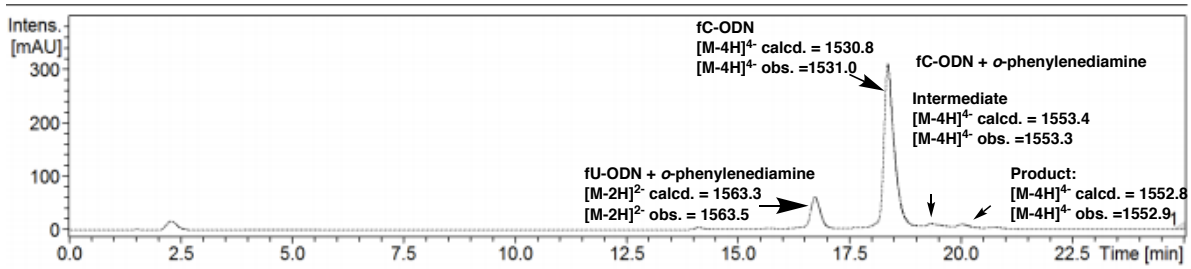


Figure 113: LC-MS trace Chapter 3, Table 6, Entry 6

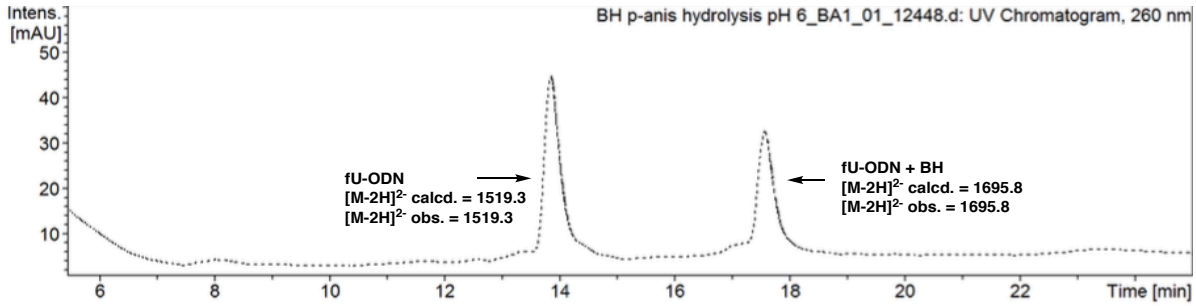


Figure 114: LC-MS trace Appendix, Table 25, Entry 1

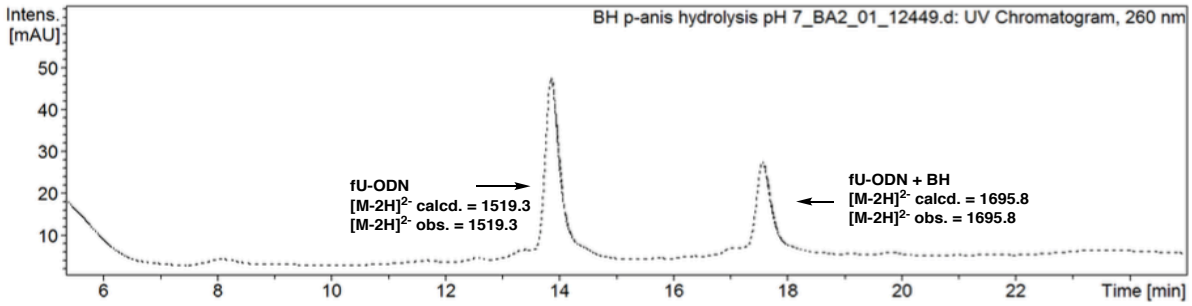


Figure 115: LC-MS trace Appendix, Table 25, Entry 2

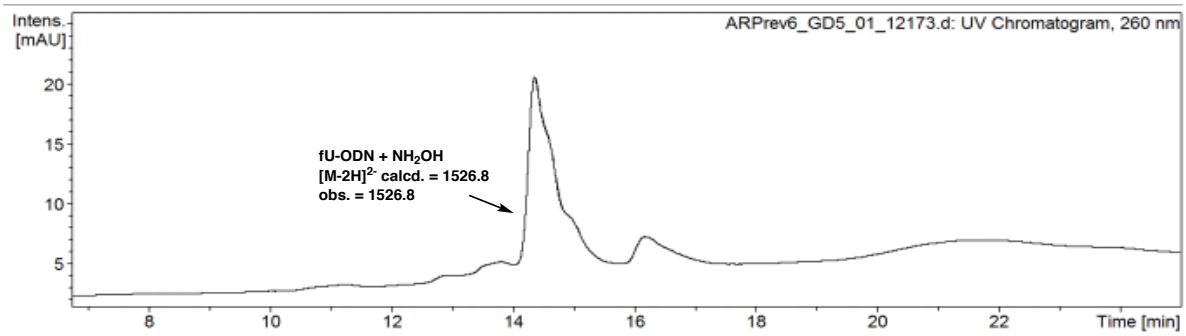


Figure 116: LC-MS trace Appendix, Table 25, Entry 1

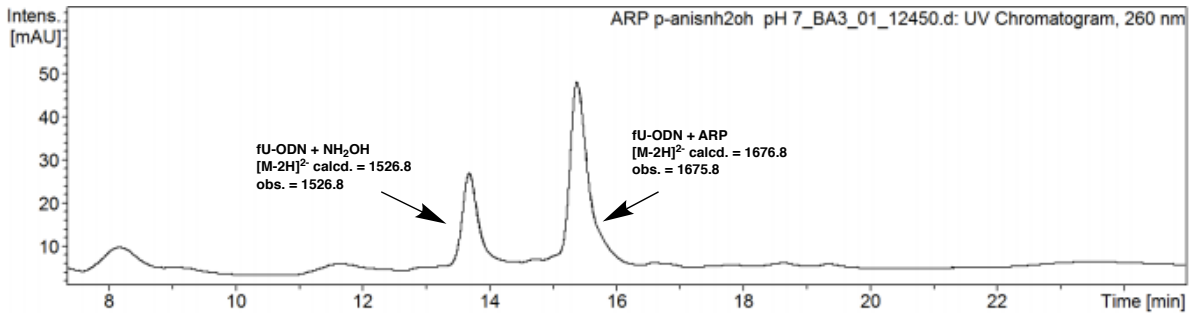


Figure 117: LC-MS trace Appendix, Table 26, Entry 2

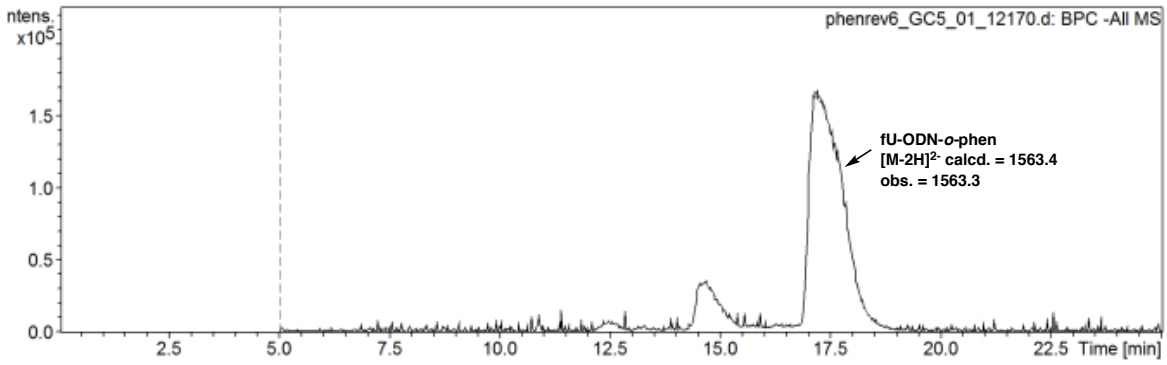


Figure 118: LC-MS trace Appendix, Table 26, Entry 3

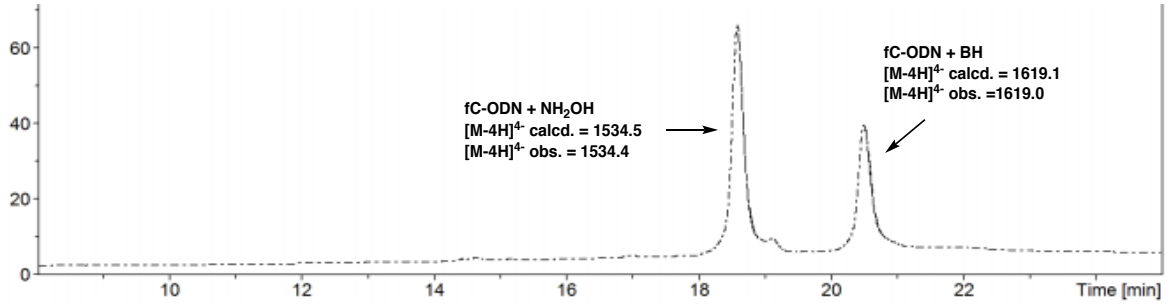


Figure 119: LC-MS trace Appendix, Table 27, Entry 1

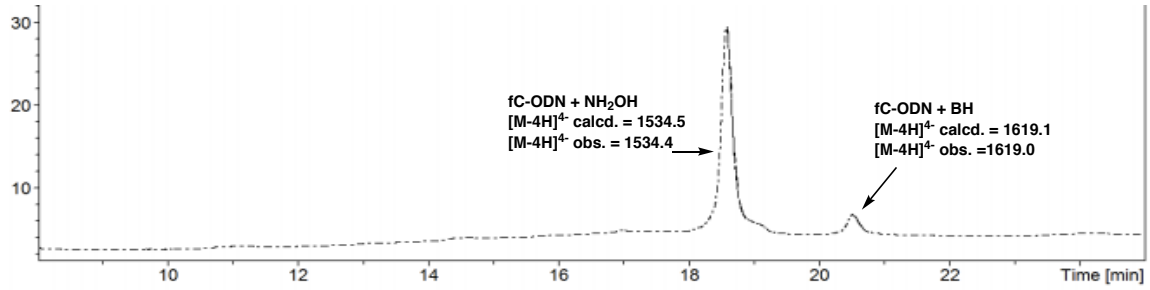


Figure 120: LC-MS trace Appendix, Table 27, Entry 2

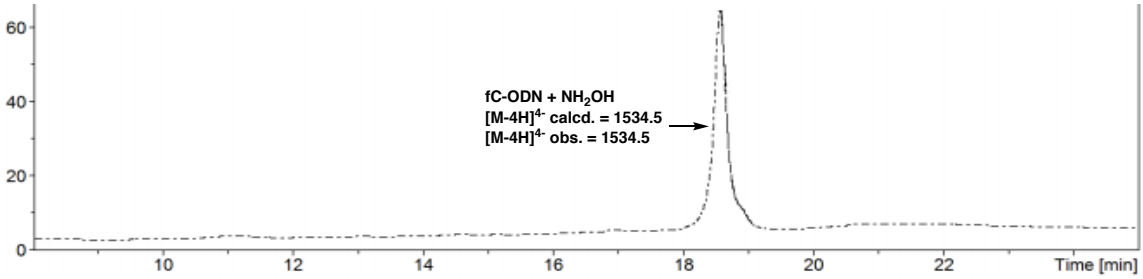


Figure 121: LC-MS trace Appendix, Table 27, Entry 3

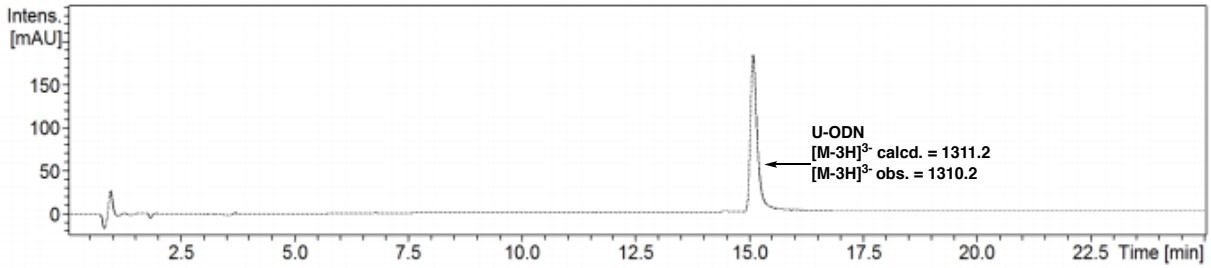


Figure 122: LC-MS trace of U-ODN starting material

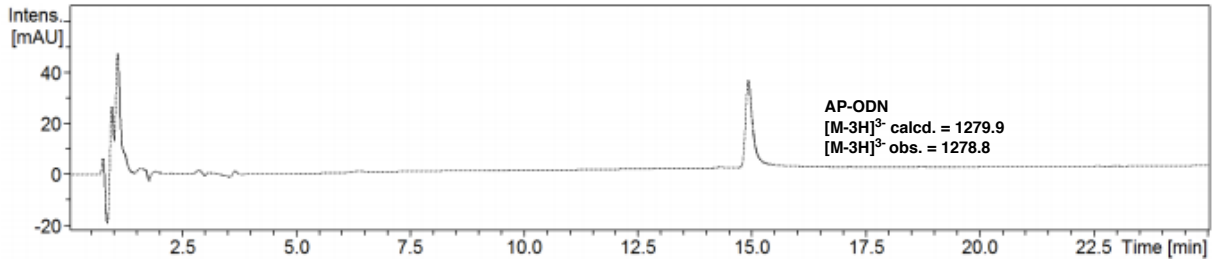


Figure 123: LC-MS trace of AP-ODN, generated by incubation of U-ODN with UNG (5U) at 37 °C for 3 hr, Chapter 3 - Figure 42

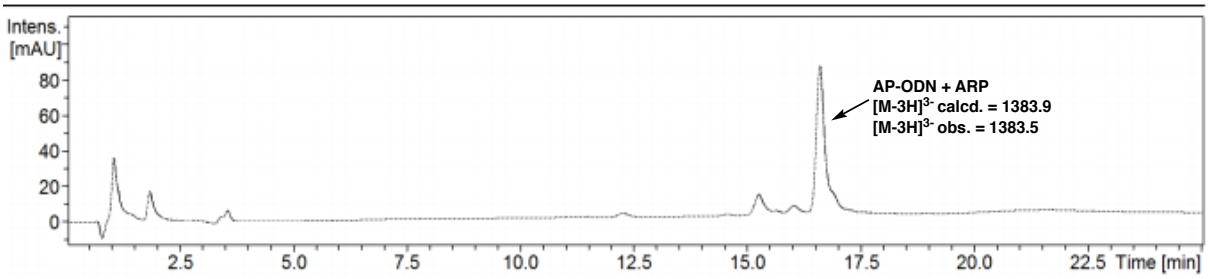


Figure 124: LC-MS trace, Chapter 3, Table 7, Entry 1

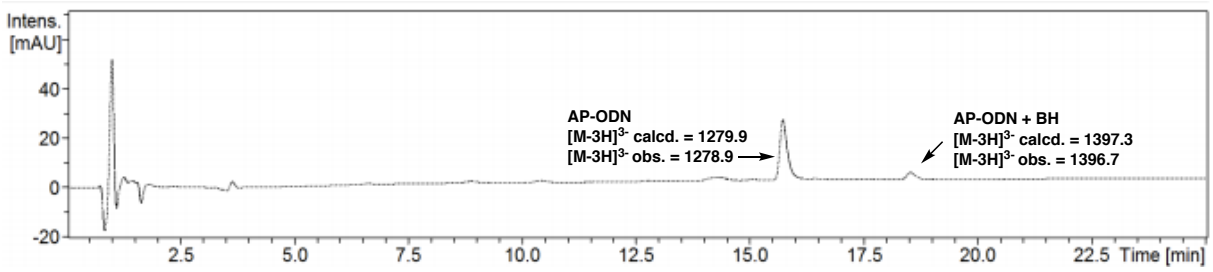


Figure 125: LC-MS trace, Chapter 3, Table 7, Entry 2

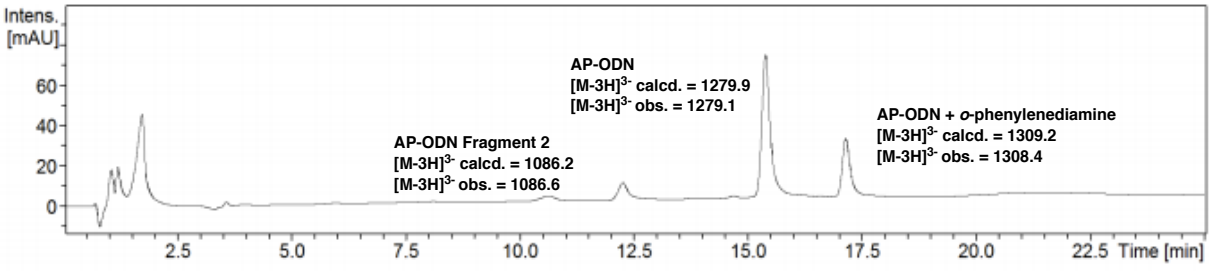


Figure 126: LC-MS trace, Chapter 3, Table 7, Entry 3

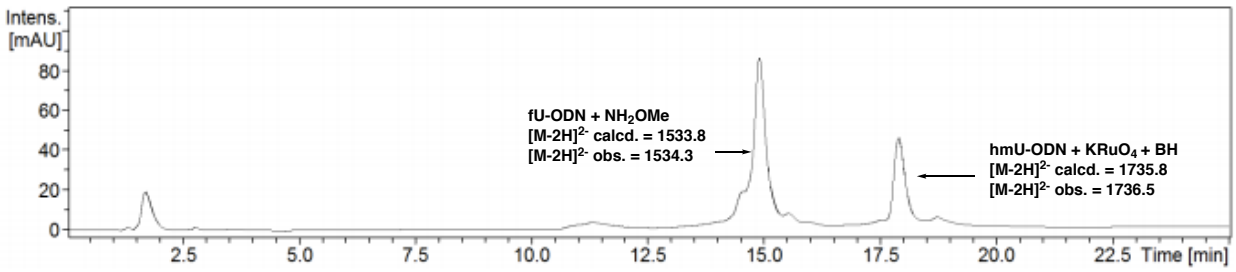


Figure 127: LC-MS trace, Chapter 3, Scheme 7

8.4. Appendix Chapter 4

Validation of hmU-DIP peaks via SMUG1 pre-treatment control

In order to rule-out a particular sequence context associated with 5-hmU antibody binding, HEK293T DNA was firstly treated with hSMUG1 (NEB) for 18 hr prior to hmU-DIP enrichment. Since hSMUG1 excises 5-hmU from the genome, this can be used as a further control to validate peaks generated from hmU-DIP sequencing via a disappearance of peak/signal after hSMUG1 treatment. Spike-in controls confirmed that 5-hmU is no longer enriched after treatment (Table 45).

Gene ontology analysis of differentially expressed genes upon SMUG1 knockdown

SMUG1 has no known function aside from being a DNA glycosylase, and can excise both U, and the oxidised T derivatives 5-hmU and 5-fU from genomic DNA. Since SMUG1 deficiency is linked with aging, cancer and disease (Introduction), it was of interest to determine the effect of SMUG1 knockdown on gene expression. Gene ontology analysis was thus performed on both upregulated and downregulated genes upon SMUG1 knockdown. Within downregulated genes, there was no clear enrichment of gene function. However, several upregulated genes were found to be involved in the inflammatory response and cell signalling, along with genes associated with a cellular response to hydrogen peroxide and a hypoxic response (Table 39).

	Genes	q-value	
Extracellular matrix organisation	17	2.30E-04	CD44, CDH1, CCDC80, COL1A1, COLD4A1, COL9A3, COLD5A2, FN1, FBLN5, ITGA11, ITGA5, ICAN1, LUM, LOX, SPIN1, TNC, VCAM1
Positive regulation of gene expression	19	2.60E-04	CITED1, CITED2, FEV, KLF4, AGO3, BMP2, CDH3, ERBB3, EPB4IL4B, FN1, HIF1A, INHBA, LAMP3, NFIL3, PIK3CD, SERPINB9, SPRY2, TNC, TLR3
Response to lipopolysaccharide	15	2.80E-04	CXCL1, CXCL10, CXCL11, CD40, CITED1, FOS, JUN, TNFRSF14, TNFRSF9, NOCT, PCK1, PLCG2, THBD, TRIB1, VCAM1
Wound healing	11	3.00E-04	ELK3, CDH3, ERBB3, EPB41L4B, FN1, LOX, PPARA, PDGFRB, SDC4, TNC
SMAD protein signal transduction	9	2.50E-03	CITED1, FOS, JUN, BMP2, BMP5, GDF10, GDF15, INHBA, INHBB
Negative regulation of cell proliferation	21	2.30E-03	ADAMTS1, CXCL1, CXCL8, CBFA2T3, ETS1, JUN, KLF4, NDRG1, RAPGEF2, SOX4, TNRFSF9, BMP2, BMP5, INHBA, IFIT3, IFNL1, IRF6, PMP22, RARRES3, RIPPLY3, SPRY2
Regulation of cell proliferation	14	2.70E-03	CXCL19, CXCL11, CD40, JUN, NDRG1, NFKBIA, RAPGEF2, TNFRSF14, ANXA1, ERBB3, JAG1, SAT1, TNC, TFRC
Defence response to virus	13	3.70E-03	OASL, BNIP3L, BNIP3, CXCL10, CD40, APOBEC3D, IFNB1, IFIT2, IFIT3, IFNL1, PMAIP1, TLR3, ZC3HAV1
Angiogenesis	15	3.50E-03	CXCL8, ELK3, JUN, TNFAIP2, ANGPTL4, CSPG4, FN1, HAND1, HMOX1, HIF1A, ITGA5, JAG1, RHOB, SAT1, SIPR1
Response to hypoxia	13	4.50E-03	BNIP3, CXCR4, CBFA2T3, CITED2, ETS1, ANGPTL4, BMP2, HMOX1, HIF1A, LDHA, PPARA, PLOD2, VCAM1
Response to drug	17	6.30E-03	HMGCS1, ABCA1, ABCG2, FOS, JUN, WFDC1, ANXA1, CDH1, CDH3, COL1A1, CCND1, LGALS1, INHBA, ICAM1, LDHA, LOX, TRPA1
Inflammatory response	19	7.30E-03	AXL, CCL5, CXCL1, CXCL10, CXCL11, CXCL8, CXCR4, CD40, FOS, TNFAIP3, TNFRSF14, TNFRSF9, ADGRE2, ANXA1, BMP2, BDKRB2, PTX3, PIK3CD, TLR3
Positive regulation of transcription from RNA polymerase II promoter	34	7.50E-03	AKNA, CXCL10, CD40, CITED1, CITED2, ELK3, ETS1, ETV1, ETV4, FOSL2, FOS, JUN, KLF4, MAFF, NFKBIA, SOX4, ABHD14B, ATF3, BMP2, BMP5, CSRNP1, EPCAM, CGA, HAND1, HIF1A, INHBA, IFNB1, JAG1, LUM, MEOX1, PPARA, SLC40A1, S1PR1, TLR3

Response to cAMP	7	1.20E-02	CITED1, FOS, JUN, COL1A1, DUSP1, LDHA, THBD
Cellular respone to hypoxia	9	2.00E-02	BNIP3L, BNIP3, NDRG1, HMOX1, HIF1A, ICAM1, PMAIP1, PCK1, STC1
Cellular response to hydrogen peroxide	7	3.40E-02	JUN, COL1A1, DUSP1, HMOX1, LDHA, PDGFRB, TRPA1

Table 39: Gene ontology analysis via DAVID using GO_Biological process of upregulated genes upon SMUG1 knockdown FDR <0.05

Chapter 4 Supplementary Tables and Figures

***T.Brucei* Differentiation experiment**	Base J/T	
	Cold-shock	Cis-aconitate/citrate
0 hr		1.91E-03
		1.33E-03
Average		**1.62E-03**
24 hr	1.17E-03	8.81E-04
	1.17E-03	1.20E-03
Average	**1.17E-03**	**1.04E-03**
48 hr	7.85E-04	4.00E-04
	1.03E-03	7.94E-04
Average	**9.07E-04**	**5.98E-04**
72 hr	5.35E-04	5.62E-05
Average	3.95E-04	9.29E-05
	4.65E-04	**7.45E-05**
96 hr	2.02E-04	6.55E-05
Average	2.92E-04	n.d
	2.47E-04	

Table 40: LC-MS/MS measurements of Base J as a proportion of T at several timepoints after differentiation initiation in *T.Brucei*.

***T.Brucei* 5-hmU spike-in experiment**	**hmU/T**	**Base J/T**
2 day	1.60E-03	3.25E-05
Tech1	1.78E-03	1.21E-05
	6.11E-05	1.14E-04
Tech2	1.21E-03	9.09E-05
Average	**1.30E-03**	**6.25E-05**
> 2 day, 2 day proliferation	9.63E-04	2.15E-04
	7.64E-04	2.02E-04
	7.38E-04	2.36E-04
	7.00E-04	1.99E-04
	1.08E-03	2.28E-04
	8.70E-04	1.81E-04
	1.21E-03	1.38E-04
	1.24E-03	1.26E-04
	8.71E-04	1.80E-04
	8.72E-04	2.26E-04
Average	**9.41E-04**	**1.93E-04**
> 2 day, 4 day proliferation	7.80E-04	1.87E-04
	7.70E-04	2.22E-04
	9.18E-04	2.38E-04
	9.72E-04	2.58E-04
	7.23E-04	3.16E-04
	7.94E-04	2.65E-04
Average	**8.26E-04**	**2.48E-04**
> 2 day, 6 day proliferation	1.10E-03	3.83E-04
	8.60E-04	2.57E-04
Average	**9.80E-04**	**3.20E-04**

Table 41: LC-MS/MS measurements of 5-hmU and Base J as a proportion of T, where PCF *T.Brucei* are cultured in the presence of 1mM 5-hmU

Actin1 fw	5'-GGATCAGCAAGCAGGAGTATG-3'
Actin1 rev	5'-AGAAAGGGTGTAACGCAACTAA-3'
Actin 2 fw	5'-GGACCTGACTGACTACCTCAT-3'
Actin 2 rev	5'-CGTAGCACAGCTTCTCCTTAAT-3'
Smug1.1 fw	5'-ACCTTTTGGCATGGCCCAGACTG-3'
Smug1.1 rev	5'-GGAGTAAGGTTGCGCCCGCT -3'
Smug1.2 fw	5'-CAATCTTTCCTTGGCCACTGC-3'
Smug 1.2 rev	5'-AACCTTACTCCTGCTGAGCTG-3'

Table 42: Actin and Smug1 Primers used for relative quantification using RT-qPCR of SMUG1 knockdown

Biological Sample	**5-hmU/N**
Average HEK293T	**3.59E-06**
Smug1 knockdown (96 hr)	5.87E-06
	5.83E-06
Smug Bio 1 Average	**5.85E-06**
	1.13E-05
Smug Bio 2 Tech 1	1.15E-05
	1.14E-05
	7.09E-06
	6.94E-06
Smug Bio 2 Tech 2	**7.02E-06**
Smug Bio 2.2 Average	**9.21E-06**
	3.54E-06
	4.34E-06
Smug Bio 3 Average	**3.94E-06**
Average SMUG1 knockdown	**6.33E-06**

Table 43: 5-hmU and 5-fU global levels in SMUG1 knockdown samples as a proportion of total nucleosides.

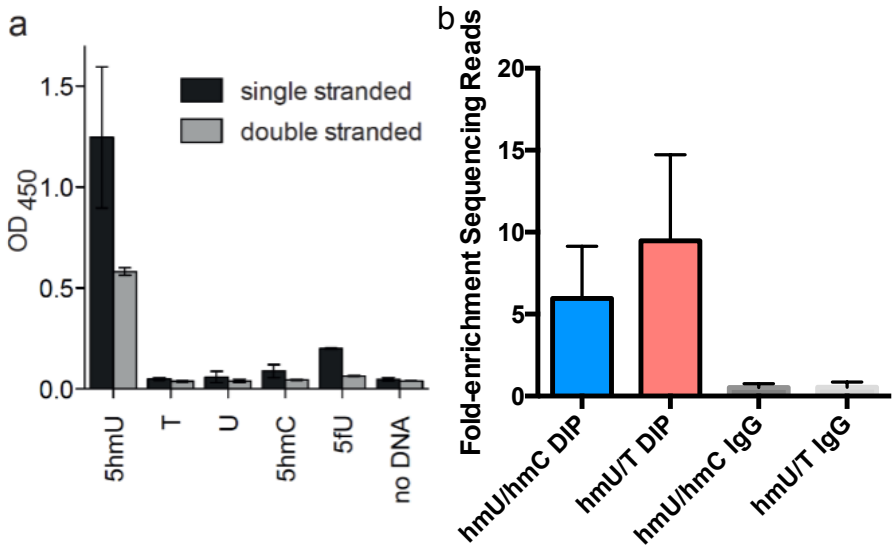


Figure 128: a) Specificity of hmU-specific antibody compared to other modified and canonical bases, determined by ELISA. Figure taken from reference [202] b) Average 5-hmU enrichment determined by sequencing ODN models (hmU-ODN, hmC-ODN2, GCAT-ODN) in hmU-DIP and control IgG libraries.

Library	Sequencing Reads	Peaks (p > 1E-05)	hmU/hmC enrichment	hmU/T enrichment
SMUG1 hmUDIP1	294071719	64848	4.37	6.09
SMUG1 hmUDIP2	151336100	53356	3.52	6.27
Rluc hmUDIP1	166330520	112635	10.00	16.08
Rluc hmUDIP2	123315707	8874	9.16	13.02
SMUG 1 Control IgG	175657640	9516	0.73	0.81
Rluc Control IgG	186341837	38886	0.31	0.24
SMUG1treatedDIP1	58874526	22551	0.24	0.10
SMUG1treatedDIP2	192234606	43156	0.67	0.16
fUchem1	185009707	151	189.41	167.09
fUchem2	44228575	26	50.63	76.50
hmU-chem	68673453	34	79.36	69.19

Table 44: Number of peaks per library determined by the MACS2 peak-caller, and extent of 5-hmU enrichment determined by spike-in sequencing controls.

Library	Consensus peaks 90%	Consensus peaks 50%	Consensus peaks 25%
SMUG1 hmUDIP	5296	29989	33484
Rluc hmUDIP	199	3225	5540
Ctrl IgG DIP	248	6829	8487
Smug1 treated DIP	457	5552	8895
SMUG1hmUDIP (50%) and RLuc hmUDIP (50%)	-	487	-
Smug1hmUDIP (50%) – Rluc1 – Rluc2	-	4488	-

Table 45: Consensus peaks at different percentage overlap between DIP-library replicates and consensus between SMUG1hmUDIP and Rluc hmUDIP common peaks (50%). (-) = not calculated

Library	Peaks – Ctrl IgG samples	Peaks – CtrlIgG – SMUG1 treated samples
SMUG1 hmUDIP (50%)	11937	7095
Rluc hmUDIP (50%)	3034	1294

Table 46: Number of peaks after Ctrl IgG peaks and SMUG1 treated peaks are subsected from the initial intersection of replicates.

Gene	Log-foldchange	FDR
MSH4	-1.89	2.03E-02
GABRB1	-1.76	8.19E-03
ZMAT1	-1.24	1.14E-03
DAB1	-1.14	6.59E-03
LUZP2	-1.12	1.11E-02
SPOCK3	-1.10	8.07E-14
KCTD8	-1.02	4.05E-02
ASTN1	-0.85	5.79E-08
AK5	-0.81	2.02E-07
SLC16A12	-0.77	1.10E-02
LINC00342	-0.74	5.25E-05
LRRIQ1	-0.73	8.06E-03
PPFIA2	-0.72	7.48E-03
RNFT2	-0.70	9.10E-04
NELL1	-0.68	2.22E-03
EBF4	-0.66	2.06E-03
SLC13A3	-0.64	1.50E-06
CFAP44	-0.64	1.23E-04
SLC35F1	-0.63	1.21E-03
TMEM265	-0.61	3.04E-05
RHPN1	-0.59	8.92E-03
RTN1	-0.53	2.50E-02
IQCH-AS1	-0.53	4.60E-02
BMPR2	-0.52	1.84E-02
CLHC1	-0.51	1.88E-02
LSAMP	-0.49	3.22E-03
LOC146880	-0.49	4.96E-03
NFIB	-0.48	1.94E-02
IFI44L	-0.43	3.54E-02
LPP	-0.42	1.94E-02
TMCO4	0.41	2.32E-02
SLC13A4	0.41	3.20E-02
ABCG2	0.43	2.52E-02
KIAA1462	0.43	8.10E-03
TBC1D5	0.45	1.84E-02
SYNPO	0.45	4.05E-02
SUSD1	0.47	4.57E-02
FAM219A	0.48	4.91E-02
COL14A1	0.52	2.23E-03
KLF8	0.53	1.16E-02
GLT8D2	0.55	4.20E-03
TNFAIP2	0.55	2.61E-02
SARDH	0.58	6.59E-03
PIK3CD	0.58	3.96E-02
CSPG4	0.60	9.50E-03
ITGA11	0.61	3.73E-02
STK31	0.62	1.09E-02
RTN4R	0.64	2.16E-02
CALCB	0.66	9.53E-04
FAM160A1	0.66	2.50E-02
LONRF3	0.70	1.16E-02
TMEM45A	0.72	4.02E-05
AMOT	0.73	4.45E-08
AXL	0.77	2.90E-03
CDH1	0.77	2.50E-02
PDGFRB	0.81	3.66E-02
GRM4	0.94	6.55E-03
CGA	1,96	1.23E-26
GFI1B	2.19	1.58E-02
VWA3B	2.52	4.88E-02

Table 47: Differentially expressed genes upon SMUG1 knockdown which contain 5-hmU.

8.5. Appendix Chapter 5

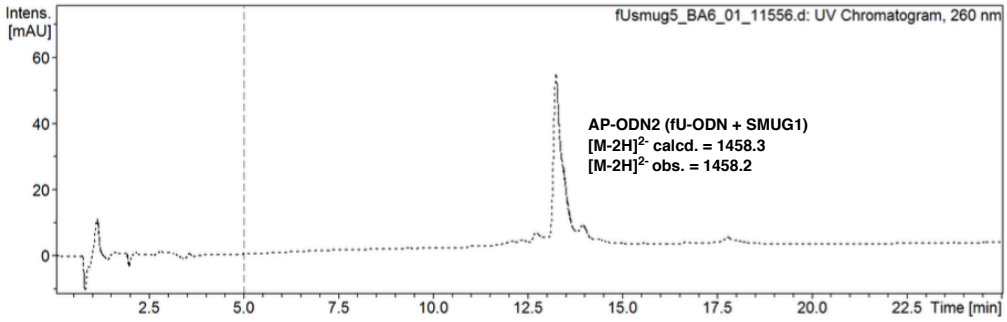


Figure 129: LC-MS trace, fU-ODN + hSMUG1 (5 U) for 3 hr.

Function	Number of proteins	p-value
Acetylation	31	5.60E-12
Phosphoprotein	41	9.10E-09
Transit peptide	9	4.10E-05
Transcription	17	1.10E-04
Nucleus	26	1.70E-04
Transcription regulation	16	2.90E-04
Mitochondrion	11	3.00E-04
Mitochondrion inner membrane	6	4.90E-04
NAD	5	8.60E-04
Cytoplasm	23	1.00E-03
Isopeptide bond	10	1.50E-03
Ubi conjugation	12	2.20E-03
Chromatin regulator	5	5.20E-03
Neurodegeneration	5	5.50E-03
Methylation	8	1.00E-02

Table 48: DAVID functional analysis of T-enriched proteins by UP_keywords function.

8.6 Appendix Chapter 6

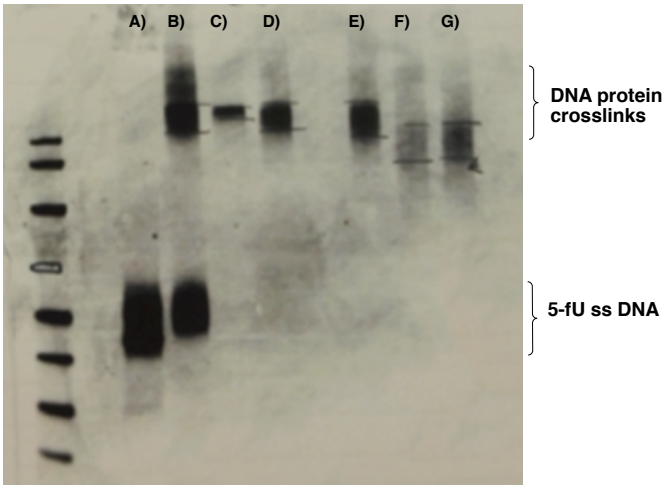


Figure 130: 12% SDS gel run in MES buffer after incubation of nucleosomes in the presence of $NaBH_3CN$. A) Free DNA; B) Free DNA + 100 mM $NaBH_3CN$; C) Nucleosome, 37 °C, 18 hr; D) 100 mM $NaBH_3CN$, Nucleosome, 4 °C, 18 hr; E) 10 mM $NaBH_3CN$, Nucleosome, 4 °C, 18 hr; F) 10 mM $NaBH_3CN$, Nucleosome, 4 °C, 3 hr; G) 1 mM $NaBH_3CN$, Nucleosome, 4 °C, 18 hr. Attempts made to suppress excessive 5-fU crosslinking included the use of 1) lower temperature reduction, 2) less reducing agent and 3) less reaction time. Extent of reduction appeared to be reduced, however bands were still > 100 Da, indicating promiscuous reactivity with 5-fU.

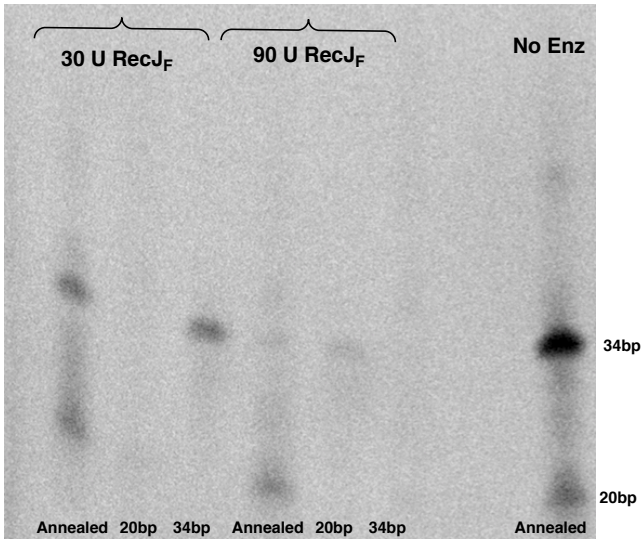


Figure 131: Optimisation and validation of $RecJ_F$ digestion with model system with a 14bp overhang. DNA with a 14bp overhang was treated by different concentrations of $RecJ_F$ for 16 hr and visualised on a TBE-Urea gel. In the presence of 90U $RecJ_F$, a band at 20bp can be observed whilst all product at 34bp is digested, confirming removal of the ss-overhang.

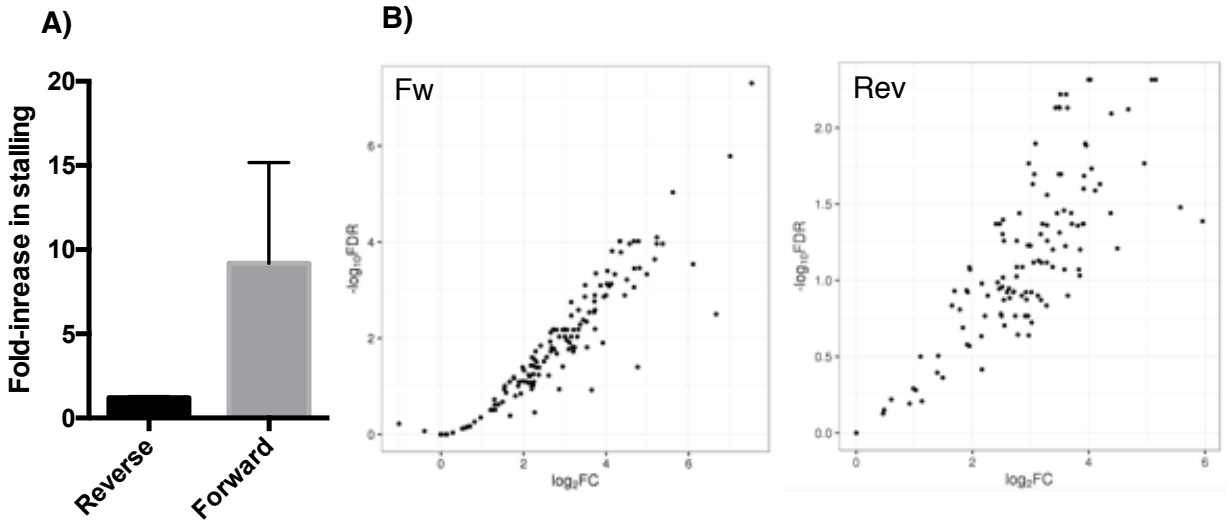


Figure 132: a) Fold-increase in truncated sequences between crosslinked and non-crosslinked polymerase extension, More stalling occurred at the majority of positions compared to the controls. b) Log_2fold stalling fC increase in stalling at each position between crosslinked and non-crosslinked polymerase extension.

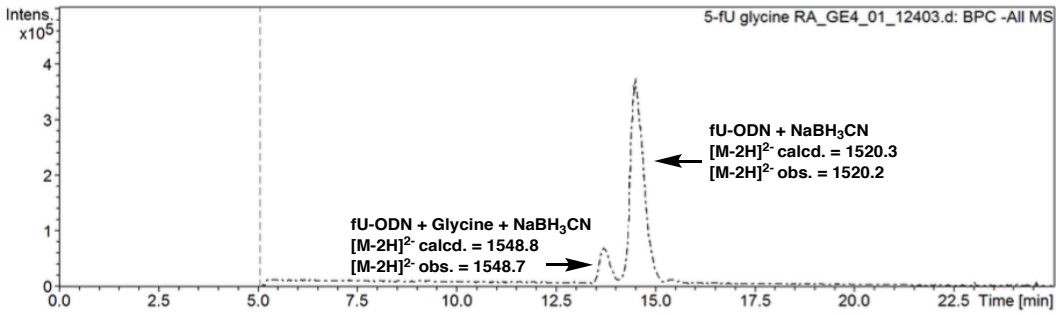


Figure 133: LC-MS trace, fU-ODN + Glycine (10 mM), $NaBH_3CN$ (25 mM), 18 hr

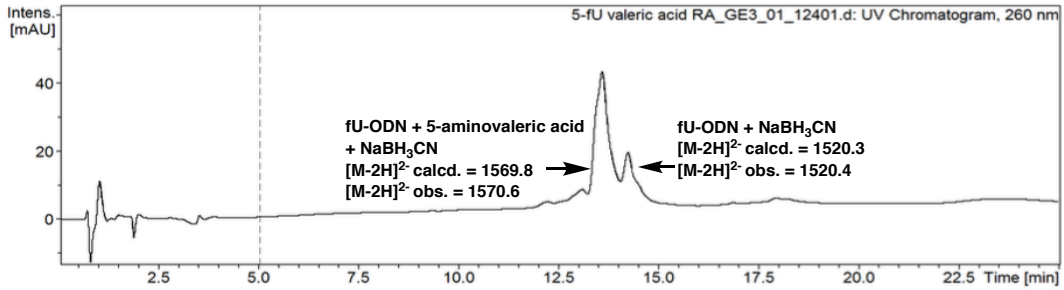


Figure 134: LC-MS trace, fU-ODN + 5-aminovaleric acid (10 mM), $NaBH_3CN$ (25 mM), 18 hr

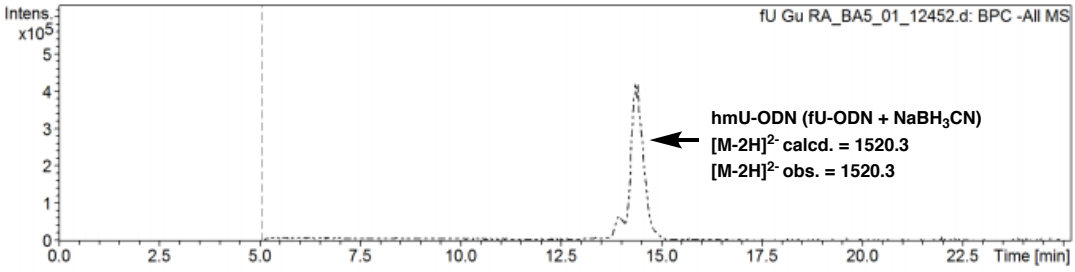


Figure 135: fU-ODN + Guanidinium hydrochloride (10 mM), $NaBH_3CN$ (25 mM), 18 hr

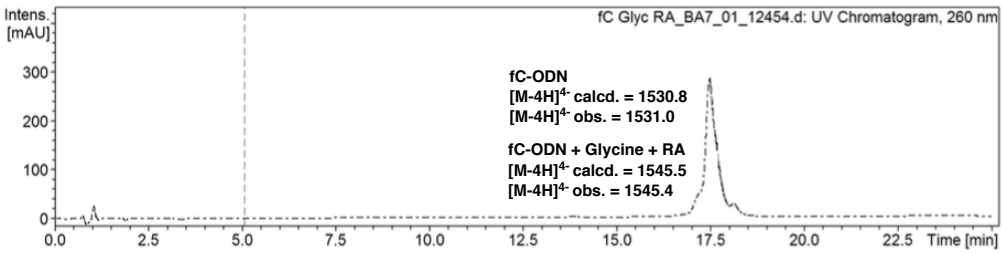


Figure 136: fC-ODN + Glycine (500 mM), $NaBH_3CN$ (100 mM), 18 hr, 37 °C

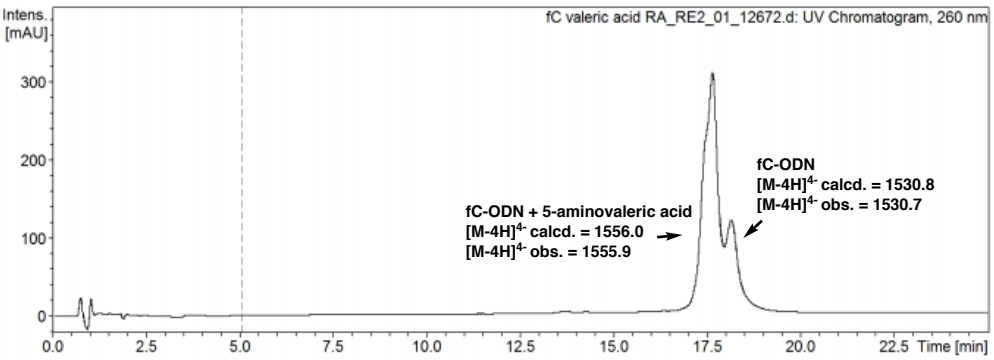


Figure 137: fC-ODN + 5-Aminovaleric acid (500 mM), $NaBH_3CN$ (100 mM), 18 hr, 37 °C

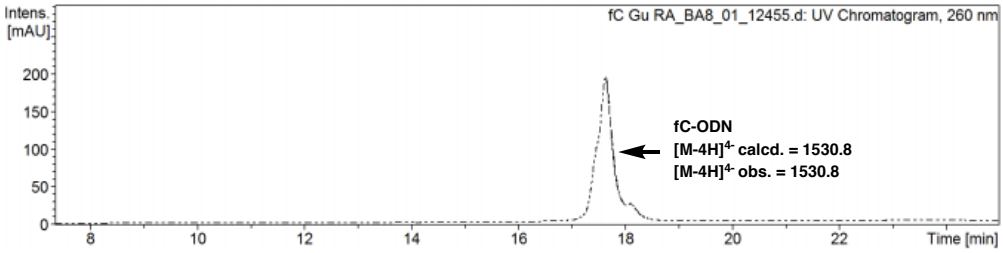


Figure 138: fC-ODN + Guanidium hydrochloride (500 mM), $NaBH_3CN$ (100 mM), 18h, 37 °C

Fw	Base	Log_2Fold Change	Rev	Base	Log_2Fold Change
35	T	4.12	113	A	4.93
36	G	3.28	112	G	4.67
37	A	3.25	111	C	3.88
38	T	3.44	110	A	4.32
39	C	3.73	109	T	4.32
40	C	3.67	108	C	4.36
41	C	3.52	107	T	4.22
42	T	3.35	106	G	4.74
43	C	3.43	105	T	4.70
44	A	3.62	104	C	4.52
45	T	3.48	103	G	4.75
46	T	3.71	102	A	4.79
47	A	3.79	101	G	4.72
48	G	4.08	100	A	3.81
49	G	4.98	99	T	4.00
50	G	4.86	98	C	4.25
51	G	4.68	97	G	5.85
52	A	2.97	96	T	5.76
53	A	2.98	95	G	5.72
54	C	3.11	94	G	3.79
55	C	3.65	93	C	4.00
56	G	3.72	92	G	3.65
57	C	3.70	91	A	3.78
58	C	3.30	90	A	3.39
59	A	3.60	89	T	2.94
60	A	4.41	88	T	3.07
61	T	4.38	87	T	3.62
62	T	4.19	86	G	4.27
63	T	3.48	85	C	4.36
64	T	3.64	84	G	4.38
65	G	3.56	83	T	3.91
66	C	3.26	82	G	3.67
67	G	2.86	81	C	3.81
68	C	2.88	80	A	4.10
69	C	3.11	79	T	3.83
70	C	3.48	78	G	3.38
71	C	3.82	77	C	3.49
72	C	3.82	76	G	4.44
73	T	3.50	75	C	4.77
74	G	3.09	74	G	4.64
75	T	2.74	73	A	3.96
76	C	3.21	72	C	3.13
77	G	3.35	71	A	3.02

78	C	3.58	70	G	2.94
79	G	3.45	69	G	2.89
80	C	3.26	68	G	2.73
81	A	3.45	67	G	2.98
82	T	3.17	66	G	3.13
83	G	3.15	65	C	3.02
84	C	2.60	64	G	2.58
85	A	3.20	63	C	3.18
86	C	3.45	62	A	3.06
87	G	3.46	61	A	3.14
88	C	2.96	60	A	2.27
89	A	3.07	59	A	2.27
90	A	3.09	58	T	2.37
91	A	3.43	57	T	2.42
92	T	3.07	56	G	3.17
93	T	2.98	55	G	3.33
94	C	2.83	54	C	3.55
95	G	2.73	53	G	3.03
96	C	2.73	52	G	2.98
97	C	3.14	51	T	2.86
98	A	3.41	50	T	3.06
99	C	3.67	49	C	2.90
100	G	3.32	48	C	2.60
101	A	3.14	47	C	2.26
102	T	2.74	46	C	2.12
103	C	3.00	45	T	2.05
104	T	2.83	44	A	2.18
105	C	2.96	43	A	2.42
106	G	2.43	42	T	2.40
107	A	2.69	41	G	2.58
108	C	2.30	40	A	2.29
109	A	2.39	39	G	2.24
110	G	2.30	38	G	1.72
111	A	2.78	37	G	1.78
112	T	3.34	36	A	2.05
113	G	3.43	35	T	2.09

Table 49: Log2Fold change of truncated sequence of crosslinked sample over control sample plotted against Widom position and base.

9. References

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953).
2. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol.* **1,** (2000).
3. Koeller, K. J. *et al.* DNA Binding Polyamides and the Importance of DNA Recognition in their use as Gene-Specific and Antiviral Agents. *Med. Chem.* **4,** 338–344 (2014).
4. Luger, K., Dechassa, M. L. & Tremethick, D. J. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* **13,** 436–447 (2012).
5. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 487–500 (2016).
6. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).
7. Verdin, E. & Ott, M. 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. *Nat. Rev. Mol. Cell Biol.* **16,** 258–264 (2015).
8. Liebers, R., Rassoulzadegan, M. & Lyko, F. Epigenetic Regulation by Heritable RNA. *PLOS Genet.* **10,** e1004296 (2014).
9. Breiling, A. & Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin* **8,** 24 (2015).
10. Gommers-Ampt, J. H. & Borst, P. Hypermodified bases in DNA. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **9,** 1034–1042 (1995).
11. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.* **10,** 2709–2721 (1982).
12. Denis, H., Ndlovu, 'Matladi N & Fuks, F. Regulation of mammalian DNA methyltransferases: a route to new mechanisms. *EMBO Rep.* **12,** 647–656 (2011).
13. Jeltsch, A. Molecular enzymology of mammalian DNA methyltransferases. *Curr. Top. Microbiol. Immunol.* **301,** 203–225 (2006).
14. Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science (80-.).* **354,** 909–912 (2016).
15. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–220 (2013).
16. Hermann, A., Goyal, R. & Jeltsch, A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J. Biol. Chem.* **279,** 48350–48359 (2004).
17. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6,** 597–610 (2005).
18. Deaton, A. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
19. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1,** 239–259 (2009).
20. Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21,** 5427–5440 (2002).
21. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69,** 915–926 (1992).
22. Liao, J. *et al.* Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet.* **47,** 469–478 (2015).
23. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99,** 247–257 (1999).
24. Dantas Machado, A. C. *et al.* Evolving insights on how cytosine methylation affects protein–DNA binding. *Brief. Funct. Genomics* **14,** 61–73 (2015).
25. Du, Q., Luu, P.-L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins:

readers of the epigenome. *Epigenomics* **7,** 1051–1073 (2015).
26. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol* **16,** 519–532 (2015).
27. Bestor, T. H., Edwards, J. R. & Boulard, M. Notes on the role of dynamic DNA methylation in mammalian development. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6796–6799 (2015).
28. Orlanski, S. *et al.* Tissue-specific DNA demethylation is required for proper B-cell differentiation and function. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 5018–5023 (2016).
29. Kagiwada, S., Kurimoto, K., Hirota, T., Yamaji, M. & Saitou, M. Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *EMBO J.* **32,** 340–353 (2013).
30. Robertson, A. B., Klungland, A., Rognes, T. & Leiros, I. DNA repair in mammalian cells: Base excision repair: the long and short of it. *Cell. Mol. Life Sci.* **66,** 981–993 (2009).
31. Wallace, S. S., Murphy, D. L. & Sweasy, J. B. Base excision repair and cancer. *Cancer Lett.* **327,** 73–89 (2012).
32. Zhang, H. & Zhu, J.-K. Active DNA Demethylation in Plants and Animals. *Cold Spring Harb. Symp. Quant. Biol.* **77,** 161–173 (2012).
33. Penn, N. W., Suwalski, R., O'Riley, C., Bojanowski, K. & Yura, R. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem. J.* **126,** 781–790 (1972).
34. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.* **14,** 341–356 (2013).
35. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science.* **324,** 930–5 (2009).
36. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science.* **333,** 1300–3 (2011).
37. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5,** e15367 (2010).
38. Munzel, M., Globisch, D. & Carell, T. 5-Hydroxymethylcytosine, the sixth base of the genome. *Angew. Chem. Int. Ed. Engl.* **50,** 6460–6468 (2011).
39. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl.* **50,** 7008–7012 (2011).
40. Pfaffeneder, T. *et al.* Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat Chem Biol* **10,** 574–581 (2014).
41. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science.* **333,** 1303–7 (2011).
42. Shen, L. *et al.* Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. *Cell* **153,** 692–706 (2013).
43. Cortazar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470,** 419–423 (2011).
44. McInroy, G. R. On Modified Nucleobases in Biological and Synthetic Systems. (University of Cambridge, 2015).
45. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol.* **11,** 555–557 (2015).
46. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem.* **6,** 1049–1055 (2014).
47. Wen, L. *et al.* Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol.* **15,** (2014).
48. Kraus, T. F. J. *et al.* Low values of 5-hydroxymethylcytosine (5hmC), the 'sixth base,' are associated with anaplasia in human brain tumors. *Int. J. cancer* **131,** 1577–1590 (2012).

49. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14,** (2013).
50. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell.* **152,** 1146–59 (2013).
51. Raiber, E. A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13,** 1146–59 (2012).
52. Song, C.-. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell.* **153,** 678–91 (2013).
53. Iurlaro, M. *et al.* In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.* **17,** 141 (2016).
54. Wang, L. *et al.* Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature* **523,** 621–625 (2015).
55. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **17,** 1195–1214 (2003).
56. Yoshihara, M., Jiang, L., Akatsuka, S., Suyama, M. & Toyokuni, S. Genome-wide profiling of 8-oxoguanine reveals its association with spatial positioning in nucleus. *DNA Res.* **21,** 603–612 (2014).
57. Allgayer, J., Kitsera, N., Bartelt, S., Epe, B. & Khobta, A. Widespread transcriptional gene inactivation initiated by a repair intermediate of 8-oxoguanine. *Nucleic Acids Res.* **44,** 7267–7280 (2016).
58. Perillo, B. *et al.* DNA Oxidation as Triggered by H3K9me2 Demethylation Drives Estrogen-Induced Gene Expression. *Science (80-.).* **319,** 202–6 (2008).
59. Fleming, A. M., Ding, Y. & Burrows, C. J. Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 2604–2609 (2017).
60. German, P. *et al.* Activation of cellular signaling by 8-oxoguanine DNA glycosylase-1-initiated DNA base excision repair. *DNA Repair (Amst).* **12,** 856–863 (2013).
61. Guz, J., Gackowski, D., Foksinski, M., Rozalski, R. & Olinski, R. Comparison of the absolute level of epigenetic marks 5-methylcytosine, 5-hydroxymethylcytosine, and 5-hydroxymethyluracil between human leukocytes and sperm. *Biol. Reprod.* **91,** 55 (2014).
62. Masaoka, A. *et al.* Mammalian 5-formyluracil-DNA glycosylase. 2. Role of SMUG1 uracil-DNA glycosylase in repair of 5-formyluracil and other oxidized and deaminated base lesions. *Biochemistry* **42,** 5003–5012 (2003).
63. Bauer, N. C., Corbett, A. H. & Doetsch, P. W. The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res.* **43,** 10083–10101 (2015).
64. Rusmintratip, V. & Sowers, L. C. An unexpectedly high excision capacity for mispaired 5-hydroxymethyluracil in human cell extracts. *Proc. Natl. Acad. Sci.* **97,** 14183–14187 (2000).
65. Miyabe, I. *et al.* Identification of 5-formyluracil DNA glycosylase activity of human hNTH1 protein. *Nucleic Acids Res.* **30,** 3443–3448 (2002).
66. Liu, P., Burdzy, A. & Sowers, L. C. Repair of the mutagenic DNA oxidation product, 5-formyluracil. *DNA Repair (Amst).* **2,** 199–210 (2003).
67. Frenkel, K. *et al.* Serum autoantibodies recognizing 5-hydroxymethyl-2'-deoxyuridine, an oxidized DNA base, as biomarkers of cancer risk in women. *Cancer Epidemiol. Biomarkers Prev.* **7,** 49–57 (1998).
68. Djuric, Z. *et al.* Levels of 5-hydroxymethyl-2′-deoxyuridine in DNA from blood as a marker of breast cancer. *Cancer* **77,** 691–696 (1996).
69. Djuric, Z. *et al.* Levels of 5-hydroxymethyl-2'-deoxyuridine in DNA from blood of women scheduled for breast biopsy. *Cancer Epidemiol. Biomarkers Prev.* **10,** 147–

149 (2001).
70. Fritz, E. L. & Papavasiliou, F. N. Cytidine deaminases: AIDing DNA demethylation? *Genes Dev.* **24,** 2107–2114 (2010).
71. Santos, F. *et al.* Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. *Epigenetics Chromatin* **6,** 39 (2013).
72. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell.* **146,** 67–79 (2011).
73. Rebhandl, S., Huemer, M., Greil, R. & Geisberger, R. AID/APOBEC deaminases and cancer. *Oncoscience* **2,** 320–333 (2015).
74. Nabel, C. S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat. Chem. Biol.* **8,** 751–758 (2012).
75. Guo, J. U., Su, Y., Zhong, C., Ming, G. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145,** 423–434 (2011).
76. Pais, J. E. *et al.* Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 4316–21 (2015).
77. Klungland, A. *et al.* 5-Formyluracil and its nucleoside derivatives confer toxicity and mutagenicity to mammalian cells by interfering with normal RNA and DNA metabolism. *Toxicol. Lett.* **119,** 71–78 (2001).
78. Privat, E. J. & Sowers, L. C. A proposed mechanism for the mutagenicity of 5-formyluracil. *Mutat. Res. Mol. Mech. Mutagen.* **354,** 151–156 (1996).
79. Rogstad, D. K. *et al.* 5-Formyluracil-Induced Perturbations of DNA Function. *Biochemistry* **43,** 5688–5697 (2004).
80. Kittaka, A. *et al.* Introduction of 5-Formyl-2 '-deoxyuridine into a k B Site : Critical Discrimination of a Base Structure in the Major Groove by NFk B p50 homodimer. *Synlett* 869–872 (1999).
81. Kemmerich, K., Dingler, F. A., Rada, C. & Neuberger, M. S. Germline ablation of SMUG1 DNA glycosylase causes loss of 5-hydroxymethyluracil- and UNG-backup uracil-excision activities and increases cancer predisposition of Ung−/−Msh2−/− mice. *Nucleic Acids Res.* **40,** 6016 (2012).
82. Krokan, H. E., Drabløs, F. & Slupphaug, G. Uracil in DNA--occurrence, consequences and repair. *Oncogene* **21,** 8935–8948 (2002).
83. Boorstein, R. J., Chiu, L. N. & Teebor, G. W. Phylogenetic evidence of a role for 5-hydroxymethyluracil-DNA glycosylase in the maintenance of 5-methylcytosine in DNA. *Nucleic Acids Res.* **17,** 7653–7661 (1989).
84. Waschke, S., Reefschlager, J., Barwolff, D. & Langen, P. 5-hydroxymethyl-2'-deoxyuridine, a normal DNA constituent in certain Bacillus subtilis phages is cytostatic for mammalian cells. *Nature* **255,** 629–630 (1975).
85. Boorstein, R. J., Chiu, L. N. & Teebor, G. W. A mammalian cell line deficient in activity of the DNA repair enzyme 5-hydroxymethyluracil-DNA glycosylase is resistant to the toxic effects of the thymidine analog 5-hydroxymethyl-2'-deoxyuridine. *Mol. Cell. Biol.* **12,** 5536–5540 (1992).
86. Abdel-Fatah, T. M. A. *et al.* Single-strand selective monofunctional uracil-DNA glycosylase (SMUG1) deficiency is linked to aggressive breast cancer and predicts response to adjuvant therapy. *Breast Cancer Res. Treat.* **142,** 515–527 (2013).
87. Xie, H., Gong, Y., Dai, J., Wu, X. & Gu, J. Genetic variations in base excision repair pathway and risk of bladder cancer: a case-control study in the United States. *Mol. Carcinog.* **54,** 50–57 (2015).
88. Ganguly, T. & Duker, N. J. Reduced 5-hydroxymethyluracil-DNA glycosylase activity in Werner's syndrome cells. *Mutat. Res.* **275,** 87–96 (1992).
89. An, Q., Robins, P., Lindahl, T. & Barnes, D. E. C - T mutagenesis and c -radiation sensitivity due to deficiency in the Smug1 and Ung DNA glycosylases. **24,** 2205–

2213 (2005).
90. Liu, S. *et al.* Quantitative Mass Spectrometry-Based Analysis of β-D-Glucosyl-5-hydroxymethyluracil in Genomic DNA of Trypanosoma brucei. *J. Am. Soc. Mass Spectrom.* **25,** 1763–1770 (2014).
91. Lopes, A. H. *et al.* Trypanosomatids : Odd Organisms , Devastating Diseases. *Open Parasitol. J.* **4,** 30–59 (2010).
92. Cliffe, L. J. *et al.* JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res.* **37,** 1452–1462 (2009).
93. Cliffe, L. J. *et al.* JBP1 and JBP2 proteins are Fe2+/2-oxoglutarate-dependent dioxygenases regulating hydroxylation of thymidine residues in trypanosome DNA. *J. Biol. Chem.* **287,** 19886–19895 (2012).
94. Bullard, W., Lopes Da Rosa-Spiegler, J., Liu, S., Wang, Y. & Sabatini, R. Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. *J. Biol. Chem.* **289,** 20273–20282 (2014).
95. Bullard, W., Cliffe, L., Wang, P., Wang, Y. & Sabatini, R. Base J glucosyltransferase does not regulate the sequence specificity of J synthesis in trypanosomatid telomeric DNA. *Mol. Biochem. Parasitol.* **204,** 77–80 (2015).
96. Borst, P. & Sabatini, R. Base J: discovery, biosynthesis, and possible functions. *Annu. Rev. Microbiol.* **62,** 235–251 (2008).
97. Hazelbaker, D. Z. & Buratowski, S. Base J: Blocking RNA Polymerase's Way. *Curr. Biol.* **22,** R960–R962 (2012).
98. van Luenen, H. G. A. M. *et al.* Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania. *Cell* **150,** 909–921 (2012).
99. Reynolds, D. *et al.* Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Res.* **42,** 9717–9729 (2014).
100. Cliffe, L. J., Siegel, T. N., Marshall, M., Cross, G. A. M. & Sabatini, R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Res.* **38,** 3923–3935 (2010).
101. Blundell, P. A., van Leeuwen, F., Brun, R. & Borst, P. Changes in expression site control and DNA modification in Trypanosoma brucei during differentiation of the bloodstream form to the procyclic form. *Mol. Biochem. Parasitol.* **93,** 115–130 (1998).
102. Bartholomeu, D. C., de Paiva, R. M. C., Mendes, T. A. O., DaRocha, W. D. & Teixeira, S. M. R. Unveiling the Intracellular Survival Gene Kit of Trypanosomatid Parasites. *PLoS Pathog.* **10,** e1004399 (2014).
103. Araújo, P. R. & Teixeira, S. M. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in Trypanosoma cruzi - A Review. **106,** 257–266 (2011).
104. Reynolds, D. *et al.* Histone H3 Variant Regulates RNA Polymerase II Transcription Termination and Dual Strand Transcription of siRNA Loci in Trypanosoma brucei. *PLOS Genet.* **12,** e1005758 (2016).
105. Rasmussen, E. M. K. *et al.* DNA base modifications in honey bee and fruit fly genomes suggest an active demethylation machinery with species- and tissue-specific turnover rates. *Biochem. Biophys. Reports* **6,** 9–15 (2016).
106. Randerath, K. Thin-layer chromatography of nucleotides on layers of cellulose ion-exchangers. *Nature* **194,** 768–769 (1962).
107. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science.* **324,** 929–30 (2009).
108. Gommers-Ampt, J., Lutgerink, J. & Borst, P. A novel DNA nucleotide in Trypanosoma brucei only present in the mammalian phase of the life-cycle. *Nucleic Acids Res.* **19,** 1745–1751 (1991).

109. van Leeuwen, F., Kieft, R., Cross, M. & Borst, P. Biosynthesis and function of the modified DNA base beta-D-glucosyl-hydroxymethyluracil in Trypanosoma brucei. *Mol. Cell. Biol.* **18,** 5643–5651 (1998).
110. Clément, G. & Benhattar, J. A methylation sensitive dot blot assay (MS-DBA) for the quantitative analysis of DNA methylation in clinical samples. *J. Clin. Pathol.* **58,** 155–158 (2005).
111. Shahal, T. *et al.* Spectroscopic Quantification of 5-Hydroxymethylcytosine in Genomic DNA. *Anal. Chem.* **86,** 8231–8237 (2014).
112. Greer, E. L. *et al.* DNA Methylation on N6-Adenine in C. elegans. *Cell* **161,** 868–878 (2015).
113. Taghizadeh, K. *et al.* Quantification of DNA damage products resulting from deamination, oxidation and reaction with products of lipid peroxidation by liquid chromatography isotope dilution tandem mass spectrometry. *Nat. Protoc.* **3,** 1287–1298 (2008).
114. Balasubramanian, S. Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun.* **47,** 7281 (2011).
115. Raiber, E.-A., Hardisty, R., van Delft, P. & Balasubramanian, S. Mapping and elucidating the function of modified bases in DNA. **1,** 69 (2017).
116. Pastor, W. A., Huang, Y., Henderson, H. R., Agarwal, S. & Rao, A. The GLIB technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat. Protoc.* **7,** 1909–1917 (2012).
117. Song, C.-X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotech* **29,** 68–72 (2011).
118. Mohn, F., Weber, M., Schubeler, D. & Roloff, T.-C. Methylated DNA immunoprecipitation (MeDIP). *Methods Mol. Biol.* **507,** 55–64 (2009).
119. Nestor, C. E. & Meehan, R. R. Hydroxymethylated DNA immunoprecipitation (hmeDIP). *Methods Mol. Biol.* **1094,** 259–267 (2014).
120. Xiao, Y. *et al.* MeSiC: A Model-Based Method for Estimating 5 mC Levels at Single-CpG Resolution from MeDIP-seq. *Sci. Rep.* **5,** 14699 (2015).
121. Horton, J. R. *et al.* Structure of 5-hydroxymethylcytosine-specific restriction enzyme, AbaSI, in complex with DNA. *Nucleic Acids Res.* **42,** 7947–7959 (2014).
122. Darst, R. P., Pardo, C. E., Ai, L., Brown, K. D. & Kladde, M. P. Bisulfite Sequencing of DNA. *Curr. Protoc. Mol. Biol.* **7,** Unit-7.917 (2010).
123. Huang, Y. *et al.* The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. *PLoS One* **5,** e8888 (2010).
124. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336,** 934–937 (2012).
125. Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem.* **6,** 435–440 (2014).
126. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–1380 (2012).
127. Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat Methods.* **12,** 1047–1050 (2015).
128. Feng, Z. *et al.* Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic. *PLOS Comput. Biol.* **9,** e1002935 (2013).
129. Song, C.-X. *et al.* Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Meth* **9,** 75–77 (2012).
130. Chavez, L. *et al.* Simultaneous sequencing of oxidized methylcytosines produced by TET/JBP dioxygenases in Coprinopsis cinerea. *Proc. Natl. Acad. Sci.* **111,** E5149–E5158 (2014).
131. Genest, P. A. *et al.* Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. *Nucleic Acids Res.* **43,** 2102–2115 (2015).

132. Wanunu, M. *et al.* Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J. Am. Chem. Soc.* **133,** 486–492 (2011).
133. Laszlo, A. H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18904–9 (2013).
134. Simpson, J. T. *et al.* Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer. *bioRxiv* (2016). doi:10.1101/047142
135. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37,** 853–862 (2005).
136. Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473,** 394–397 (2011).
137. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473,** 398–402 (2011).
138. Robertson, A. B., Dahl, J. A., Ougland, R. & Klungland, A. Pull-down of 5-hydroxymethylcytosine DNA using JBP1-coated magnetic beads. *Nat. Protoc.* **7,** 340–350 (2012).
139. Sérandour, A. A. *et al.* Single-CpG resolution mapping of 5-hydroxymethylcytosine by chemical labeling and exonuclease digestion identifies evolutionarily unconserved CpGs as TET targets. *Genome Biol.* **17,** 56 (2016).
140. Sun, Z. *et al.* High-Resolution Enzymatic Mapping of Genomic 5-Hydroxymethylcytosine in Mouse Embryonic Stem Cells. *Cell Rep.* **3,** 567–576 (2013).
141. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol Cell.* **57,** 750–61 (2015).
142. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32,** 1231–1240 (2014).
143. Zhu, C. *et al.* Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **20,** 720–731 (2017).
144. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res.* **25,** 386–389 (2015).
145. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32,** 1231–1240 (2014).
146. Ding, Y., Fleming, A. M. & Burrows, C. J. Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **139,** 2569–2572 (2017).
147. Torres, I. O. & Fujimori, D. G. Functional coupling between writers, erasers and readers of histone and DNA methylation. *Curr. Opin. Struct. Biol.* **35,** 68–75 (2015).
148. Hendrich, B. & Bird, A. Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. *Mol. Cell. Biol.* **18,** 6538–6547 (1998).
149. Choy, J. S. *et al.* DNA Methylation Increases Nucleosome Compaction and Rigidity. *J. Am. Chem. Soc.* **132,** 1782–1783 (2010).
150. Ngo, T. T. M. *et al.* Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* **7,** 10813 (2016).
151. Lee, J. Y. & Lee, T.-H. Effects of DNA Methylation on the Structure of Nucleosomes. *J. Am. Chem. Soc.* **134,** 173–175 (2012).
152. Mendonca, A., Chang, E. H., Liu, W. & Yuan, C. Hydroxymethylation of DNA influences nucleosomal conformation and stability in vitro. *Biochim. Biophys. Acta* **1839,** 1323–1329 (2014).
153. Teif, V. B. *et al.* Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res.* **24,** 1285–

1295 (2014).
154. Collings, C. K., Waddell, P. J. & Anderson, J. N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* **41,** 2918–2931 (2013).
155. Collings, C. K. & Anderson, J. N. Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenetics Chromatin* **10,** 18 (2017).
156. Van Eeckhaut, A., Lanckmans, K., Sarre, S., Smolders, I. & Michotte, Y. Validation of bioanalytical LC–MS/MS assays: Evaluation of matrix effects. *J. Chromatogr. B* **877,** 2198–2207 (2009).
157. Sargent, M. Guide to achieving reliable quantitiative LC-MS measurements. *ISBN 978-0-948926-27-3* **RSC Analyt,** (2013).
158. Taghizadeh, K. *et al.* Quantification of DNA damage products resulting from deamination, oxidation and reaction with products of lipid peroxidation by liquid chromatography isotope dilution tandem mass spectrometry. *Nat. Protoc.* **3,** 1287–1298 (2008).
159. Wagner, M. *et al.* Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. *Angew. Chem. Int. Ed. Engl.* **54,** 12511–12514 (2015).
160. Gackowski, D., Zarakowska, E., Starczak, M., Modrzejewska, M. & Olinski, R. Tissue-Specific Differences in DNA Modifications (5-Hydroxymethylcytosine, 5-Formylcytosine, 5-Carboxylcytosine and 5-Hydroxymethyluracil) and Their Interrelationships. *PLoS One* **10,** e0144859 (2015).
161. Zarakowska, E., Gackowski, D., Foksinski, M. & Olinski, R. Are 8-oxoguanine (8-oxoGua) and 5-hydroxymethyluracil (5-hmUra) oxidatively damaged DNA bases or transcription (epigenetic) marks? *Mutat. Res. Toxicol. Environ. Mutagen.* **764–765,** 58–63 (2014).
162. Liu, S. *et al.* Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. *Nucleic Acids Res.* **41,** 6421–9 (2013).
163. Guo, P. *et al.* Synthesis and spectroscopic properties of fluorescent 5-benzimidazolyl-2[prime or minute]-deoxyuridines 5-fdU probes obtained from o-phenylenediamine derivatives. *Org. Biomol. Chem.* **11,** 1610–1613 (2013).
164. Soldatenkov, A. T., Polyanskii, K. B., Kolyadina, N. M. & Soldatova, S. A. Oxidation of heterocyclic compounds by manganese dioxide (Review). *Chem. Heterocycl. Compd.* **45,** 633 (2009).
165. Liguori, A., Napoli, A. & Sindona, G. Determination of substituent effects on the proton affinities of natural nucleosides by the kinetic method. *Rapid Commun. Mass Spectrom.* **8,** 89–93 (1994).
166. van Leeuwen, F., Kieft, R., Cross, M. & Borst, P. Biosynthesis and Function of the Modified DNA Base β-d-Glucosyl-Hydroxymethyluracil in Trypanosoma brucei. *Mol. Cell. Biol.* **18,** 5643–5651 (1998).
167. Dipaolo, C., Kieft, R., Cross, M. & Sabatini, R. Regulation of Trypanosome DNA Glycosylation by a SWI2 / SNF2-like Protein. **17,** 441–451 (2005).
168. Militello, K. T. *et al.* African Trypanosomes Contain 5-Methylcytosine in Nuclear DNA . *Eukaryot. Cell* **7,** 2012–2016 (2008).
169. Valentine, E., Smindak, R., Militello, K. T. The Detection of 5-hydroxymethylcytosine in Trypanosoma brucei DNA. *FASEB J.* **27.1 Suppl,** 981–987 (2013).
170. Amoroso, A. *et al.* Oxidative DNA damage bypass in Arabidopsis thaliana requires DNA polymerase lambda and proliferating cell nuclear antigen 2. *Plant Cell* **23,** 806–822 (2011).
171. Nilsen, H. *et al.* Excision of deaminated cytosine from the vertebrate genome: role of the SMUG1 uracil–DNA glycosylase. *EMBO J.* **20,** 4278–4286 (2001).
172. Hong, H. & Wang, Y. Derivatization with Girard Reagent T Combined with LC–MS/MS for the Sensitive Detection of 5-Formyl-2‘-deoxyuridine in Cellular DNA. *Anal. Chem.* **79,** 322–326 (2007).

173. Jewett, J. C. & Bertozzi, C. R. Cu-free click cycloaddition reactions in chemical biology. *Chem. Soc. Rev.* **39,** 1272–1279 (2010).
174. Barbin, A. *et al.* Endogenous deoxyribonucleic Acid (DNA) damage in human tissues: a comparison of ethenobases with aldehydic DNA lesions. *Cancer Epidemiol. Biomarkers Prev.* **12,** 1241–1247 (2003).
175. Sugiyama, H. *et al.* Chemistry of Thermal Degradation of Abasic Sites in DNA. Mechanistic Investigation on Thermal DNA Strand Cleavage of Alkylated DNA. *Chem. Res. Toxicol.* **7,** 673–683 (1994).
176. Ji, D. & Wang, Y. Facile Enzymatic Synthesis of Base J-Containing Oligodeoxyribonucleotides and an Analysis of the Impact of Base J on DNA Replication in Cells. *PLoS One* **9,** e103335 (2014).
177. Yu, M., Song, C.-X. & He, C. Detection of mismatched 5-hydroxymethyluracil in DNA by selective chemical labeling. *Methods* **72,** 16–20 (2015).
178. Xu, X. *et al.* One-step to get 5-azidomethyl-2′-deoxyuridine from 5-hydroxymethyl-2′-deoxyuridine and detection of it through click reaction. *Tetrahedron* **69,** 9870–9874 (2013).
179. Noguchi, M. *et al.* A dimethoxytriazine type glycosyl donor enables a facile chemo-enzymatic route toward [small alpha]-linked N-acetylglucosaminyl-galactose disaccharide unit from gastric mucin. *Chem. Commun.* **48,** 5560–5562 (2012).
180. Akhlaghinia, B. & Samiei, S. A Novel and highly selective conversion of alcohols, thiols, and silyl ethers to azides using the 2,4,6-trichloro[1,3,5]triazine/n-Bu4NN3 system. *J. Braz. Chem. Soc.* **18,** 1311–1315 (2007).
181. Thompson, S. K. *et al.* Structure-Based Design of Cathepsin K Inhibitors Containing a Benzyloxy-Substituted Benzoyl Peptidomimetic. *J. Med. Chem.* **41,** 3923–3927 (1998).
182. Hardisty, R. E., Kawasaki, F., Sahakyan, A. B. & Balasubramanian, S. Selective Chemical Labeling of Natural T Modifications in DNA. *J. Am. Chem. Soc.* **137,** 9270–9272 (2015).
183. Armstrong, V. W., Dattagupta, J. K., Eckstein, F. & Saenger, W. The base catalysed anomerisation of β-5-formyluridine; crystal and molecular structure of α-5-formyluridine. *Nucleic Acids Res.* **3,** 1791–1810 (1976).
184. Munzel, M. *et al.* Improved synthesis and mutagenicity of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. *Chemistry* **17,** 13782–13788 (2011).
185. Jia, X.-F., Liu, N., Fang, L.-L. & Zhang, X.-Y. [(2,3,5)-3-Acetoxy-5-(5-formyl-2,4-dioxo-1,2,3,4-tetrahydropyrimidin-1-yl)-2,3,4,5-tetrahydrofuran-2-yl]methyl acetate. *Acta Crystallogr. Sect. E* **67,** o2951--o2952 (2011).
186. He, Y., Cui, L.-Y. & Zhang, X.-Y. 2,4-Dioxo-1-(prop-2-ynyl)-1,2,3,4-tetrahydropyrimidine-5-carbaldehyde. *Acta Crystallogr. Sect. E* **67,** o2350 (2011).
187. Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* **23,** 1833–1840 (1955).
188. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **5,** 129–145 (1984).
189. Besler, B. H., Merz, K. M. & Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **11,** 431–439 (1990).
190. Reed, A. E. & Weinhold, F. Natural bond orbital analysis of near-Hartree-Fock water dimer. *J. Chem. Phys.* **78,** 4066–4073 (1983).
191. Foster, J. P. & Weinhold, F. Natural Hybrid Orbitals. *J. Am. Chem. Soc.* **102,** 7211–7218 (1980).
192. Ide, H. *et al.* Synthesis and damage specificity of a novel probe for the detection of abasic sites in DNA. *Biochemistry* **32,** 8276–8283 (1993).
193. Grúz, P. *et al.* Processing of DNA lesions by archaeal DNA polymerases from Sulfolobus solfataricus. *Nucleic Acids Res.* **31,** 4024–4030 (2003).
194. Gates, K. S. An Overview of Chemical Processes That Damage Cellular DNA:

Spontaneous Hydrolysis, Alkylation, and Reactions with Radicals. *Chem. Res. Toxicol.* **22,** 1747–1760 (2009).
195. Oyola, S. O. *et al.* Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* **13,** (2012).
196. Head, S. R. *et al.* Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* **56,** 61–passim (2014).
197. McInroy, G. R., Raiber, E. A. & Balasubramanian, S. Chemical biology of genomic DNA: minimizing PCR bias. *Chem Commun* **50,** 12047–12049 (2014).
198. Dirksen, A., Dirksen, S., Hackeng, T. M. & Dawson, P. E. Nucleophilic catalysis of hydrazone formation and transimination: implications for dynamic covalent chemistry. *J. Am. Chem. Soc.* **128,** 15602–15603 (2006).
199. Rashidian, M., Song, J. M., Pricer, R. E. & Distefano, M. D. Chemoenzymatic reversible immobilization and labeling of proteins without prior purification. *J. Am. Chem. Soc.* **134,** 8455–8467 (2012).
200. Bullard, W., Kieft, R. & Sabatini, R. A method for the efficient and selective identification of 5-hydroxymethyluracil in genomic DNA. *Biol. Methods Protoc.* **2,** 1–10 (2017).
201. Liu, C. *et al.* Enrichment and fluorogenic labelling of 5-formyluracil in DNA. *Chem. Sci.* **8,** 4505–4510 (2017).
202. Kawasaki, F. *et al.* Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biol.* **18,** 23 (2017).
203. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** (2008).
204. Engstler, M. & Boshart, M. Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in Trypanosoma brucei. *Genes Dev.* **18,** 2798–2811 (2004).
205. Overath, P., Czichos, J. & Haas, C. The effect of citrate/cis-aconitate on oxidative metabolism during transformation of Trypanosoma brucei. *Eur. J. Biochem.* **160,** 175–182 (1986).
206. Sultan, M. *et al.* A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science (80-.).* **321,** 956–60 (2008).
207. Alsoe, L. *et al.* Uracil Accumulation and Mutagenesis Dominated by Cytosine Deamination in CpG Dinucleotides in Mice Lacking UNG and SMUG1. *Sci. Rep.* **7,** 7199 (2017).
208. Theis, M. & Buchholz, F. MISSION esiRNA for RNAi Screening in Mammalian Cells. *J. Vis. Exp.* 2008 (2010).
209. Blau, J. A. & McManus, M. T. Renewable RNAi. *Nat Biotech* **31,** 319–320 (2013).
210. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–1831 (2012).
211. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36,** e105 (2008).
212. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27,** 1696–1697 (2011).
213. Deplus, R. *et al.* TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *EMBO J.* **32,** 645–655 (2013).
214. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473,** 389–393 (2011).
215. Amouroux, R., Campalans, A., Epe, B. & Radicella, J. P. Oxidative stress triggers the preferential assembly of base excision repair complexes on open chromatin regions. *Nucleic Acids Res.* **38,** 2878–2890 (2010).
216. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* 123–131 (2006).
217. Lan, L. *et al.* Novel method for site-specific induction of oxidative DNA damage reveals differences in recruitment of repair proteins to heterochromatin and

euchromatin. *Nucleic Acids Res.* **42,** 2330–2345 (2014).
218. Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24,** 2537–2538 (2008).
219. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics.* **29,** 2046–8 (2013).
220. Jin, S.-G., Wu, X., Li, A. X. & Pfeifer, G. P. Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic Acids Res.* **39,** 5015–5024 (2011).
221. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8,** R183 (2007).
222. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* **4,** 44–57 (2009).
223. van der Meer, L. T., Jansen, J. H. & van der Reijden, B. A. Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia* **24,** 1834–1843 (2010).
224. Rice, D. S. *et al.* Disabled-1 acts downstream of Reelin in a signaling pathway that controls laminar organization in the mammalian brain. *Development* **125,** 3719–3729 (1998).
225. Amente, S. *et al.* LSD1-mediated demethylation of histone H3 lysine 4 triggers Myc-induced transcription. *Oncogene* **29,** 3691–3702 (2010).
226. Price, M. A. *et al.* CSPG4, a potential therapeutic target, facilitates malignant progression of melanoma. *Pigment Cell Melanoma Res.* **24,** 1148–1157 (2011).
227. Liebler, D. Introduction to Proteomics: Tools for the New Biology. 50 (2002).
228. Mann, M. Origins of mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol* **17,** 678 (2016).
229. Rivlin, N., Brosh, R., Oren, M. & Rotter, V. Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes Cancer* **2,** 466–474 (2011).
230. Searle, B. C. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **10,** 1265–1269 (2010).
231. Freund, D. M. & Prenni, J. E. Improved Detection of Quantitative Differences Using a Combination of Spectral Counting and MS/MS Total Ion Current. *J. Proteome Res.* **12,** 1996–2004 (2013).
232. Miyamoto, K. N. *et al.* Comparative proteomic analysis of Listeria monocytogenes ATCC 7644 exposed to a sublethal concentration of nisin. *J. Proteomics* **119,** 230–237 (2015).
233. Robison, F. M., Heuberger, A. L., Brick, M. A. & Prenni, J. E. Proteome Characterization of Leaves in Common Bean. *Proteomes* **3,** 236–248 (2015).
234. Aqrawi, L. A. *et al.* Identification of potential saliva and tear biomarkers in primary Sjögren's syndrome, utilising the extraction of extracellular vesicles and proteomics analysis. *Arthritis Res. Ther.* **19,** 14 (2017).
235. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
236. Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37,** D412-6 (2009).
237. Tesfazghi, M., Riman, S., Alexander, K., Rizkallah, R. & Hurt, M. The Recruitment of the Transcription Factor YY1 to DNA Damage Sites in Human Cells. *FASEB J.* **29,** no. 1 Supplement LB186 (2015).
238. Azam, S. *et al.* Human Glyceraldehyde-3-phosphate Dehydrogenase Plays a Direct Role in Reactivating Oxidized Forms of the DNA Repair Enzyme APE1. *J. Biol.*

Chem. **283,** 30632–30641 (2008).
239. de Peralta, M. S. P. *et al.* Cnbp ameliorates Treacher Collins Syndrome craniofacial anomalies through a pathway that involves redox-responsive genes. *Cell Death Dis.* **7,** e2397 (2016).
240. Duan, X., Kelsen, S. G., Clarkson, A. B., Ji, R. & Merali, S. SILAC Analysis of Oxidative Stress-Mediated Proteins in Human Pneumocytes: New Role for Treacle. *Proteomics* **10,** 2165–2174 (2010).
241. Lubos, E., Loscalzo, J. & Handy, D. E. Glutathione Peroxidase-1 in Health and Disease: From Molecular Mechanisms to Therapeutic Opportunities. *Antioxid. Redox Signal.* **15,** 1957–1997 (2011).
242. Li, L., Monckton, E. A. & Godbout, R. A role for DEAD box 1 at DNA double-strand breaks. *Mol. Cell. Biol.* **28,** 6413–6425 (2008).
243. Shih, J.-W. & Lee, Y.-H. W. Human DExD/H RNA helicases: emerging roles in stress survival regulation. *Clin. Chim. Acta.* **436,** 45–58 (2014).
244. Lockman, K. A. *et al.* Proteomic profiling of cellular steatosis with concomitant oxidative stress in vitro. *Lipids Health Dis.* **15,** 114 (2016).
245. Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* **25,** 1125–1142 (2006).
246. Iben, S. *et al.* TFIIH Plays an Essential Role in RNA Polymerase I Transcription. *Cell* **109,** 297–306 (2002).
247. Baek, H. J., Kang, Y. K. & Roeder, R. G. Human Mediator enhances basal transcription by facilitating recruitment of transcription factor IIB during preinitiation complex assembly. *J. Biol. Chem.* **281,** 15172–15181 (2006).
248. Brand, M. *et al.* UV-damaged DNA-binding protein in the TFTC complex links DNA damage recognition to nucleosome acetylation. *EMBO J.* **20,** 3187–3196 (2001).
249. Silverstein, R. A. & Ekwall, K. Sin3: a flexible regulator of global gene expression and genome stability. *Curr. Genet.* **47,** 1–17 (2005).
250. Lalmansingh, A. S., Urekar, C. J. & Reddi, P. P. TDP-43 Is a Transcriptional Repressor: The Testis-Specific Mouse acrv1 Gene is a TDP-43 Target In Vivo. *J. Biol. Chem.* **286,** 10970–10982 (2011).
251. Kume, T., Jiang, H., Topczewska, J. M. & Hogan, B. L. M. The murine winged helix transcription factors, Foxc1 and Foxc2, are both required for cardiovascular development and somitogenesis. *Genes Dev.* **15,** 2470–2482 (2001).
252. Rountree, M. R., Bachman, K. E. & Baylin, S. B. DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nat. Genet.* **25,** 269–277 (2000).
253. Volk, A. & Crispino, J. D. The role of the chromatin assembly complex (CAF-1) and its p60 subunit (CHAF1b) in homeostasis and disease. *Biochim. Biophys. Acta* **1849,** 979–986 (2015).
254. Steward, M. M. *et al.* Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nat. Struct. Mol. Biol.* **13,** 852–854 (2006).
255. Bian, C. *et al.* Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. *EMBO J.* **30,** 2829–2842 (2011).
256. Schormann, N., Ricciardi, R. & Chattopadhyay, D. Uracil-DNA glycosylases—Structural and functional perspectives on an essential family of DNA repair enzymes. *Protein Sci.* **23,** 1667–1685 (2014).
257. Jiang, Y. L., Gao, X., Zhou, G., Patel, A. & Javer, A. Selective Recognition of Uracil and Its Derivatives Using a DNA Repair Enzyme Structural Mimic. *J. Org. Chem.* **75,** 324–333 (2010).
258. Doseth, B. *et al.* Uracil-DNA glycosylase in base excision repair and adaptive immunity: species differences between man and mouse. *J. Biol. Chem.* **286,** 16669–16680 (2011).
259. Millioni, R. *et al.* High abundance proteins depletion vs low abundance proteins

enrichment: comparison of methods to reduce the plasma proteome complexity. *PLoS One* **6,** e19603 (2011).
260. Cole, A. R. *et al.* Architecturally diverse proteins converge on an analogous mechanism to inactivate Uracil-DNA glycosylase. *Nucleic Acids Res.* **41,** 8760–8775 (2013).
261. Mol, C. D. *et al.* Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. *Cell* **80,** 869–878 (1995).
262. Handa, P., Acharya, N. & Varshney, U. Effects of mutations at tyrosine 66 and asparagine 123 in the active site pocket of Escherichia coli uracil DNA glycosylase on uracil excision from synthetic DNA oligomers: evidence for the occurrence of long-range interactions between the enzyme and subs. *Nucleic Acids Res.* **30,** 3086–3095 (2002).
263. Liu, P., Burdzy, A. & Sowers, L. C. Substrate Recognition by a Family of Uracil-DNA Glycosylases: UNG, MUG, and TDG. *Chem. Res. Toxicol.* **15,** 1001–1009 (2002).
264. Mauro, D. J., De Riel, J. K., Tallarida, R. J. & Sirover, M. A. Mechanisms of excision of 5-fluorouracil by uracil DNA glycosylase in normal human cells. *Mol. Pharmacol.* **43,** 854–857 (1993).
265. Radman-Livaja, M. & Rando, O. J. Nucleosome positioning: how is it established, and why does it matter? *Dev Biol* **339,** (2010).
266. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10,** 161–172 (2009).
267. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
268. Kawasaki, F., Murat, P., Li, Z., Santner, T. & Balasubramanian, S. Synthesis and biophysical analysis of modified thymine-containing DNA oligonucleotides. *Chem. Commun.* **53,** 1389–1392 (2017).
269. Raiber, E.-. A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat Struct Mol Biol.* **22,** 44–49 (2015).
270. Hardwick, J. S. *et al.* 5-Formylcytosine does not change the global structure of DNA. *Nat Struct Mol Biol* **24,** 544–552 (2017).
271. Fenouil, R. *et al.* CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* **22,** 2399–2408 (2012).
272. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).
273. Dohno, C., Okamoto, A. & Saito, I. Stable, Specific, and Reversible Base Pairing via Schiff Base. *J. Am. Chem. Soc.* **127,** 16681–16684 (2005).
274. Knocel, P.; Molander, G. Comprehensive Organic Synthesis. *Elsevier* **2nd Edition,** 97 (2014).
275. Wolff, E. C., Folk, J. E. & Park, M. H. Enzyme-substrate intermediate formation at lysine 329 of human deoxyhypusine synthase. *J. Biol. Chem.* **272,** 15865–15871 (1997).
276. Raindlova, V., Pohl, R. & Hocek, M. Synthesis of aldehyde-linked nucleotides and DNA and their bioconjugations with lysine and peptides through reductive amination. *Chemistry* **18,** 4080–4087 (2012).
277. Yang, B. *et al.* Bioconjugation of Oligodeoxynucleotides Carrying 1,4-Dicarbonyl Groups via Reductive Amination with Lysine Residues. *Bioconjug. Chem.* **26,** 1830–1838 (2015).
278. Johnson, K. M. *et al.* On the formation and properties of interstrand DNA-DNA cross-links forged by reaction of an abasic site with the opposing guanine residue of 5'-CAp sequences in duplex DNA. *J. Am. Chem. Soc.* **135,** 1015–1025 (2013).
279. Wickramaratne, S., Mukherjee, S., Villalta, P. W., Schärer, O. D. & Tretyakova, N. Y. Synthesis of Sequence-Specific DNA–Protein Conjugates via a Reductive Amination Strategy. *Bioconjug. Chem.* **24,** 1496–1506 (2013).

280. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning1. *J. Mol. Biol.* **276,** 19–42 (1998).
281. Olsen, J. V, Ong, S.-E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614 (2004).
282. Keil, B. Specificity of proteolysis. *Springer-Verlag Berlin-Heidelberg-NewYork* 335 (1992).
283. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* **11,** 993–1006 (2016).
284. Cousin, D. *et al.* Antitumor imidazo[5,1-d]-1,2,3,5-tetrazines: compounds modified at the 3-position overcome resistance in human glioblastoma cell lines. *Med. Chem. Commun.* **7,** 2332–2343 (2016).
285. Yeo, J. E. *et al.* Synthesis of Site-Specific DNA–Protein Conjugates and Their Effects on DNA Replication. *ACS Chem. Biol.* **9,** 1860–1868 (2014).
286. Shu, X., Xiong, X., Song, J., He, C. & Yi, C. Base-Resolution Analysis of Cisplatin–DNA Adducts at the Genome Scale. *Angew. Chemie Int. Ed.* **55,** 14246–14249 (2016).
287. Collepardo-Guevara, R. *et al.* Chromatin Unfolding by Epigenetic Modifications Explained by Dramatic Impairment of Internucleosome Interactions: A Multiscale Computational Study. *J. Am. Chem. Soc.* **137,** 10205–10215 (2015).
288. Li, F. *et al.* 5-Formylcytosine Yields DNA–Protein Cross-Links in Nucleosome Core Particles. *J. Am. Chem. Soc.* **139,** 10617–10620 (2017).
289. Ji, S., Shao, H., Han, Q., Seiler, C. L. & Tretyakova, N. Reversible DNA-protein cross-linking at epigenetic DNA marks. *Angew. Chemie Int. Ed.* (2017). doi:10.1002/anie.201708286
290. Kuang, Y., Sun, H., Blain, J. C. & Peng, X. Hypoxia-Selective DNA Interstrand Cross-Link Formation by Two Modified Nucleosides. *Chem. – A Eur. J.* **18,** 12609–12613 (2012).
291. Schiesser, S. *et al.* Deamination, Oxidation, and C–C Bond Cleavage Reactivity of 5-Hydroxymethylcytosine, 5-Formylcytosine, and 5-Carboxycytosine. *J. Am. Chem. Soc.* **135,** 14593–14599 (2013).
292. Häser, M. Møller-Plesset (MP2) perturbation theory for large molecules. *Theor. Chim. Acta* **87,** 147–173 (1993).
293. Head-Gordon, M., Pople, J. A. & Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **153,** 503–506 (1988).
294. Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90,** 1007 (1989).
295. Peterson, K. A., Woon, D. E. & Dunning, T. H. Benchmark calculations with correlated molecular wave functions. IV. The classical barrier height of the H+H2→H2+H reaction. *J. Chem. Phys.* **100,** 7410 (1994).
296. M. J. Frisch, G. W. T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G; M. J. Frisch, G. W. T., H. B. Sc, J. A. P. Gaussian 03 Revision E1. (2004).
297. C.~Peng & H.~B.~Schlegel. Combining Synchrous Transit and Quasi-Newton Methods to Find Transition States. *Isr. J. Chem.* **33,** 449 (1993).
298. Ziegelbauer, K., Quinten, M., Schwarz, H., Pearson, T. W. & Overath, P. Synchronous differentiation of Trypanosoma brucei from bloodstream to procyclic forms in vitro. *Eur. J. Biochem.* **192,** 373–378 (1990).
299. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).
300. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).

301. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13,** 952–962 (1992).
302. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **4,** 116–122 (2008).
303. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20,** 786–798 (1999).
304. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 14101 (2007).
305. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).
306. Best, R. B. & Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix–Coil Transition of Polypeptides. *J. Phys. Chem. B* **113,** 9004–9015 (2009).
307. Pérez, A. *et al.* Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys. J.* **92,** 3817–3829 (2007).
308. Smith, D. E. & Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **100,** 3757–3766 (1994).
309. Collepardo-Guevara, R. *et al.* Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: A multiscale computational study. *J. Am. Chem. Soc.* **137,** 10205–10215 (2015).
310. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution. *J. Mol. Biol.* **319,** 1097–1113 (2002).
311. Hong, H. & Wang, Y. Derivatization with Girard Reagent T Combined with LC-MS/MS for the Sensitive Detection of 5-Formyl-2′-deoxyuridine in Cellular DNA. *Anal. Chem.* **79,** 322–326 (2007).
312. Ho, N. W. & Gilham, P. T. The reversible chemical modification of uracil, thymine, and guanine nucleotides and the modification of the action of ribonuclease on ribonucleic acid. *Biochemistry* **6,** 3632–3639 (1967).