High-resolution mapping of abasic sites and pyrimidine modifications in DNA

Zheng Liu

Pembroke College

Department of Chemistry

University of Cambridge

August 2019

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

High-resolution mapping of abasic sites and pyrimidine modifications in DNA

Zheng Liu

The maintenance of genomic stability is critical for the growth and survival of cells. Cellular DNA is constantly subject to both endogenous and exogenous sources of damage, leading to the formation of damage products. Abasic sites occur when a nucleobase is lost from DNA by hydrolysis and can lead to mutations and genomic instability. This thesis focuses on the development and application of methodology to map the location of abasic sites in DNA by next-generation sequencing.

Chemically, abasic sites are reactive entities due to the aldehyde moiety in the ring-open form of the deoxyribose. Many studies on abasic sites have focused on targeting this aldehyde, however, a major drawback to this approach is the cross-reactivity of nucleophilic probes with other aldehyde-containing modifications that naturally occur within genomic DNA. In this thesis, a chemical method was developed and demonstrated to allow the sequencing of abasic sites at single-nucleotide resolution by affinity enrichment. Crucially, this method was shown to selectively target abasic sites in the presence of other reactive sites in DNA.

Glycosylase enzymes excise base modifications from DNA to generate an abasic site. By treating isolated DNA with a glycosylase *in vitro*, abasic site sequencing methodology can be widely applied to study a range of DNA base modifications. This approach was utilised to investigate the distribution of the modification 5-hydroxymethyluracil at single-nucleotide resolution in the DNA of trypanosomatids. This study provided proof-of-concept of the sequencing methodology in a genomic context and also revealed the genomic features and sequence motifs at which these sites accumulate.

The distribution of endogenous abasic sites was also explored in the human genome. The genomic location of abasic sites was mapped following depletion of the key repair enzyme, APE1, where an increase was observed in the number of enriched loci compared to control cells. A relationship was also revealed between this form of DNA damage and coding and regulatory regions of the genome upon knockdown of the APE1 protein.

The genomic location of the base modification uracil was studied in the mouse genome. This base has been implicated in a pathway proposed to regulate the epigenetic DNA marker, 5-methylcytosine during embryonic development. Sequencing of uracil was achieved by enzymatic conversion into abasic sites followed by genome-wide mapping to investigate the extent to which the proposed mechanism contributes towards epigenetic reprogramming.

Overall, a versatile method has been developed that can reveal the location of endogenous abasic sites within DNA at single-nucleotide resolution, as well as abasic sites generated *in vitro* from base modifications. This method is useful for studies on both DNA damage and DNA base modifications.

Acknowledgements

Firstly, I would like to thank my supervisor Professor Sir Shankar Balasubramanian for giving me the opportunity to carry out my PhD in such an exciting group. I am very grateful for his continued support and advice throughout my time here.

My special thanks go to Dr Pieter van Delft, to whom I am extremely grateful for all the help and day to day guidance he has given me since I first joined the Balasubramanian group. A huge thank you also goes to Dr Sergio Martínez Cuesta. Thank you for constantly providing results throughout the copious amounts of data, and for your endless patience. Thanks also to Dr Fumiko Kawasaki and Dr Robyn Hardisty for their constant advice and guidance.

I am extremely grateful for the wonderful collaborators I have had the opportunity to work with during my studies. Special thanks go to Professor Wolf Reik, Dr Fátima Santos and Dr Poppy Gould from the Babraham Institute, and Professor Mark Carrington from the Department of Biochemistry for sharing their expertise, providing samples and many useful discussions.

Other members of the Balasubramanian group, past and present, have all made my time here so enjoyable. It has been a pleasure to work alongside so many wonderful and friendly people, many thanks go to Kim, Robyn, Zhe, Areeb, Euni, Fumi, Piet, Marco, Alex, Antanas, Christina, Matt, Max and Chloe. A special mention to Chris, Jo and David for all their day to day support. Thanks also to everyone at CI for being so helpful; a special thank you to Barbara and Robert for all the cell culture and sequencing advice. I am also very grateful to Shankar, Chris, Alex, Areeb, Antanas and Ben for their proofreading of this thesis and useful comments.

I would like to thank my family for their constant support and encouragement. Thanks also to Ben for being so supportive and making my time in Cambridge so much fun. Finally, I would like to acknowledge Pembroke College and the Herchel Smith Fund for funding my studies.

Publications

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

- Liu, Z. J., Martínez Cuesta, S., van Delft, P. & Balasubramanian, S. Sequencing abasic sites in DNA at single-nucleotide resolution. *Nat. Chem.* **11**, 629-637 (2019).

- Hofer, A., Liu, Z. J. & Balasubramanian, S. Detection, structure and function of modified DNA bases. *J. Am. Chem. Soc.* **141**, 6420-6429 (2019).

- Raiber, E-A. Portella, G., Martínez Cuesta, S., Hardisty, R., Murat, P., Li, Z., Iurlaro, M., Dean, W., Spindel, J., Beraldi, D., Liu, Z., Dawson, M. A., Reik, W. & Balasubramanian, S. 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat. Chem.* **10**, 1258-1266 (2018).

Abbreviations

5-caC	5-Carboxycytosine
5-fC	5-Formylcytosine
5-FU	5-Fluorouracil
5-fU	5-Formyluracil
5-hmC	5-Hydroxymethylcytosine
5-hmU	5-Hydroxymethyluracil
5-hoC	5-Hydroxycytosine
5-hoU	5-Hydroxyuracil
5-mC	5-Methylcytosine
6-mA	*N*6-methyladenine
7-mG	*N*7-methylguanine
8-oxoG	8-Oxoguanine
A	Adenine
AID	Activation induced cytosine deaminase
AP	Apurinic/apyrimidinic, abasic
APE1	AP endonuclease 1
APE2	AP endonuclease 2
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
ARP	Aldehyde reactive probe, *O*-biotinylcarbazoylmethyl hydroxylamine
Base J	5-(β-Glucopyranosyl)hydroxymethyluracil
BER	Base excision repair
bp	Base pairs
C	Cytosine
ChIP-seq	Chromatin immunoprecipitation sequencing
CuAAC	Copper catalysed azide-alkyne Huisgen cycloaddition
dA	2'-Deoxyadenosine
dATP	2'-Deoxyadenosine-5'-triphosphate
dCTP	2'-Deoxycytidine-5'-triphosphate
dfCTP	2'-Deoxy-5-formylcytidine-5'-triphosphate
dfUTP	2'-Deoxy-5-formyluridine-5'-triphosphate
dGTP	2'-Deoxyguanosine-5'-triphosphate
dhmUTP	2'-Deoxy-5-hydroxymethyluridine-5'-triphosphate
dhpU	5-Dihydroxypentyluracil

DIP-seq	DNA immunoprecipitation sequencing
DMOG	Dimethyloxalylglycine
dN	Deoxyribonucleotide
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	Deoxyribonucleotide triphosphate
DSB	Double-strand break
dsDNA	Double-stranded DNA
dTTP	Thymidine-5'-triphosphate
dUTP	2'-Deoxyuridine-5'-triphosphate
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
FEN1	Flap endonuclease 1
G	Guanine
HIPS	Hydrazino-*iso*-Pictet-Spengler
HMCES	5-Hydroxymethylcytosine binding ESC-specific enzyme
HPLC	High performance liquid chromatography
JBP	J binding protein
JGT	J-associated glucosyltransferase
kb	Kilobase
L. major	*Leishmania major*
LC-MS	Liquid chromatography mass spectrometry
LC-MS/MS	Liquid chromatography tandem mass spectrometry
M	Molar
MBD4	Methyl-CpG binding domain 4, DNA glycosylase
mESC	Mouse embryonic stem cell
MMR	Mismatch repair
mol	Moles
MX	Methoxyamine
NER	Nucleotide excision repair
NGS	Next generation sequencing
NMR	Nuclear magnetic resonance
NPC	Nucleosome core particle
ODN	Oligodeoxyribonucleotide
OGG1	8-Oxoguanine DNA glycosylase 1
OQS	Observed G-quadruplex sequence

PARP1	Poly ADP-ribose polymerase-1
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PQS	Predicted G-quadruplex sequence
putT	α-Putrescinylthymine
qPCR	Real-time polymerase chain reaction
RER	Ribonucleotide excision repair
RNA	Ribonucleic acid
RNAP II	RNA polymerase II
ROS	Reactive oxygen species
RPKM	Reads per kilobase million
RT-qPCR	Reverse transcription real-time polymerase chain reaction
S.E.M.	Standard error of the mean
SAM	S-adenosyl methionine
SIL	Stable isotopic label
siRNA	Small interfering RNA
SMRT	Single-molecule real-time sequencing
SMUG1	Single-strand selective monofunctional uracil DNA glycosylase
snAP-seq	Single-nucleotide AP site sequencing
SSB	Single-strand break
ssDNA	Single-stranded DNA
ssODN	Single-stranded oligodeoxynucleotide
SSR	Strand-switch region
T	Thymine
T. brucei	*Trypanosoma brucei*
TCEP	Tris(2-carboxyethyl)phosphine
TDG	Thymine DNA glycosylase
TET	Ten-eleven translocase
Tg	Thymine glycol
U	Uracil
UGI	Uracil glycosylase inhibitor
UNG	Uracil DNA glycosylase
UTR	Untranslated region
UV	Ultraviolet
VEGF	Vascular endothelial growth factor

Table of contents

Chapter 1

Introduction

1.1 Deoxyribose nucleic acid

The genetic information of living organisms is stored in the biopolymer deoxyribose nucleic acid (DNA). Passed on through generations, the order in which DNA bases are arranged within a genome dictates the development, function and reproduction of an organism. In 1869, Johannes Friedrich Miescher isolated a substance from the nuclei of leucocytes[1]. Consisting of carbon, hydrogen, nitrogen, oxygen and phosphorus, this substance appeared to be distinct from proteins as it was found to be resistant to protease digestion. This was the first known crude isolation of DNA and Miescher termed this new class of compound 'nuclein'.

Although the significance of nuclein to heritable traits was not yet clear, Ludwig Albrecht Kossel made key observations on its chemical composition and properties. Kossel isolated the four DNA nucleobases, adenine (A), guanine (G), thymine (T) and cytosine (C) as well as the RNA nucleobase uracil (U)[2] (**Figure 1.1**). Phoebus Levene determined that in addition to the nucleobases, nuclein also consisted of deoxyribose sugar units, as well as phosphate groups[3]. A nucleoside is formed when a nucleobase is linked to the deoxyribose sugar, whilst the further addition of a phosphate at the 5'-OH of deoxyribose generates a nucleotide. Many structures of nuclein were proposed, including a tetranucleotide structure by Levene in which G, C, A and T nucleotides were joined together in discrete tetramer units[4]. This structure assumed that the four nucleotides occurred in equal ratios and was later disproven.

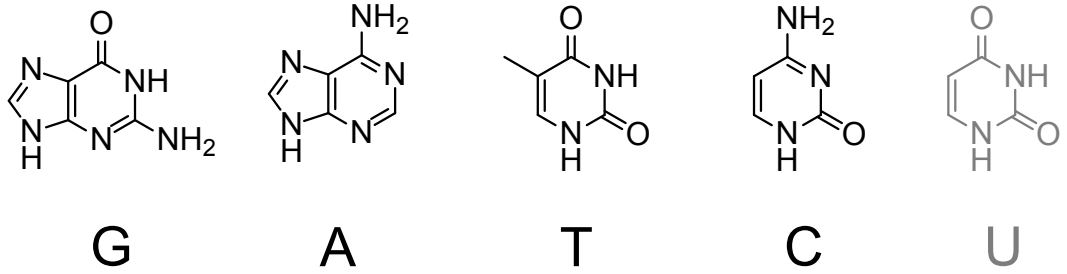


Figure 1.1: Structures of the four nucleobases found in DNA, guanine (G), adenine (A), thymine (T) and cytosine (C), and uracil (U) found in RNA.

It was not until 1944 that Oswald Avery, Collin MacLeod and Maclyn McCarty showed that it was DNA, and not proteins as previously believed, which carried the hereditary genetic information within cells[5]. They verified that the 'transforming principle' first reported by Frederick Griffith that allowed nonvirulent pneumococcus to become virulent again was DNA. This transforming principle was found to precipitate in ethanol, and was resistant to protease, lipase and RNase digestion, but not DNase digestion. Due to its high molecular weight and reactivity with enzymes, the transforming principle was concluded to be DNA. Inspired by Avery, MacLeod and McCarty's work, Erwin Chargaff also made key observations on the properties of DNA which led to his proposal that DNA obeyed two key rules. Firstly, that the amount of adenine in DNA samples was equivalent to that of thymine, and the amount of guanine to that of cytosine, and secondly Chargaff noted that the relative ratios of adenine and thymine, to guanine and cytosine, varied between species[6].

1.1.1 The double helix

By the mid 20th century, the composition of DNA and many of its properties had already been documented[1,6]. Further critical experiments carried out by Rosalind Franklin, Raymond Gosling and Maurice Wilkins led to James Watson and Francis Crick's proposal of the structure of DNA[7]. X-ray crystallography data collected by Gosling and Franklin demonstrated that DNA adopts a double-stranded helical structure[8,9] (**Figure 1.2a**). Together with the body of work already reported on the composition and ratios of nucleosides, Watson and Crick ultimately concluded that DNA exists as two strands joined together by hydrogen bonds, where a strict pattern was in place between opposing base pairs. Adenine was base-paired with thymine forming two inter-strand hydrogen bonds, whilst three hydrogen bonds were formed between guanine and cytosine (**Figure 1.2b**).

The two strands of DNA in a duplex are antiparallel with one running in a 5'- to 3'- orientation and the other 3'- to 5'-. The base-pairs between the strands are stacked, forming π-π and van der Waals interactions between adjacent pairs[10,11]. The resultant duplex is twisted into a right-handed helix with 10.4 base-pairs in every complete turn[12] (**Figure 1.2c**). In this structure, the highly charged phosphate backbone is positioned on the outer surface facing the aqueous phase, whilst the nucleobases are shielded within the centre. The exposed regions of the nucleosides between the phosphate backbones form grooves, which due to the helical structure are separated into a wider major groove, and a smaller minor groove. The grooves enable proteins to access and recognise part of the DNA strands without

the need to fully unwind the duplex. This B form of DNA is accepted as the dominant form of DNA in solution and in a cellular context. However, alternate structures including a more compact A form and a left-handed Z form have been observed under specific conditions[13,14], whilst secondary structures can also exist[15,16].

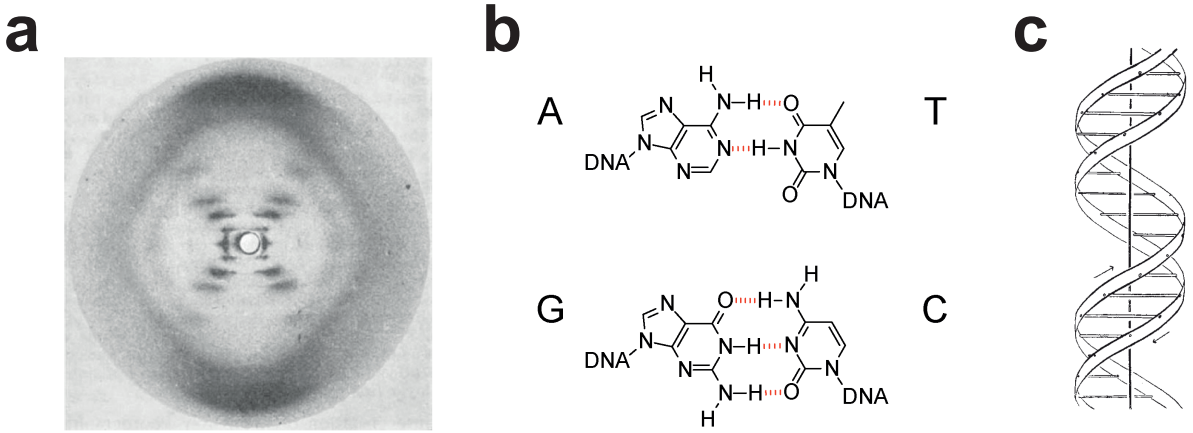


Figure 1.2: Double helix structure of DNA. **a**) 'Photograph 51', showing the diffraction pattern of sodium deoxyribose nucleate from calf thymus by x-ray crystallography. **b**) Hydrogen bonding pattern between Watson-Crick base-pairs. **c**) Diagram of helical structure proposed by Watson and Crick. Figures have been adapted from Franklin *et al*.[8] and Watson *et al*.[7].

1.1.2 Cellular packaging of DNA

Each human nucleus contains around 2 m of DNA, if stretched out end to end. In order to fit the entire genome within a cell nucleus with a width of a few microns, DNA must be highly compacted. Within nuclei, double-stranded DNA (dsDNA) is first wrapped around individual histone protein octamers forming a nucleosome core particle (NPC). DNA within an NPC consists of around 147 base pairs[17], and individual NPCs are joined by histone-free linker DNA consisting of around 80 base pairs. Nucleosomes are further condensed into chromatin fibres, which are then finally condensed into chromosomes (**Figure 1.3**).

Histone proteins are rich in amino acids such as lysine and arginine, which at physiological pH are positively charged. These residues can form favourable ionic interactions with the negatively charged phosphate backbone and overcome the charge repulsion of phosphate groups within DNA. NPCs that are loosely arranged, resembling a 'beads on a string' arrangement form euchromatin, or open chromatin, which is easily accessible for transcription factors and other cellular machinery. In contrast, regions

containing highly condensed nucleosomes form heterochromatin, or closed-chromatin; these are generally associated with repressed genes and centromeres[18]. The arrangement and accessibility of chromatin is dynamic within cells and can be controlled by a variety of factors including some histone modifications that are discussed below (section 1.2.3).

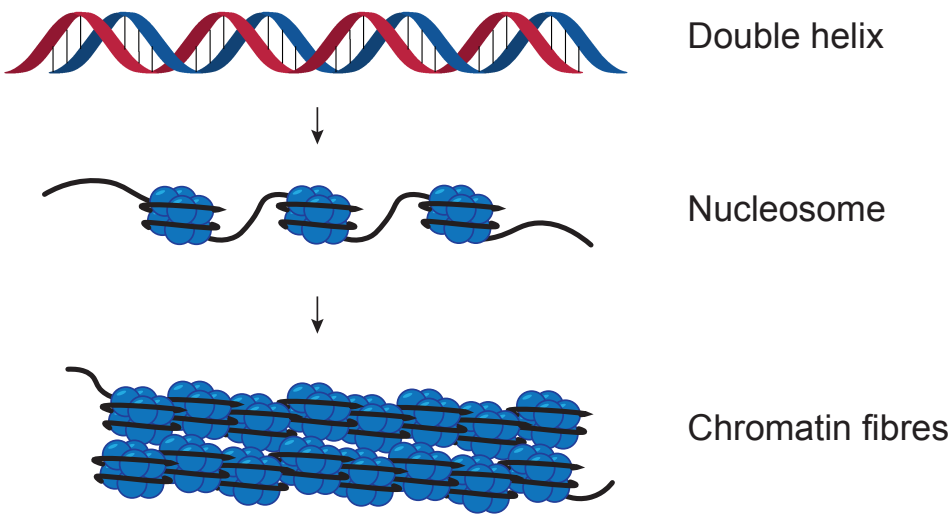


Figure 1.3: Organisation of DNA in the nucleus. The double helix is wrapped around histone octamers to form nucleosomes in a 'beads on a string' model, before condensing further into chromatin fibres that make up the chromosomes. Thick black lines represent double-stranded DNA, blue spheres represent histone units.

1.2 DNA modifications

Within eukaryotic cells, DNA is closely associated with histone octamers and condensed into chromatin. As such, two classes of modifications that can affect DNA function, structure or accessibility exist; those on the nucleic acid itself, and those on the histone proteins. Many well-studied modifications that naturally occur in DNA are located on the nucleobase, thus expanding the DNA alphabet beyond the four canonical nucleobases, whilst modifications on the backbone are also known[19]. Within histones, covalent modifications to amino acids can be introduced by post-translational modification. Like many cellular proteins, modified amino acids include lysine, serine, threonine and arginine[20]. Lysine, and to a lesser extent arginine, modifications in particular have been studied extensively due to their role in gene regulation[21].

Epigenetics can be defined as the heritable changes in gene function that are not due to changes in the primary DNA sequence[22]. Therefore, it is only those modifications to either DNA itself or histone proteins that have an effect on gene function or expression that can be considered epigenetic; lesions or intermediates that have no direct effect do not fall under this definition.

1.2.1 Cytosine modifications

Often considered the 'fifth' base in DNA, 5-methylcytosine (5-mC) was first detected as a nucleobase in the DNA of Tubercle bacillus[23] and was later confirmed to also exist in the genomes of other species including mammals[24,25]. Much lower in abundance than the four canonical bases, global levels of 5-mC are generally around a few percent of cytosine in vertebrates. For example, levels in mouse embryonic stem cells (mESCs) have been measured at 4-5% of C[26,27]. 5-mC is the most abundant base modification in many vertebrates and is a key epigenetic marker that can control the expression of genes.

In mammalian genomes, 5-mC occurs largely in the context of CpG dinucleotides, locations where cytosine is immediately followed by a guanine, affecting up to 90% of these sites[28]. Methylation of cytosine has been associated with the silencing and inactivation of transposable elements of the genome[29,30] and X-chromosome inactivation[31]. CpG islands, regions in which CpG dinucleotides are densely clustered, occur within over 70% of gene promoters[32]. Despite their high CpG content, CpG islands are generally depleted of 5-mC and are hypomethylated. Methylation of promoters containing CpG islands leads to a downregulation of transcription, thus providing a way of silencing genes[33,34]. This type of epigenetic regulation involving DNA methylation is particularly important during embryonic development and differentiation, and alterations in methylation patterns can lead to developmental defects[35]. Aberrant methylation is also observed during cancer, where many studies report a global hypomethylation status where methylation is depleted[36,37]. However, a recent study in which 5-mC was analysed accurately using a more sophisticated approach suggested an increase of 5-mC in glioblastoma tissue[38], whilst localised hypermethylation at specific tumour suppressor genes has also been associated with human tumorigenesis[39].

5-mC is installed into the genome by methylation of cytosine, catalysed by the DNA methyltransferase family of enzymes (DNMTs). Mechanistically, DNMT enzymes attack the C6 position of cytosine *via* a nucleophilic cysteine residue, breaking the aromaticity of the

pyrimidine ring. A methyl group is donated from S-adenosyl methionine (SAM) to the C5 position, after which cysteine is removed by elimination to restore aromaticity (**Figure 1.4**). Within the DNMT family, DNMT1 preferentially methylates cytosines at hemi-methylated CpG sites[40,41], resulting in a symmetrically methylated DNA duplex. DNMT1 is therefore largely considered a maintenance methyltransferase, whilst DNMT3a and DNMT3b can methylate DNA *de novo* and do not require hemi-methylated DNA substrates[35].

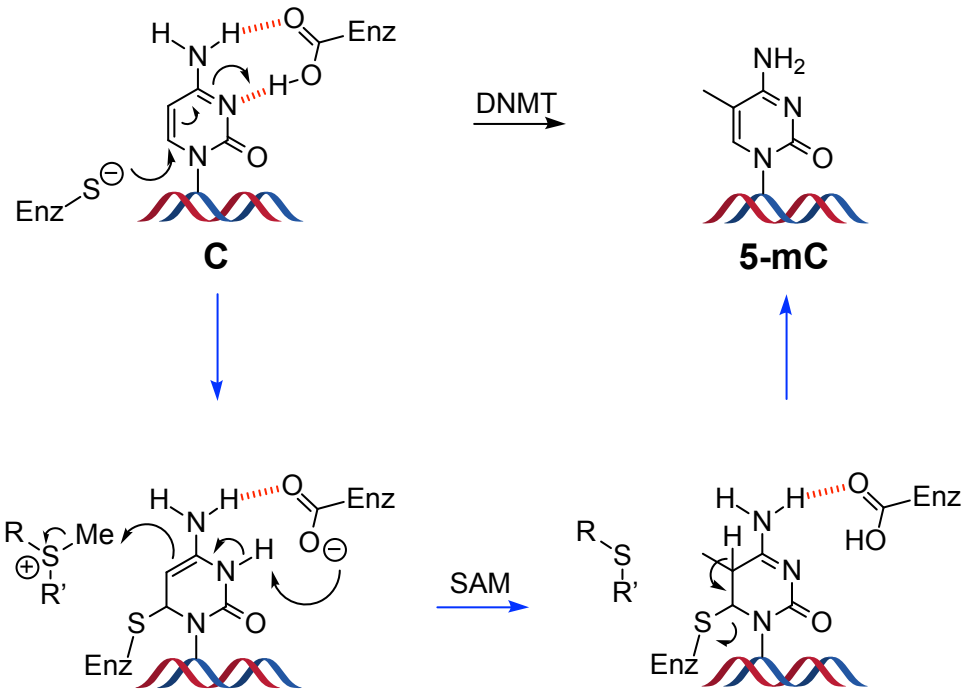


Figure 1.4: Mechanism of cytosine methylation by DNMT enzymes.

While passive dilution of 5-mC during DNA replication without enzymatic restoration of methylation can result in DNA demethylation[42], a number of enzymatic demethylation pathways have also been proposed. Some experimental evidence has supported an active demethylation pathway, dependent on the ten eleven translocase (TET) and thymine DNA glycosylase (TDG) enzymes (**Figure 1.5a**). In this proposed mechanism, 5-mC is oxidised stepwise to 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC) and 5-carboxycytosine (5-caC) by the TET enzymes. TDG recognises and excises both 5-fC and 5-caC, generating an abasic site which is further processed by the base excision repair (BER) pathway to restore cytosine back into the genome. An overall demethylation of 5-mC is thus observed. First recognised as a glycosylase that removes thymine and uracil when mismatched with guanine, TDG has since been shown in *in vitro* studies to also excise 5-fC

and 5-caC[43,44]. *In vivo* evidence in support of this mechanism is provided by observations that the combined deletion of TET1/2/3 leads to a loss of 5-hmC and impairs differentiation of mESCs[45], whilst depletion of TDG results in the accumulation of 5-fC and 5-caC[44]. Furthermore, reconstitution of TET, TDG, along with key proteins in the BER pathway *in vitro* has been shown to result in the demethylation of DNA[46]. Together, these studies provide some support for this mechanism. However, in a study quantifying the global levels of abasic sites in mESCs, no accumulation of abasic sites was seen in correlation with the removal of 5-fC or 5-caC[47]. This may indicate that the abasic site intermediate is short lived and difficult to detect at steady-state, or that demethylation does not significantly occur *via* this route. The true significance of this proposed pathway in a cellular context therefore remains unclear.

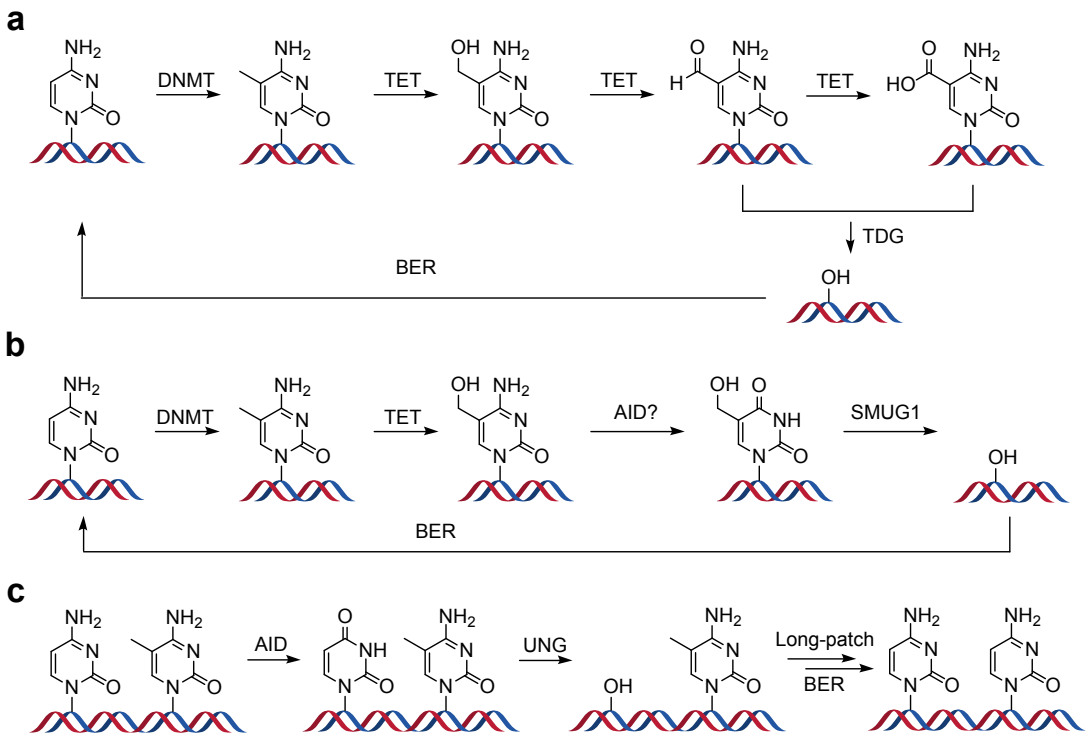


Figure 1.5: Proposed mechanisms of 5-methylcytosine demethylation. **a**) Active demethylation, where successive oxidation of 5-mC by TET enzymes generates 5-hmC, 5-fC and 5-caC. The latter two are substrates for TDG, which initiates DNA repair by the BER pathway to restore cytosine. **b**) The oxidation of 5-mC to 5-hmC, followed by deamination forms 5-hmU, a substrate for SMUG1. An abasic site is generated, followed by BER to restore cytosine. **c**) Processive demethylation, where deamination of an unmodified cytosine a few base pairs 5'- to 5-mC, followed by UNG excision generates an abasic site. When repaired by long-patch BER, both the uracil and 5-mC sites are restored with cytosine.

Alternative mechanisms of demethylation have also been proposed, although these are less well explored in comparison to active demethylation. It has been suggested that 5-hmC, generated by TET-mediated oxidation of 5-mC, is a possible substrate for deaminase enzymes such as activation induced cytosine deaminase (AID). The deamination product, 5-hydroxymethyluracil (5-hmU) is excised by single-stranded monofunctional uracil-DNA glycosylase 1 (SMUG1)[48,49] to generate an abasic site, which is then repaired by BER to restore cytosine[50] (**Figure 1.5b**). In support of this hypothesis, it was found using isotope labelling of 5-mC that upon knockdown of TDG in mESCs, 7% of 5-hmU was derived from methylated cytosine instead of thymine. However, 5-mC-derived 5-hmU could not be detected in the presence of TDG, suggesting that this pathway may not be significant in wild-type cells.

A further alternative pathway, processive demethylation, is also dependent on deamination[51,52]. Instead of enzymatic activity directly on the 5-mC site itself, AID is proposed to deaminate an unmethylated cytosine a few base pairs 5'- to a 5-mC site. The resultant uracil is recognised and excised by uracil DNA glycosylase (UNG)[53] to generate an abasic site. When repaired by long-patch BER (section 1.2.4), up to 13 nucleotides downstream of the deamination event may also be replaced. Therefore, if the deamination event and 5-mC are in close proximity, this mechanism is able to lead to overall demethylation (**Figure 1.5c**). *In vitro* experiments in which xenopus egg extracts were incubated with AID showed elevated levels of 5-mC demethylation, with this effect minimised in the presence of an UNG inhibitor (UGI)[54]. This pathway has been proposed to be particularly important during the wave of demethylation that occurs in the paternal genome during embryonic development[51], which was supported by immunostaining studies showing demethylation defects in the absence of AID and UNG enzymes in zygotes. However, as for the other proposed demethylation pathways, the extent to which each mechanism contributes towards demethylation *in vivo* remains unclear. Further investigation, as well as the development of accurate methodology to study the proposed intermediates within these pathways, is still required in order to delineate the exact mechanism of DNA demethylation in specific biological contexts.

Cellular levels of 5-hmC are generally lower than those of 5-mC, measured at around 0.07% of deoxynucleotides (dN) in mESCs. Levels of 5-fC and 5-caC are lower still, at around 10 and 0.6 per million dN, respectively[50]. In addition to the roles of 5-mC oxidation products as possible intermediates during DNA demethylation, there is emerging evidence that these modifications have biological significance in their own right. Global measurements in mESCs have revealed that both 5-hmC and 5-fC can be stable modifications that are maintained,

rather than transient intermediates[55,56], whilst depletion of 5-hmC is a hallmark of cancers such as melanoma[38,57]. Evidence that may suggest a function of 5-fC independent of demethylation includes findings that the aldehyde group in 5-fC can react with lysine tails in histone proteins, forming a Schiff base. This chemical species has been identified by trapping with sodium cyanoborohydride reduction[58,59]. 5-fC has also been suggested to promote nucleosome formation and influence nucleosome organisation, based on further observations that nucleosomes associated with 5-fC are elevated in gene expression[59].

1.2.2 Thymine modifications

Beyond cytosine modifications, many modifications of thymine have also been detected in a range of organisms. Early observations were in bacteriophage genomes, where a large proportion of all thymines in a genome may be replaced by a modified thymine base[60,61]. Often, these modifications consist of large or bulky groups attached to the C5 position of T, for example α-putrescinylthymine (putT)[62], 5-dihydroxypentyluracil (dhpU)[63] and α-glutamylthymine (gluT) (**Figure 1.6**). These modifications may themselves be further functionalised. For example, 62% of T in bacteriophage SP-15 is replaced by modified dhpU residues, in which one hydroxyl group of the dihydropentyl moiety is glucosylated and the other is attached to a phosphoglucuronate group[64]. Many bacteriophage hypermodifications are believed to provide protection from restriction-endonucleases released by the host upon infection of bacteria whilst putT is also believed to facilitate the packaging of phage DNA[65].

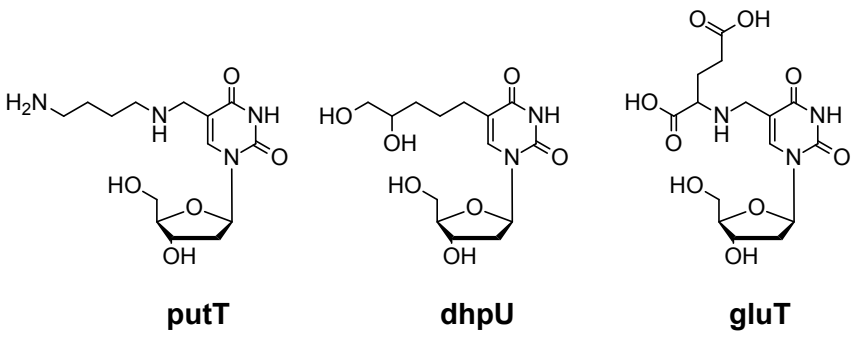


Figure 1.6: Structures of some thymine modifications found in bacteriophage genomes.

Thymine modifications are also important in the genomes of a number of eukaryotes. Organisms within the Trypanosomatidae family contain two thymine modifications in their

DNA; 5-hmU and 5-(β-glucopyranosyl)hydroxymethyluracil (base J). These bases are involved in transcriptional control in trypanosomatid genomes, where in contrast to mammalian systems, C modifications are not present at detectable levels.

Species of trypanosomatids where T modifications have been studied in detail include leishmania, such as *Leishmania major* and *Leishmania donovani*, and trypanosomes such as *Trypanosoma brucei* and *Trypanosoma cruzi*. Parasitic species of leishmanias cause cutaneous leishmaniasis in mammals, presenting as skin lesions on infected individuals[66] and species of trypanosomes cause trypanosomiasis, or African sleeping sickness[67]. Both parasites are transmissible between insects and mammals through insect bites and are classified by the WHO as neglected tropical diseases. Parasitic trypanosomatids have two stages in their life cycles, which in *Leishmania major* take the form of procylic promastigotes in their insect host, female sandflies, as well as amastigotes in the human bloodstream[68] (**Figure 1.7a**).

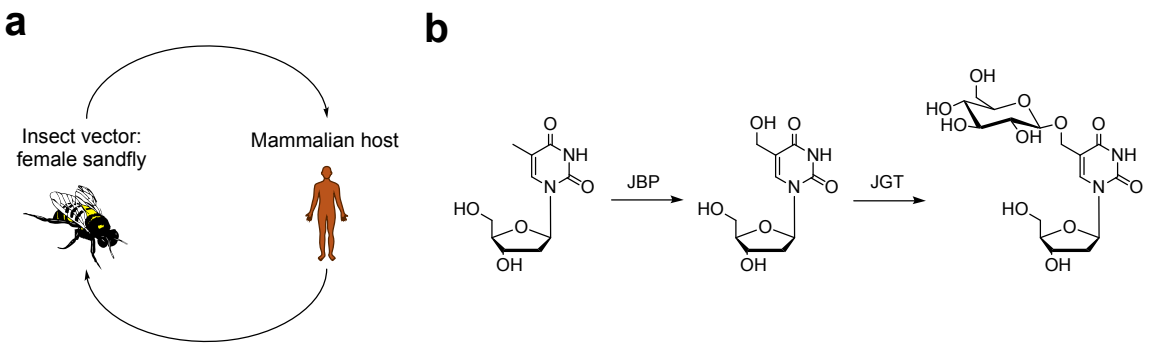


Figure 1.7: Thymine modifications in the *Leishmania major* genome. **a**) Life cycle of *L. major*. **b**) Biosynthetic pathway of 5-hmU and base J in trypanosomatids.

In trypanosomatids, 5-hmU can be generated enzymatically through oxidation of thymine by the J binding protein family of enzymes, JBP1 and JBP2, which are homologues of the TET enzymes[69]. A portion of these sites are glucosylated by J-glucosyl transferase (JGT), forming base J (**Figure 1.7b**). Global quantification of these modifications by mass spectrometry has revealed that 5-hmU and base J replace 0.01% and 0.08% of thymine respectively in *Leishmania major*, and 0.02% and 0.5% in *Trypanosoma brucei* (bloodstream form)[70,71]. Depletion of JBP proteins is found to reduce levels of 5-hmU by three-fold and render base J no longer detectable[72].

Members of the Trypanosomatidae family have polycistronic genomes, in which groups of genes are clustered together[73]. Rather than transcriptional initiation as in many mammalian systems, it is the accurate termination of transcription at gene clusters that heavily influences gene expression. Antibody-based detection and mapping of base J has shown that this glucosylated modification is strongly enriched within telomeric regions[74], with around 50% of base J in *T. brucei*, and up to 99% in *L. major* found to be localised at telomeres. The remaining non-telomeric base J sites in *L. major* have been found to occur at RNA polymerase II (RNAP II) termination sites, suggesting that this bulky modification is important in mediating the correct transcriptional termination of polycistronic genes[75]. RNAP II transcription is initiated and terminated largely at strand-switch regions (SSRs) in *Leishmania major*, as well as head-tail sites that are located between adjacent gene clusters on the same DNA strand. Inhibition of the Fe^{2+} and 2-oxoglutarate dependent JBP enzymes is possible by treatment of cultures with dimethyloxalylglycine (DMOG), an analogue of 2-oxoglutarate. DMOG treatment of *L. major* reduces the levels of base J, in both telomeric regions and within chromosomes[76]. Defects in transcriptional termination were observed upon DMOG-induced loss of base J at both SSRs and head-tail sites, where the degree of readthrough past transcriptional termination sites correlated with the degree of base J loss. A similar effect was observed upon JBP2 knockout in *Leishmania tarentolae*, which became hypersensitive towards 5-bromouracil treatment[75]. In contrast to leishmania, the effect of DMOG suppression of base J in *T. brucei* did not affect transcriptional termination in the majority of convergent SSRs analysed. Instead, base J was found to affect the termination of RNAP II at a small number of genomic loci within polycistronic gene clusters, leading to the proposal that this hypermodification regulates the expression of specific genes in a more specialised manner in *T. brucei* than in *L. major*.

Genome-wide mapping of both 5-hmU and base J in *L. major* has further confirmed that both 5-hmU and base J are enriched in SSRs[70]. In the same study, a small number of 5-hmU loci were detected in which base J was not detectable, suggesting the existence of base J-independent 5-hmU. These regions were found to be largely depleted within genomic features and occurred mainly in intergenic regions, whilst sites that were unique to base J that were depleted in 5-hmU were enriched within tRNA genes. Together, the body of work on thymine modifications in trypanosomatids reveals potentially important roles of these bases in parasite genomes and further investigation is required to elucidate the function of these modifications.

Thymine modifications 5-hmU and 5-fU have also been detected at low levels in higher eukaryotes including mammalian genomes. Levels of 5-hmU are around 0.5 per million dN, and levels of 5-fU are approximately 2.5 per million dN in mESCs[50]. Specific reader proteins have been identified that interact with the 5-hmU:A base pair including some transcription factors. This raises the possibility that this modification can have biological significance beyond trypanasomatids; however, the evidence of this is currently very limited.

Although commonly a deamination product of cytosine, uracil can be considered a thymine modification due to the similarities in Watson-Crick base-pairing. Uracil derived from cytosine is therefore mutagenic by subsequent mismatch with adenine. In addition to spontaneous deamination, enzymatic deamination of cytosine can also be mediated by AID and apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like (APOBEC) enzymes. This process is essential for antibody diversification, during somatic hypermutation and class-switch recombination of DNA in B cells[77,78], as well as the proposed processive pathway of 5-mC demethylation discussed above[52] (section 1.2.1).

The aberrant expression of AID has been associated with cancer. Elevated levels of AID can lead to cytosine deamination at genomic loci independent of antibody generation, as well as in cell types in which this mechanism is not normally active, leading to development of lymphomas[79]. Furthermore, deficiencies in UNG may also lead to carcinogenesis as suggested by the 22-fold increased incidence of lymphoma in mice lacking UNG compared to wild-type mice[80]. Interestingly, the UNG gene is absent in *Drosophila melanogaster* where uracil is tolerated in genomic DNA, and global levels are highest during the larval stages of development[81]. Uracil has thus been suggested to play a role during the development of this species. The quantification of global levels of uracil is technically challenging. Due to the high rate of cytosine deamination particularly in single-stranded DNA, and the large quantities of cytosine in comparison to basal uracil levels, artificial elevation of measurements through additional deamination is possible. Measurements vary significantly between reports and can differ by over three orders of magnitude. Towards the lower end of this range, uracil has been measured at less than 0.2 per million dN in human and murine genomes by LC-MS/MS[82].

Genome-wide mapping of uracil in human cells has revealed a non-random distribution of this modification across multiple human cell lines, where an accumulation was observed in centromeres, intergenic regions and satellite repeats[83]. The tolerance of uracil at specific genomic loci may indicate a biological role of this modification beyond being a product of DNA damage.

1.2.3 Histone modifications

Like many proteins, amino acids within histone units can be heavily modified. In particular, the methylation and acetylation of lysine residues have been shown to influence gene regulation. Nucleosomes consist of a central histone fold domain, along with two histone tails that are rich in lysine[84]. The acetylation of lysine is carried out by histone acetyltransferases and removed by histone deacetylases[85]. The methylation of lysine is stepwise, where mono-, di- and tri-methylated species of lysine can all exist on histone proteins[86]. This methylation is mediated by histone lysine methyltransferase enzymes and removed by demethylases. The methylation of arginine residues in histones can also be considered epigenetic, and similarly is installed and removed by specific enzymes[87]. In general, the acetylation of histone lysines is associated with an open-chromatin state that is accessible to transcriptional machinery, and the acetylated histone sites such as H3K9ac and H3K27ac are considered to be activating histone markers[88]. The methylation of lysine in histones can be associated with both the activation and repression of transcription. For example, H3K4me3 is associated with active transcription, whilst H3K27me3 is a marker for repressed transcription[89].

1.2.4 DNA damage

The faithful replication and transcription of DNA is essential in order to maintain genomic stability. Cellular DNA is constantly subject to a variety of sources of damage, both endogenous and exogenous to cells and therefore the efficient repair of DNA damage is of great biological importance. Damage to DNA can occur on the nucleobase or the backbone; it also manifests as cleavage of DNA to form either single-strand breaks (SSB) or double-strand breaks (DSB). Many of these products of DNA damage have the potential for severe biological consequences, such as polymerase stalling and mutagenesis.

The repair of a wide range of DNA base damage is through the base excision repair (BER) pathway[90]. This pathway is initiated by glycosylase enzymes, which recognize and excise base lesions to generate an abasic site. At least 11 distinct glycosylases have been identified to date in mammals[91], and more are known that are unique to other organisms including bacteria and plants. The substrates of a glycosylase commonly comprise a single nucleobase, or a small number of nucleobases that share some structural similarity[91–93]

(**Table 1.1**). Some glycosylases may also have overlapping substrates. After removal of the damaged DNA base by hydrolysis, the first common intermediate in the BER pathway is an abasic (apurinic, AP) site. DNA glycosylases can be divided into two main classes; monofunctional and bifunctional. Monofunctional glycosylases are responsible only for the hydrolysis of a nucleobase substrate, leaving behind an intact abasic site as the product. In contrast, bifunctional glycosylases also have an associated AP lyase functionality, and cleave the abasic site after formation at the 3'- end, by forming a Schiff-base followed by β-elimination of the sugar backbone (**Figure 1.8a**).

Glycosylase	*Substrates*	*Mono-/Bi-functional*
UNG1	U, 5-FU (mitochondrial)	Mono
UNG2	U, 5-FU (nuclear)	Mono
SMUG1	U, 5-hmU, 5-fU, 5-hoU	Mono
TDG	U:G, T:G, 5-fC, 5-caC	Mono
MBD4	U:G, T:G, 5-hmU	Mono
MPG	3-mA, 7-mG, 3-mG, hypoxanthine	Mono
OGG1	8-oxoG, FapyG	Bi
MUTYH	A:8oxoG, A:G, A:C	Mono
NTH1	Tg, FapyG, 5-hoC, 5-hoU	Bi
NIEL1	Tg, FapyG, FapyA, 8-oxoG, 5-hoU, dhU, Sp, Gh	Bi
NIEL2	Tg, FapyG, FapyA, 8-oxoG, 5-hoU, dhU, Sp, Gh	Bi
NIEL3	FapyG, FapyA, Sp, Gh	Bi

Table 1.1: List of mammalian glycosylases and substrates. Abbreviations: 5-fluorouracil (5-FU), 5-hydroxyuracil (5-hoU), 5-hydroxycytosine (5-hoC), 2,6-diamino-4-hydroxy-5-formamidopyrimidine (FapyG), 4,6-diamino-5-formamidopyrimidine (FapyA), thymine glycol (Tg), dihydrouridine (dhU), spiroiminodihydantoin (Sp) and guanidinohydantoin (Gh). Mismatches are shown with the substrate base first. Table adapted from Krokan *et al*.[92].

For both an abasic site and its β-elimination product, AP endonuclease 1 (APE1) continues BER by creating a nick in the DNA backbone 5'- to the abasic site to generate a SSB. One strand contains the abasic site on the DNA end as a 5'-phosphorylated sugar and the other strand released contains a 3'-OH group (**Figure 1.8a**). The BER pathway then divides into two sub-pathways; short-patch repair and long-patch repair[90,94]. In short patch repair, DNA Pol β fills in a single-nucleotide to base-pair with the nucleobase opposite the

abasic site on the complementary DNA strand (**Figure 1.8b**). In the event of a monofunctional glycosylase, DNA Pol β also has inherent AP lyase activity that removes the 5'-phosphorylated deoxyribose unit whilst in the event of bifunctional glycosylase activity this step is already completed. The SSB, with the abasic site now replaced is then sealed by a DNA ligase. For long-patch repair, DNA Pol β continues DNA synthesis for 2-13 nucleotides beyond the original abasic site, leaving an overhang with a 5'-phosphorylated deoxyribose end. This is removed by Flap endonuclease 1 (FEN1), and the SSB is again sealed by a DNA ligase. The identity of the DNA ligase involved is not fully understood, however both DNA LIG1 and LIG3 have been implicated in both short-patch and long-patch repair.

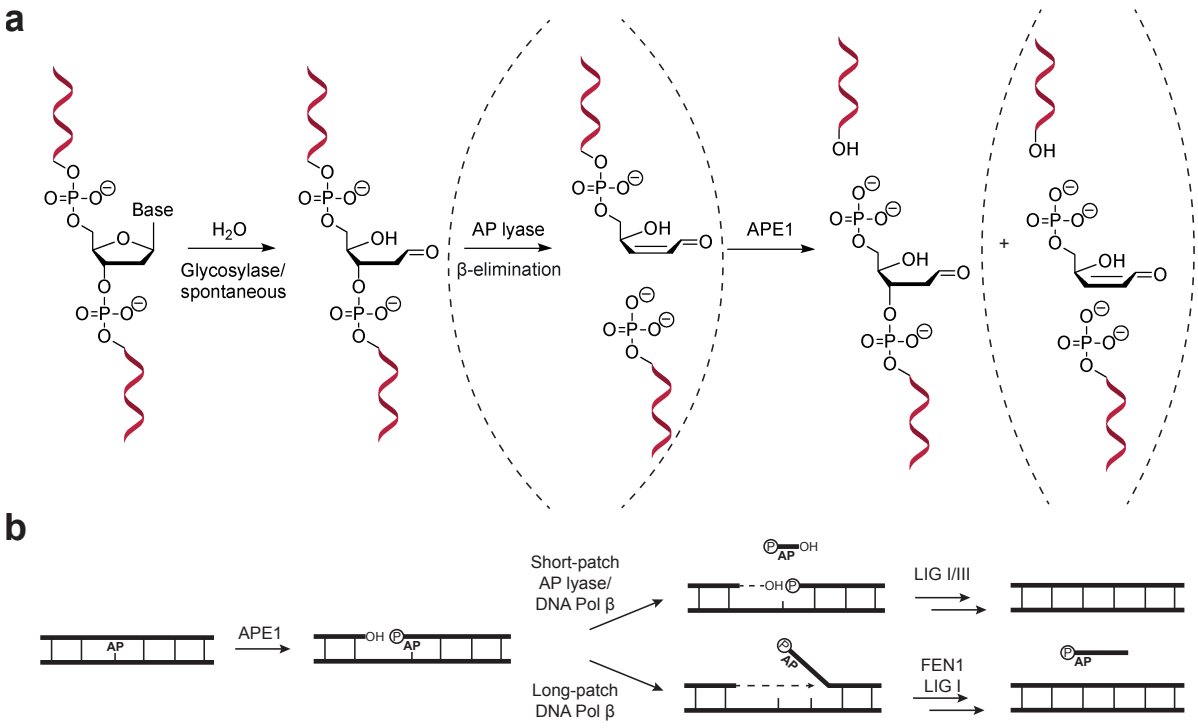


Figure 1.8: Abasic site formation and BER pathway. **a**) After abasic site formation, bifunctional glycosylases have an additional AP lyase step. Both the β-elimination product and intact abasic sites are recognised and cleaved by APE1 at the 5'-end. **b**) Processing of abasic sites by the short-patch and long-patch BER pathways.

A major source of DNA damage is from oxidative stress. This may be caused by reactive oxygen species (ROS) that can be endogenous to cells, with levels further elevated under exogenous oxidative stress. Amongst the four canonical bases, guanine has the lowest oxidation potential and its oxidation product, 8-oxoguanine (8-oxoG) is one of the most

abundant DNA damage products with levels measured at around 5 per million dN in mESCs[50]. Beyond its role as a product of DNA damage, there is some evidence that 8-oxoG may function as an epigenetic marker that can sense ROS under specific cellular conditions, capable of controlling gene expression. Under hypoxic and inflammatory conditions, the formation of 8-oxoG in promoter sequences has been associated with an upregulation of gene expression, possibly mediated by the 8-oxoG glycosylase OGG1[95,96]. Similarly, whilst the majority of 5-hmU and 5-fU may be attributed to oxidative damage of T in mammalian systems, there is emerging evidence that these modifications may also have biological function. Up to 80% of 5-hmU in mESC DNA has been found to be generated in a ROS-independent manner[50]. Together, these findings offer some support that bases traditionally associated solely with oxidative damage may be further harnessed in a cellular context to mediate biological processes.

Alternative pathways are in place to repair a range of DNA lesions alongside BER. Bulky adducts that arise in the event of UV damage such as cyclopyrimidine dimers and 6-4 photoproducts, as well as adducts from some chemotherapy drugs including cisplatin, are removed instead by the nucleotide excision repair (NER) pathway[97]. Here, a preincision complex consisting of a number of proteins that recognise and bind to damage substrates is recruited, followed by dual incision around the site by two endonucleases, ERCC1-XPF and XPG, to release a ssDNA fragment of roughly 25-30 nucleotides in length. DNA polymerases δ and ε and DNA ligase are then amongst the key proteins responsible for DNA synthesis and repair to complete NER. Some mismatched bases in DNA that do not fall within the substrates of BER can be repaired by mismatch repair (MMR)[98], whilst misincorporated ribonucleotides are removed by the ribonucleotide repair (RER) pathway[99]. As in NER, and in contrast to BER, these repair pathways depend on direct endonuclease activity after recognition of the damage site, and do not proceed *via* glycosylase activity or an abasic site intermediate.

Strand breaks in DNA can be cytotoxic and the repair of these types of damage is usually very efficient. The repair of single-strand breaks depends on the mechanism of SSB formation. SSBs are intermediates formed during BER and are efficiently processed after formation in this pathway by ligases. SSBs can also arise as result of direct damage to the DNA sugar backbone, or through enzymatic activity such as cleavage by DNA topoisomerase 1. These lesions are primarily detected by poly ADP-ribose polymerase-1 (PARP1), which then initiates the recruitment of protein complexes including DNA Pol β and DNA LIG3, in a pathway that closely resembles the latter stages of BER[100].

Double-strand breaks occur when both strands of DNA are broken around the same genomic location. They can be repaired in one of two processes; homologous recombination and non-homologous end joining (NHEJ). Homologous recombination requires a homologous donor, such as part of a sister chromatid to template DNA synthesis before the two DNA ends can be annealed together and joined by ligation[101]. In contrast, during NHEJ Ku protein recognises and binds the broken DNA ends. After activity from a series of signalling molecules and repair enzymes to blunt the DNA ends, these are directly ligated together without annealing[102]. DSBs in DNA are detrimental to cellular processes and are typically processed quickly. To prevent activation of DNA repair pathways, DSBs at the ends of telomeres are closely protected by shelterin[103]. Independent of telomeric ends, an estimated 50 DSBs are further generated per cell, per day[104]. Genome-wide mapping of DSBs has revealed a correlation with genomic regions of elevated transcription and nucleosome-depleted chromatin[105]. Despite the negative consequences of these lesions, a detectable accumulation occurs in genomic DNA, which appears to be tolerated by cells at steady-state.

1.2.5 DNA abasic sites

The loss of a nucleobase in DNA by hydrolysis leads to the formation of an abasic (apurinic, AP) site (**Figure 1.9a**). When left unrepaired, abasic sites can lead to strand breaks as well as mutations due to their non-coding nature[106–108]. A number of high-fidelity polymerases are not able to efficiency pass an AP site and lead instead to stalling during replication[106,107], whilst some polymerases are capable of translesional synthesis past AP sites. In *Escherichia coli* and some mammalian systems, an 'A-rule' is observed, where adenine is installed preferentially opposite the non-instructional AP site whilst cytosine can be preferred in a 'C-rule' in yeast[109]. The misincorporation of a nucleotide, or deletions that may occur at abasic sites lead to mutations and this type of DNA damage must be removed efficiently to maintain genomic stability.

The spontaneous hydrolysis of nucleobases alone generates an estimated 10,000 abasic sites per mammalian cell, per day[110]. Weaker than in RNA, the *N*-glycosidic bond linking the backbone to each nucleobase in DNA is susceptible to hydrolysis, particularly under acidic conditions[111]. *In vitro* studies have shown that for DNA, the rate of spontaneous depurination is much higher than depyrimidination under aqueous conditions[112]. In addition to spontaneous hydrolysis, AP levels can be further elevated in the presence of exogenous

damage. For example, ROS can lead directly to the formation of abasic sites[113]. Certain damaged DNA bases such as *N*7-methylguanine (7-mG) can also increase the rate of depurination compared to the canonical bases[114]. A number of DNA damaging agents associated with oxidative damage such as bleomycin and calicheamicin generate oxidised abasic sites, rather than true abasic sites, such as C4'-oxidised and 2-deoxyribonolactone abasic sites (**Figure 1.9b**)[115].

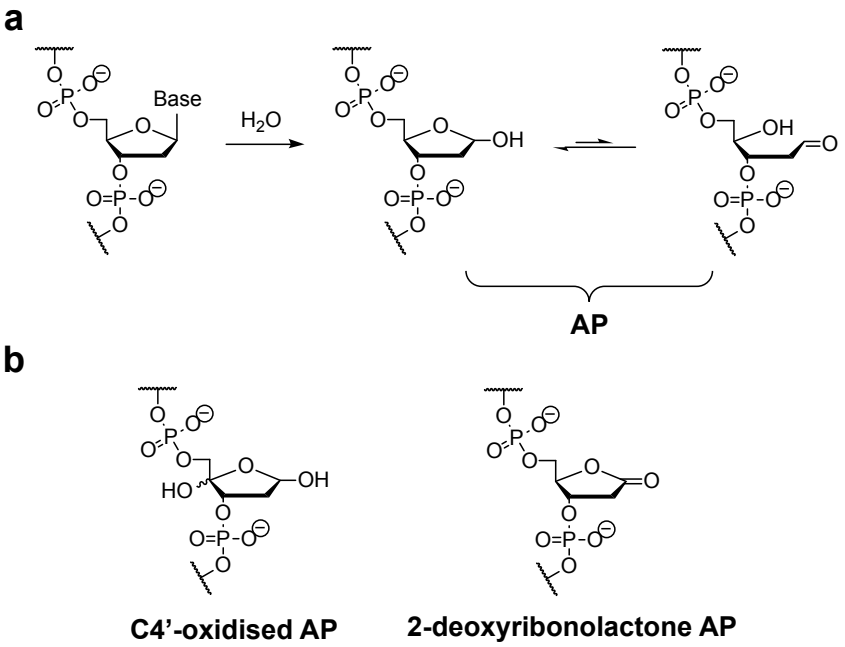


Figure 1.9: Structure of DNA abasic site. **a**) Formation of a DNA abasic (AP) site by hydrolysis of a nucleobase. **b**) Structures of oxidised derivates of an abasic site.

Hydrolysis of the *N*-glycosidic bond can also be catalysed enzymatically *in vivo*. As discussed above (section 1.2.4), a range of glycosylases have been identified that form abasic sites as a common intermediate (**Figure 1.10**). Bifunctional glycosylases are suggested to immediately process abasic sites upon generation, to form the β-elimination product. However, the dual functionalities of OGG1, glycosylase and AP lyase, have been successfully decoupled under specific conditions including those that contain a resemblance to the magnesium concentration *in vivo*[116]. This suggests that although AP lyase activity is possible for this enzyme, this step may not be carried out significantly *in vivo*.

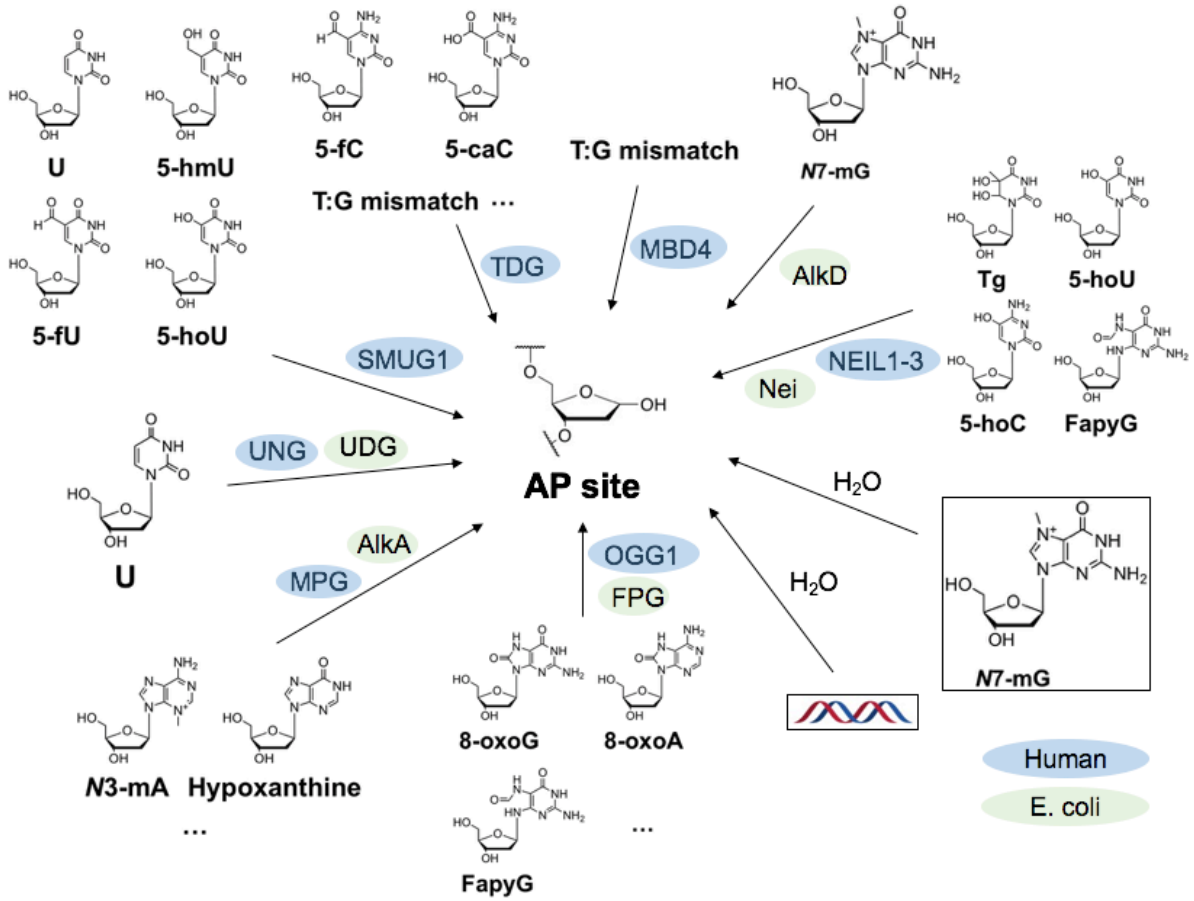


Figure 1.10: Example of routes to abasic site formation in DNA. Direct formation by spontaneous hydrolysis is outlined in black boxes, whilst all other routes shown are enzyme mediated. Examples of glycosylases are shown that are found in humans (blue) and *E. coli* (green). Substrate lists are not exhaustive.

DNA abasic sites are in equilibrium between the ring-closed and ring-open form (**Figure 1.9a**). Although the equilibrium is pushed towards the ring-closed form, the aldehyde exposed in the ring-open form is highly reactive. This aldehyde group leaves the sugar susceptible to β-elimination, possibly followed by further δ-elimination, leading to cleavage of the DNA backbone and the formation of strand breaks (**Figure 1.11**). The rate of this elimination is enhanced under alkaline conditions as well as elevated temperatures[117]. Under physiological conditions, the half-life of an abasic site towards elimination in a duplex context has been found to vary between 273 to 974 hours, depending on the base on the opposing strand[118]. Therefore, although chemically labile, abasic sites can be stable for weeks in the absence of DNA repair machinery.

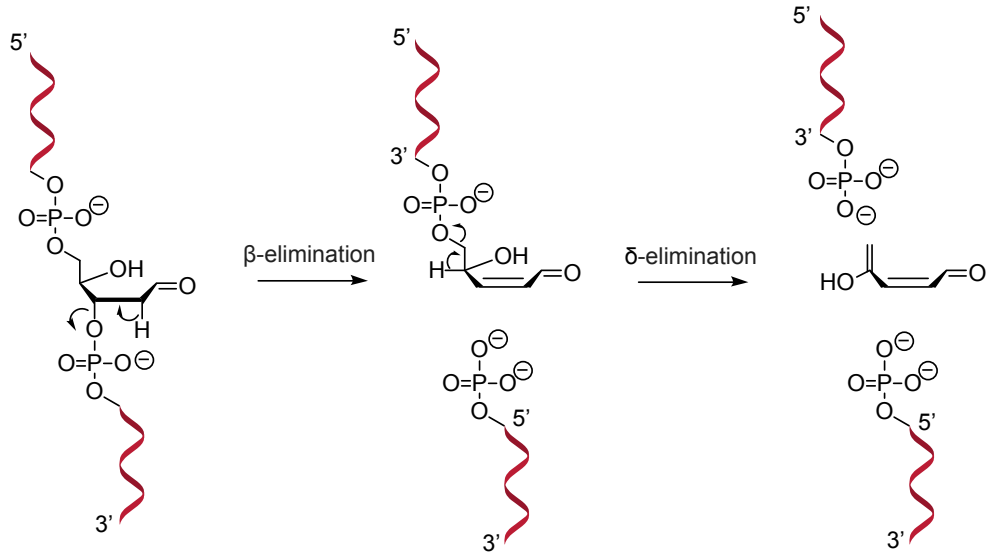


Figure 1.11: Elimination at DNA abasic sites. Both β- and δ-elimination are base-catalysed.

The aldehyde functionality of abasic sites has been suggested to facilitate the formation of crosslinks with DNA in an opposing strand, as well as histone proteins. Interstrand crosslinks between DNA abasic sites and the exocyclic amine of a guanine or adenine base in the opposing DNA strand have been detected upon incubation of DNA under physiologically relevant conditions *in vitro*; however, long reaction times (120 h) were required to observe detectable levels of crosslinking (15%)[119]. DNA-protein crosslinks at abasic sites have also been reported in nucleosome core particles[120], with 10% of abasic site-containing DNA found to be crosslinked with histones within 1 h incubation under physiological conditions[121]. The Schiff base formed between lysine residues and the abasic site was also suggested to accelerate DNA backbone elimination, resulting in a reduced half-life of abasic sites of around 24 h, compared to weeks for free abasic sites. In contrast, the formation of SSBs at abasic sites by APE1 enzyme activity is reduced in certain nucleosome contexts. The orientation of abasic sites within a nucleosome core particle influences the efficiency of cleavage by APE1, with those that have their phosphate backbone oriented towards the histone protein being cleaved less efficiently[122]. Together, these studies highlight the changes to AP reactivity and biology once nucleosomes and chromatin are taken into account.

Measurement of genomic levels of abasic sites by mass spectrometry detection after functionalisation with a hydroxylamine probe (**Figure 1.12a**) has shown that these lesions

occur at around 0.88 lesions per million dN in mESCs, whilst the β-elimination products of abasic sites are slightly more abundant, occurring at 1.7 per million dN[47]. The levels measured in other cell types, including neural stem cells and somatic HEK293T cells were comparable. An increase in AP levels was found when APE1 activity was inhibited, by either siRNA knockdown, or treatment with the small-molecule inhibitor CRT0044867.

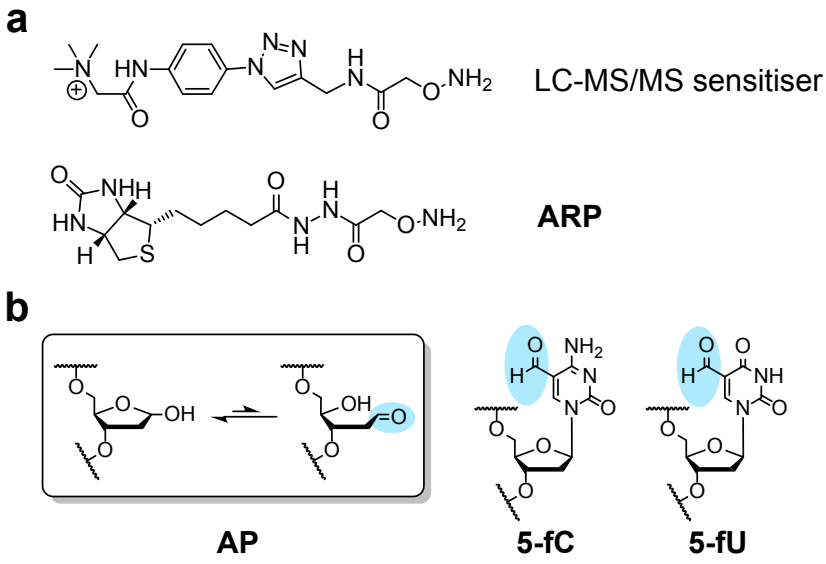


Figure 1.12: Aldehyde reactivity in DNA. **a**) Structure of abasic site sensitiser probe used for LC-MS/MS quantification[47] and ARP. **b**) Structures of reactive aldehydes found in DNA.

Analysis of DNA fibres using an aldehyde reactive probe (ARP, *O*-biotinylcarbazoylmethyl hydroxylamine)[123–125] (**Figure 1.12a**) has revealed that aldehyde-containing damage within DNA is non-random[126]. 68% of detected sites were spaced less than 500 base-pairs (bp) apart from another site, suggesting that they preferentially cluster. In a further study using the same detection method, ARP-labelled lesions were found to preferentially occur at regions of DNA undergoing replication[127]. Whilst these ARP-labelled sites were interpreted as abasic sites, the occurrence and relative abundance of other aldehyde-containing DNA modifications that may also react with ARP calls into questions the true identity of these sites. Two naturally-occurring DNA nucleobases, 5-fC and 5-fU (**Figure 1.12b**), have previously been shown to react with ARP and have been detected at higher levels than AP sites in genomic DNA[128,129]. A study has also been reported in which the genome-wide distribution of both abasic sites and 8-oxoG, which were marked as abasic sites by *in vitro* treatment of isolated DNA with OGG1, were mapped in human HepG2 cells[130]. In both cells treated with x-rays to induce oxidative damage and control cells, a non-random

distribution of these two oxidative damage products was found, as well as a correlation of these lesions with open-chromatin, transposable elements and repetitive regions of the genome. However, this study also utilised the molecule ARP to target abasic sites and the contribution from alternative reactive sites in DNA remains unclear. With the exception of mass spectrometry studies in which aldehyde modifications in DNA can be easily distinguished from one another by molecular weight, many of the studies on abasic sites that utilise ARP or a similar nucleophilic probe may be confounded by the presence of formylpyrimidines in DNA. Therefore, the accurate study of abasic sites remains a challenge and the development of accurate methodology to detect these sites is crucial for further investigation.

Abasic sites are a common putative intermediate in many biological processes, particularly those that involve BER. It has been suggested that specifically in the context of 4-stranded DNA G-quadruplex structures, abasic sites and the APE1 enzyme can be associated with transcriptional control. Single-stranded nucleic acids rich in guanine can form G-quadruplexes as an alternative structure to B-form duplex DNA. These structures consist of stacks of guanine quartets that are connected by hydrogen-bonds using both the Watson-Crick and Hoogsteen faces[131]. In support of earlier proposals that 8-oxoG can act as an epigenetic marker[95], this hypothesis is also based on the generation of oxidative damage in the form of 8-oxoG in specific gene promoters that contain a sequence capable of forming a G-quadruplex. The VEGF promoter has been used as an example in a plasmid-based system introduced to mammalian cells, where a G-quadruplex sequence consisting of five runs of guanine exists, while only four are required for quadruplex folding[132]. Upon generation of 8-oxoG, the base lesion is excised by OGG1 to form an abasic site. The destabilising effect of the loss of base-pairing at this site shifts the equilibrium from duplex to denatured DNA, which can then refold into the quadruplex structure. To exclude the damaged site, the fifth run of guanines is used instead during folding. APE1 binds the AP substrate; however, in the quadruplex context cleavage is inefficient, and the prolonged binding of APE1 is associated with transcriptional upregulation. Both knockdown and small-molecule inhibition of APE1 decreased this upregulation, offering some support for this hypothesis.

Elevated levels of abasic sites in DNA have been associated with cancer. Infection of gastric epithelial cells with *Helicobacter pylori*, which has been associated with gastric cancer, elevated the levels of intracellular ROS and abasic sites[133]. The repair of oxidatively damaged DNA bases such as 8-oxoG with OGG1, as well as abasic sites, was accompanied by genomic instability that has been suggested to play a role in cancer development.

The repair of abasic sites in mammalian systems is largely through BER, for both spontaneously and enzymatically generated sites. APE1 accounts for up to 95% of endonuclease activity at abasic sites[134], thus initiating the BER pathway. Mechanistically, APE1 cleaves DNA at the 5'- side of an abasic site *via* a Mg^{2+} dependent one-step hydrolysis[135]. Key residues include Tyr171 and Glu96, which are involved in guiding the DNA substrate and coordinating the Mg^{2+} ion, whilst His309 activates the nucleophile. AP endonuclease 2 (APE2) is also present in humans, however, much less is currently known about this endonuclease. It is suggested that although AP endonuclease activity is possible for this enzyme, its main function is as a 3'-5' exonuclease[136]. In yeast it has also been shown that in addition to BER, the nucleotide excision repair pathway can also be involved in abasic site repair[137–139].

APE1 is a multifunctional enzyme with a range of possible functions. At the C-terminus, in addition to AP endonuclease activity there is also evidence of weaker activity of this protein as a 3'-5'-exonuclease, phosphodiesterase and 3'-phosphatase, as well as RNase H activity[140,141]. At the N-terminus, APE1 is more commonly referred to as Ref-1 and has redox function to activate transcription factors by reduction at cysteine residues[142,143]. A number of transcription factors that are activated by APE1/Ref-1 have been identified that are involved in apoptosis, inflammation, angiogenesis and survival pathways amongst others. The activities at the two termini can be independent of each other. The small molecule lucanthone has been shown to inhibit the endonuclease, but not the redox or exonuclease activities of APE1[144]. Whilst knockout of APE1 is embryonic lethal, with deaths occurring within 6.5 days post implantation[145], APE1 knockout has been successfully carried out in a murine cell line[134].

Much of the work on DNA abasic site repair has focused on this lesion in the context of double-stranded DNA. More recently, 5-hydroxymethylcytosine binding ESC-specific enzyme (HMCES) has been reported to act as a sensor of abasic sites in single-stranded DNA[146]. HMCES, originally identified as a binder and potential reader of 5-hmC in mESCs[147], was found to bind DNA at replication forks and crosslink single-stranded DNA at abasic sites, which was proposed to shield the abasic site from translesion synthesis during replication. Instead, HMCES acts as a suicide enzyme to prevent repair by BER. In double-stranded DNA, the identity of the base-pair at an abasic site is retained in the opposing base, which can guide BER to restore the correct nucleotide in place of the missing nucleobase. In the event of newly-synthesised DNA opposite a single-stranded abasic site, the opposing base is likely to be incorrectly installed and further translesional synthesis by BER is error-prone. The downstream resolution of the HMCES-abasic site complex, resulting in the proposed

error-free repair of abasic site has not been fully determined. However, this study highlights the potential need for alternative repair pathways when abasic sites are not accompanied by an opposing base-pair to direct DNA repair.

1.3 Methods of detecting DNA modifications

In order to elucidate the role of DNA modifications, it is important to develop methodology to accurately detect them. Quantitative detection of global levels across a genome is an important and powerful tool, whilst the advent of high-throughput NGS has allowed the mapping of DNA modifications in different biological contexts.

1.3.1 Global quantification of DNA modifications

Quantification of the levels of a given DNA modification globally across a genome can be achieved in different ways. Earlier work in this area relied on the ability to bind to modifications through antibody recognition and quantify levels by colorimetric means, often using dot-blot or ELISA assays[148]. Genomic DNA is first immobilised and then incubated with a primary antibody of interest followed by removal of unbound antibody by washing. A secondary antibody specific to the primary antibody is then introduced, which is often capable of amplifying a colorimetric signal such as by incorporation of a horseradish peroxidase unit. This allows the catalytic oxidation of peroxide to water, which in turn catalyses the redox reaction of a suitable fluorogenic substrate. Colorimetric quantification allows the amount of bound antibody and therefore original modification to be calculated, often alongside calibration against known quantities of standard. This type of quantification has been applied using antibodies against 5-mC, as well as a biotinylated probe to target abasic sites[123,149].

Accurate global quantification of DNA modifications can also be achieved by mass spectrometry detection[150–152]. This can be a highly sensitive technique, capable of detecting low abundance modifications. First, DNA is digested using a cocktail of nuclease and phosphatase enzymes into a mixture of individual nucleosides which can then be detected at high accuracy by tandem LC-MS/MS[153,154] (**Figure 1.13**). It is essential to spike-in stable isotopically labelled synthetic standards (SILs) in samples to normalise for differences in ionisation efficiency between species in the mass spectrometer that may arise due to the

presence of salts, contaminating species and technical variation in instrumentation. By recording the ratio of analyte/SIL signal instead of absolute ion abundance, accurate quantification is possible down to femtomolar concentrations, or less than one modification per million canonical bases. It is also essential to generate a calibration curve using known amounts of synthetic standard in order to accurately quantify the levels of modification. By quantifying both the modification of interest and at least one of the canonical bases, measurements can be represented as a proportion of the genome.

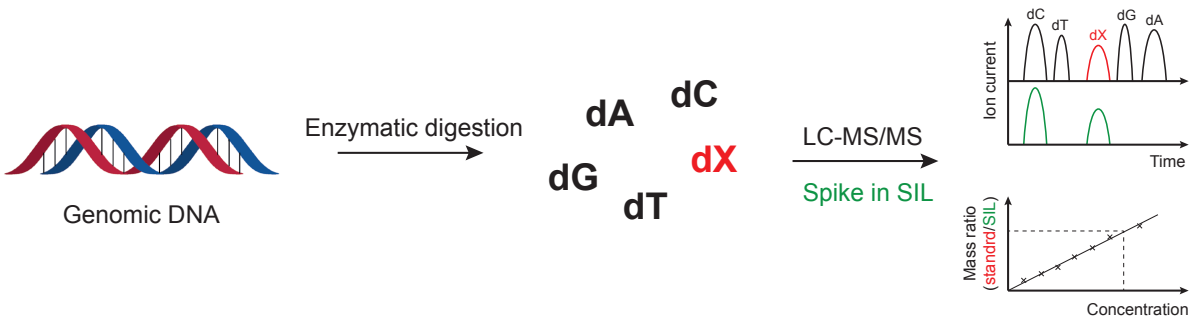


Figure 1.13: Quantification of base modifications by LC-MS/MS. A calibration line using synthetic standards must be generated to carry out accurate quantification. The ion current of known amounts of synthetic standard is measured alongside an internal isotopically labelled standard of fixed concentration and expressed as a standard/SIL ratio to obtain the calibration line. Genomic DNA is digested into nucleosides, and the SIL is also added. The concentration of the modification of interest (dX, red) can then be calculated from the calibration line. At least one of the canonical nucleosides must be quantified in the same way with a further set of synthetic standards, to express the quantification results as a proportion of the genome.

For relatively high-abundance modifications, such as 5-mC and also 5-hmC in some contexts where levels are up to a few percent of cytosine, LC-MS/MS quantification is easily achieved without the need for highly sensitive instrumentation. For rare modifications that are around the parts per million (ppm) levels of the genome or lower, an enhancement of the mass spectrometry signal may be required. Chemical derivatisation of modifications such as the reaction of Girard's reagents T and P[155] with the aldehyde groups in 5-fC or 5-fU allows the formation of ionic species that can improve the sensitivity of LC-MS/MS detection by up to 750-fold[156,157]. Abasic sites are also difficult to directly quantify by mass spectrometry, due to ionisation difficulties of the deoxyribose unit obtained after digestion. Derivatisation of deoxyribose using reactive probes has similarly been used to enhance detection limits to less than 1 abasic site per million dN[47,158,159].

1.3.2 DNA sequencing

DNA sequencing is a powerful tool that has been essential in the furthering of our understanding of genetics over the past four decade. Techniques developed by Frederick Sanger[160], as well as Allan Maxam and Walter Gilbert[161] gave rise to a first generation of DNA sequencing methods. Although low-throughput and costly on a large scale, these techniques were used to successfully complete the human genome project where the vast majority of the human genome was decoded[42].

Since the initiation of the human genome project, a second generation of DNA sequencing, or next-generation sequencing (NGS) techniques, has been made possible[162,163] allowing the speed of sequencing to be dramatically increased, along with higher throughput and lower costs[164]. Individual techniques include Illumina (previously Solexa) sequencing[165], Ion-torrent[166], Roche 454[167] and SOLiD sequencing (Applied Biosystems)[168]. Due to the high throughput and reduced cost of Illumina sequencing, this technology in particular has been widely adopted in the past decade.

Like many other sequencing platforms, Illumina sequencing is based on the concept of sequencing by synthesis. In order to decode the order of bases in a given DNA template, the identity of the bases installed by a polymerase on a newly synthesised, complementary DNA strand is recorded. This information can then be used to reveal the original template sequence. To identify each incorporated base in a controlled and stepwise manner, non-natural triphosphates are used in place of natural ones in which two key modifications are present; a dye unique to each triphosphate attached to the base allowing identification by fluorescent imaging, and a removable group on the sugar ring to temporarily inhibit further polymerase extension after each incorporation cycle[169] (**Figure 1.14**). Specifically, the *O*-azidomethyl group is used at both the 3'-OH of the sugar, and within the linker between the nucleobase and dye. Treatment with TCEP reduces the *O*-azidomethyl group into a hemiaminal, which under aqueous conditions rapidly hydrolyses into a free hydroxyl. This generates both an uncapped 3'-OH, and a truncated linker on the base free of the fluorophore that does not significantly hinder further polymerase activity. This method of reversible termination allows successive cycles of DNA synthesis to occur, with enough time to allow the collection of data by imaging between cycles. Another key advantage of Illumina sequencing is the use of massively parallel processing technology. By sequencing millions of unique DNA fragments in parallel on a single flow-cell, throughput is vastly increased, and the cost of sequencing is reduced. As the detection of single fluorophores is challenging and

costly, clones are generated on the flow-cell prior to sequencing by a process known as bridge amplification. This allows the amplification of fluorescence signal, as the fluorophores within each clonal cluster are identical.

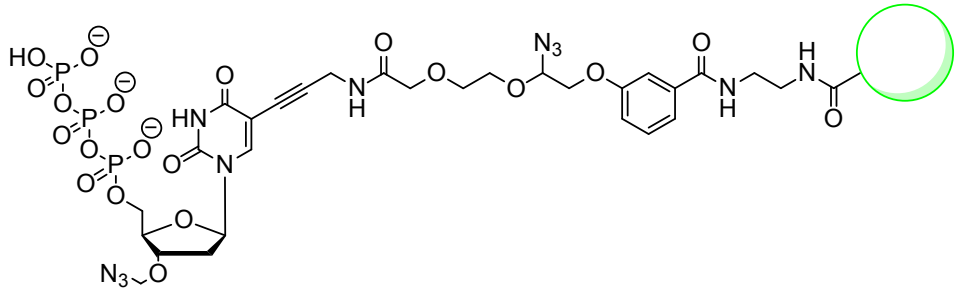


Figure 1.14: Structure of a modified triphosphate used during Illumina sequencing. The 3'-OH is capped with a removable azido group, which is also incorporated into the linker between the dye (green circle) and the base. In the presence of a reducing agent such as TCEP, the 3'-OH is uncapped and the dye linker is cleaved, allowing another round of polymerase synthesis to occur.

Before DNA can be loaded onto a sequencing flow cell, a sequencing library must be generated (**Figure 1.15**). As the quality of individual reads obtained by Illumina sequencing rapidly drops beyond a few hundred base pairs, it is necessary to break genomic DNA into shorter pieces. By computationally aligning millions of such shorter reads together against a known reference genome, coverage is possible for the majority of the genome. Fragmentation of genomic DNA can be achieved enzymatically by controlled partial degradation of non-specific nucleases, or more commonly by mechanical or acoustic shearing *via* sonication. The ends of the resultant DNA fragments are generally uneven with partial overhangs, and an enzymatic end-repair step is often required to create blunt ends and also phosphorylate the 5'- end.

In order for sequencing libraries to be recognised and amplified on the flow-cell, sequencing adapters are introduced onto both ends of DNA fragments. The adapters contain a universal priming region to allow for amplification of the library, as well as a sequencing primer region on both ends adjacent to the fragment insert at which sequencing can be initiated. Often, one or more barcodes of typically 6 nucleotides in length are also included in the adapter to label all the DNA in a library from a given sample. This allows multiple libraries to be sequenced in parallel, and the combined data can be demultiplexed later based on the identity of the barcode[170,171]. Enzymatic ligation is typically used to introduce adapters onto

DNA fragments. The adapters are synthesised with a 3'-dT overhang, whilst library preparation incorporates a 3'-dA tailing step to generate a single base-pair complementary overhang with the adapter to enhance ligation efficiency. Optional library amplification then completes library preparation, which is ready for bridge-amplification on the flow-cell followed by successive rounds of sequencing by synthesis. Sequencing from only one of the two ends is known as single-end sequencing, thereby generating data only in 'read 1'. Alternatively, clusters can be regenerated from the opposing DNA ends mid-way through sequencing; 'read 2' data is also collected to obtain sequencing data initiated at the complement sequence at the other end. This is known as paired-end sequencing. The barcodes are usually read in a separate step, using a further primer specific to the barcode region within the adapter.

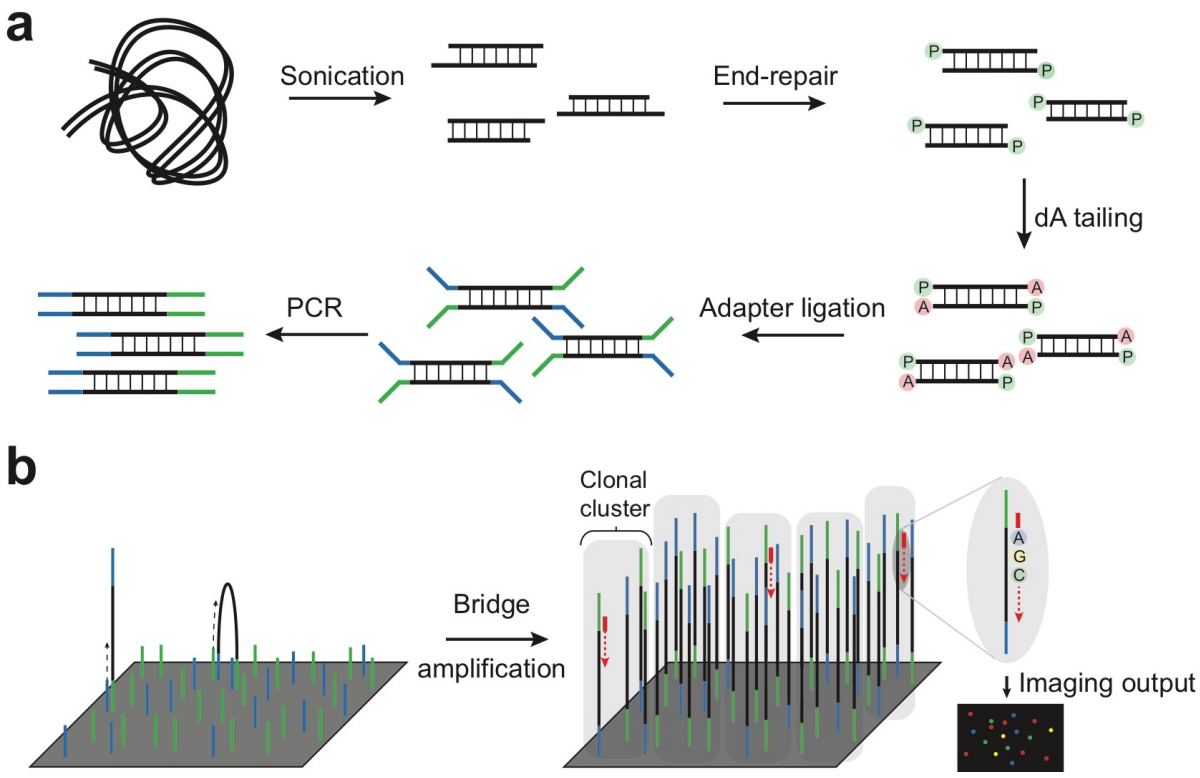


Figure 1.15: Illumina sequencing technology. **a**) Key steps during sequencing library preparation. High molecular weight genomic DNA is fragmented by e.g. sonication into short (100-1000 bp) fragments. DNA ends are blunted and 5'-ends are phosphorylated during end-repair, followed by dA-tailing of the 3'- ends. Y-shaped sequencing adapters are then introduced by ligation, followed by PCR to straighten out DNA ends. **b**) Flow-cell during sequencing. Adapter-ligated libraries are introduced to flow-cells, which are amplified into clusters *via* bridge-amplification. Sequencing primers are introduced (red) to the flow-cell, complementary to part of the adapters and cycles of sequencing by synthesis using reversible terminator dyes occurs. The output is a series of images, when can be decoded computationally.

1.3.3 Mapping DNA modifications by sequencing

Currently, NGS methods can only decode the four canonical bases. PCR amplification of DNA is generally required both during library preparation and on the flow-cell, during which DNA modifications are erased. Therefore, to map the location of DNA modifications, an alternative readout is required.

For low resolution DNA modification mapping, affinity enrichment has been widely used for a range of base modifications (**Figure 1.16**). After sonication, DNA fragments containing a feature of interest are isolated preferentially over background DNA fragments. Sequencing of libraries enriched in this way results in the pile-up of reads at genomic locations where the modification accumulates, which can be detected bioinformatically as peaks with higher sequencing coverage than the background. Specific primary antibodies have been raised against a number of modifications, allowing the enrichment mapping of a number of modifications by DNA immunoprecipitation (DIP-seq) including 5-mC[172], 5-hmC[173] and *N*6-methyladenine (6-mA)[174]. The requirements for input DNA quantities for DIP-seq are relatively low, and peaks can be called even when sequencing depth is kept relatively low. However, potential drawbacks of DIP-seq include a density bias of some antibodies[175], whilst IgG antibodies have been observed to bind preferentially to short tandem repeats in mammalian DNA and therefore show a false enrichment in these regions[176]. This can be overcome to some extent by the careful use of control antibodies alongside enrichment libraries. A similar approach is also routinely used to identify features within chromatin during chromatin immunoprecipitation sequencing (ChIP-seq).

An alternative to DIP-seq is the use of a specific chemical probe to introduce a biotin moiety to DNA modifications, which can then be enriched using streptavidin pulldown[70,128,129,177,178]. Depending on the functional groups available for tagging, this approach can be more specific than antibody recognition. Suitable candidates include the aldehyde groups at 5-fC[128,177] and 5-fU[129,179,180], where the latter can also be used to map 5-hmU when a selective oxidation step is included[70]. The relative enrichment of modified to unmodified DNA achieved by chemical pulldown can often be much higher than that offered by antibodies[128], thus improving the sensitivity and reliability of enrichments. Despite being a non-covalent interaction, the binding interaction between biotin and streptavidin is one of the strongest known in nature, with a K_D of around 10^{-15} M[181,182]. Together with the strong covalent linkages often used to tag biotin at target DNA modifications, this allows the use of extensive washing steps to minimise non-specific binding of background DNA during

pulldown and can greatly improve specificity. Chemical pulldown approaches are, however, limited to modifications for which selective chemistry can be designed, whilst modifications such as 5-mC remain challenging to target in the presence of canonical bases.

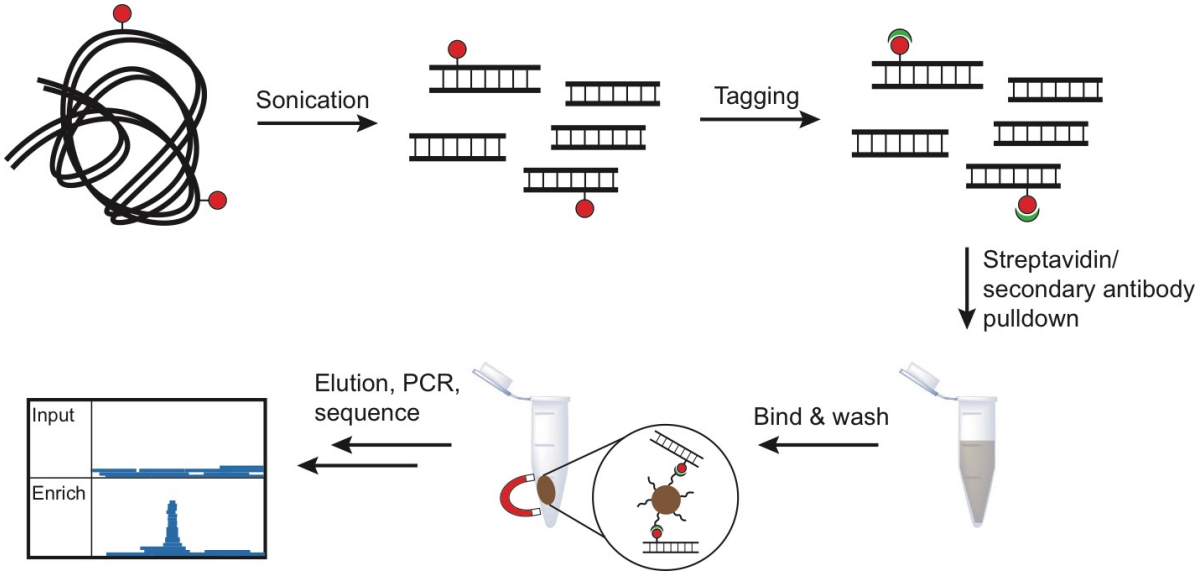


Figure 1.16: DNA modification mapping by enrichment. High-molecular weight genomic DNA is sonicated into shorter fragments. The modification of interest (red) is tagged using e.g. an antibody or biotinylated probe (green), then captured using either a secondary antibody or streptavidin beads, respectively (brown). The sample is washed extensively to remove non-bound DNA, then the captured fragments are eluted, amplified by PCR and sequenced. The sequencing output is compared to an input sample without enrichment, in order to identify loci of read pile-ups as peaks.

NGS has also been used to map a number of modifications at single-nucleotide resolution. One widely used technique is the mapping of 5-mC by bisulfite sequencing[183,184]. This approach is based on the differences in reactivity of 5-mC and C towards bisulfite. Whilst canonical cytosine is quantitatively deaminated in the presence of bisulfite at low pH, the rate of deamination of 5-mC is up to 100 times slower. The deamination of cytosine to uracil changes its base-pairing pattern, which upon PCR amplification will be replaced by thymine. Therefore, it is possible to replace all cytosine sites in a genomic sample with thymine *via* uracil, leaving 5-mC as the only sites that will still be decoded as cytosine during sequencing (**Figure 1.17a,b**). This powerful technique is also quantitative, as the amount of residual cytosine detected by sequencing at a given site in the genome can be compared to the reads in which the same location is read as a T, and thus a ratio of 5-mC to C can be calculated. A major drawback of bisulfite sequencing is the depth required to obtain sites reliably, which has an impact on cost. For example, more than 100X sequencing coverage would be required to consistently detect 5-mC levels of 1% at a given site. To reduce the sequencing depth

required, reduced representation bisulfite sequencing (RRBS) has been developed as a way to cover a subset of the genome in which 5-mC is most relevant[185]. Genomic DNA is first digested with a methylation insensitive restriction enzyme, MspI[186], which recognises and cuts in the middle of a CCGG motif. The DNA fragments generated in this way end in a CpG site, and thus is enriched for these locations. Around 1% of the genome is covered by this method, allowing the information from those regions to be selectively obtained using 100-fold less sequencing power than during whole genome bisulfite sequencing.

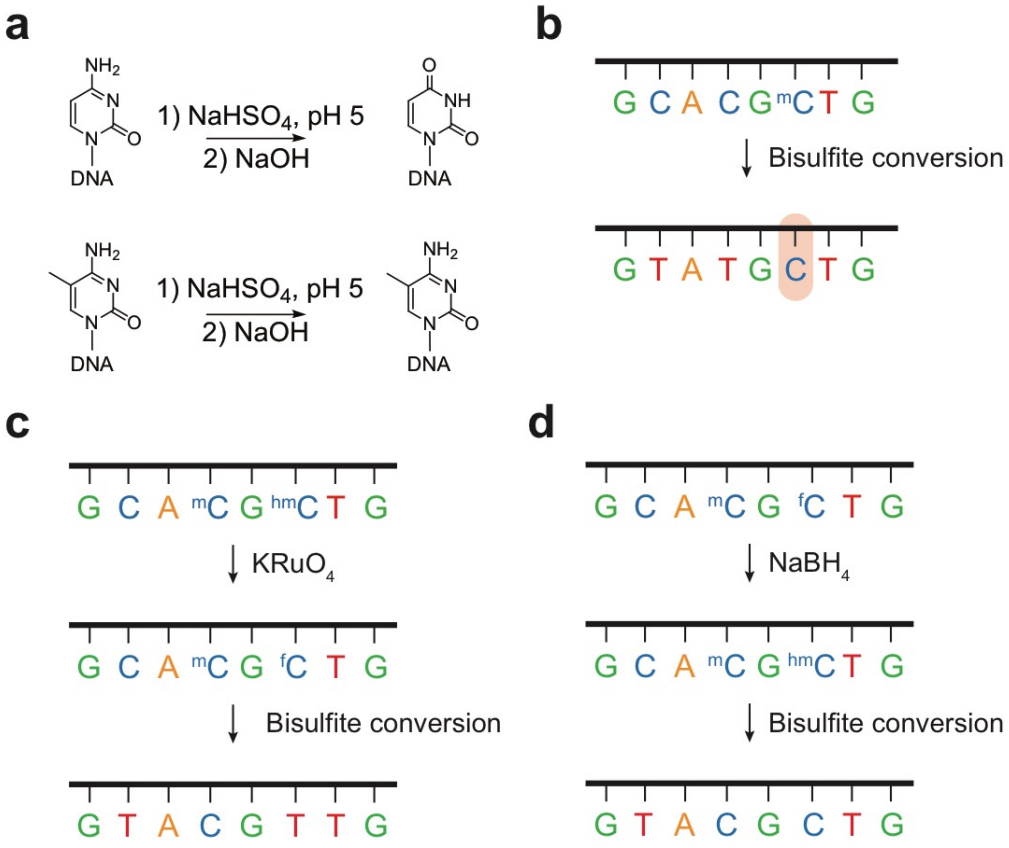


Figure 1.17: Mapping DNA modifications at base-resolution. **a**) Cytosine is deaminated by bisulfite to form uracil, while 5-mC is not. **b**) Schematic representing the expected base-calling pattern before and after bisulfite conversation. **c**) Key steps during ox-bisulfite sequencing. **d**) Key steps during red-bisulfite sequencing.

After the discovery of the importance of the oxidised derivatives of 5-mC, it was found that standard bisulfite sequencing was not specific to 5-mC. 5-hmC is also resistant to bisulfite conversion, and therefore all datasets generated using bisulfite alone map both 5-mC and 5-hmC simultaneously. To differentiate between 5-mC and 5-hmC, a selective oxidation of 5-hmC to 5-fC using the water soluble oxidant potassium perruthenate was

developed and combined with bisulfite sequencing[187] (ox-bisulfite sequencing). As 5-fC is deaminated in the presence of bisulfite, 5-hmC can be identified as the sites that are read as a C after bisulfite treatment, but T after ox-bisulfite treatment. The true 5-mC sites are then those that persist as C after both bisulfite and ox-bisulfite treatment. Using a similar concept, 5-fC sites can be sequenced after sodium borohydride reduction in a reduced-bisulfite (red-bisulfite) treatment[188] when compared with standard bisulfite sequencing. The base-resolution mapping of other modifications, including 5-fU, has also been made possible by similarly detecting a mutational signature, which in the case of 5-fU is selectively introduced at 5-fU sites under specific PCR conditions[189].

A further way to obtain nucleotide-resolution mapping information by NGS is to mark sites with the start or end of sequencing reads. This can be achieved by inducing polymerase-stalling at the modification site, or through chemically induced DNA fragmentation at the modified nucleotide. The pattern of polymerase-stalling has been used to map both DNA and RNA secondary structure, due to the natural tendency of polymerases to halt DNA synthesis at non-canonical structures[190,191]. This approach can also be applied to certain DNA modifications. The intra-strand bulky adduct formed between cisplatin and guanine bases in DNA, for example, can be used to cause polymerase-stalling during primer extension (**Figure 1.18a**). Sequencing of the truncated DNA strands synthesised allows the stall site, and therefore site of the cisplatin adduct, to be identified as the genomic position after which a pile-up of sequencing reads begin[192]. Endogenous DNA double-strand breaks (DSBs) can also be sequenced at base-resolution in a similar approach (**Figure 1.18b**). Here, the DNA is already truncated at the site of interest, so the *in vitro* stalling of a polymerase is not required to generate a signature. In the DSBcapture method[105], a biotinylated adapter is introduced to genomic DSBs *in situ*, and isolated using streptavidin to enrich the signal. Upon sequencing, peaks appear that are centred around the DSB site. A similar concept has also been used to map modifications in RNA; ribose sugar (2'-*O*)-methylation sites were mapped using Nm-seq by inducing RNA fragmentation at this modification[193] (**Figure 1.18c**). Iterative oxidation–elimination–dephosphorylation (OED) cycles were used to cleave the RNA backbone at the 3'- position of 2'-*O*-methylation sites, whilst leaving unmethylated sites intact. Sequencing of the resultant fragments reveals a pile-up of reads directly adjacent to methylated sites of the transcriptome, thus mapping this modification at base-resolution. The read depth required to identify polymerase stalling or alternative fragmentations can be high; a significant proportion of the genome must be marked by start sites as background, before a positive signal can be detected. Often, these methods can be coupled with a biotinylated

tag as in DSBcapture[105] or a pre-selection of genomic DNA using a specific antibody[192], to initially focus reads to specific regions of the genome.

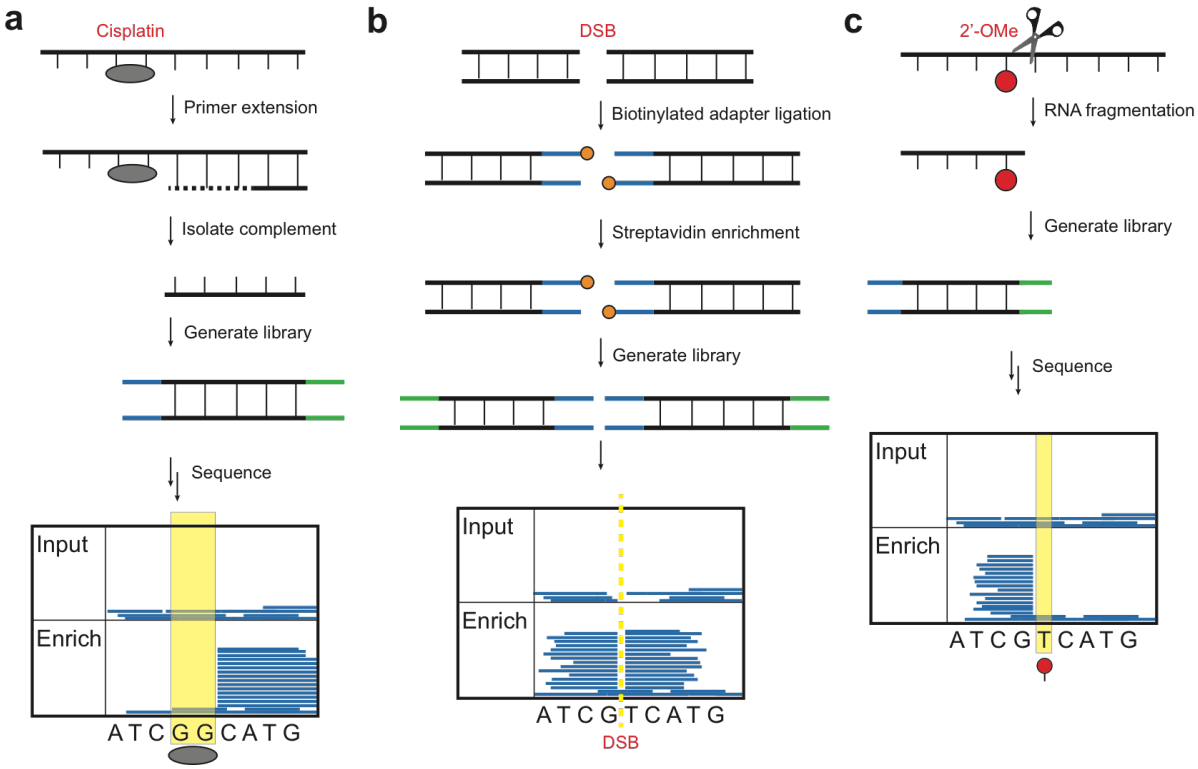


Figure 1.18: Single-nucleotide resolution mapping of DNA modifications by marking sequencing read start sites. **a**) Polymerase-stalling at DNA-cisplatin adducts (grey) causes termination of primer extension. After sequencing libraries generated from the truncated DNA products, the cisplatin adduct site is identified as the position directly preceding sequencing read start sites. **b**) Mapping of endogenous double-strand breaks (DSBs) in DNA. A biotinylated adapter is introduced at the ends of DSBs, which are isolated and used to generate an enriched library. The centre of read pile-ups after sequencing is identified as the original captured DSB site. **c**) Mapping of 2'-*O*-methylated (red) RNA sites. RNA is fragmented preferentially at methylated sites. The resultant RNA fragments are prepared for sequencing, and the 2'-*O*-methylated site is identified as the position directly 3'- to sequencing read start sites.

Third-generation DNA sequencing techniques have been in development in recent years that hold great potential for the direct detection of DNA modifications. Single-molecule real-time (SMRT) sequencing[194] (PacBio) collects data not only on the identity of an incorporated triphosphate during sequencing by synthesis, but also on the amount of time that the polymerase is associated with its substrates during each cycle. This temporal information is often enough to distinguish between modifications in the template that lead to the same Watson-Crick base-pairing patterns. Alternatively, Nanopore sequencing technology (Oxford Nanopore Technologies)[195] does not rely on the traditional sequencing

by synthesis approach. Instead, the ionic current of a piece of DNA passing base by base through a nanopore is recorded, generating a unique signature for each of the four canonical nucleobases and often for different modifications as well. Both of these third-generation sequencing methodologies have the additional advantage of much longer read lengths compared to Illumina sequencing, up to 10s or 100s of kilobases. Highly repetitive DNA in parts of some genomes is difficult to align using short reads from current Illumina technology and is challenging to study, but it may be possible to align to these regions using much longer reads. Together, these advantages of third-generation sequencing are promising, however further development is still needed, particularly regarding the accuracy of base calling. The high error rates currently limit these methods to smaller genomes such as bacterial DNA. With future developments, third-generation sequencing is likely to provide exciting opportunities in the field of genomics and DNA modifications.

1.4 Objectives

The removal and repair of a range of DNA modifications is dependent on the BER pathway. As such, abasic sites are common intermediates that are central to the regulation of a number of DNA modifications. Independent of enzymatic generation, abasic sites are also formed directly in the event of endogenous and exogenous DNA damage. Furthermore, there is emerging evidence that abasic sites may in their own right able to influence biological processes, beyond simply being a passive intermediate. Despite the many ways in which abasic sites have biological significance, both as an intermediate and as a potential DNA modification, there is currently no methodology to accurate locate these sites in DNA. With the exception of LC-MS/MS detection, many studies on abasic sites that rely solely on the reactivity of the aldehyde functionality may be confounded by cross-reactivity with alternative sources of aldehydes in DNA such as 5-fC and 5-fU, which can be more abundant than abasic sites. The main focus of this thesis was therefore to develop and apply methodology to accurately detect the location of DNA abasic sites by next-generation sequencing.

The first key objective was to develop a chemical method to selectively target abasic sites and combine this with sequencing on the Illumina platform. A chemical pulldown approach is described in Chapter 2, involving the targeting of the aldehyde functionality of abasic sites. Extensive controls were used to ensure chemical selectivity over alternative sources of aldehydes in DNA before application of the method to synthetic DNA containing

an abasic site at a known location, as a proof of concept that abasic sites can be mapped at single-nucleotide resolution.

Further objectives in this thesis were to apply this sequencing method in different biological contexts. Methodology to map abasic sites is useful in wider applications due to the availability of glycosylase enzymes that can be used to treat isolated DNA *in vitro* to replace specific modifications with an abasic site. As such, any modification for which a glycosylase is available can also be mapped using AP mapping technology. As a validation of the sequencing method in a genomic context, studies on thymine modifications in the *Leishmania major* genome are described in Chapter 3; these modifications were converted to abasic sites by treatment of DNA with the SMUG1 enzyme. A low-resolution map of 5-hmU has been generated previously in this genome and was compared to the data generated here, allowing validation of the new method. The high-resolution data was also used to further analyse the distribution and sequence-context of 5-hmU, to further shed light on the potential role of this modification where low-resolution data may be limited.

In Chapter 4, the objective was to map the distribution of endogenous abasic sites in the human genome. Knockdown of the repair protein APE1 was carried out using siRNA control, and a map of endogenous AP sites was generated in these cells as well as control cells to assess whether DNA damage accumulates preferentially in certain parts of the genome. The data was also analysed relative to genomic features, as well as the transcriptional levels of mRNA to investigate whether AP damage can influence gene expression.

Finally, in Chapter 5 the significance of the uracil-dependent processive demethylation pathway[51,52] is described. The location of uracil was mapped genome-wide at single-nucleotide resolution by pre-treatment of DNA with UNG to mark these locations as abasic sites, followed by abasic site sequencing. This approach was first explored and optimised in the *E. coli* genome which was easier to investigate due to its small size. Uracil mapping was then carried out on mESC DNA, where levels of the key components of the proposed pathway have been manipulated to investigate whether deamination of cytosine can be detected to support this hypothesis.

Chapter 2

Chemical tagging of DNA abasic sites

2.1 Background

Amongst the many different possible products of DNA damage, abasic sites are frequent lesions with levels measured at around 1 per million dN in mESCs by mass spectrometry[47]. In addition to their role as markers of DNA damage that may arise in the event of either endogenous or exogenous damage to genomic DNA, abasic sites are also central intermediates in the base excision repair (BER) pathway. At least 11 distinct mammalian glycosylases have been identified to date that initiate the repair of a wide range of DNA base lesions as part of BER, during which abasic sites are the first common intermediate[91,92]. Therefore, the ability to accurately detect and investigate abasic sites is useful both in the study of DNA damage and in further understanding a number of important biological pathways that rely on BER, such as the removal and regulation of epigenetic DNA modifications[44,46,196].

In addition to decoding the primary sequence of DNA, sequencing is also a powerful tool for the study of DNA modifications. Despite the fact that next-generation sequencing (NGS) techniques are generally limited to readouts that decipher between only the four canonical bases, methods to detect base modifications have been developed on these platforms at both low (100s of bp) and single-nucleotide resolution. These approaches typically rely either on the detection of a distinct mutational signature that can be induced at specific modifications by chemical treatment[183,184,187], or the affinity enrichment of DNA fragments containing specific modifications to selectively increase sequencing coverage at locations where modifications accumulate. The latter can be achieved by affinity capture using an antibody or by selective covalent chemistry using a functionalised probe. Third generation sequencing methods such as SMRT or Nanopore sequencing show promise in mapping DNA modifications[197,198]; however, due to high error rates these platforms are not yet suitable for the routine analysis of large mammalian genomes.

Despite the prevalence and severe biological consequences of abasic sites, there is still a lack of understanding of the distribution and dynamics of these features within genomic DNA. Many methods of detecting abasic sites that are currently available utilise *O*-biotinylcarbazoylmethyl hydroxylamine, also known as ARP. This biotinylated nucleophile reacts with the aldehyde group exposed in the ring-open form of abasic sites to form an oxime linkage. This approach has been coupled with ELISA or dot-blot assays to quantify the global levels of AP sites in DNA[123,124], as well as to map the genomic location of AP sites by NGS. In AP-seq[130], fragments of genomic DNA that have been treated with ARP were enriched by isolation using magnetic streptavidin beads, and eluted by enzymatic treatment (PreCR repair mix, NEB). The enriched DNA was then used to generate a sequencing library. A major drawback to the use of ARP in labelling abasic sites is that naturally occurring DNA modifications such as 5-fC and 5-fU also contain reactive aldehydes that can be targeted by nucleophilic probes[125,128,129]. As these modifications have been measured at higher levels than AP sites in mammalian DNA (mESCs)[47,50], it is imperative that methods used for the detection and mapping of abasic sites do not cross-react with these alternate base modifications. The aim of this chapter was therefore to develop a method to enrich for DNA abasic sites in the presence of genomic DNA, including aldehyde-bearing bases. This affinity enrichment strategy was then combined with sequencing on the Illumina platform, to develop a method of mapping abasic sites that can be applied widely to a range of genomes and biological contexts. Some antibodies used to map the location of base modifications as part of DIP-seq have been found to display density bias, whilst IgG antibodies have been suggested to preferentially bind to short tandem repeats and lead to false positives in these genomic regions[175,176]. Due to these drawbacks associated with antibody detection, a chemical probe that can react selectively at abasic sites was explored instead.

The biotin-streptavidin linkage is one of the strongest non-covalent interactions found in nature. Chemical probes that contain a biotin moiety have been utilised in the affinity enrichment of a number of DNA modifications, effected by isolation using magnetic streptavidin beads[70,128,177]. It was therefore envisaged that an abasic site probe should either directly contain a biotin moiety, or a reactive handle through which a biotin moiety can be introduced at a later stage. The further functionalisation of chemically modified DNA has previously been demonstrated with high efficiency using click reactions[177,199], including the copper catalysed azide-alkyne Huisgen cycloaddition (CuAAC).

The rate of depurination, and to a lesser extent depyrimidination, of DNA is enhanced under acidic conditions and elevated temperatures[111,200]. In order to avoid the generation of

AP site artefacts, it is essential that such conditions are avoided during chemical tagging reactions. Furthermore, the rate of strand-cleavage at pre-existing AP sites *via* β- or β-δ-elimination at the 3'- and 5'- phosphate groups of the backbone is increased under alkaline conditions[117]. Both the artefactual generation of AP sites, and AP removal by base-catalysed elimination is likely to change the distribution of AP sites and affect sequencing results. Therefore, any probes developed in this chapter should react cleanly with abasic sites under near-neutral conditions, without the need for elevated temperatures.

Abasic sites cause stalling during DNA synthesis for many polymerases, including the high-fidelity polymerases that are generally required to prepare samples for NGS[106,201]. Whilst it is sometimes possible to use PCR-free methods to generate a library, amplification of DNA on the sequencing flow-cell still remains essential during Illumina sequencing. Polymerases that carry out translesional synthesis, such as Vent from *Thermococcus litoralis*, or Dpo4 from *Sulfolobus solfataricus* are able to partially bypass abasic sites[202]. However, the efficiencies of synthesis past abasic sites are still hundreds of times lower than that past the canonical bases. PCR bias against a target modification is detrimental to enrichments by underrepresenting modified DNA. The inefficient polymerase bypass of biotin-tagged 5-fC has been shown to hinder PCR amplification of 5-fC DNA after affinity capture. This bias has been removed by introducing a cleavable linker into the functionalised probe, so that the bulky biotin moiety could be removed after streptavidin pulldown, prior to PCR amplification[203]. A further desirable feature of an abasic site probe is therefore the ability to allow efficient PCR amplification of enriched DNA prior to sequencing, possibly *via* a cleavable linker or a ligation that is reversible under controlled conditions.

Affinity enrichment mapping of DNA modifications is largely achieved at low resolution. The resolution is often determined by the size of DNA fragments obtained upon the breaking of high-molecular weight genomic DNA, which is generally limited to a few hundred base pairs. An increase in this resolution has been demonstrated during the affinity enrichment of some modifications. For example, in cisplatin-seq, cisplatin-treated DNA is first immunoprecipitated using an antibody that recognises cisplatin-DNA adducts. By utilising the subsequent polymerase stalling at cisplatin-DNA crosslink sites to generate the sequencing library, an accumulation of truncated reads allows the crosslink site to be identified at single-nucleotide resolution[192]. A similar strategy was also utilised in this chapter, to improve the resolution of the mapping method beyond hundreds of base pairs. Together, the points highlighted above outline the requirements for a suitable chemical probe for DNA abasic sites.

2.2 Results and discussion

2.2.1 Aldehyde reactive probes

Amine nucleophiles are used widely in organic synthesis to form imines with aldehyde groups. In particular, hydroxylamines and hydrazines are suitable candidates for reaction at aldehydes due to the increased nucleophilicity of these functional groups, provided by the α-effect[204,205]. The condensation reaction between hydroxylamines and carbonyl groups typically requires acidic conditions, whilst hydrazines react more readily under near-neutral conditions[206]. However, the resultant hydrazones are less stable than oximes and are labile to reversal by hydrolysis under aqueous conditions[207]. The relatively slower formation of oximes can be catalysed by nucleophiles such as aniline or *p*-anisidine[128]. Whilst such nucleophiles have been used to catalyse the reaction between hydroxylamines and 5-fC, amine catalysts have been shown to degrade DNA at abasic sites by strand-cleavage and are therefore less suitable for these purposes[208].

In order to study the reactivity of abasic sites, short synthetic DNA obtained by solid-phase synthesis was used. The direct incorporation of deoxyribose as an abasic site into synthetic DNA is incompatible with solid-phase synthesis as the exposed aldehyde is susceptible to degradation. Whilst it is possible to use a protected abasic phosphoramidite to obtain synthetic AP sites[209], a commonly used method is to install uracil residues into DNA and generate an abasic site by enzymatic treatment with uracil DNA glycosylase (UNG)[53]. All AP oligodeoxynucleotides (ODNs) used in this thesis were generated this way, *via* excision of uracil from the corresponding uracil-ODNs (**Figure 2.1a**). The UNG reaction on U-ODN1 was followed by LC-MS, which showed that quantitative base excision was achieved within 2 h with no degradation of the abasic site product by elimination detected under the conditions used. The retention times of the uracil ODN and AP product were very similar, however, the identity of the product was confirmed by mass spectrometry (**Figure 2.1b**). (See **Table 7.3** for sequences of all ODNs used.)

ARP has been previously used to study abasic sites, as well as 5-fC and 5-fU in DNA[128]. It has been assumed in AP-seq for example, that acidic conditions (pH 5) are required for efficient oxime formation between ARP and 5-fC, whilst the reaction with abasic sites is favoured around neutral pH[130]. The further cross-reactivity of ARP with 5-fU has largely been ignored in previous studies, despite evidence that this modification also exists in genomic

DNA. The reactivity of ARP was therefore compared between these three aldehyde modifications. Short, single-stranded ODNs (ssODNs) containing either an AP site, 5-fC or 5-fU were incubated with ARP (5 mM), at pH 7.4 (phosphate buffer, 40 mM) for 2 h at 37 °C. Quantitative conversion was observed for both AP-ODN1 and fU-ODN2; however no product formation was detected for fC-ODN3 (**Figure 2.2**). This result corroborates reports on the reduced reactivity of 5-fC towards nucleophiles compared to 5-fU which has also been confirmed by *ab initio* quantum mechanical calculations[129].

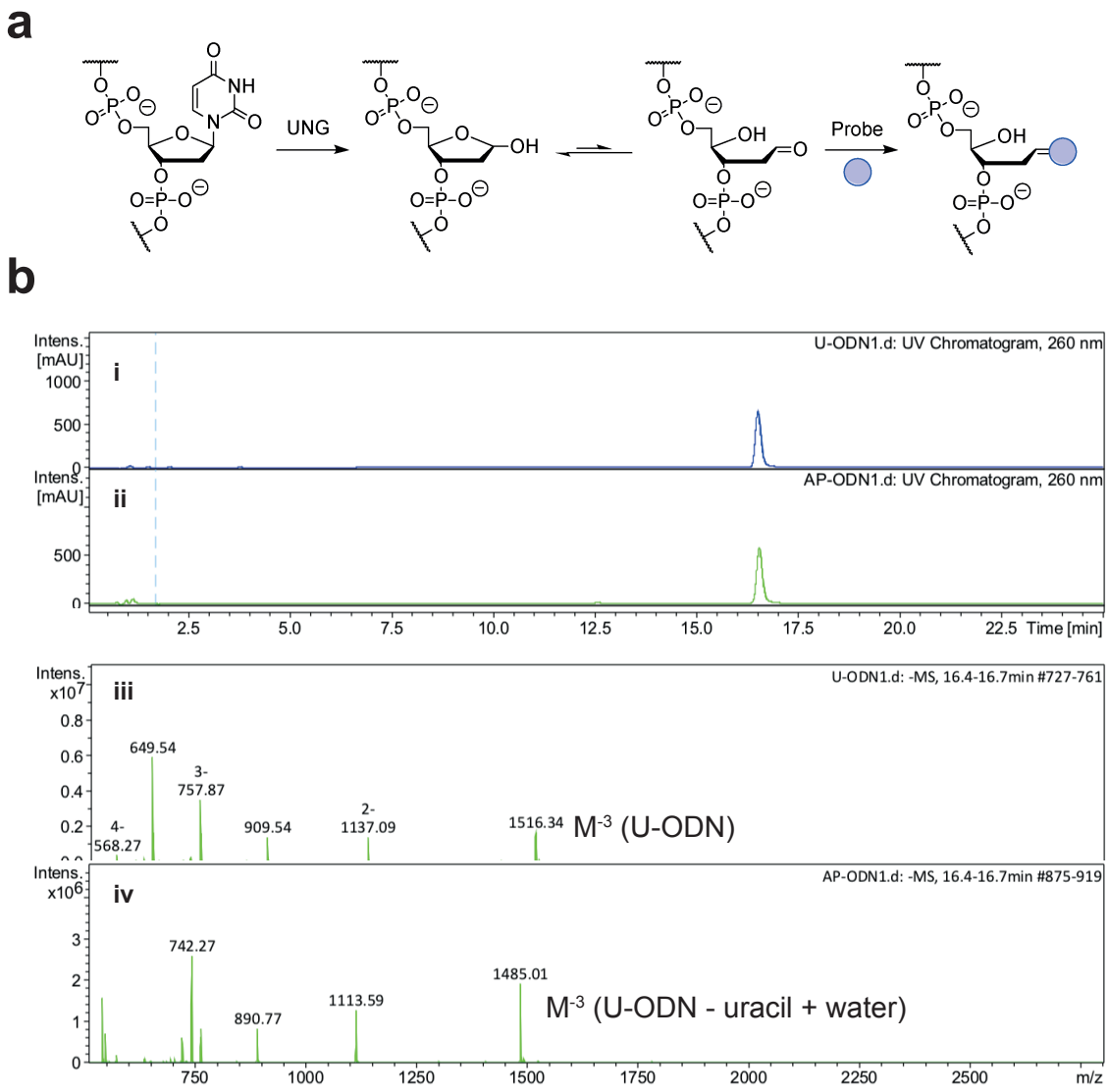


Figure 2.1: Generation of an AP site from uracil in DNA. **a**) Scheme showing the excision of uracil in synthetic DNA by UNG to generate an AP site, which is then used to study the reactivity of abasic site probes. **b**) LC-MS UV trace of **i**) U-ODN1 and **ii**) U-ODN1 after treatment with UNG to generate AP-ODN1**.** Mass spectrum of **iii**) U-ODN1 and **iv**) AP-ODN1. The Elution times of starting material and product are very similar at around 16.5 min; however, from the mass traces **iii**) and **iv**) a loss of 31 is observed for the M^{-3} ion, corresponding to the mass of uracil minus water (112-18 = 94). See **Table 7.3** for sequences of all ODNs used.

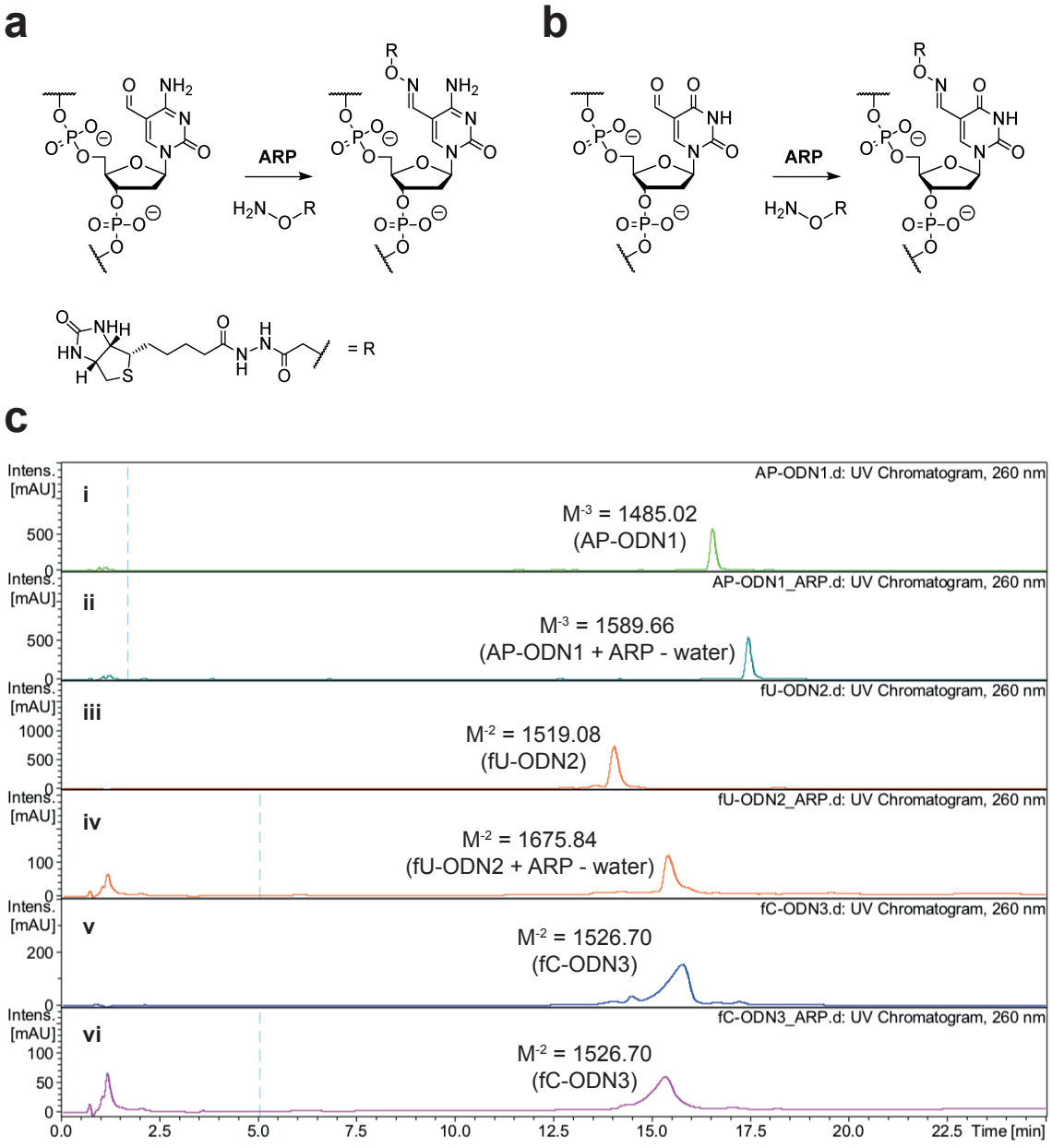


Figure 2.2: Reactivity of DNA modifications with ARP. Scheme showing reaction of ARP with **a**) 5-fC and **b**) 5-fU. **c**) LC-MS UV traces of ODNs before and after reaction with ARP, **i**) and **ii**) correspond to AP-ODN1, **iii**) and **iv**) to fU-ODN2, **v**) and **vi**) to fC-ODN3. ODNs (10 µM) were incubated with ARP (5 mM) in phosphate buffer (pH 7.4, 40 mM) for 2 h at 37 °C. UV absorption at 260 nm is shown.

Interestingly, when the reaction buffer was replaced with PBS (pH 7.4) in conditions that reflected those in AP-seq[130], quantitative conversion was also observed for fC-ODN3 (**Figure 2.3**). This somewhat surprising result was confirmed on a further 5-fC containing ODN (fC-ODN4). The two reaction conditions differ in the concentration of phosphate, and the inclusion of sodium chloride in PBS. A possible explanation for the change in reactivity is that the buffer capacity is exceeded in PBS when using high probe concentration (5 mM),

due to the relatively low phosphate concentration (4 mM). Measurement of the reaction pH in PBS confirmed that this drops to around pH 6 in the presence of 5 mM ARP, whilst the pH in 40 mM phosphate buffer remains above 7. Together, these results confirm that 5-fU and abasic sites are both good targets for ARP and also show that 5-fC may be labelled under specific conditions. Therefore, this probe does not appear to offer the selectivity required and measurements relying on this probe alone are likely to be confounded by the presence of both 5-fC and 5-fU.

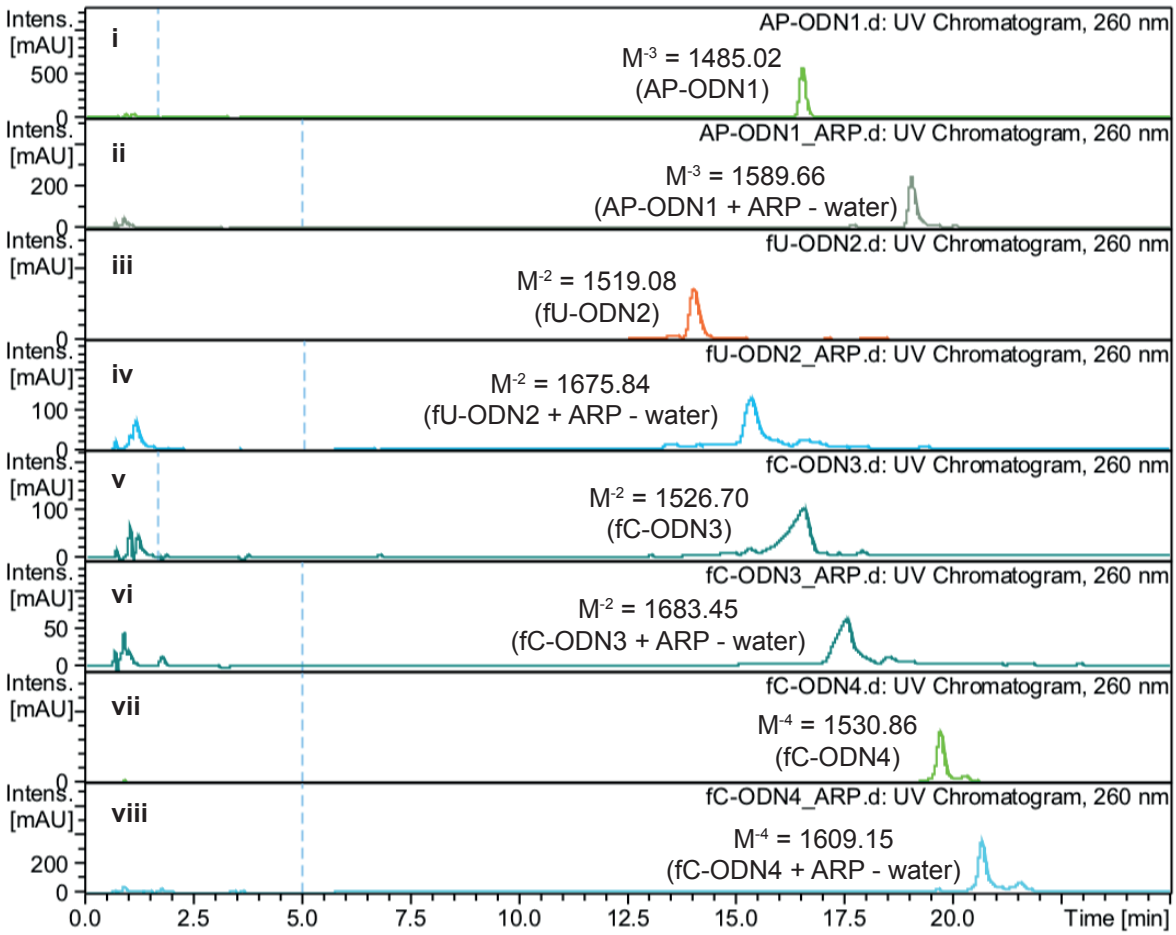
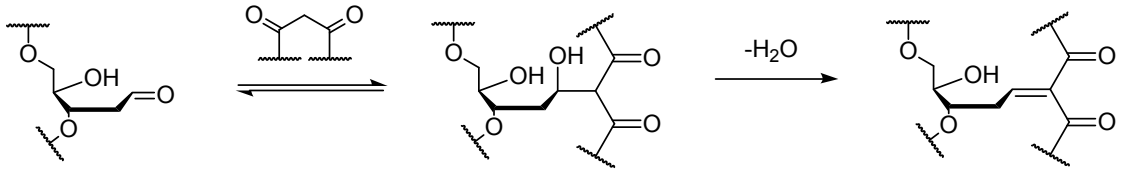


Figure 2.3: LC-MS UV traces of ODNs before and after reaction with ARP. **i**) and **ii**) Correspond to AP-ODN1, **iii**) and **iv**) to fU-ODN2, **v**) and **vi**) to fC-ODN3, **vi**) and **viii**) to fC-ODN4. ODNs (10 μM) were incubated with ARP (5 mM) in PBS (pH 7.4) for 2 h at 37 °C. UV absorption at 260 nm is shown.

1,3-diketones are nucleophilic compounds that can react with aldehydes through the aldol condensation to form an α,β-unsaturated diketone (**Scheme 2.1**). The condensation between reducing sugars and diketone nucleophiles has previously been reported under aqueous conditions[210–212] and a functionalised indandione probe has also been used to react

with the aldehyde group in 5-fC[177]. The reactivity of this class of compounds was therefore tested on abasic sites. Mildly basic conditions, achieved by addition of sodium hydrogen carbonate, as well as elevated temperatures are generally required for the efficient condensation of diketone compounds with glucose. These conditions are likely to lead to β-elimination at abasic sites and were avoided. The condensation between pentane-2,4-dione and mannose, however, proceeds in the absence of sodium hydrogen carbonate[210], suggesting that neutral conditions may still be compatible for this ligation. The reactivity of three diketone compounds including 1,3-indandione **3** and pyrazolone **4** were tested on an AP-ODN, where reaction conditions were kept mild at pH 7 with incubation at room temperature.



Scheme 2.1: Reaction mechanism of aldol condensation of diketone compounds at an AP site.

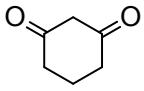
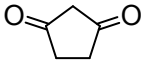
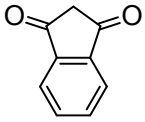
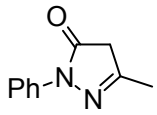
Probe	***1*** 	***2*** 	***3*** 	***4*** 
Conversion 2 h	None	None	None	Trace
Conversion 24 h / %	n.d.	n.d.	53%	11% (48% cleavage)

Table 2.1: Reactivity of diketone probes with AP sites. AP-ODN5 (10 μM) was incubated with probes (10 mM) in phosphate buffer (pH 7.0, 40 mM) for 2-24 h at room temperature. Reactions were followed by LC-MS, and conversions were calculated by integration of UV signal at 260 nm. n.d. = not determined.

Out of the four compounds tested, only compound **4** showed traces of reactivity after 2 h (**Table 2.1**). The reaction time was therefore extended to 24 h. Although the yield of ligated product increased slightly (11%), cleavage of the abasic site started to also occur. A moderate yield for the reaction with indandione **3** was observed after 24 h reaction, however, further extension of the reaction time was impractical as long incubations are likely to subject DNA to further damage. Raising the reaction pH to 9.0 did not further improve yields. Given that yields were modest even after 24 h, this class of compounds was not further explored.

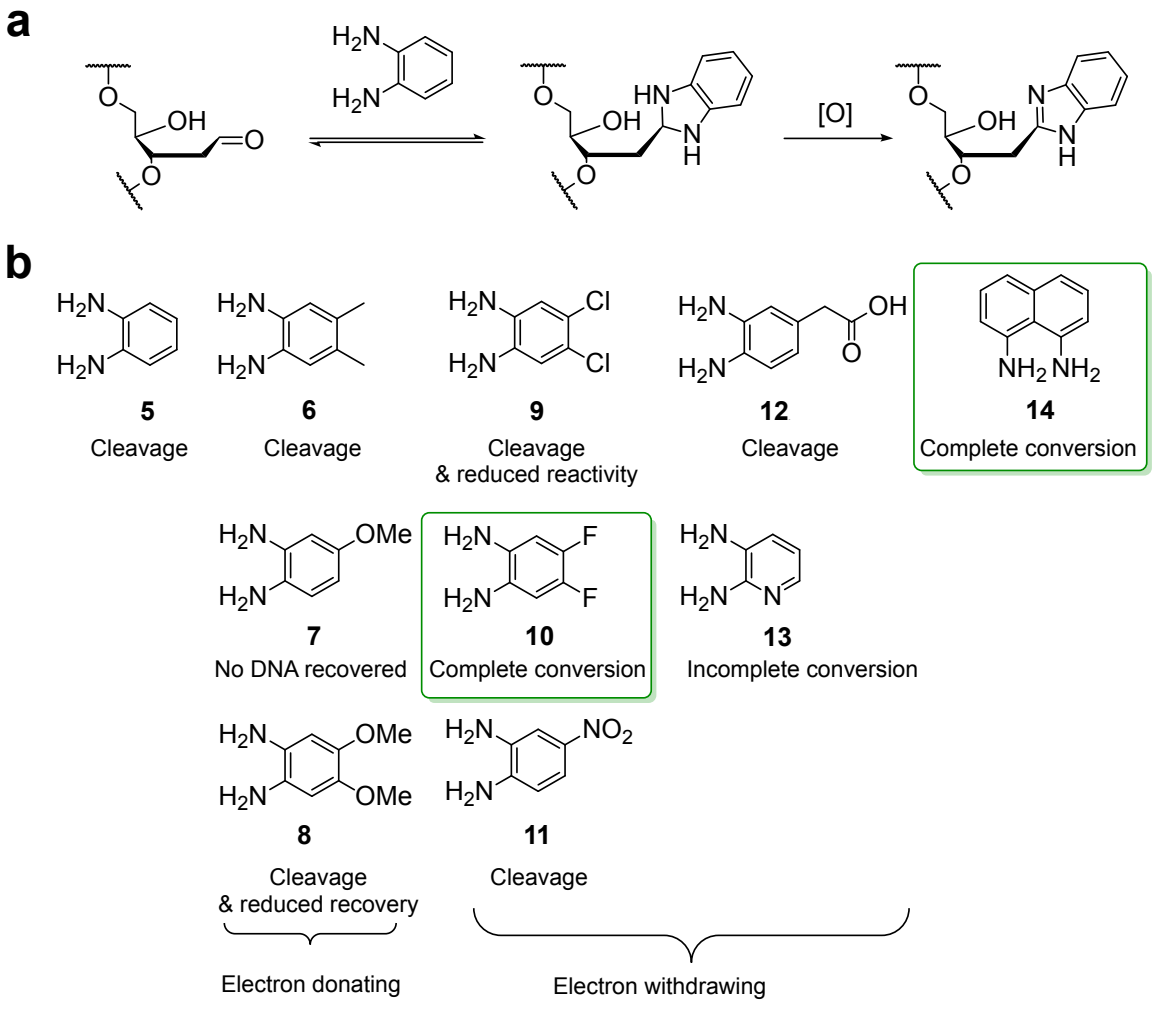


Figure 2.4: Reactivity of diamine probes with AP sites. **a**) Reaction mechanism of *o*-phenylenediamine with an abasic site. **b**) Outcomes of reactions of AP-ODN5 with derivatised diamine probes. AP-ODN5 (10 μM) was incubated with probes (10 mM) in phosphate buffer (pH 6.0, 40 mM) for 2 h at room temperature. Reactions were followed by LC-MS, and conversions were calculated by integration of UV signal at 260 nm.

o-Phenylenediamine has previously been used as a probe to react at 5-fU sites in DNA[129,179,213]. The 1,2-diamine group cyclises with aldehydes, and the resultant aminals readily oxidise in air to form a stable benzimidazole ring (**Figure 2.4a**). Reaction of *o*-phenylenediamine **5** with an AP-ODN generated the expected benzimidazole adduct. However, cleavage of the AP site through β-elimination was also observed. A number of functionalised diamines were therefore screened for reactivity without causing DNA cleavage. Of the compounds tested, the difluoro derivative **10**, and diaminonaphthalene **14** showed good reactivity without significant elimination (< 5%) and were chosen as potential candidates for an abasic site probe. Some of the more reactive diamines functionalised with electron donating groups were found to reduce DNA recovery, possibly by non-specific degradation (**Figure 2.4b**). Although less reactive, electron-withdrawing groups on the aromatic ring were found to be more suitable.

2.2.2 Hydrazino-*iso*-Pictet-Spengler reaction

The Pictet-Spengler reaction was first reported in 1911, in which β-arylethylamines condense with carbonyl groups to form cyclic adducts[214] (**Figure 2.5a**). The reaction is slow for phenylethylamine, requiring reflux under acidic conditions. The initial condensation to form an iminium ion is accelerated in the presence of acid, after which an intramolecular cyclisation leads to the formation of a carbon-carbon bond. In nature, tryptamine undergoes an enzyme-catalysed Pictet-Spengler reaction during the biosynthesis of alkaloids[215] (**Figure 2.5b**). The reaction has also been used synthetically as a route to form tetrahydro-β-carbolines, including asymmetric versions in which the stereochemistry of the cyclic product can be controlled[216].

Whilst the original Pictet-Spengler reaction requires harsh conditions, a biocompatible version of this reaction has also been reported[217]. In the Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction, a C2 functionalised indole is used instead of tryptamine where the carbon chain is located on C3, and the amine nucleophile exchanged for a more nucleophilic hydrazine moiety (**Figure 2.5c**). The traditional Pictet-Spengler reaction with tryptamine has been suggested to involve cyclisation by attack from the more nucleophilic C3 carbon, forming a spirocyclic compound, which requires migration of the carbon-carbon bond to the C2 position before re-aromatisation can occur[218]. The rate of reaction proceeds more efficiently under biocompatible conditions when a stable 6-membered ring is formed directly after attack from the C3 carbon. Indole **18** has been shown to be a good substrate for the HIPS reaction with

aldehyde residues within proteins at near-neutral conditions. It was therefore reasoned that a similar reaction could also occur at an abasic site, and HIPS probe **19** was designed (**Figure 2.6**). This probe contained the same reactive scaffold as **18** and the linker at the indole nitrogen was switched for a propargyl handle to allow biotinylation at this position at a later stage *via* a CuAAC reaction.

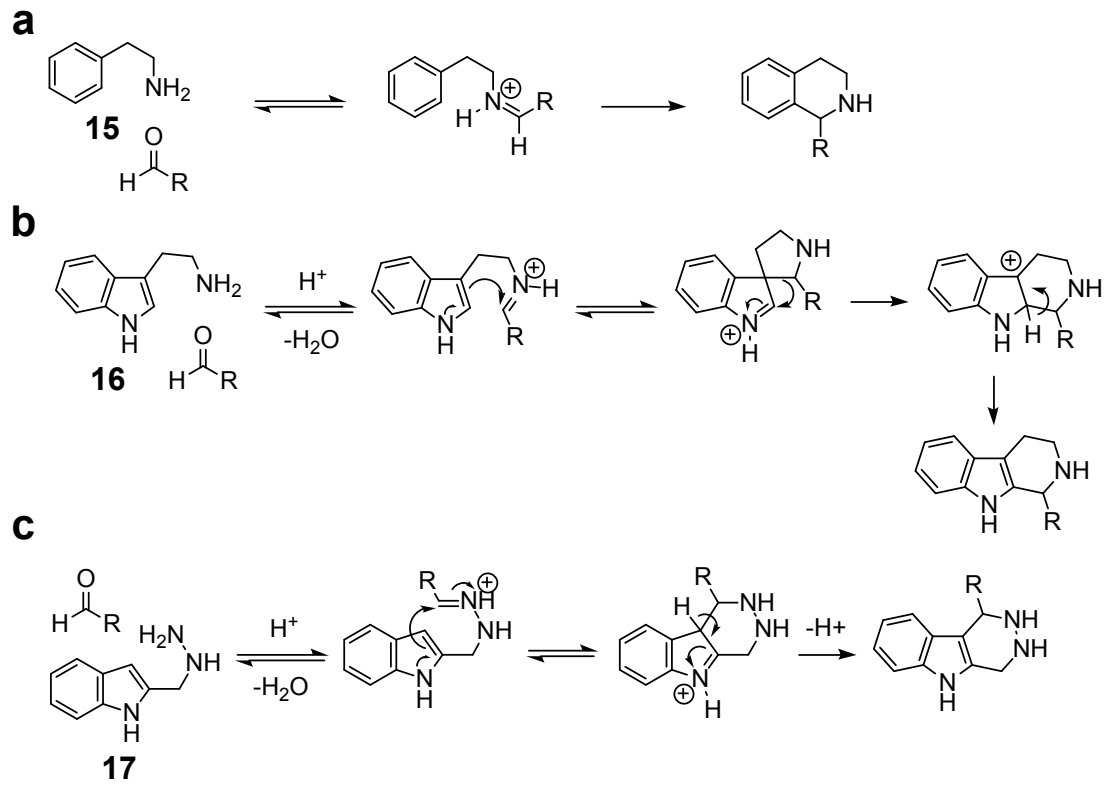


Figure 2.5: Mechanism of Pictet-Spengler reaction. **a**) Pictet-Spengler reaction between phenylethylamine and an aldehyde. **b**) Pictet-Spengler reaction between tryptamine and an aldehyde. **c**) Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction.

Compound **19** was obtained in six steps in a synthesis modified from that reported by Agarwal *et al*.[217] (**Figure 2.6b**). Propargylation of indole **21** at the endocyclic nitrogen was achieved in the presence of sodium hydride and propargyl bromide, forming indole **22** in good yield. The ester group in **22** was then reduced by lithium aluminium hydride to generate **23**, followed by oxidation with Dess-Martin periodinane to form aldehyde **24**. 1,2-dimethylhydrazine **25** was Fmoc-protected on one end. Here, a mixture of mono- and bis-protected hydrazines was obtained and a modest yield of the mono-protected compound **26** was isolated. Protected hydrazine **26** was then coupled to aldehyde **24** by reductive amination using sodium triacetoxyborohydride. A final Fmoc deprotection affords **19**.

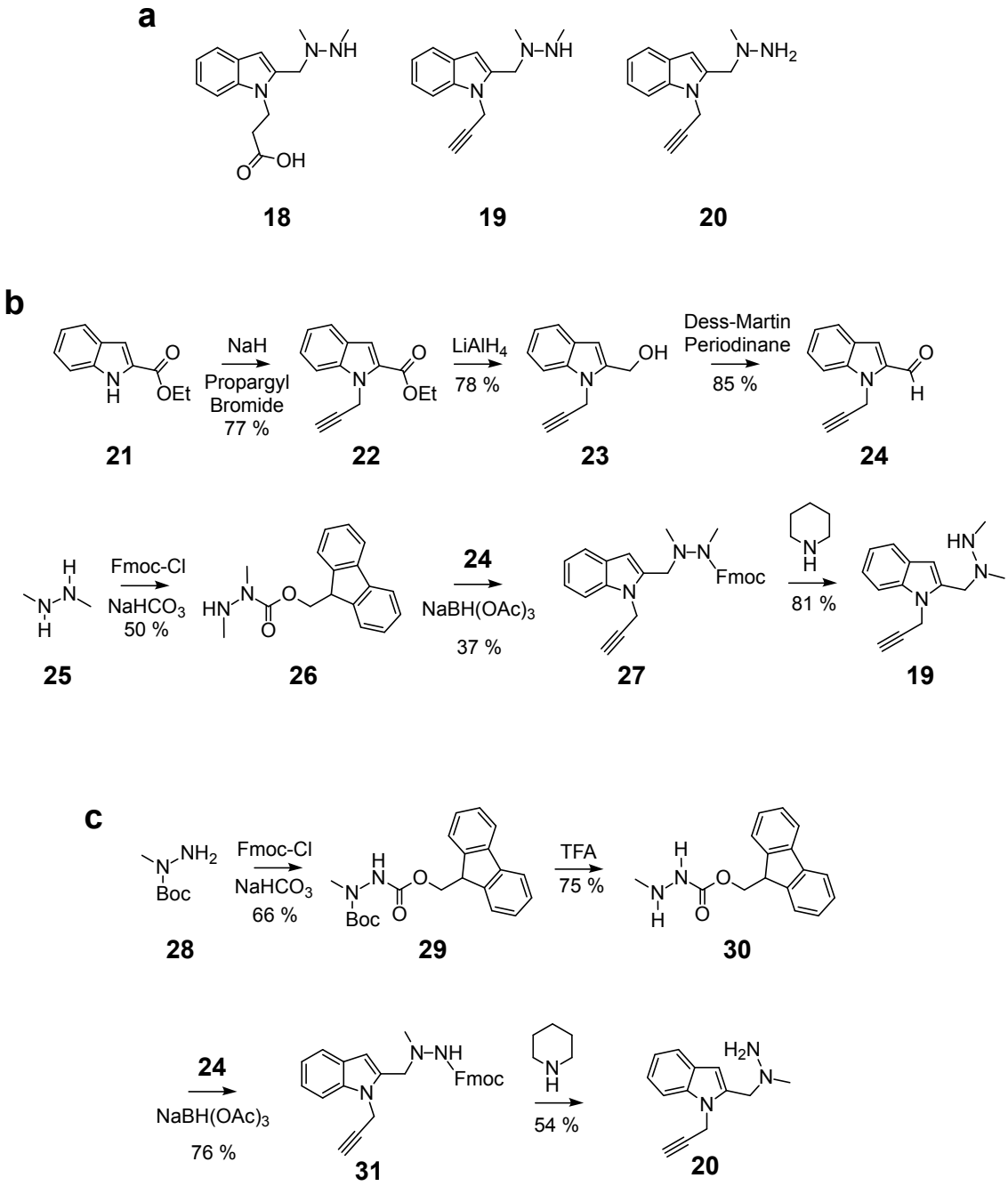


Figure 2.6: HIPS probes for aldehyde tagging. **a**) Structures of probes capable of HIPS ligation. **18** has been shown to react with formylglycine residues in proteins[217], and **19** and **20** were designed for reactivity with abasic sites. **b**) Synthetic route for HIPS probe **19**. **c**) Synthetic route for HIPS probe **20**.

Incubation of HIPS probe **19** with an AP-ODN afforded only a small amount of the expected adduct (30%). A major side product was formed instead (70%), for which the mass was 14 Da lower than expected (**Figure 2.7**). As this mass difference corresponds to the loss of a methyl group, it was reasoned that demethylation of the probe may occur during the HIPS reaction. HIPS probe **20** was therefore synthesised, in which the terminal nitrogen is unmethylated. Whilst it was possible that demethylation of the non-terminal nitrogen was occurring, a probe lacking a methyl group at this position would cause problems with potential imine formation from both ends of the hydrazine. This would lead to a mixture of products forming both 5- and 6-membered rings upon cyclisation. The synthetic route to **20** was similar to **19** with the exception that a Boc-protected mono-methylated hydrazine **28** was used instead of 1,2-dimethyl hydrazine (**Figure 2.6c**). This protected hydrazine was further protected with Fmoc on the exposed nitrogen, then treated with acid to remove the Boc group. The resultant hydrazine was coupled to aldehyde **24** as before, then deprotected to generate **20**. Reaction of HIPS probe **20** with AP-ODN affords a single product corresponding to the expected mass, and demethylation at of the hydrazine did not appear to significantly hamper the HIPS reaction (**Figure 2.7**).

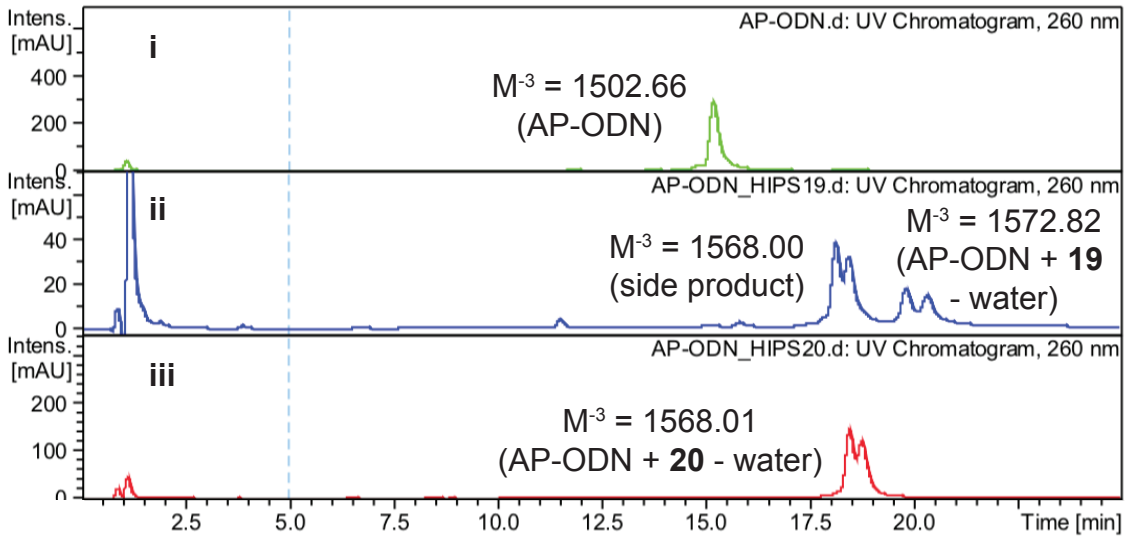


Figure 2.7: LC-MS UV traces of AP-ODN5. **i**) Untreated, **ii**) after reaction with **19** and **iii**) after reaction with **20**. ODN (10 μM) was incubated with **19** or **20** (10 mM) in sodium phosphate buffer (40 mM, pH 7.4) for 2 h at room temperature. UV absorption at 260 nm is shown. The expected mass for product with **19** is 1573 (M^{-3}), whilst the side product with a mass shift of around 4.8 for the M^{-3} ion corresponds to loss of methyl group (14 Da).

2.2.3 Comparison of probes

Two diamine probes, **10** and **14** and HIPS probe **20** have shown good reactivity towards abasic sites to generate adducts without forming side products or leading to DNA degradation. The relative reactivity of these probes, along with the ARP molecule was therefore compared on an AP-ODN and followed by LC-MS. As reaction pH is an important consideration when studying abasic sites in order to preserve their distribution within DNA, reactions were tested over a range of conditions around neutral pH (5.0-7.4). Given that acidic conditions should be avoided to prevent depurination, and strongly alkaline conditions should be avoided to prevent elimination, physiological pH (7.4) was selected as an appropriate compromise.

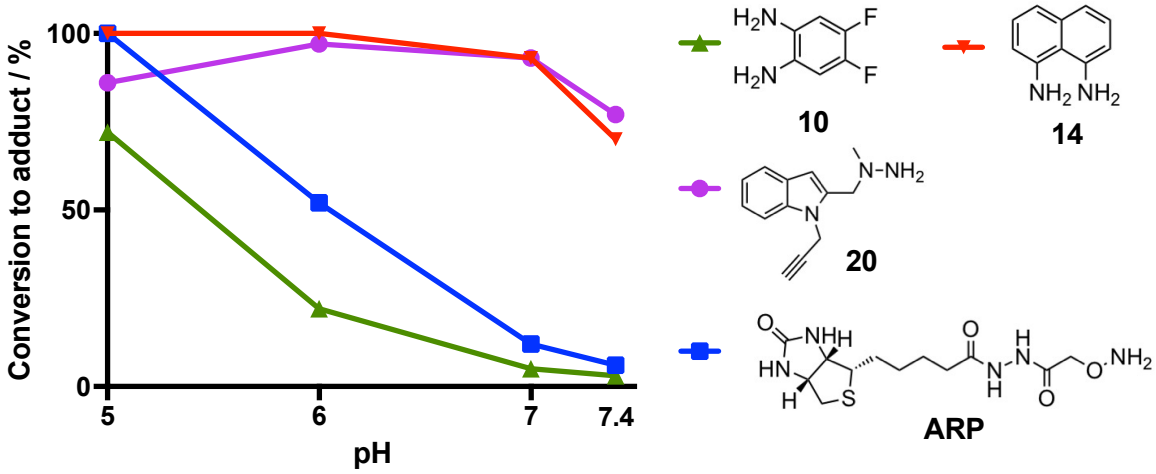


Figure 2.8: Reactivity of abasic site probes at varying pH. AP-ODN5 (10 μM) was incubated with each probe (1 mM, except **14** which was at 0.5 mM) for 2 h at room temperature. Reactions were buffered by sodium acetate (pH 5.0) or sodium phosphate (pH 6.0-7.4) buffers at 40 mM. % conversion was calculated by integration of ODN species at 260 nm UV absorption.

The reactivity of diamine **10** and ARP were found to drop with increasing pH (**Figure 2.8**). Oxime formation is catalysed by acid, and low reactivity at and above neutral pH has previously been observed with these types of probes[217]. As expected, HIPS probe **20** retains reactivity well at up to pH 7.4, due to the hydrazine nucleophile. Whilst a hydrazone linkage would subsequently suffer from hydrolytic instability, the combination of a reactive hydrazine and stable HIPS adduct makes this probe particularly suitable for use under biocompatible conditions. Diamine **14** was found to be exceptionally reactive across the pH range tested and the concentration of this probe had to be reduced (0.5 mM) compared to the other probes tested (1 mM) in order to observe detectable differences over the pH range. At this lower

concentration, reactivity is still maintained well up to pH 7.4. Overall, these results suggest that **14** and **20** are the most suitable for labelling abasic sites at physiological pH.

As the ultimate goal for an abasic site probe is to enable enrichment mapping, it is important that probes form a stable adduct with an AP site that will not reverse or degrade during the DNA treatment steps required during library preparation. The functionalised AP-ODNs with each probe were treated with methoxyamine to assess the stability of adducts towards displacement by a nucleophile, and sodium hydroxide to assess the stability of adducts towards DNA degradation by backbone cleavage. Probes **10**, **14** and **20** were all stable in the presence of methoxyamine (**Figure 2.9**). In contrast, 36% of the ARP adduct reacted with methoxyamine to form an oxime, suggesting that this adduct is susceptible to nucleophilic attack. Whilst the conditions used here are much harsher than DNA is generally subject to during library preparation, the results suggest the possibility that some ARP adducts may be lost during the processing of labelled DNA. For HIPS probe **20**, it was difficult to determine by LC-MS alone whether the obtained DNA adduct had undergone cyclisation or was halted after hydrazone condensation as the masses of the two ODN products were difficult to distinguish. The inability of methyoxyamine to displace **20** suggests that cyclisation had occurred as the hydrazone would be expected to be unstable in the presence of a stronger nucleophile. All the adducts were relatively stable towards DNA degradation when heated under alkaline conditions (100 mM sodium hydroxide), with only the oxime formed with ARP showing small amounts of elimination (< 10% after 15 min at 70 ˚C).

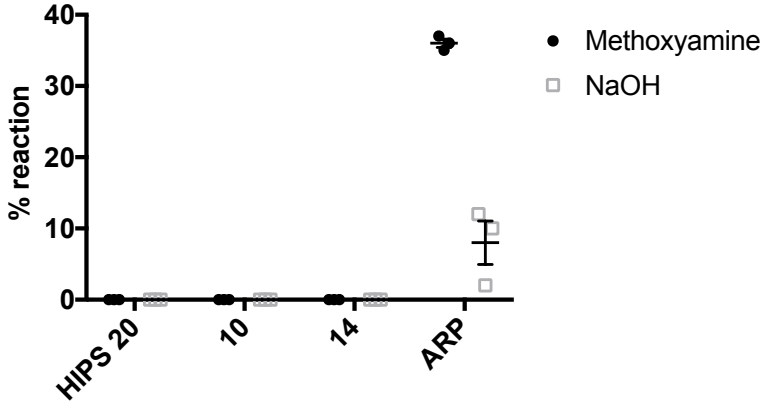


Figure 2.9: Stability of functionalised AP-ODN5 in the presence of a nucleophile (methoxyamine) and towards backbone elimination in the presence of sodium hydroxide. Reactions conditions were 100 mM sodium hydroxide at 70 ˚C for 15 min, and 100 mM methoxyamine, pH 6.0 for 2 h at room temperature. % product was calculated from integrated area of UV absorption at 260 nm. Mean and S.E.M. of three independent reactions are shown.

2.2.4 Removable biotinylation of abasic sites

A desirable feature for an abasic site probe is the ability to remove the biotin moiety after pulldown to minimise the hindering of polymerase-mediated synthesis during subsequent PCR amplification. It was found that for the adduct between HIPS probe **20** and an AP site, a unique cleavage reaction could occur that resulted in removal of the biotin moiety from the tagged DNA, which was not possible for the other probes. Probe **20** bears an alkyne handle which can be functionalised *via* a CuAAC reaction. This was designed to allow DNA adducts formed with this probe to be biotinylated at a later stage under mild, biocompatible conditions that will not interfere with other sites in DNA. Treatment of HIPS-functionalised AP-ODN with biotin-PEG3-azide in the presence of copper (I) bromide and the ligand tris(3-hydroxypropyltriazolylmethyl)amine (THPTA) was confirmed to generate a single, biotinylated species. The inclusion of THPTA in CuAAC reactions on DNA has been shown to protect DNA from copper damage[199,219]. Interestingly, the mass of the biotinylated product was lower than the calculated mass by 4 Da, suggesting that oxidation of the product also occurs alongside the click reaction. The most likely site of this oxidation is at the six-membered ring formed during HIPS cyclisation, leading to aromatisation of the adduct. A cationic pyridazinoindole structure was therefore proposed for this adduct (**Figure 2.10**). This slightly unusual aromatic species has previously been reported[220]. Whilst the initial HIPS adduct was found to be stable at high pH (**Figure 2.9**), this oxidised adduct becomes labile towards elimination. Similar to an unfunctionalised abasic site, the oxidised, biotinylated HIPS-AP adduct undergoes quantitative elimination upon heating at 70 °C in 100 mM sodium hydroxide (**Figure 2.10**). Analysis of the fragmentation products by LC-MS showed that β-δ-elimination dominates under the conditions used, whilst traces of the incomplete β-elimination product could also be detected. Although relatively harsh, these alkaline-cleavage conditions do not show signs of leading to non-specific degradation and should be otherwise compatible with DNA.

The base-catalysed elimination of biotinylated HIPS-AP adducts generates two fragments of DNA. Depending on whether β- or β-δ-elimination has occurred, the ends of the fragment generated on the 5'- side of the initial abasic site will differ. In the case of β-elimination, the biotin moiety is still attached to this fragment, forming a non-natural DNA end. If δ-elimination then follows, the biotin moiety is cleaved also from the 5'- fragment, leaving a 3'-phosphorylated end (**Figure 2.10a**). In contrast, for both β- and β-δ-elimination, a single fragment species is released from the 3'- side of the abasic site that is free of modification and no longer biotinylated. Therefore, this DNA cleavage presents an opportunity to remove

biotin from abasic sites after they have been isolated by streptavidin pulldown. It was envisaged that the fragment released from the 3'- side of the abasic site could be recovered for sequencing. Similar approaches have been successfully carried out to map cleavage-sensitive sites in RNA at single-base resolution when the alignment of the sequencing start site was considered[193]. As such, base-catalysed elimination simultaneously allows the removal of biotin to avoid the hindering of PCR amplification, as well as the position of abasic sites to be marked at single-nucleotide resolution.

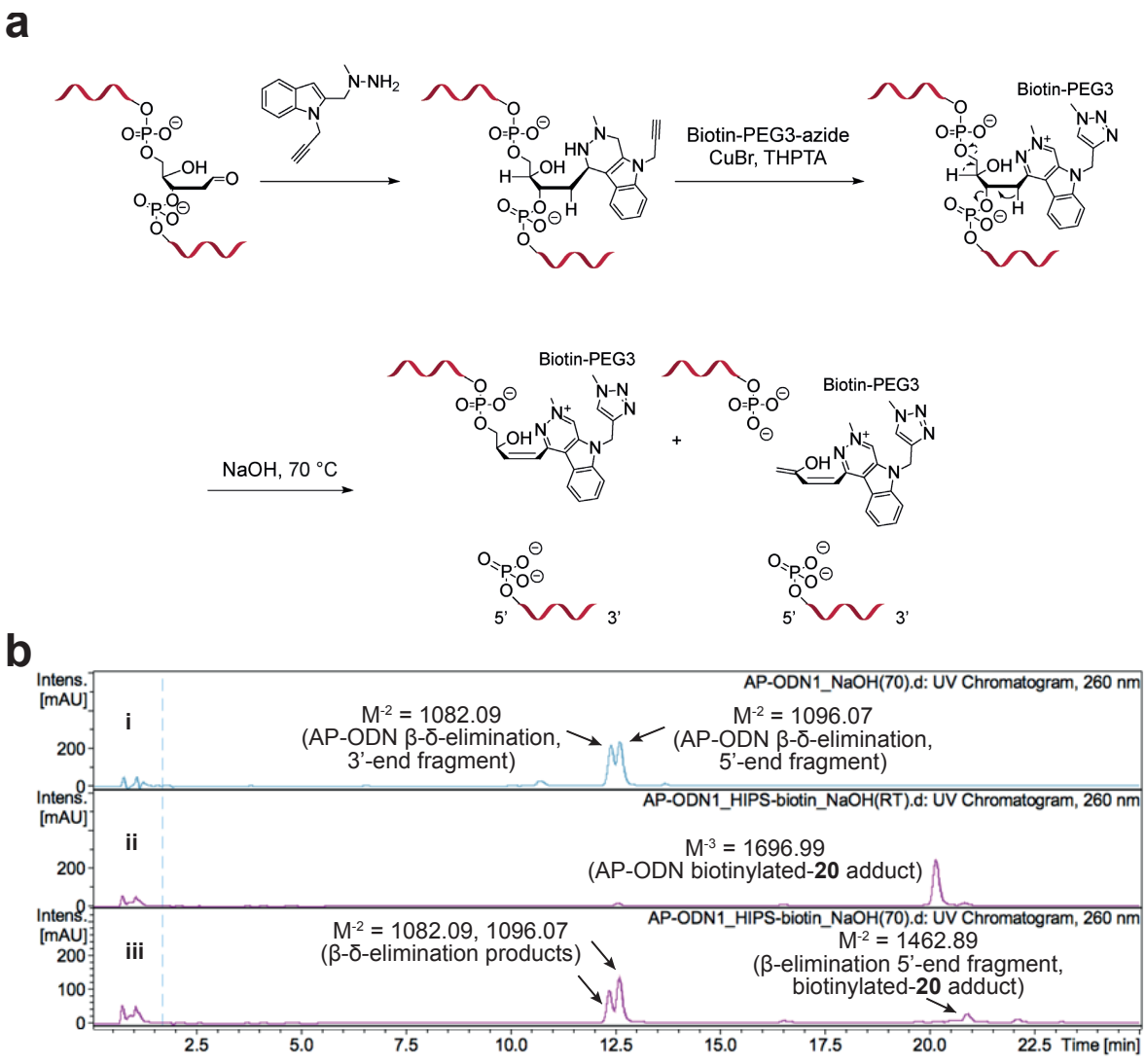


Figure 2.10: Elimination of the DNA backbone at biotinylated HIPS-AP adducts. **a**) Reaction mechanism of HIPS probe **20** with abasic sites, followed by biotinylation *via* the CuAAC reaction. The adduct generated is labile towards β- and β-δ-elimination in the presence of base at 70 °C. **b**) LC-MS UV traces of AP-ODN1. **i**) Unfunctionalised ODN after treatment with sodium hydroxide (100 mM, 70 °C, 15 min), **ii**) oligo after HIPS and biotinylation reaction treated with sodium hydroxide (100 mM, room temperature, 15 min), and **iii**) oligo after HIPS and biotinylation reaction treated with sodium hydroxide (100 mM, 70 °C, 15 min). The minor products between 20.0 and 22.5 min correspond to β -elimination, and the products around 12.5 min to β-δ-elimination.

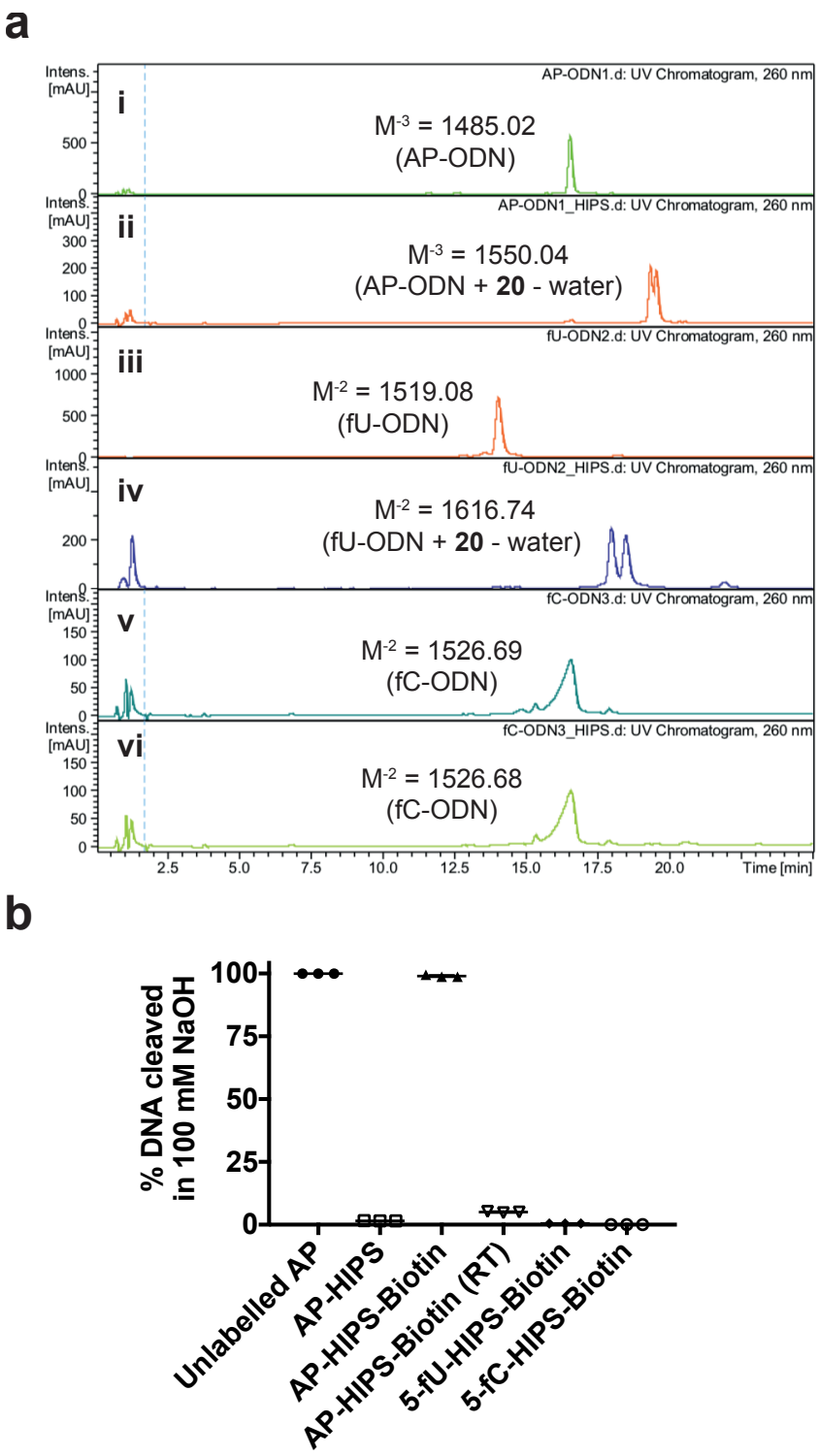


Figure 2.11: Comparison of HIPS probe reactivity and stability on AP-, fU- and fC-ODNs. **a**) LC-MS UV trace of ODNs (10 μM) before and after reaction with HIPS probe **20** (10 mM, pH 7.4, 2 h at room temperature). **i**) and **ii**) Correspond to AP-ODN1, **iii**) and **iv**) to fU-ODN2, **v**) and **vi**) to fC-ODN3. **b**) Stability of ODNs towards DNA cleavage after treatment with 100 mM sodium hydroxide for 15 min. All reactions are carried out at 70 °C, except where labelled 'RT', which was carried out at room temperature. Reactions were followed by LC-MS; % cleavage was determined by integration of the UV signal at 260 nm. Mean values of three independent reactions are plotted.

Comparison of the initial HIPS reaction between **20** and AP-, 5-fU- or 5-fC-containing ODNs showed that similarly to ARP, reactivity at both an abasic site and 5-fU is high in phosphate buffer (pH 7.4), whilst there was minimal adduct formation with 5-fC (**Figure 2.11a**). Therefore, selectivity relative to 5-fC is easily achieved by maintaining the pH of reactions above 7.0. Furthermore, the stability of the HIPS adducts formed with these three aldehydes was found to differ drastically. Upon biotinylation of the HIPS adduct with 5-fU, a product also reduced by 4 Da was obtained. This biotinylated adduct showed no signs of degradation after heating in 100 mM sodium hydroxide (**Figure 2.11b**), in contrast to the same adduct formed with an AP site. Although reactivity of **20** with 5-fC was already low at pH 7.4 and cross-reactivity with this base was minimal, the stability of the trace amounts of adduct expected to form with 5-fC was also explored. To obtain quantitatively tagged 5-fC ODN, the HIPS reaction pH was lowered to 5.0 and incubation was carried out over 24 h. After the CuAAC reaction a pure biotinylated 5-fC species was obtained, which was confirmed to also be stable towards sodium hydroxide treatment. Whilst the oxidised HIPS adduct renders the deoxyribose at an AP site unstable under basic conditions, no similar effect was observed when the adduct forms with a nucleobase. Together, these results show that the combined HIPS biotinylation and subsequent removal by DNA cleavage can be considered a highly chemoselective method for the capture and release of abasic sites. Combined with streptavidin pulldown, this can be utilised as a selective elution in which only abasic site-derived DNA is recovered. A degree of cross-reactivity of the initial HIPS probe on formylpyrimidines is tolerable when these can be distinguished from abasic sites during the recovery of DNA. Furthermore, weakly electrophilic sites in DNA such as the C6 position of pyrimidines also have the potential to react with nucleophiles. Although much less reactive than aldehydes, these sites within the canonical bases are substantially higher in abundance. The two-step process of biotinylation and cleavage is likely to further distinguish abasic sites from these possible sites of cross-reactivity. Although labile, the biotinylated HIPS-AP adduct is stable at room temperature even in the presence of 100 mM sodium hydroxide for 15 min (5% cleavage, **Figure 2.11b**). Therefore, any substantial loss of this adduct prior to streptavidin pulldown and controlled alkaline-cleavage remains unlikely.

The removal of the biotin moiety from abasic sites after streptavidin capture is also beneficial for PCR amplification of enriched DNA. PCR bias against biotinylated DNA adducts is likely to reverse enrichments, resulting ultimately in the underrepresentation of abasic site DNA. As the fragments released from the 3'- end of the AP site upon sodium hydroxide treatment consist of canonical DNA bases and an unmodified 5'-phosphorylated end, these pieces of DNA are particularly suitable for recovery to generate a sequencing library, where

efficient PCR amplification can occur. Many methods that involve the enrichment of DNA fragments through biotin capture incorporate a cleavable site in the linker region[177,203], or otherwise the biotinylated probe may be displaced by a smaller group such as hydroxylamine[70] to minimise the chemical scarring of DNA. Following alkaline-cleavage, the 3'- end fragment is free of chemical modification.

Compared to the other probes that also displayed good reactivity towards abasic sites, DNA cleavage under alkaline conditions was unique to the oxidised HIPS adduct. Although not functionalised, derivatives of **10** and **14** could be synthesised to contain biotin or another intermediate handle to allow chemical pulldown. However, the adducts formed between an abasic site and either **10** or **14** were found to be extremely stable and did not show signs of cleavage (**Figure 2.9**). A small amount of DNA cleavage was observed for the ARP adduct, however conditions harsher than those currently used would be required for quantitative elimination, which may begin to lead to non-specific DNA degradation. Interestingly, another hydrazone adduct, between biotin hydrazide and AP-ODN, also undergoes quantitative elimination when heated in 100 mM sodium hydroxide. However, due to the use of a hydrazide rather than hydrazine nucleophile, the initial condensation is much slower here and low conversion rates (< 50%) were achieved within 2 h at pH 7.4. HIPS probe **20** has the combined advantages of high reactivity and controllable cleavage under DNA compatible conditions, thus providing a way to discern abasic sites from other electrophilic sites in DNA.

2.2.5 Enrichment of abasic sites in synthetic double-stranded DNA

HIPS probe **20** reacts with DNA at abasic sites and can be used to tag these locations with biotin. Although the initial HIPS reaction is not selective for AP sites alone and **20** was also found to react with 5-fU in DNA, the selective removal of the biotin tag can be achieved at abasic sites after streptavidin pulldown, allowing specific recovery of DNA derived solely from abasic sites.

To test whether this combined pulldown-elution strategy can enrich for DNA abasic sites in the presence of DNA comprising the canonical bases as well as 5-fC and 5-fU, longer double-stranded DNA (dsDNA) models that were more representative of genomic DNA were used. DNA sequences between 100-105 bp in length containing either a single AP site, 5-fU or 5-fC on only one strand of the duplex were designed. To represent unmodified, bulk genomic DNA, a GCAT DNA sequence consisting of only canonical bases was also included.

Duplex AP DNA was obtained by primer extension of a uracil-containing ODN, followed by UNG treatment to generate an abasic site (**Figure 2.12**). GCAT DNA was obtained by PCR amplification of the template sequence using unmodified dNTPs, and 5-fU and 5-fC DNA were obtained by a single-primer extension where dTTP or dCTP had been replaced respectively by the formylated analogue. By careful design of the template DNA, it was ensured that a single 5-fC or 5-fU site was installed into the synthesised strand. For 5-fU DNA, a single adenine base was designed in the template or 'reverse' strand after the position of primer annealing, and a similar constraint was in place in the 5-fC template for guanine. To minimise an imbalance in base composition due to this constraint, primer lengths were kept long. This resulted in the position of modification for 5-fC and 5-fU DNA being slightly further along the DNA strand than for AP DNA (position 64, compared to position 28).

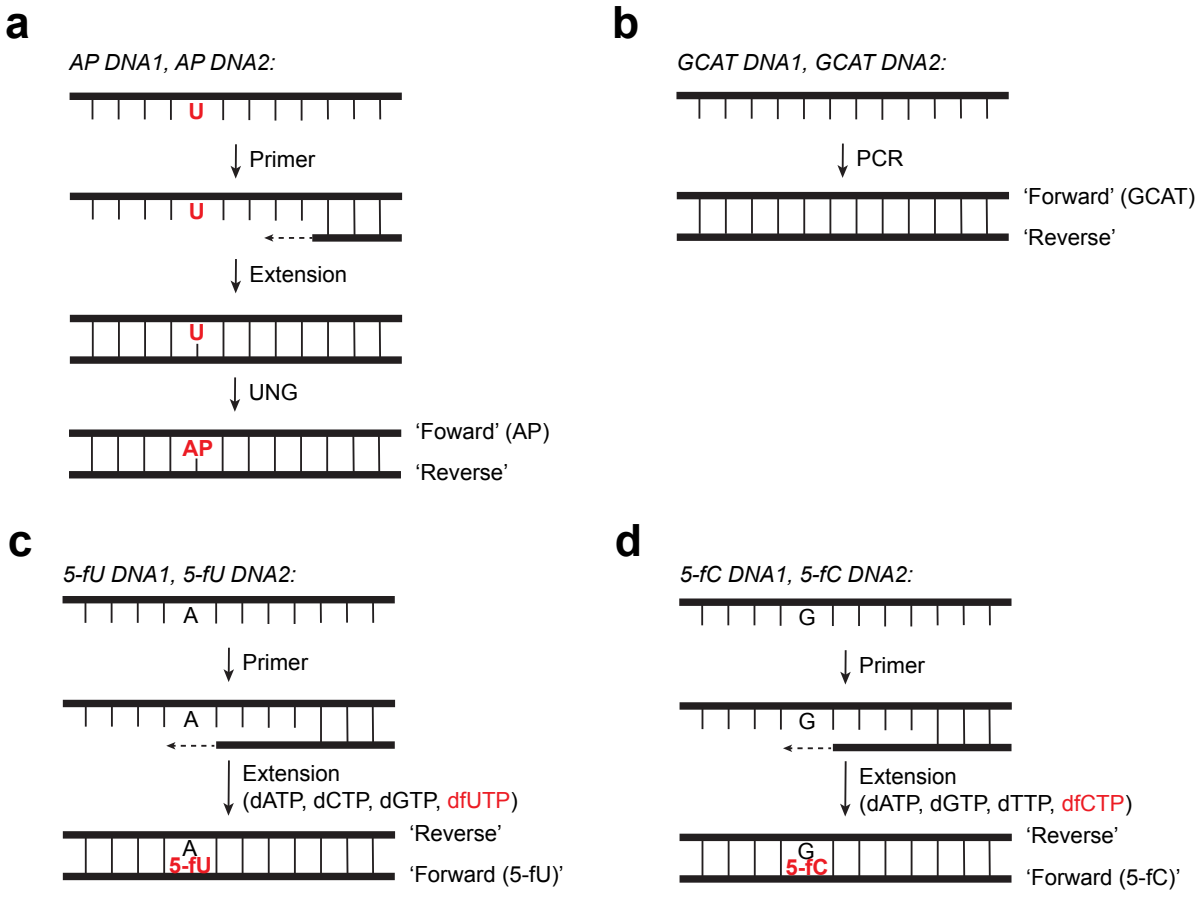


Figure 2.12: Method of obtaining dsDNA models by primer extension. **a**) AP DNA is obtained by primer extension of uracil-containing DNA which is then treated with UNG to reveal the abasic site. **b**) GCAT DNA is obtained by PCR amplification of an unmodified template. **c**) 5-fU DNA is obtained by primer extension of an unmodified template in the presence of dfUTP instead of dTTP. **d**) 5-fC DNA is obtained by primer extension of an unmodified template in the presence of dfCTP instead of dCTP.

The dsDNA models were pooled together and treated with HIPS probe **20** followed by CuAAC-catalysed biotinylation. The purified DNA was then incubated with magnetic streptavidin beads followed by extensive washing to remove unbound DNA. The captured AP DNA was then eluted by elimination, and recovered fragments were quantified by qPCR. This was achieved by DNA amplification using sets of primers specific to each DNA sequence, and the extent of qPCR amplification was compared to that of an input sample that had not undergone streptavidin enrichment (**Figure 2.13**). Primers used for qPCR were designed 3'-to modification sites, so that any potential DNA cleavage would not affect amplification.

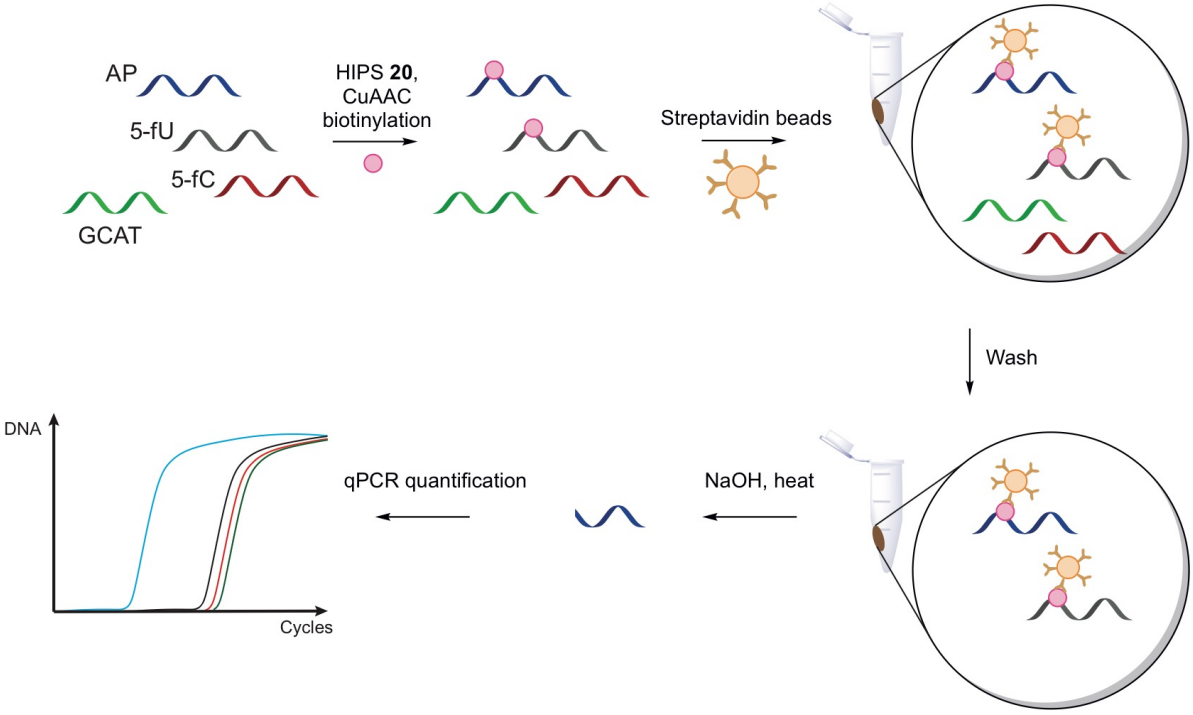


Figure 2.13: Method of quantification of DNA enrichment. dsDNA models were subjected to HIPS reaction followed by biotinylation (pink circles) and enriched on streptavidin beads (orange circles). After washing, only AP DNA is eluted from beads by alkaline-cleavage, and the DNA recovered from beads is quantified by qPCR.

To first ensure successful recovery of the tagged AP sites, the enrichment of AP DNA was compared relative to that of unmodified GCAT DNA. Differing wash and elution conditions were also explored. In the absence of a cleavable linker or adduct, biotinylated DNA is typically released from streptavidin beads by heating in formamide[128]. Under these conditions, around 50-fold enrichment was achieved for AP relative to GCAT DNA (**Figure**

2.14a). As the dsDNA sequences used here were designed with only a single modification on one strand within the duplex, it was possible to remove complementary, unbiotinylated DNA strands by denaturation in aqueous sodium hydroxide. The biotinylated HIPS-AP adduct was confirmed on a shorter ODN model to be largely stable under these denaturation conditions (100 mM sodium hydroxide, room temperature, **Figure 2.11b**) and the inclusion of this denaturation step was found to improve the enrichment further to around 100-fold. This is likely due to further removal of non-specific DNA binding. When DNA was eluted under alkaline-cleavage conditions (100 mM sodium hydroxide, 70 °C, 15 min) after denaturation, the enrichment was moderately improved to an average of 110-fold over three independent replicates. In a control experiment where DNA was not treated with **20** prior to pulldown, no enrichment of AP DNA was observed, demonstrating that the results were not due to qPCR amplification artefacts, for example.

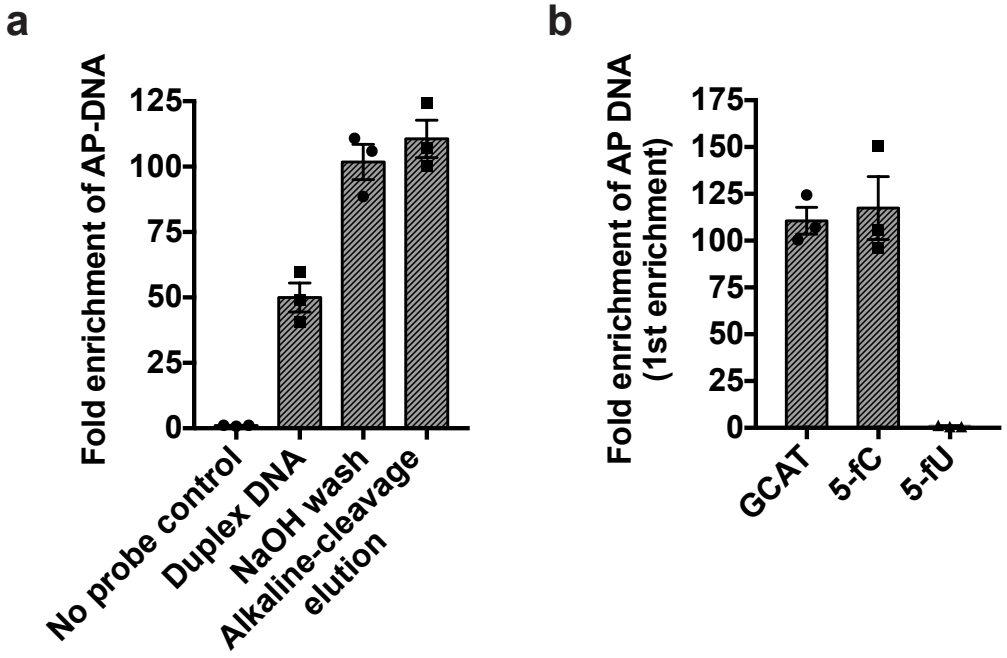


Figure 2.14: Enrichment of AP DNA relative to 5-fC-, 5-fU-, GCAT DNA. Pooled DNA samples were treated with HIPS probe **20** and biotinylated, then incubated with magnetic streptavidin beads. The beads were washed six times with wash buffer (see methods) for all conditions. **a**) Enrichment of AP DNA relative to GCAT DNA under differing wash and elution conditions. For duplex DNA, after 6 washes samples were eluted directly by heating in 95% formamide solution. For sodium hydroxide wash, DNA was further denatured in 100 mM sodium hydroxide at room temperature, then eluted in 95% formamide solution. For alkaline-cleavage elution, after DNA denaturation samples were eluted by heating in 100 mM sodium hydroxide for 15 min at 70 ˚C. **b**) Enrichment of AP DNA relative to sequences after denaturation and alkaline-cleavage elution. DNA recovery was quantified by qPCR by comparison against input samples, and fold enrichment was calculated by comparing recovery of DNA sequences relative to AP DNA. Mean and S.E.M. of three independent reactions are shown.

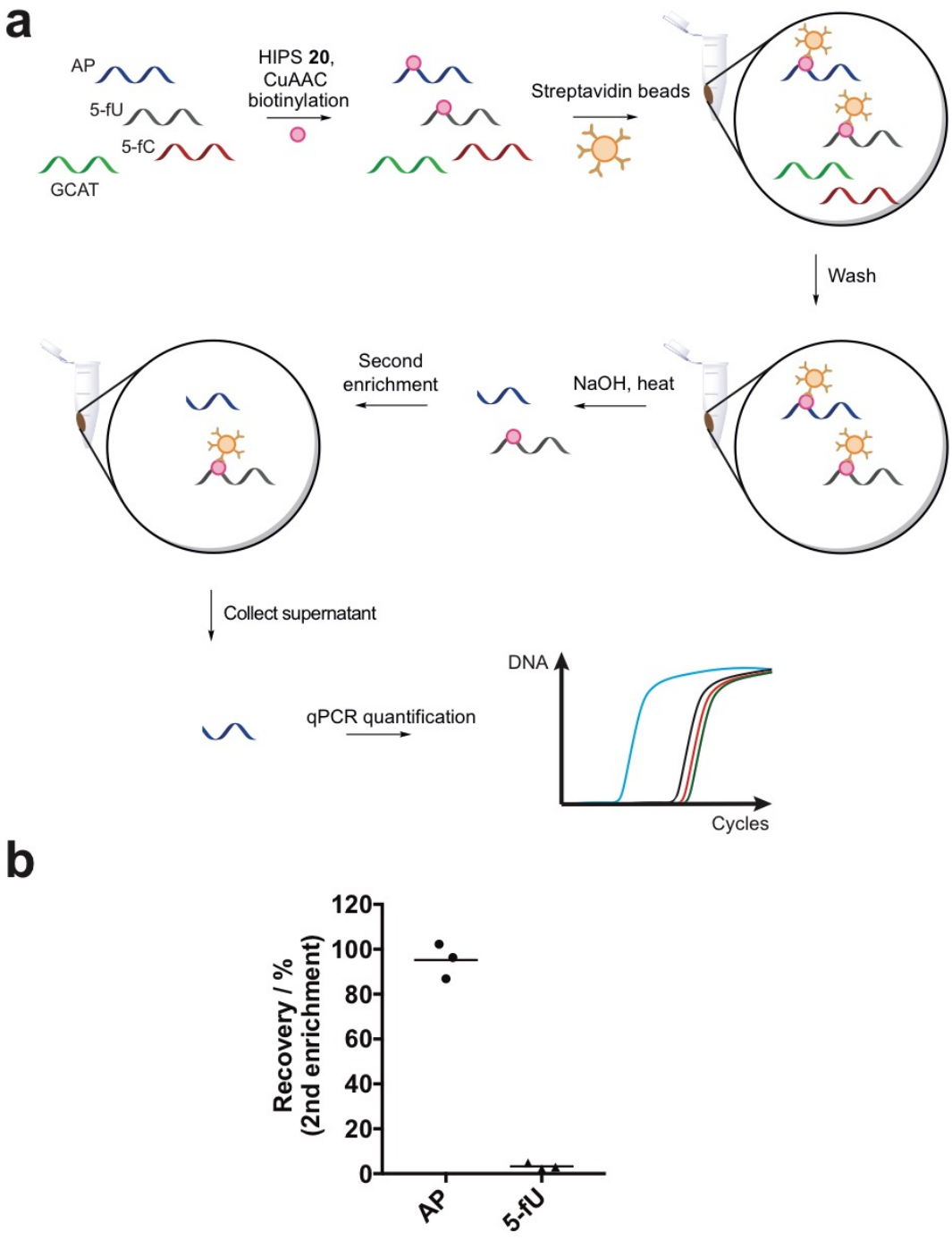


Figure 2.15: Enrichment of DNA using HIPS probe **20**. **a**) DNA is treated with HIPS **20** and biotinylated using the CuAAC reaction. Tagged DNA is isolated on streptavidin beads, and unbound DNA is removed by extensive washing. The captured AP site DNA is eluted by alkaline-cleavage (100 mM sodium hydroxide, 70 ˚C). The eluted DNA is further enriched in a second round using fresh streptavidin beads, where the eluent containing truncated DNA after alkaline-cleavage at AP sites is collected, and non-specific release of DNA such as 5-fU is recaptured on the beads. DNA is then quantified by qPCR to calculate enrichments. **b**) Recovery of AP and 5-fU DNA after second incubation with streptavidin (second enrichment). High recovery of AP DNA is found in the supernatant, whilst 5-fU is removed by binding to beads once again. DNA was quantified by qPCR and represented as a percentage of samples prior to second enrichment. Results from three replicates are shown.

Elution of DNA by alkaline-cleavage also resulted in good enrichment of AP DNA relative to 5-fC DNA (> 100-fold). This suggests that as found by LC-MS for the shorter ODN model, reactivity of probe **20** at 5-fC is minimal and comparable to that at the canonical bases. However, despite DNA cleavage being specific to the HIPS-AP adduct, high recovery of 5-fU DNA was also observed. The adduct between 5-fU and **20** was confirmed to be stable under the cleavage conditions on a shorter DNA model (**Figure 2.11b**) and therefore it is more likely that the biotin-streptavidin interaction is disrupted under the harsh alkaline conditions used. Although a relatively strong interaction, the biotin-streptavidin complex can be broken by heating in water at 70 ˚C[221]. Importantly, however, any 5-fU DNA released in this way remains biotinylated. To recapture these pieces of DNA, the eluent after the first round of enrichment was incubated with a further sample of fresh streptavidin beads to undergo a second round of 'reverse' enrichment. As the AP DNA is no longer biotinylated, these fragments should remain in the supernatant during the 5-fU recapture (**Figure 2.15a**). Quantification of DNA purified from the supernatant after the second treatment with streptavidin beads confirmed that near-quantitative amounts of AP DNA were recovered during this step, whilst less than 5% of 5-fU DNA remained (**Figure 2.15b**). Therefore, AP sites can be selectively enriched after two successive rounds from both unmodified DNA and formylpyrimidines sites. The use of this method of distinguishing AP sites from DNA bases in conjunction with Illumina sequencing was explored next, to enable the sequencing of DNA abasic sites.

2.2.6 Design of snAP-seq library preparation

The selective cleavage at oxidised adducts between abasic sites and **20** generates two DNA fragments that comprise of bases directly adjacent to captured abasic sites. It was reasoned that upon sequencing, one of these fragments can be aligned to the bases directly 3'- to the abasic site, and therefore the original abasic site can be located at single-nucleotide resolution. The mapping method thus developed was termed single-nucleotide AP-sequencing (snAP-seq).

The preparation of sequencing libraries compatible with the Illumina platform using DNA enriched from the HIPS strategy was explored next. Illumina sequencing libraries consist of short DNA inserts (~100s of bp) that are decoded during sequencing, flanked by two adapter sequences. The P5 adapter is on the 5'- end, and P7 on the 3'-end. DNA recovered after two rounds of streptavidin enrichment using the HIPS approach is single-

stranded, due to denaturation during the alkaline wash and cleavage steps. Whilst it is possible to prepare libraries using ssDNA, these methods are generally less convenient and efficient than using dsDNA[222]. As the amount of DNA recovered after streptavidin enrichment with stringent washing is expected to be low, further loss of DNA by inefficient enzymatic activity should be avoided. Therefore, a custom library preparation scheme was designed.

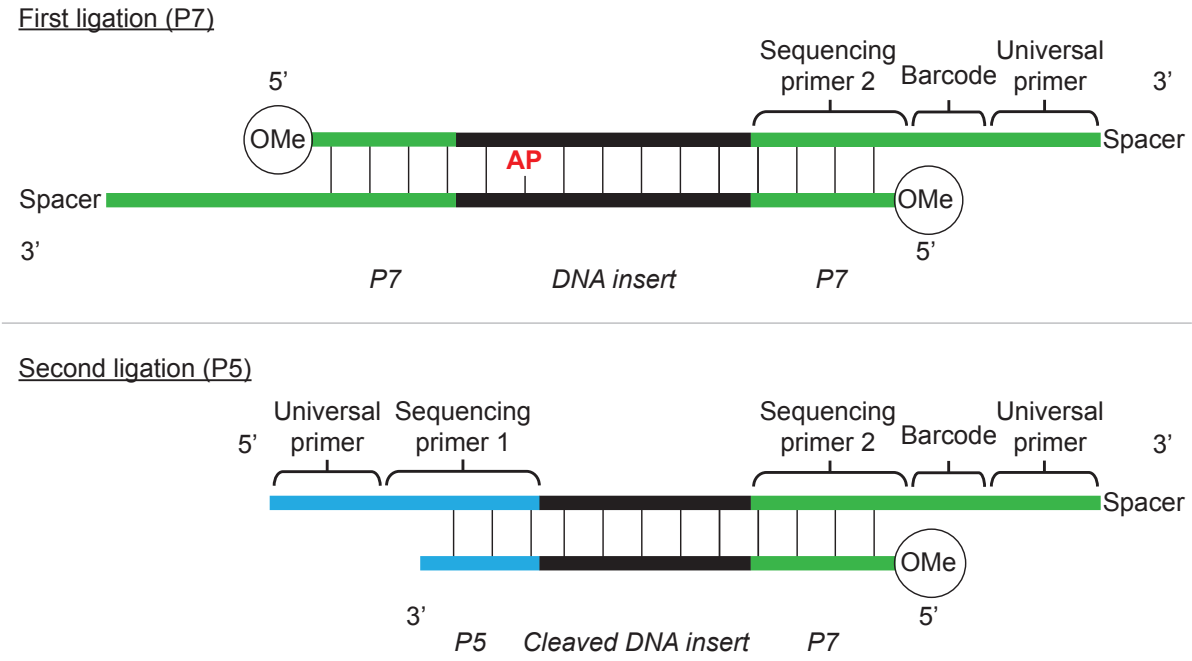


Figure 2.16: Design of sequencing adapters. Top: A custom P7 adapter is introduced first onto duplex DNA fragments at both ends. Bottom: A P5 adapter is then introduced adjacent to positions of abasic sites. Only DNA containing both the P5 and P7 adapters can be amplified and sequenced. Barcodes and regions corresponding to primer sequences used on the flow-cell are labelled.

For the most streamlined library preparation protocols, the two sequencing adapters, P5 and P7, are introduced simultaneously in a single step during the construction of Illumina sequencing libraries. During snAP-seq, the P7 adapter is introduced first in a separate step onto both ends of DNA prior to streptavidin pulldown. This is carried out after sonication, HIPS treatment and biotinylation, whilst DNA is still in the duplex form. As such, traditional methods of adapter ligation using T4 DNA ligase with high efficiency can be used. A custom version of the P7 adapter is used, which contains 5'-OMe and 3'-spacer modifications to avoid self-ligation (**Figure 2.16**). As this adapter is designed to ligate to both ends of the DNA insert, it can also act as a protecting group so that only the free DNA ends generated during alkaline-cleavage are able to accept the P5 adapter later on in the protocol. The sequence

design of this adapter follows that used in Illumina TruSeq technology, with a 6-nucleotide barcode for library multiplexing (**Figure 2.16**). The P7 ligated DNA is then enriched in two rounds as optimised using streptavidin and eluted by alkaline-cleavage. To generate dsDNA from the recovered ssDNA, the P7 adapter is used for priming during a polymerase extension step. Prior to adapter ligation, DNA ends are typically blunted, and the 5'- ends are phosphorylated whilst the 3'- end is adenylated with a dA-overhang to be complementary with a dT-overhang designed within the adapter sequence. As the 5'-end of DNA fragments after alkaline-cleavage are already phosphorylated due to the mechanism of elimination, further treatment of this end is not required before adapter ligation. A polymerase with an inherent dA-tailing property is used to add a dA tail onto the synthesised duplex, after which a P5 adapter is introduced by ligation. The resultant DNA can be amplified by PCR to generate the final library (**Figure 2.17**).

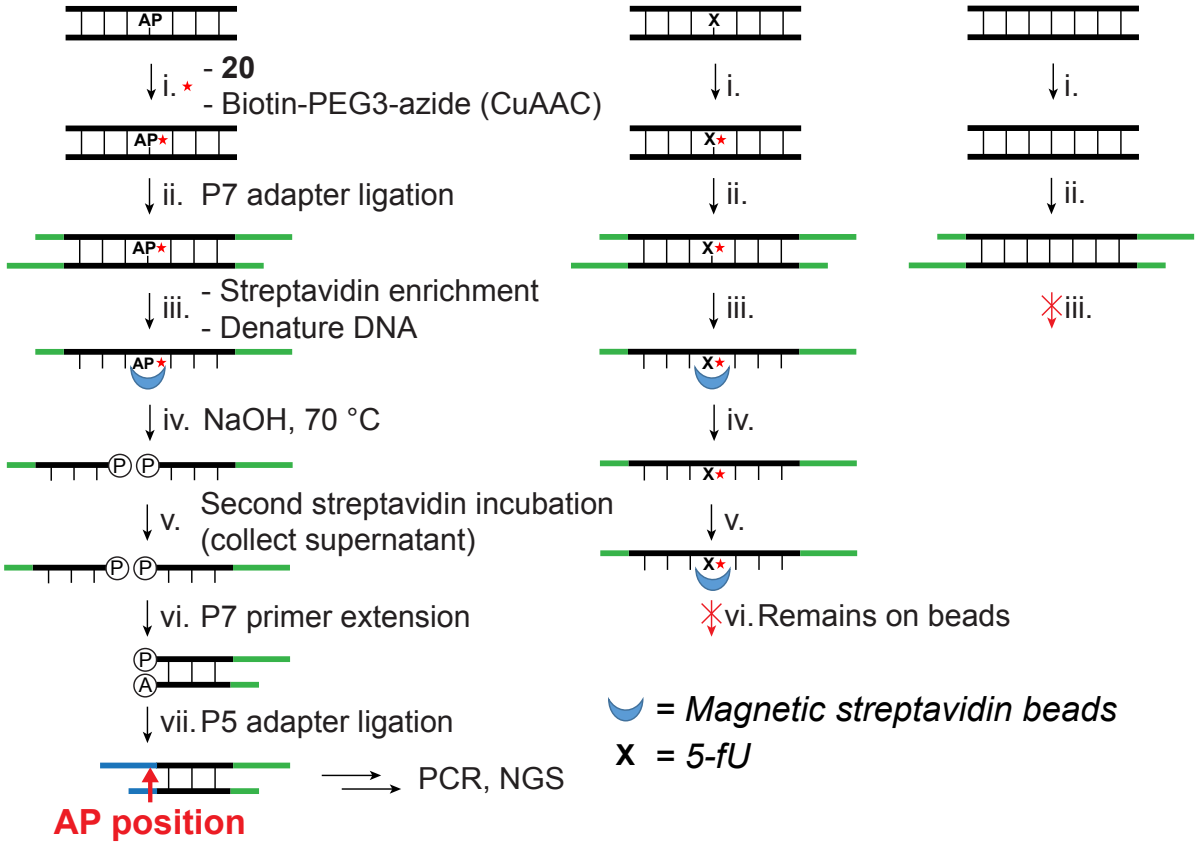


Figure 2.17: DNA library preparation during snAP-seq. After HIPS treatment and biotinylation (i), a P7 adapter is introduced onto duplex DNA by ligation (ii). Samples are then enriched in two rounds on streptavidin beads (iii, v), during which unmodified and formylpyrimidine DNA is lost. AP DNA is eluted and cleaved in a combined step by heating in sodium hydroxide (iv). The P7 adapter is used to prime polymerase synthesis to generate dsDNA (vi). The P5 adapter is then introduced by ligation (vii), after which PCR amplification and sequencing can be carried out.

A phosphatase treatment was also included in the strategy to deactivate any remaining DNA ends that had not received the first P7 adapter on both ends. This approach has been demonstrated to improve ligation specificity in AlkAniline-seq[223], in which RNA abasic sites generated from the hydrolysis of *N*7-methylguanine and *N*3-methylcytosine residues in RNA were enriched. Aniline-catalysed elimination at RNA abasic sites resulted in RNA fragmentation and the 5'-phosphorylated ends that were generated at fragments directly 3'- to RNA abasic sites were used to selectively ligate adapter sequences, whilst unphosphorylated ends were unable to undergo ligation. Even in the absence of prior enrichment by an alternative method such as streptavidin pulldown, this ligation step alone was found to offer sufficient selectivity to map cleavage sites in RNA. The importance of an extensive phosphatase treatment of RNA prior to enrichment was found to be particularly beneficial for enrichment, by removal of non-specific adapter ligation at residual phosphorylated ends.

2.2.7 Sequencing results

The work described in this section was carried out in collaboration with Dr Sergio Martínez Cuesta, Balasubramanian group, who performed the bioinformatical analysis of sequencing data.

An input library was prepared in which equal amounts of the four dsDNA models were pooled together and standard library preparation was carried out using Y-shaped Illumina adapters in a single step. Analysis of the sequencing reads revealed an underrepresentation of AP DNA (**Figure 2.18a**). For analysis, the number of reads aligning to the modified, forward DNA strands were normalised against the unmodified reverse strands to further ensure equal input of the four duplexes. The underrepresentation of the AP strand was expected here, due to the inability of high-fidelity polymerases including the Q5 polymerase used to amplify DNA containing an abasic site. An input library generated with the same DNA with additional HIPS **20** and biotinylation treatment resulted in a similar distribution of reads after sequencing (**Figure 2.18b**), showing that the chemical treatment alone does not alter this pattern.

Application of snAP-seq on the same pool of DNA dramatically changed the distribution of sequencing reads. Over 98% of reads aligned to the AP sequence, with each of the unmodified GCAT, 5-fU and 5-fC sequences accounting for less than 1% of the total reads (**Figure 2.18c**). These results suggest that good enrichment is obtainable by snAP-seq

relative to bulk genomic DNA, as well as known aldehyde modifications that naturally occur in DNA. Furthermore, over 98% of these reads start at the nucleotide immediately adjacent to the AP position (**Figure 2.18d**). Therefore, by considering the position directly 5'- to the start of sequencing read starts, snAP-seq allows the location of abasic sites to be identified at single-nucleotide resolution (**Figure 2.18e**).

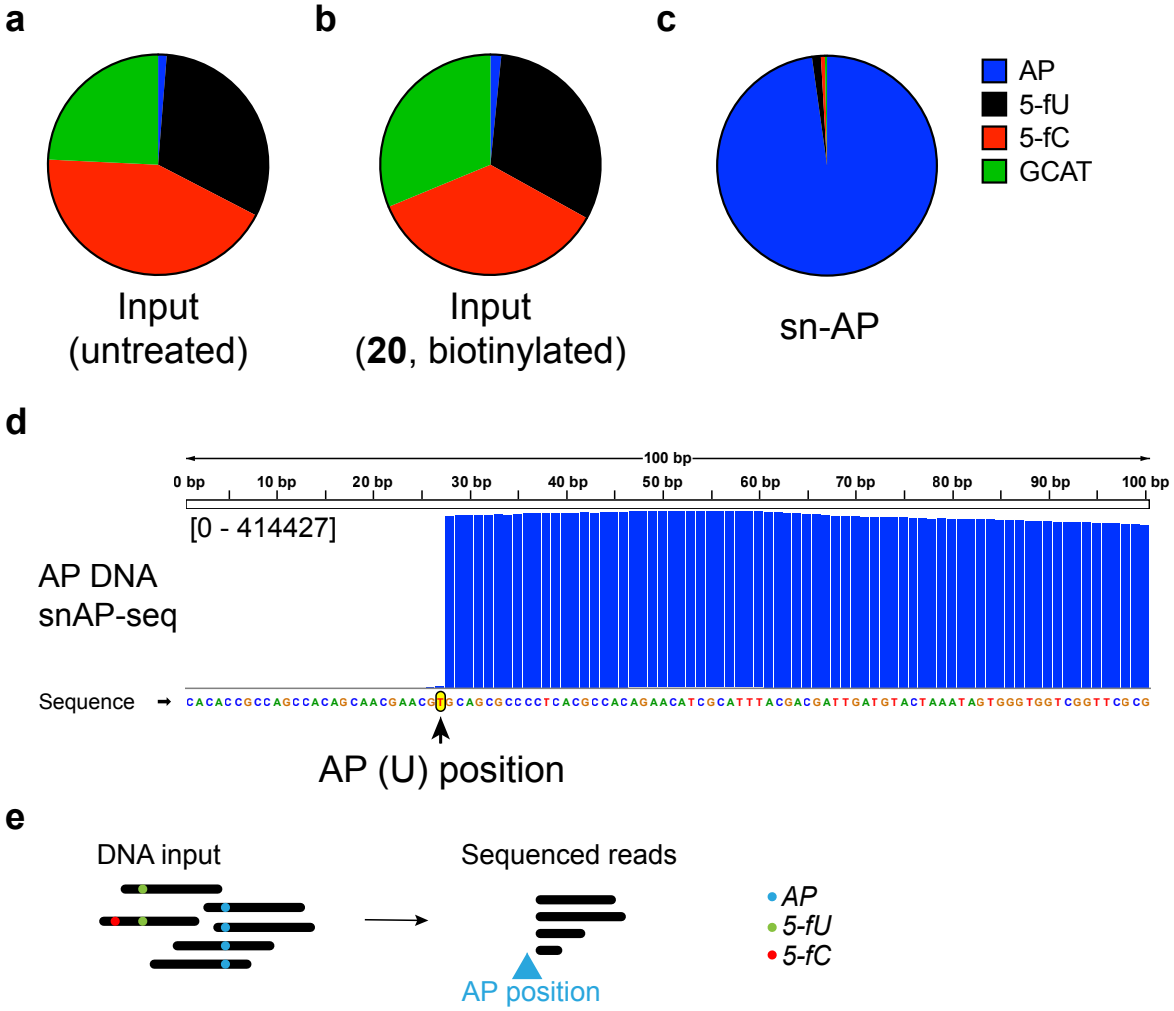


Figure 2.18: Analysis of sequencing reads after different library preparation methods. **a**) Input library without HIPS treatment, **b**) input library after HIPS treatment and CuAAC biotinylation, and **c**) snAP-seq. For 'input' libraries, standard Y-shaped adapters (TruSeq, Illumina) were introduced onto DNA and were sequenced without enrichment. For all three libraries, reads aligning to the forward strand only are represented and were normalised to the number of reads of the corresponding unmodified, reverse strand in the input (untreated) sample. **d**) Sequencing coverage across the 100 nucleotide AP DNA sequence after snAP-seq. A sharp increase is seen after the AP position (labelled T in the forward sequence, representing U in the original template), demonstrating that the AP site can be mapped at single-nucleotide resolution. The total number of raw reads covering each position is shown. **e**) Schematic illustrating the expected sequencing reads after snAP-seq.

To confirm that the observed enrichments using synthetic DNA were not due to potential sequence biases, snAP-seq was applied to a further set of dsDNA models with differing sequence. Furthermore, the initial dsDNA models were designed largely at random and consistency in the DNA length and position of modification was not purposefully designed. To account for these differences, a further set of sequences was designed. AP DNA2, which matched the length and position of modification of 5-fC and 5-fU in the first set of models was generated, along with 5-fC DNA2 and 5-fU DNA2, which matched the length and position of AP DNA1. A further GCAT sequence, GCAT DNA2 was also designed at random. snAP-seq libraries generated from the eight combined oligos were dominated by AP DNA1 and AP DNA2, with underrepresentation of the remaining sequences (**Figure 2.19a**). Therefore, the enrichments seen in snAP-seq were not explainable by sequence biases.

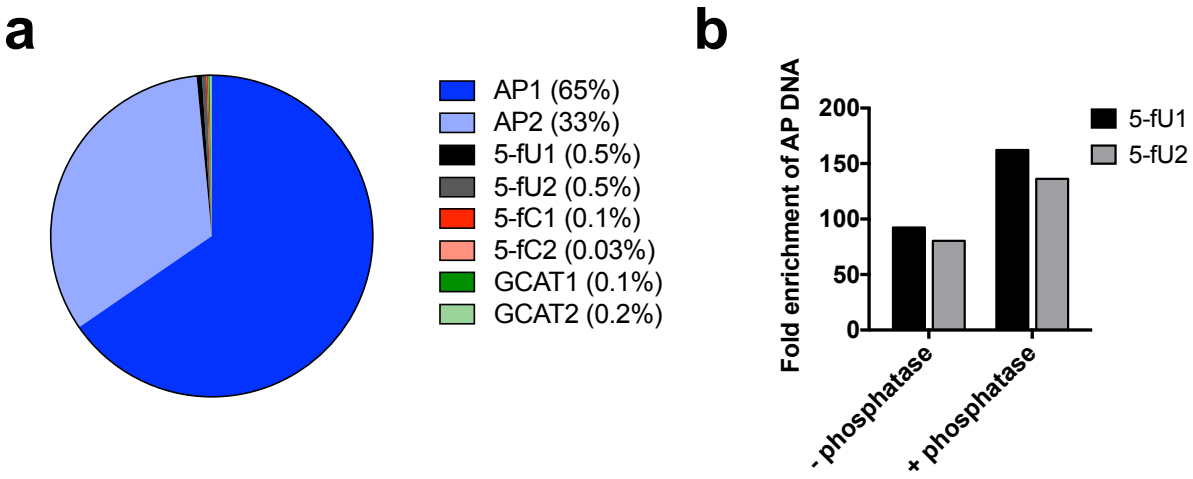


Figure 2.19: Analysis of dsDNA sequences after snAP-seq. **a**) Normalised number of reads aligning to the forward strand of dsDNA models. **b**) Enrichment of AP DNA relative to 5-fU DNA (1 and 2) after snAP-seq, with and without phosphatase treatment (shrimp alkaline) prior to streptavidin pulldown. An improvement in enrichment is seen when 5-fU DNA ends are deactivated by dephosphorylation prior to P5 adapter ligation.

Despite good retention of 5-fU DNA on streptavidin beads during the second round of enrichment (**Figure 2.15**), the overall recovery of 5-fU was still slightly higher than 5-fC or GCAT DNA after the streptavidin enrichment steps, resulting in a weaker AP/5fU enrichment (around 20-fold, compared to > 100-fold AP/GCAT and AP/5-fC). It was found that the alkaline phosphatase treatment of DNA ends prior to enrichment and alkaline-cleavage was important in enhancing the enrichment of AP DNA, by increasing the specificity of P5 adapter ligation. As this step requires a 5'-phosphorylated end, which is generated by alkaline-

cleavage for AP DNA, deactivation of remaining DNA ends by dephosphorylation could further improve selectivity. This was most evident for 5-fU DNA, where a further reduction in recovery could be seen (**Figure 2.19b**).

As all AP DNA models used in this study were obtained through excision of uracil by UNG, a control experiment was carried out to confirm that enrichments observed were due to abasic sites and not residual uracil. UNG treatment was carried out only on AP DNA2, leaving AP DNA1 as a uracil-containing sequence. The two oligos were pooled together with 5-fC, 5-fU and GCAT sequences and snAP-seq was applied. The sequencing results show that only AP DNA2, which had undergone UNG treatment, was enriched and the uracil-bearing AP DNA1 was underrepresented (**Figure 2.20a**). This confirms that the method is specific for abasic sites, and not uracil.

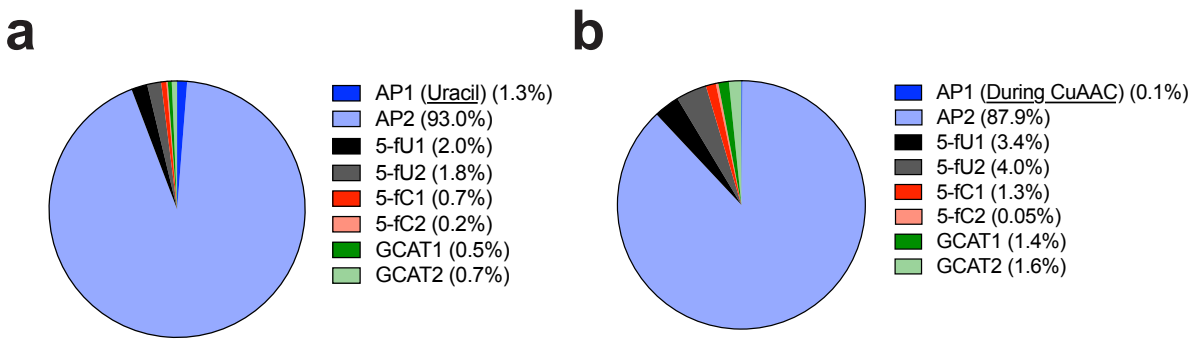


Figure 2.20: Normalised number of reads aligning to the forward strand of sequences in control libraries. **a**) UNG treatment was carried out for AP DNA2 only and not AP DNA1 which remains as a uracil-containing duplex. Enrichment was observed for AP DNA2 and not AP DNA1, confirming that enrichment is specific to abasic sites and not the uracil precursor used to generate this species. **b**) AP DNA1 was added to pooled samples after the treatment of other DNA sequences with HIPS probe **20**, thus representing the AP sites that may be generated in the presence of copper during the CuAAC biotinylation reaction. Enrichment of AP DNA2 and not AP DNA1 was observed, suggesting that any potential oxidative damage generated during the CuAAC reaction is not likely to affect sequencing results.

The treatment of DNA with copper has been suggested to generate oxidative damage through generation of ROS[224], some of which may take the form of abasic sites. A possible way to avoid the use of copper would have been to synthesise an azide-functionalised HIPS probe and introduce the biotin moiety through a strain-promoted alkyne-azide cycloaddition (SPAAC) reaction. However, this approach would not have allowed the oxidation of the HIPS adduct to occur as a side reaction in the presence of copper, which is an essential step in

enabling the subsequent elimination reaction in snAP-seq. Although the copper ligand THPTA has been shown to offer protection of DNA[199] and has been supplemented at the optimum THPTA to copper ratio of 5:1[219] in snAP-seq, the possible effects of this source of AP site artefacts was explored. AP DNA2, along with 5-fC, 5-fU and GCAT sequences were subject to reaction with **20**. After purification of DNA, AP DNA1 was then added during the CuAAC reaction, and therefore represents the AP sites that may be generated in the presence of copper. The remaining steps of snAP-seq were carried out as described and this control library was sequenced. The aligned data showed no enrichment of AP DNA1 (**Figure 2.20b**). Therefore, in the event of AP site generation during the CuAAC step, these sites are unlikely to be captured by snAP-seq. Any artefacts from copper treatment, if present, are unlikely to appear in sequencing data.

2.3 Conclusions and future directions

Using short, synthetic ssODNs, a number of probes were tested for their reactivity at abasic sites. The LC-MS data showed that the widely used probe ARP is susceptible to cross-reactivity at alternative aldehyde sites in addition to abasic sites in DNA. The aldehyde in 5-fU was found to be comparable in reactivity to abasic sites, whilst under certain conditions the aldehyde in 5-fC can also be highly reactive towards nucleophilic probes. As such, a single-step labelling using an aldehyde probe is unlikely to easily distinguish between these three sites and selectively label abasic sites.

Two diamine probes, along with HIPS probe **20** were demonstrated to display good reactivity at DNA abasic sites. In particular, the adduct between **20** and a DNA abasic site was found to undergo an elimination reaction under alkaline-cleavage conditions, thus removing the biotin moiety from one or both of the resultant DNA fragments, which did not occur for the analogous adducts on either 5-fU or 5-fC. This alkaline-cleavage reaction was therefore used as a way to discern tagged abasic sites from tagged formylpyrimidines, in a selective recovery step after capture of the biotin tag. This approach was combined with a library preparation strategy to provide a method of mapping abasic sites at single-nucleotide resolution, termed snAP-seq. The specificity of the chemistry was demonstrated on short ssODNs (≤ 20 bp), whilst the overall snAP-seq approach was demonstrated on longer dsDNA models (100-105 bp). A strong enrichment (> 100-fold) was obtained for AP DNA relative to 5-fC and 5-fU containing DNA and unmodified DNA upon sequencing on the Illumina platform.

Whilst the ability of snAP-seq to enrich for a known abasic site in synthetic DNA has been demonstrated, the applicability of this method to map modifications at unknown locations in genomic DNA will be explored in the following chapters. As snAP-seq is an enrichment-based method, the limit of detection will be dependent on the frequency of modification at specific genomic loci across a population of cells. The sequencing depth and coverage obtained will also likely affect the detection limit, which will also be discussed in the remainder of this thesis.

A further possible extension of the snAP-seq approach is to modify this method to target specific regions of the genome. For some applications, it may be desirable to focus sequencing at chosen locations without the need to explore the entire genome; this may also improve the limit of detection for low frequency modifications. One way to selectively increase sequencing coverage is to pre-select for regions of interest in the genome in a targeted approach. Whilst PCR amplification is a common way to target for parts of the genome, this is not possible when studying modifications to the primary DNA sequence. It may be possible instead in future work to focus on the incorporation of a primer extension step that is selective for targeted regions of the genome within the snAP-seq protocol to selectively increase sequencing coverage. This is further explored in Chapter 5.

Finally, the snAP-seq method may be adapted to study base modifications by pre-treating DNA with a glycosylase. A number of natural and engineered glycosylases are known that can excise a wide range of base modifications[53,225–228]. UNG has already been used to generate abasic sites from synthetic uracil sites in this chapter and glycosylase-coupled snAP-seq will be further discussed in Chapters 3 and 5 of this thesis.

Chapter 3

Mapping thymine modifications in parasite genomes

3.1 Background

In the genomes of a number of species within the Trypanosomatidae family, a small proportion of thymine is replaced by two modified bases, 5-hmU and base J. Organisms within this family include the unicellular parasites *Leishmania major* and *Trypanosoma brucei* that are responsible for causing leishmaniasis and trypanosomiasis, respectively, when transmitted into the human bloodstream. 5-hmU can arise biosynthetically in these organisms from oxidation of thymine by the Fe(II) and 2-oxoglutarate dependent enzymes JBP1 and JBP2[69]. Further glucosylation of 5-hmU by the JGT enzyme generates the hypermodification, base J[72] (**Figure 3.1a**).

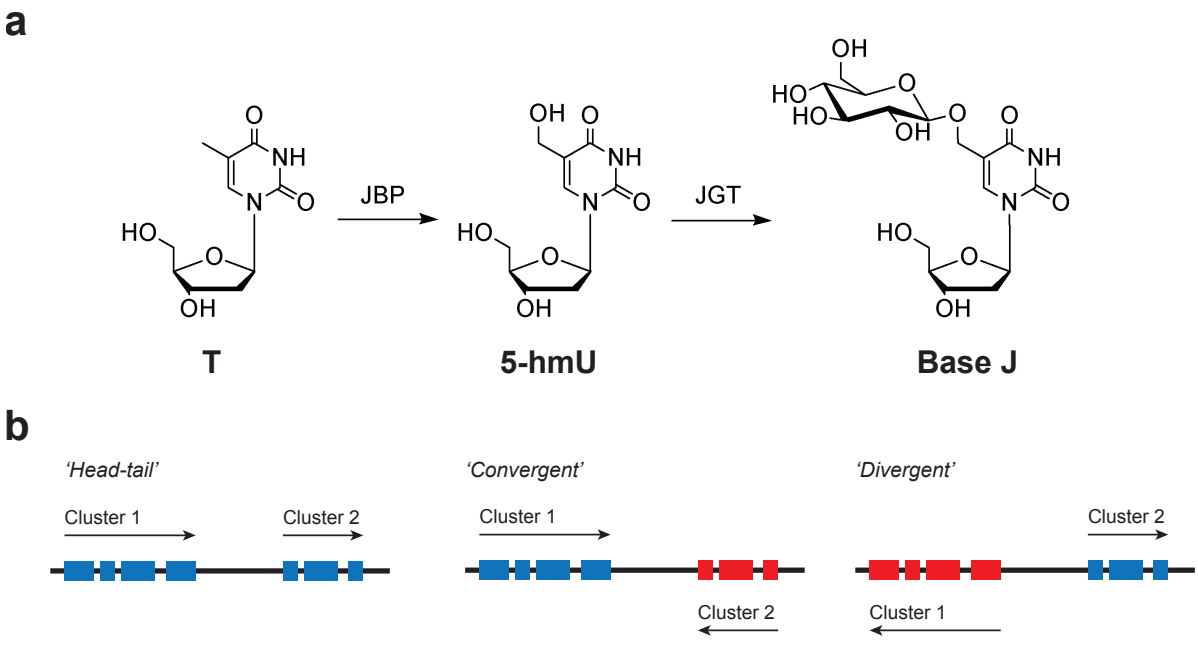


Figure 3.1: Base modifications and features in trypanosomatid genomes. **a**) Biosynthesis of thymine modifications, 5-hmU and base J. **b**) Boundaries between polycistronic clusters can be separated into head-tail and divergent or convergent strand-switch regions. Black arrows represent the direction of transcription.

Global levels of 5-hmU and base J in the *L. major* genome have been measured by LC-MS/MS at around 0.01% and 0.08% of total thymine, respectively[70]. Detection of base J using anti-J antiserum in DIP-seq experiments has revealed that up to 99% of this modification exists within telomeric repeats in *L. major* DNA, whilst the remaining 1% is internal to chromosomes and has been associated with RNA pol II termination sites[75]. The genomic structure of trypanosomatids differs from most eukaryotes in that groups of genes are organised within large polycistronic clusters[229]. Adjacent clusters that are oriented in the same direction with transcription occurring on the same coding DNA strand are separated by head-tail regions. Alternatively, the coding DNA strand may switch between two adjacent clusters, which are then separated by strand-switch regions (SSRs). The SSRs are classified as convergent when the two clusters meet in a tail-to-tail orientation, or divergent when they start from a common region in a head-to-head configuration (**Figure 3.1b**). The 1% of base J that is internal within chromosomes has been found in both types of SSRs in *L. major*. The correct termination of transcription is particularly important in convergent SSRs, to prevent readthrough into the non-coding strand of an adjacent polycistronic cluster. Loss of base J by knockout of the JBP2 enzyme in *Leishmania tarentolae* has been associated with such transcriptional readthrough, where transcription of the non-coding strand was subsequently detected by RNA-seq[75].

Whilst the role of base J has been explored in detail in the *L. major* genome, the significance of the likely precursor to this hypermodification, 5-hmU is less well understood. Recently, two chemical methods capable of independently mapping the locations of 5-hmU and base J sites were reported and applied to study these modifications in *L. major*[70]. In both cases, the modification was oxidised to a reactive aldehyde that was enriched by treatment with a biotinylated hydrazide probe followed by streptavidin pulldown. Oxidation of 5-hmU to 5-fU was achieved by treatment with potassium perruthenate[230], whilst the sugar moiety in base J was oxidised to a dialdehyde using sodium periodate[178]. As these two oxidations are selective without detectable cross-reactivity at the other modification, genomic maps of 5-hmU and base J were generated independently. Over 80% of 5-hmU and base J peaks were overlapping. In concordance with DIP-seq data[75], peaks that were in common between 5-hmU and base J were enriched in SSRs, and peaks unique to base J were enriched in telomeric regions. Interestingly, 19% of the detected 5-hmU peaks were unique to 5-hmU and not enriched in base J. These sites were also enriched in telomeric regions.

Sequencing data obtained by both antibody DIP-seq and chemical pulldown to study T-modifications have been limited to low resolution, typically down to ~200 bp. As such, the

exact location of thymine modifications in the *L. major* genome and the precise sequence contexts in which they occur remain unknown. High-resolution sequencing data can aid in further elucidating the roles that these modifications play. The objective of this chapter was to sequence 5-hmU at single-nucleotide resolution. The human glycosylase SMUG1 excises 5-hmU, in addition to other thymine modifications including 5-fU, uracil and 5-hydroxyuracil (5-hoU), from DNA to generate an abasic site[49]. Therefore, by treating isolated genomic DNA with the SMUG1 enzyme, 5-hmU locations can be marked with an AP site. Using snAP-seq, these sites can then be detected at base-resolution. SMUG1 treatment and snAP-seq are combined in the SMUG1-snAP-seq strategy to map the location of all SMUG1 substrates simultaneously. As 5-hmU is not the only possible substrate of SMUG1, control experiments were also carried out to assess the degree to which SMUG1-sensitive sites were specifically due to 5-hmU.

In contrast to *L. major*, up to 50% of base J in the *T. brucei* genome is internal within chromosomes[231]. The global levels of T modifications change within the *T. brucei* life cycle; whilst base J is absent in the procyclic form, levels increase to around 0.4% of total thymine in the bloodstream form. 5-hmU levels also increase from the procyclic to the bloodstream form, from 0.008% to 0.04% of thymine (measurements by Dr Fumiko Kawasaki, University of Cambridge). Whilst deletion of both JBP enzymes is lethal in *L. major*, thymine modifications do not appear to be essential in *T. brucei* as knockout of both JBP enzymes is possible without obvious phenotypic defects[232]. DIP-seq experiments investigating the genome-wide distribution of base J has revealed an enrichment in both SSRs and telomeric regions[233]. Despite this, the relationship between base J and transcriptional termination is not as clear in this organism. Although base J occurs in SSRs, inhibition of 2-oxoglutarate-dependent enzymes including the JBP family with DMOG does not lead to genome-wide defects in transcription termination and no transcriptional readthrough was detected at SSRs[76]. Instead, a small number of genes were upregulated that were internal to polycistronic clusters. A number of these genes were expressed at low levels in wild-type cells and located towards the ends of clusters. It was therefore suggested that base J may be involved in terminating transcription at locations before the end of polycistronic clusters and plays a more specialised role compared to in *L. major*.

In this chapter, the role of 5-hmU was explored in trypanosomatid genomes by sequencing. A genome-wide map of 5-hmU in the *L. major* genome has previously been reported that has been generated by direct chemical pulldown[70] and was compared to the data obtained here by SMUG1-snAP-seq in the same genome. The SMUG1-snAP-seq data

was also used to further explore the genomic significance of 5-hmU sites in *L. major* at high resolution. In the *T. brucei* genome, base J and possibly therefore 5-hmU have already been determined to have non-essential roles. The distribution of 5-hmU was also explored in this genome, as well as the enzymatic pathways of 5-hmU formation, to further elucidate the role of this base in the *T. brucei* genome.

3.2 Results and discussion

This project was carried out in collaboration with Dr Sergio Martínez Cuesta, Balasubramanian group, who performed the bioinformatical analysis of sequencing data generated in this chapter.

3.2.1 Development of SMUG1-snAP-seq

In Chapter 2, the specificity of snAP-seq was demonstrated on abasic sites within synthetic DNA that were obtained by pre-treatment of uracil-containing DNA with UNG, when compared to a number of control DNA models. To demonstrate that this approach can be extended to other modifications by using different glycosylases, a synthetic ssODN containing a 5-hmU site was treated with SMUG1. To ensure efficient conversion, the enzymatic treatment was carried out over an extended period of 18 h. LC-MS analysis showed successful base excision from the ODN to generate an AP site (**Figure 3.2**). No side products were detected and despite the long incubation time at 37 °C, no degradation of the AP site was observed in the reaction buffer. Further reaction of this product with probe **20** resulted in HIPS ligation and the CuAAC reaction formed the oxidised biotinylated adduct. This oxidised product was confirmed to undergo quantitative fragmentation under alkaline-cleavage conditions (100 mM sodium hydroxide, 70 °C, 15 min), with the LC-MS UV trace matching that of the alkaline-cleavage of the unfunctionalised 5-hmU-excised (abasic) ODN.

5-fU is a known substrate of SMUG1 and therefore is likely to influence SMUG1-snAP-seq results if present in DNA samples. However, 5-fU has not been detected in *L. major* DNA by LC-MS/MS analysis (below limit of detection)[70] and therefore contributions from this base may not be significant when compared to the more abundant 5-hmU sites (0.01% of T). Base J contains a sugar moiety and is present in relative abundance in *L. major* and therefore the

reactivity of **20** on this modification was also investigated. LC-MS analysis of a short ssODN containing three base J sites did not show a change in mass after treatment with **20** or with further CuAAC-catalysed biotinylation from the expected starting ODN (**Figure 3.3**). Consistent with reports[234], SMUG1 was also not found to display any activity on the base J ODN. Together, these results show that base J is not expected to be detected by SMUG1-snAP-seq. The extent to which other substrates of SMUG1 may affect the specificity of SMUG1-snAP-seq for analysing 5-hmU are more difficult to predict and is discussed in further detail below (section 3.2.5).

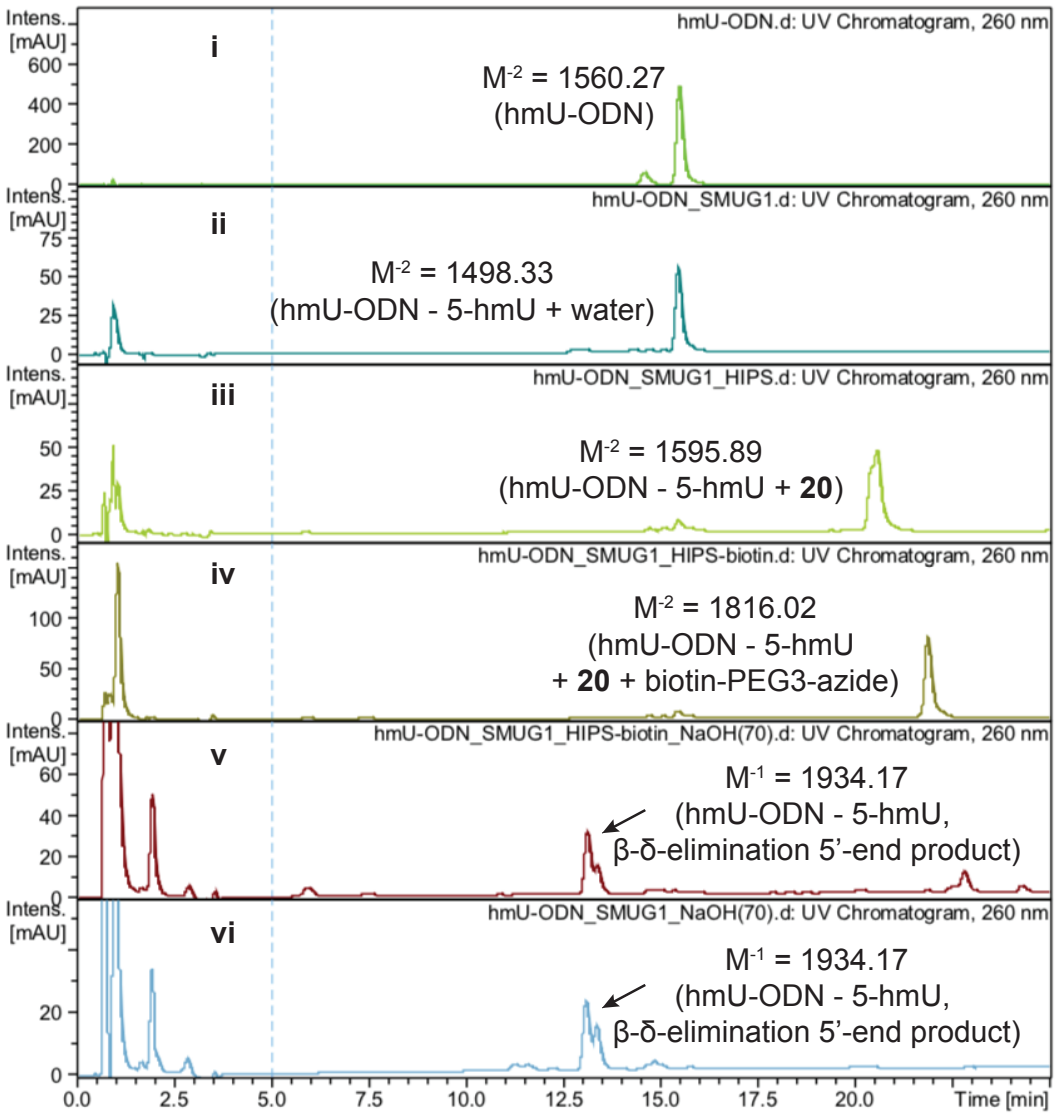


Figure 3.2: LC-MS UV trace of 5-hmU ODN. **i**) untreated starting material, **ii**) after SMUG1 treatment (18 h, 37 °C), **iii**) after SMUG1 and HIPS **20** treatment, **iv**) after SMUG1, HIPS **20** and CuAAC biotinylation treatment, **v**) alkaline-cleavage (100 mM sodium hydroxide, 70 °C, 15 min) of biotinylated ODN after SMUG1 and HIPS treatment, and **vi**) alkaline-cleavage of SMUG1-treated (abasic) ODN. Details of ODNs and corresponding masses are given in section 7.6.

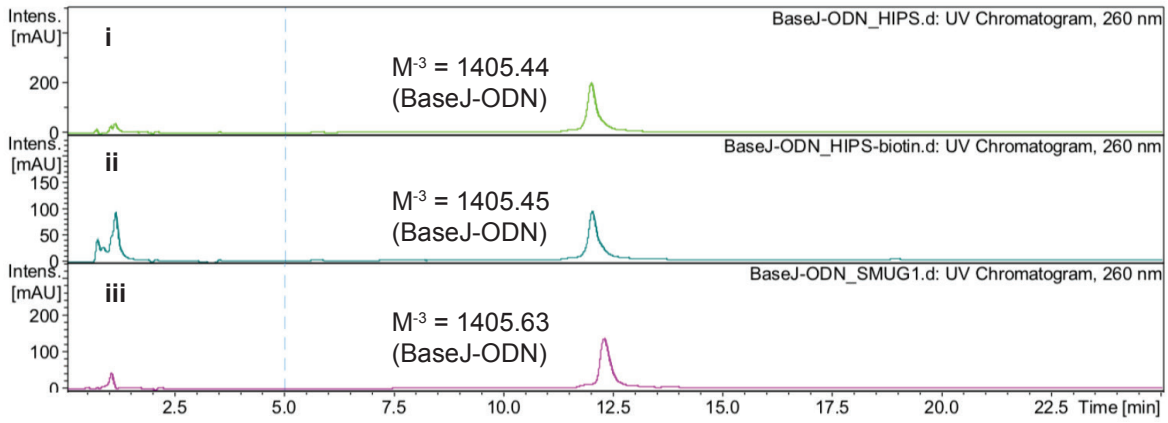


Figure 3.3: LC-MS UV trace of base J ODN. **i**) After treatment with HIPS probe **20**, **ii**) after HIPS **20** and CuAAC biotinylation reactions and **iii**) after SMUG1 treatment (18 h, 37 °C). Calculated mass of the untreated ODN is 4220 (M^{-3}=1406).

During the affinity enrichment of modifications within genomic samples, it is important to include spike-in sequences as positive and negative controls to follow the extent of enrichment. These can be used to check the enrichment within a library prior to sequencing, for example by qPCR quantification of the spike-in sequences. Alternatively, reads aligning to the spike-ins can be retrieved from sequencing data by alignment to the known sequences, to confirm enrichments prior to the detailed analysis of the remaining genomic DNA within the same library. In the absence of these controls, it is difficult to ensure the success of the pulldown and therefore the quality of the library. The synthetic dsDNA models used for method optimisation in Chapter 2 were designed also for this purpose. In addition to the sequences used in Chapter 2, a 5-hmU spike-in was also obtained by primer extension using dhmUTP in place of dTTP, in the same design as for the generation of 5-fU DNA (section 2.2.5 and **Figure 2.12**). To follow the pulldown efficiency, the pool of spike-in DNA was added to DNA from *L. major* and treated with SMUG1 followed by snAP-seq (**Figure 3.4a**).

The sequencing reads obtained after NGS were first aligned to each of the known spike-in sequences. An enrichment was observed for both 5-hmU DNA and 5-fU DNA, in addition to the uracil-derived AP DNA representing endogenous AP sites that are pre-existing in DNA prior to SMUG1 excision, relative to unmodified GCAT DNA (**Figure 3.4c**). In contrast, in the absence of SMUG1 treatment enrichment was observed only for the spike-in representing endogenous AP sites. Therefore, any signals observed by snAP-seq alone can be subtracted from SMUG1-snAP-seq data, to discern SMUG1-generated AP sites from endogenous AP sites that may be present in the genomic samples due to DNA damage.

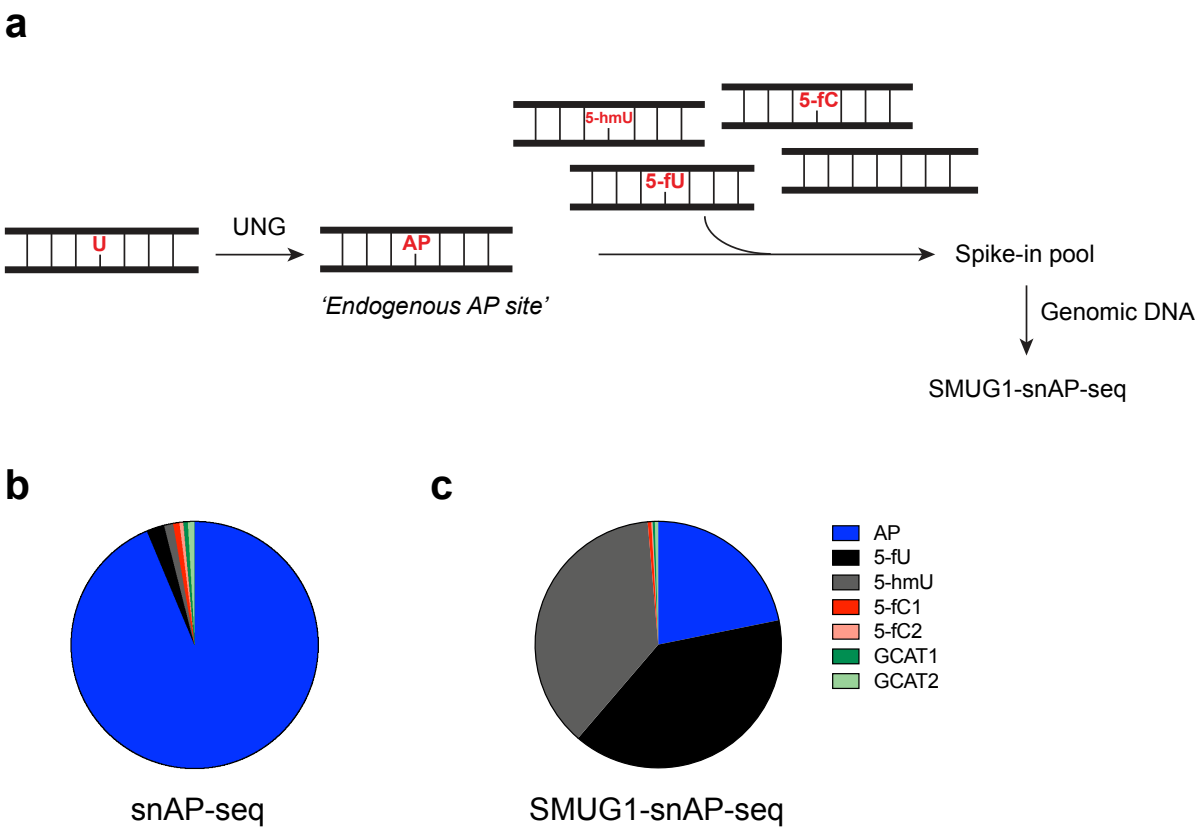


Figure 3.4: Analysis of sequencing reads aligning to synthetic dsDNA models that were added as spike-ins during SMUG1-snAP-seq of genomic DNA. **a**) Workflow for the preparation and use of spike-ins during SMUG1-snAP-seq. **b**) Normalised reads aligned after snAP-seq without SMUG1 treatment, showing enrichment of AP DNA only. **c**) Normalised reads aligned after SMUG1-snAP-seq shows enrichment of AP DNA, along with 5-hmU and 5-fU DNA. The total number of sequencing reads aligning to the forward strand of each DNA model was normalised against the number of reads aligning to the corresponding reverse strands in the input library, which were prepared without chemical pulldown.

3.2.2 Detecting SMUG1-snAP-seq sites in the *Leishmania major* genome

Using the spike-in DNA as an internal control, it was confirmed that enrichment for 5-hmU was achieved using SMUG1-snAP-seq. The *L. major* DNA within the same sequencing libraries was therefore processed for further analysis. Reads were aligned to the reference *L. major* genome (Sanger Institute assembly), for two technical replicates of both SMUG1-snAP-seq and input libraries. Although the detection of abasic sites by snAP-seq only requires single-end sequencing, paired-end sequencing was performed here in order to identify PCR duplicates. For PCR-amplified libraries, duplicates arise when individual sequencing reads correspond to the exact same DNA insert. These are considered artefacts and are removed computationally prior to genomic analysis. Typically, single-end data is

sufficient to identify PCR duplicates as reads that begin at the same genomic location. However, as snAP-seq relies on the enrichment of reads beginning at the same position, paired-end sequencing is required to also obtain information on the read end site, generated during sequencing read 2. Inserts that both start and end at the same position can then be identified as true PCR duplicates.

To identify genomic positions that precede the pile-up of sequencing reads, genome-wide modelling and comparative assessment of read counts was performed at each genomic position. Defining each nucleotide in the reference genome as position '0', the total number of reads that begin at the base adjacent in the 3'- direction, at position '1', was compared between the enriched and input libraries. Each sequencing read is therefore assigned to the location of the expected AP site, immediately adjacent to the fragmentation site. For simplicity, reference genomes are only represented using the sequence of a forward DNA strand. As SMUG1-snAP-seq data retains strand specificity, the reverse strand of the genome can be obtained as the reverse complement of the reference sequence. For each genomic location in the reverse reference strand (position 0), the number of reads beginning at the position '-1' was used.

The analysis of individual sites was represented in a volcano plot (**Figure 3.5**), showing the fold-change in the number of reads obtained in the two SMUG1-snAP-seq replicates compared to the two input libraries at individual genomic loci and the associated FDR values. A positive $\log_2$ fold-change indicates enrichment after SMUG1-snAP-seq, which is expected for captured abasic sites, whilst any loci detected with negative $\log_2$ fold-change indicates a greater pile-up of reads in the input library. As input libraries represent the entire genome without enrichment, these negative enrichments are attributed to sequencing artefacts. A skew is observable in the data, with more sites identified that have positive $\log_2$ fold-change, suggesting that sites have been successfully enriched by SMUG1-snAP-seq. The artefacts that appear as loci with negative $\log_2$ fold-change can be removed by setting a stringent probability threshold when detecting sites. For example, an FDR threshold of 10^{-10} ensures that only sites with positive $\log_2$ fold-change are detected in these libraries without detection of sites with negative $\log_2$ fold-change (**Figure 3.5**). A highly stringent threshold was chosen here to ensure that only high-confidence sites were obtained for further analysis, which results in the loss of some sites with weaker enrichment that are likely to correspond to a weaker accumulation of 5-hmU.

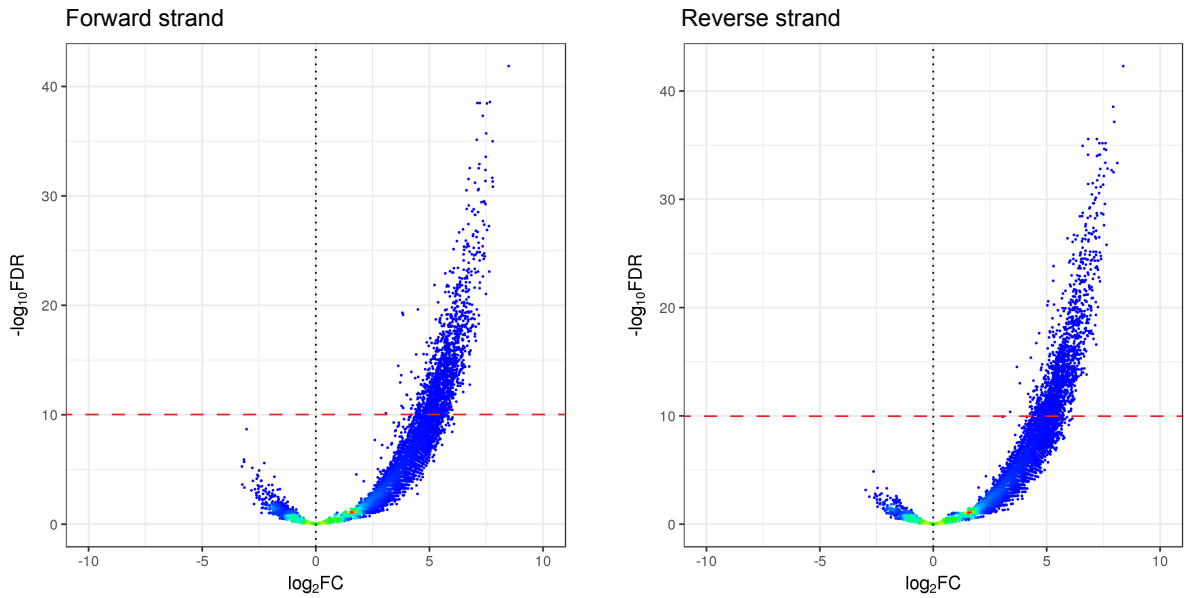


Figure 3.5: Volcano plots representing genomic loci with differential coverage in SMUG1-snAP-seq and input libraries. Reads were split between those that align to the forward and reverse strands based on the reference *L. major* genome; the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (SMUG1-snAP-seq vs. input libraries). Analysis was carried out on two replicates of each library preparation strategy in parallel.

At the stringent threshold of FDR < 10^{-10}, 3,200 high-confidence sites were detected in the SMUG1-snAP-seq data. A visual representation of example sites is depicted in **Figure 3.6**. Interestingly, some of the captured SMUG1-snAP-seq sites are spaced very closely together. Within the snAP-seq protocol, it is expected that multiple abasic sites located on the same individual piece of sonicated DNA will ultimately be detected as a single site. When simultaneous fragmentation occurs at multiple positions along an individual DNA strand during the alkaline-cleavage step, the smallest possible fragment product is expected to be obtained, corresponding to the AP site closest to the 3'- end of the DNA strand. These are then the fragments that are successfully processed by the remaining steps in snAP-seq to generate the final library. Therefore, the ability to detect closely spaced sites may suggest the possibility of non-quantitative efficiency of the enzymatic reaction by SMUG1 or the HIPS reaction when sites are in close proximity. Whilst these steps were individually optimised at high efficiency on synthetic DNA models representing a single site on a short piece of DNA, efficiency may decrease in the presence of nearby reactive sites. Alternatively, it is also possible that at the single-cell level there is a degree of heterogeneity in the distribution of 5-hmU in the *L. major* genome. Closely spaced sites may represent sub-populations in which 5-hmU is located in closely clustered but differing positions. Such sites should be detected simultaneously by snAP-seq enrichment. Due to the lack of understanding of the prevalence

of 5-hmU in a population of *L. major* at the single-nucleotide level, it is difficult to distinguish between the two possible scenarios with confidence.

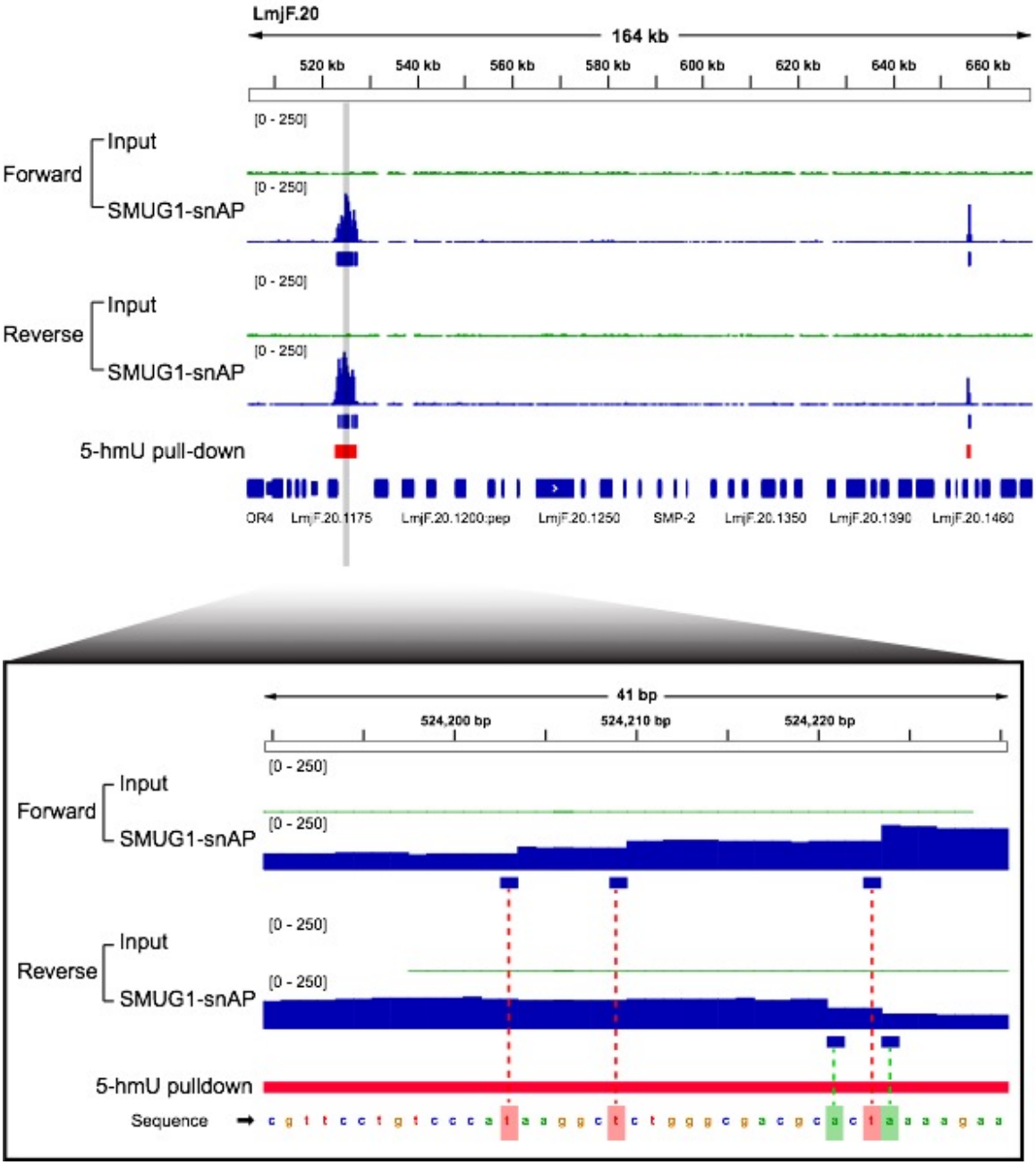


Figure 3.6: Representative genome browser (IGV) view of high-confidence sites detected by SMUG1-snAP-seq. A pile-up of reads is observed in SMUG1-snAP-seq libraries (blue) compared to input libraries (green), which can be used to detect individual nucleotides after alignment to the reference genome sequence. Blue bars represent detected sites (FDR < 10^{-10}), red bars represent loci previously detected as 5-hmU peaks[70].

To assess the degree to which the SMUG1-snAP-seq signal was due to endogenous abasic sites that were present in the DNA samples prior to SMUG1 treatment, two replicates of snAP-seq libraries were generated in the absence of SMUG1 treatment. The data was analysed using the same bioinformatic pipeline as described above. At the chosen threshold of FDR < 10^{-10}, no sites were detected in these control libraries (**Figure 3.7**). These results suggest that endogenous AP sites, if present, are not significantly contributing to the high-confidence SMUG1-snAP-seq sites. This analysis also shows that the non-specific pile-up of reads as artefacts during snAP-seq is low, and that the SMUG1-snAP-seq signal is highly specific to SMUG1 activity.

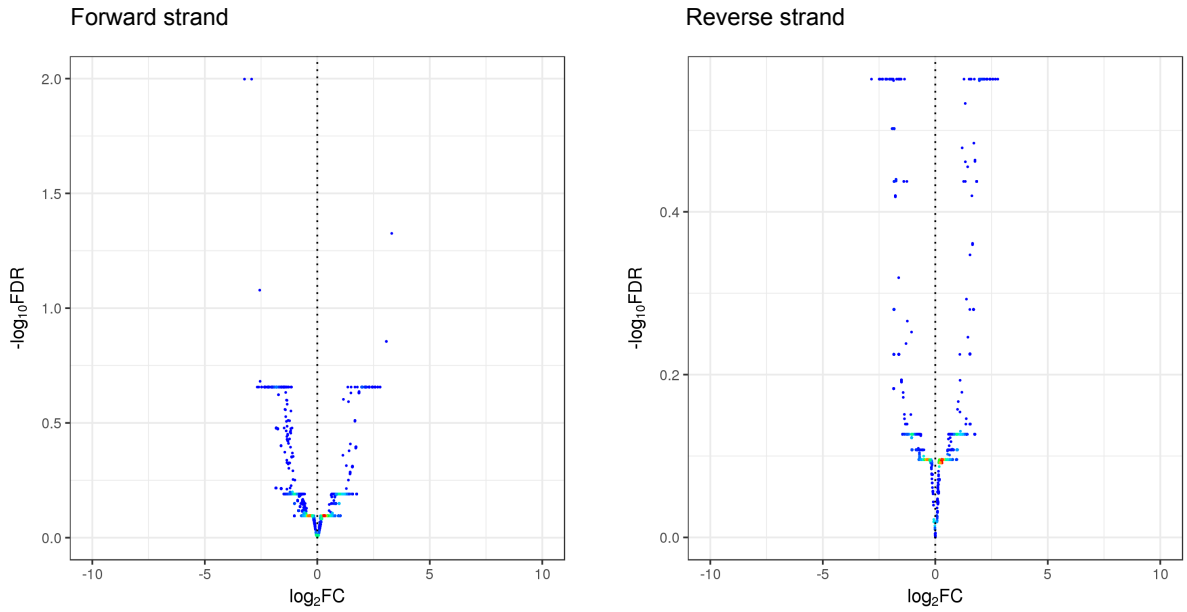


Figure 3.7: Volcano plot representing genomic loci with differential coverage in snAP-seq and input libraries. Reads were split between those aligning to the forward and reverse strands based on the reference *L. major* genome; the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (snAP-seq vs. input libraries). Analysis was carried out on two replicates of each library preparation strategy in parallel.

Comparison of the normalised read counts averaged across the 3,200 high-confidence sites showed that the snAP-seq libraries prepared without SMUG1-treatment closely resembles that of the input library and is largely unchanged from the background, whilst a sharp increase in read depth is seen in the SMUG1-snAP-seq libraries (**Figure 3.8**). This further confirms that there is no clear accumulation of AP sites in the absence of SMUG1 excision in these samples and the sites detected are generated only in the presence of SMUG1.

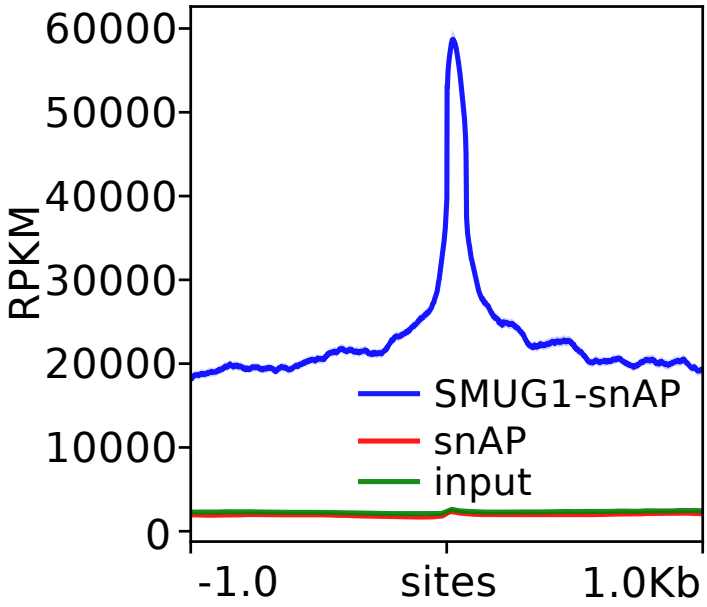


Figure 3.8: Normalised coverage in reads per kilobase million (RPKM) at high-confidence SMUG1-snAP-seq sites (3,200) in SMUG1-snAP-seq, snAP-seq and input libraries.

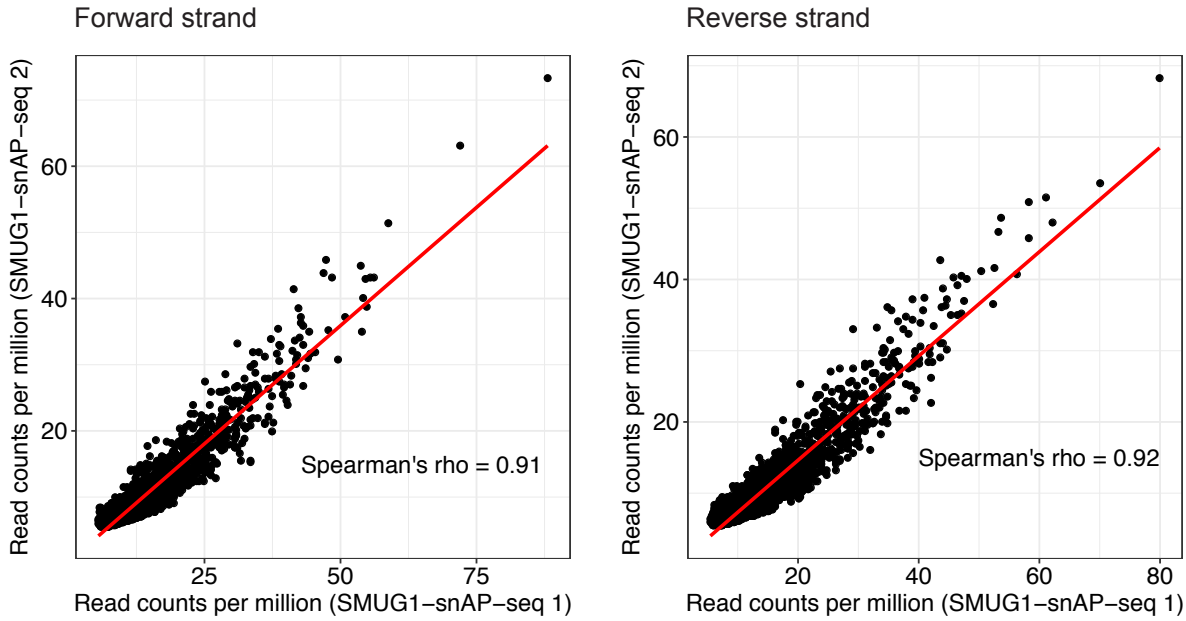


Figure 3.9: Correlation of normalised read counts in two technical replicates of SMUG1-snAP-seq, at the 3,200 consensus sites detected using both datasets in parallel. Analysis was divided into sites that correspond to the forward and reverse strand according to the reference genome.

To also assess the technical variability of SMUG1-snAP-seq libraries, the two technical replicate libraries were separated and the correlation in normalised read counts at the 3,200 consensus sites was calculated. A correlation coefficient of 0.91 and 0.92 for sites in the forward and reverse strand, respectively, was found, indicating that these have good technical reproducibility (**Figure 3.9**).

3.2.3 Genomic analysis of SMUG1-snAP-seq sites

In *L. major*, 5-hmU is expected to be derived biosynthetically from the oxidation of thymine. As such, 5-hmU sites would be expected to align exclusively to thymine positions in the genome. Analysis of the sequence identity of the 3,200 SMUG1-snAP-seq sites showed that 98% of called sites correspond to thymine in the reference genome (**Figure 3.10**). As the forward strand of the reference genome was used here, sites in the reverse strand appear as the complement, adenine.

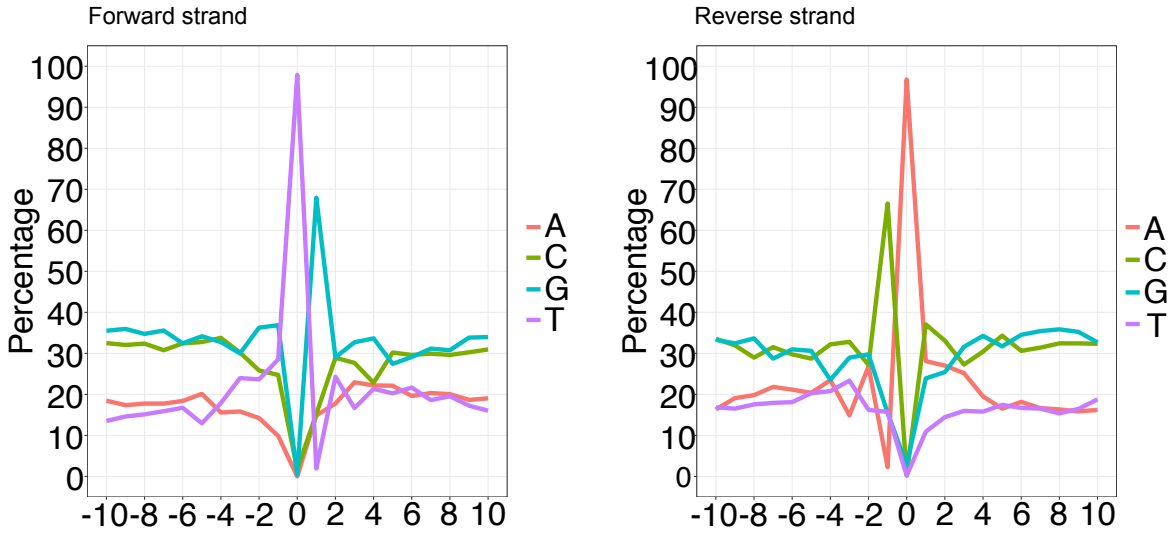


Figure 3.10: Plot of base composition around identified SMUG1-snAP-seq sites in the forward and reverse strands of the reference genome. Position '0' corresponds to the called SMUG1-AP site.

When the '0' position of all reads, corresponding to the nucleotide directly preceding the aligned read start site, were directly analysed without focusing on the called sites, no global enrichment in T could be detected (**Figure 3.11**). These results suggest that although the site-specific accumulation of modifications at distinct thymine locations is strong and

detectable with high-confidence using SMUG1-snAP-seq, the abundance of these SMUG1-sensitive sites on a global level across this genome is relatively low. Global quantities have been measured at 0.01% of T, which equates to around 32 per million dN and therefore 5-hmU remains a relatively low-abundance modification on a global scale.

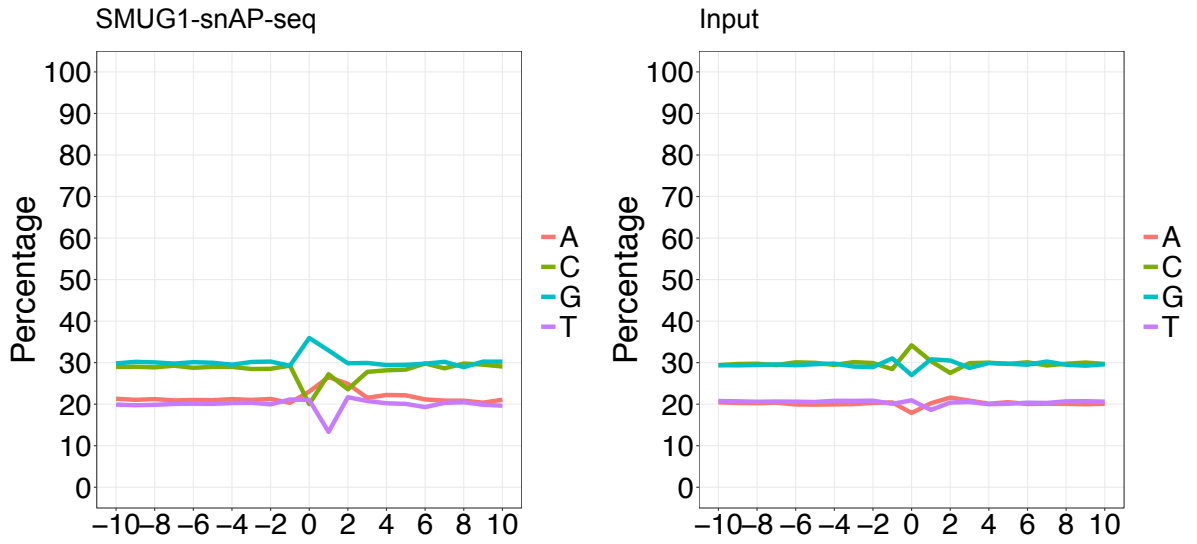


Figure 3.11: Plot of base composition around the start sites of total reads aligned to the *L. major* reference genome, in two combined replicates of SMUG1-snAP-seq and input libraries. Reads begin at position '1', and therefore captured sites are expected at position '0'. Only reads aligning to the forward strand of the reference genome are shown.

The high-confidence SMUG1-snAP-seq sites were also compared to both 5-hmU and base J peaks that have been detected by enrichment sequencing upon oxidation and chemical pulldown[70] (**Figure 3.12**). A high degree of overlap was observed with both datasets, with 96% and 97% of the SMUG1-snAP-seq sites detected within 5-hmU and base J peaks, respectively. Whilst both of these methods of mapping 5-hmU rely on affinity enrichment using a functionalised chemical probe, the chemical approaches are different. 5-hmU chemical pulldown utilises the oxidation of 5-hmU, followed by selective reaction with 5-fU, whilst SMUG1-snAP-seq is dependent on the activity of SMUG1 and abasic site reactivity. The large overlap in loci detected by these two orthogonal methods provides validation for both approaches and increases confidence that 5-hmU is being accurately detected.

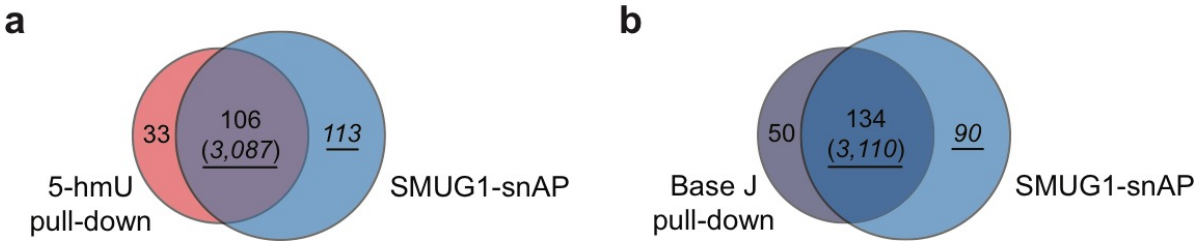


Figure 3.12: Overlap of SMUG1-snAP-seq sites (underlined) with **a**) 5-hmU and **b**) base J chemical pulldown peaks[70].

The high-resolution data generated by SMUG1-snAP-seq also revealed that within the broad stretches that are identified by 5-hmU chemical pulldown, these modifications are densely clustered (**Figure 3.13**). The resolution of peaks obtained by affinity enrichment is generally limited by the length of DNA fragments obtained by sonication, which in the case of 5-hmU chemical pulldown was an average of around 200 bp. The peaks obtained were, however, typically kilobases long. From the SMUG1-snAP-seq results it can be seen that these broad pulldown peaks correspond to many closely spaced individual sites that largely occur on both strands of DNA.

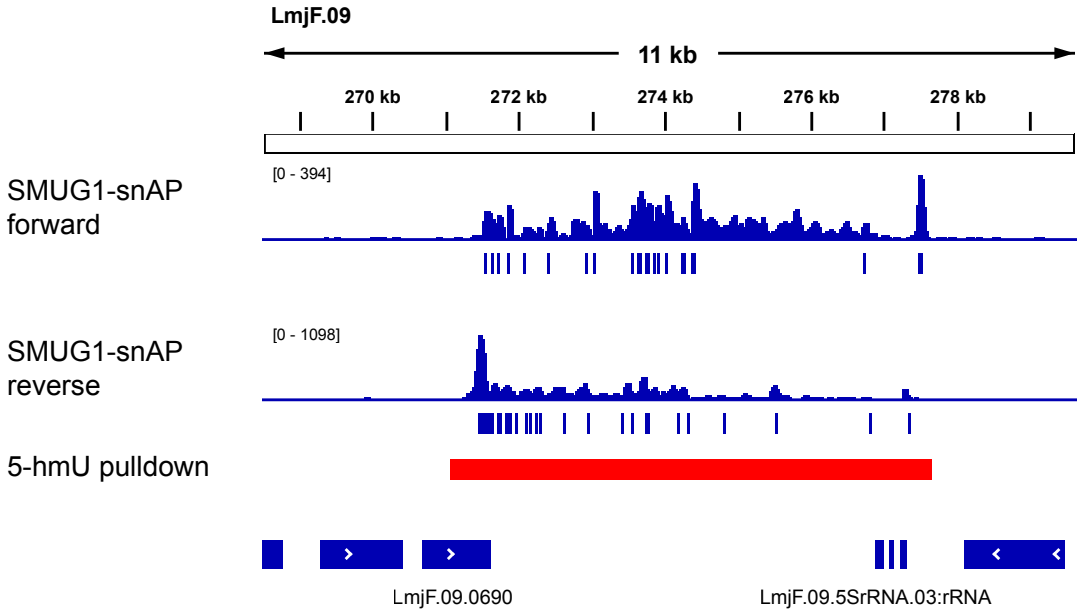


Figure 3.13: Representative IGV browser view of SMUG1-snAP-seq sites (blue). Genomic loci corresponding to 5-hmU peaks obtained by chemical pulldown sequencing are also shown (red). SMUG1-snAP-seq data show that long stretches of 5-hmU identified as peaks from chemical pulldown are densely modified by clusters of individual 5-hmU sites.

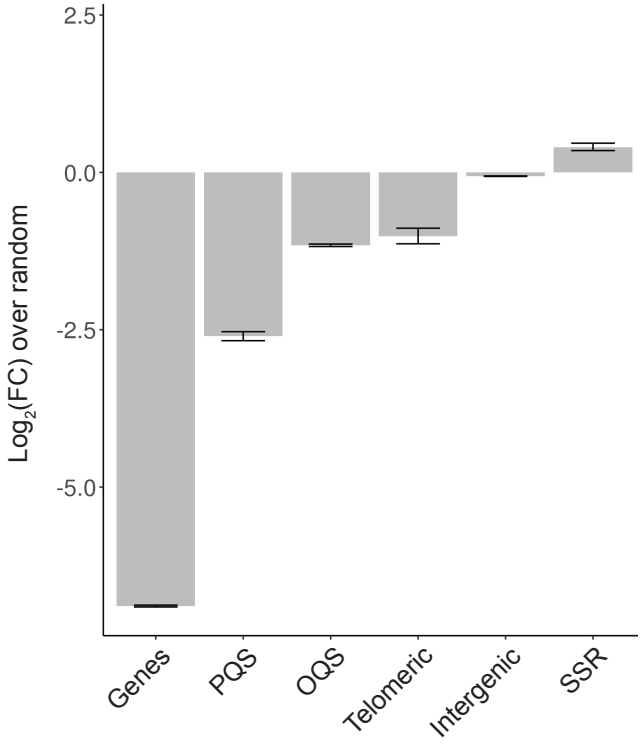


Figure 3.14: Relative enrichment of SMUG1-snAP-seq sites in different genomic regions, expressed as $\log_2$ fold-change when compared to randomised sets of peaks obtained through simulation (*N*=10,000). Error bars represent 95% confidence intervals.

The individual nucleotides detected by SMUG1-snAP-seq were also analysed in the context of genomic features. The genomic location of detected sites was compared to the distribution of sets of sites that were shuffled at random, to calculate an enrichment within selected genomic regions. This analysis showed that SMUG1-senstive sites were enriched within SSRs and depleted within gene bodies, telomeric and intergenic regions (**Figure 3.14**). This supports previous findings that thymine modifications are enriched within SSRs in *L. major*[70,75]. The vast majority of base J occurs in telomeric regions[75], and it may be expected that its precursor, 5-hmU, is also overrepresented in these loci. However, standard NGS is not typically suitable for the analysis of telomeres as sequencing reads from these highly repetitive regions are difficult to align. Telomeric regions were defined here as 5 kilobases at the start and end of each chromosome, which does not necessarily correspond directly to the TTAGGG telomeric repeats that base J and possibly 5-hmU are expected to occur in. Furthermore, highly clustered 5-hmU sites may remain challenging to detect by SMUG1-snAP-seq due to excessive fragmentation of DNA during the site-selective cleavage step. In trypanosomatids, it has also been suggested that G-quadruplexes are associated with the occurrence of base J[235]. To further investigate this relationship, the enrichment of SMUG1-

snAP-seq sites was investigated in both computationally predicted quadruplex sequences (PQS)[236] and locations confirmed experimentally to be able to form quadruplexes by G4-seq (OQS)[190]. No enrichment was observed for SMUG1-senstive sites in either dataset, in contrast to that suggested for base J. Overall, this genomic analysis further confirms the correlation between thymine modifications and SSRs in the *L. major* genome, whilst no enrichment of SMUG1-snAP-seq sites was found within the other features analysed.

3.2.4 Sequence context of SMUG1-snAP-seq sites

In *L. major*, JBP enzymes can oxidise thymine to 5-hmU, which can then be glucosylated to form base J. Whilst the bulky base J modification has been suggested to control transcriptional termination, it is not known whether 5-hmU plays a role beyond a biosynthetic intermediate. The localisation of base J to specific regions in the *L. major* genome may be controlled at either the oxidation or glucosylation level, or a combination of these events. The ability to detect enriched sites by SMUG1-snAP-seq at steady-state suggests that the oxidation of thymine does not occur randomly across the genome and also shows that a detectable amount of 5-hmU is not immediately processed to form base J. To investigate whether there is an underlying sequence preference for these SMUG1-snAP-seq sites, motif analysis was performed using the DREME package[237]. For this analysis, motifs of a definable length are searched for within a set of sequences, relative to either shuffled sequences or a control dataset. For ChIP-seq or affinity enrichment data, the exact position of a modification or feature within a broad peak is not generally known and therefore the direct relationship between identified motifs and modifications is not clear. In contrast, single-nucleotide data allows the direct analysis of the local sequence context of sites. 5-mers centred around the 3,200 sites were used as input for DREME analysis and compared to 5-mers centred around all thymine positions across the *L. major* genome. The most strongly enriched motif was identified as GGTGB, where B is G, C or T (**Figure 3.15**). An enrichment in the TpG dinucleotide was also observed within the top motifs, in addition to TpT where the second T corresponds to 5-hmU. Peak-based motif analysis using chemical pulldown data also found an enrichment of 5-hmU within G-rich stretches and TpT dinucleotides, in agreement with the results here[70]. The enrichment of cytosine rather than thymine at the SMUG1-AP site (motif 3) is likely due to the restriction of the control dataset to thymine loci, leading to statistical significance when analysing the small amounts of non-thymine sites. These may arise from single-nucleotide mutations in these samples from the reference genome, or inaccuracies in the reference.

Motif	E-value
	2.0e-066
	1.9e-033
	1.8e-016
	5.2e-015
	6.7e-011

Figure 3.15: Enriched motifs around high-confidence SMUG1-snAP-seq sites obtained by DREME[237]. The detected 5-hmU site is centred at position '3' in each motif.

In an alternative way to assess for motif preferences, the base composition of the 3,200 sites and flanking regions were averaged and represented as a sequence logo (**Figure 3.16**). An enrichment for G in the '1' position is seen, corresponding again to a TpG motif.

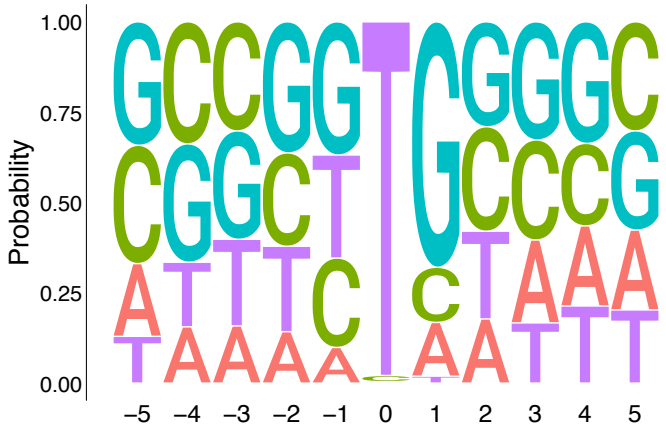


Figure 3.16: Sequence logo plot of nucleotides 5 bases upstream and downstream of high-confidence SMUG1-snAP sites (base '0').

To be confident that the enriched motifs are accurately detected, it is important to assess whether there are significant biases in SMUG1 activity that will ultimately influence the identity of SMUG1-snAP-seq sites. A set of sequencing experiments were designed using a 100-mer ODN that contained a single 5-hmU site on one strand, flanked by a randomised 10-mer on either side (*N*-oligo) (**Figure 3.17**). This double-stranded ODN was directly subjected to sequencing without any enrichment, to first determine the relative amount of each of the canonical bases at each position within the two random 10-mers. SMUG1-snAP-seq was also performed, to assess whether the distribution of bases is changed. As SMUG1-snAP-seq fragments the DNA, causing the 10-mer on the 5'- end of the 5-hmU site to be lost, a further enriched complement library was generated in which DNA was treated with SMUG1 and **20**, then biotinylated and enriched using streptavidin. The complement was recovered by denaturation and sequenced without inducing DNA fragmentation. The identity of the forward strand was then calculated from this complement to assess the composition of bases in both of the 10-mers after enrichment.

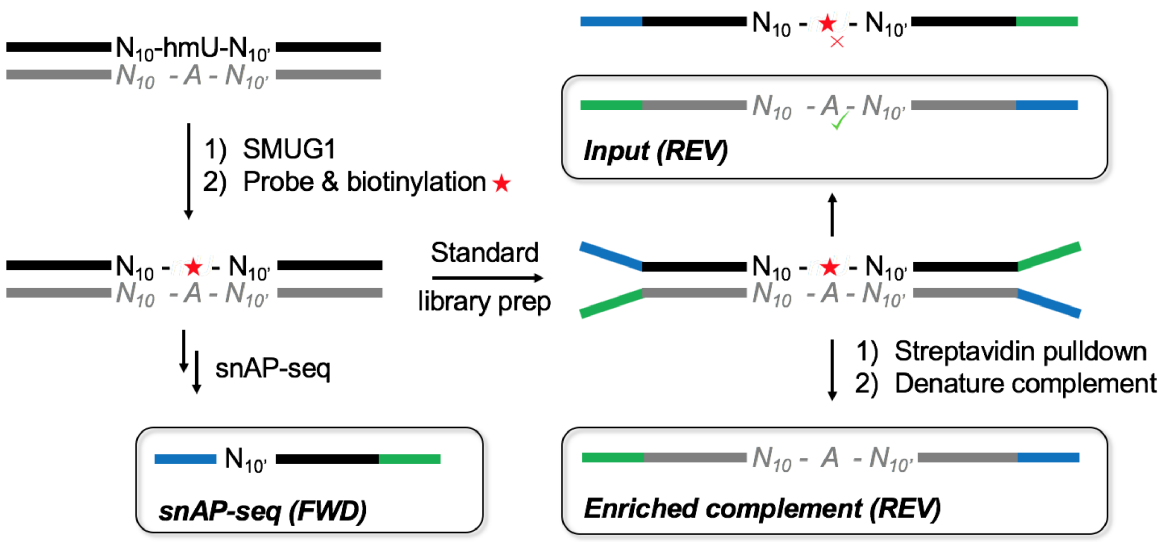


Figure 3.17: Workflow of generating libraries to assess the sequence context preference of the SMUG1 enzyme. A synthetic ODN (*N*-oligo) was used that contained a 5-hmU site flanked by randomised 10-mers (N_{10} and $N_{10'}$). This dsDNA was subjected to SMUG1-snAP-seq, input library preparation and also library preparation of the complement strand following streptavidin pulldown (enriched complement).

The randomised 10-mers are expected to contain roughly equal proportions of the four bases at all positions. The input library without enrichment reveals the basal levels of each nucleotide, where for example a slight overrepresentation of guanine was consistently

observed across all 20 positions, accounting for more than the expected 25% of bases (**Figure 3.18a**). For the enriched complement, a small enrichment in TpG is observed at the 5-hmU position, which increases slightly in the SMUG1-snAP-seq library. This suggests that some technical bias is introduced by the sequencing method that favours the TpG motif. To assess the extent of this bias, the relative enrichment in TpG was compared between the *N*-oligo and *L. major* libraries. The observed genomic enrichment was still significant when compared to the potential technical bias introduced by SMUG1-snAP-seq ($p < 0.05$), suggesting that some of the observed enrichment is expected to be biological (**Figure 3.18b**).

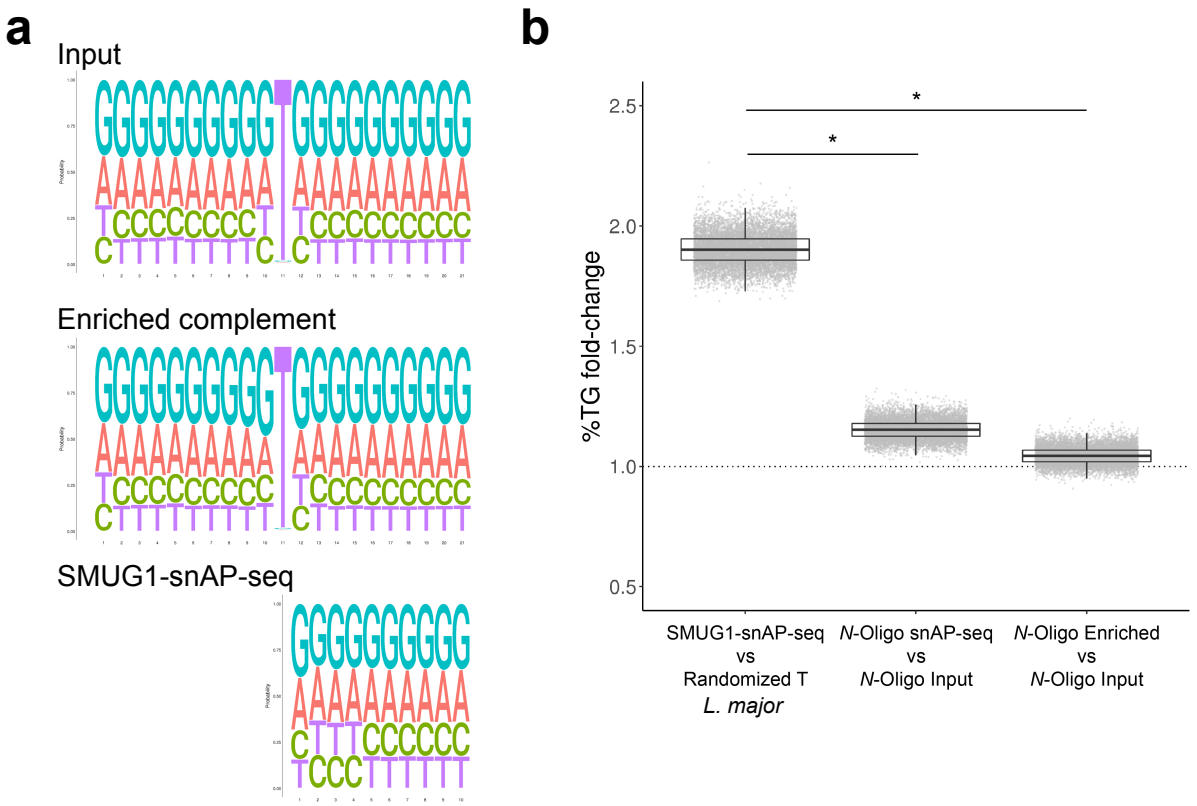


Figure 3.18: Analysis of sequencing results of 5-hmU randomised ODN (*N*-oligo). **a**) Sequence logos of 5-hmU and flanking region after input, enriched complement and SMUG1-snAP-seq library preparation. A weak enrichment for the GpT and TpG motifs is observed after enrichment. **b**) Comparison of TpG enrichment in SMUG1-snAP-seq sites identified in *L. major*, with technical bias detected from *N*-oligo experiment (*$p < 0.05$).

Overall, these results suggest that within the *L. major* genome, the accumulation of 5-hmU is non-random both in terms of genomic location and the sequence context of the thymine positions at which the modification is installed. This may reflect an inherent preference during the generation of these bases, or suggest that 5-hmU is better tolerated in these specific contexts without further processing to form base J.

3.2.5 Specificity of SMUG1-snAP-seq

The SMUG1 enzyme has a number of substrates that can be excised from DNA, including uracil, 5-fU and 5-hydroxyuracil (5-hoU), in addition to 5-hmU. In contrast, UNG displays exceptional specificity for uracil in DNA. Mechanistically, this is due to the exclusion of steric bulk at the C5 pyrimidine position in the active site of UNG such that only a small substituent such as hydrogen is tolerated[238]. The only other known substrate of UNG, 5-fluorouracil, is not generally considered to be a naturally-occurring DNA modification. To assess the contribution specifically of uracil to the 3,200 sites detected by SMUG1-snAP-seq, SMUG1 was replaced by UNG in the protocol to carry out UNG-snAP-seq on *L. major* DNA. The enzymatic treatment conditions were kept consistent at 18 h, which was confirmed by LC-MS to be sufficient to excise uracil from a ssODN. Analysis of the sequencing data showed that no sites were detected above the chosen threshold (FDR < 10^{-10}) (**Figure 3.19**). These results suggest that uracil is not expected to contribute significantly to the sites detected using SMUG1.

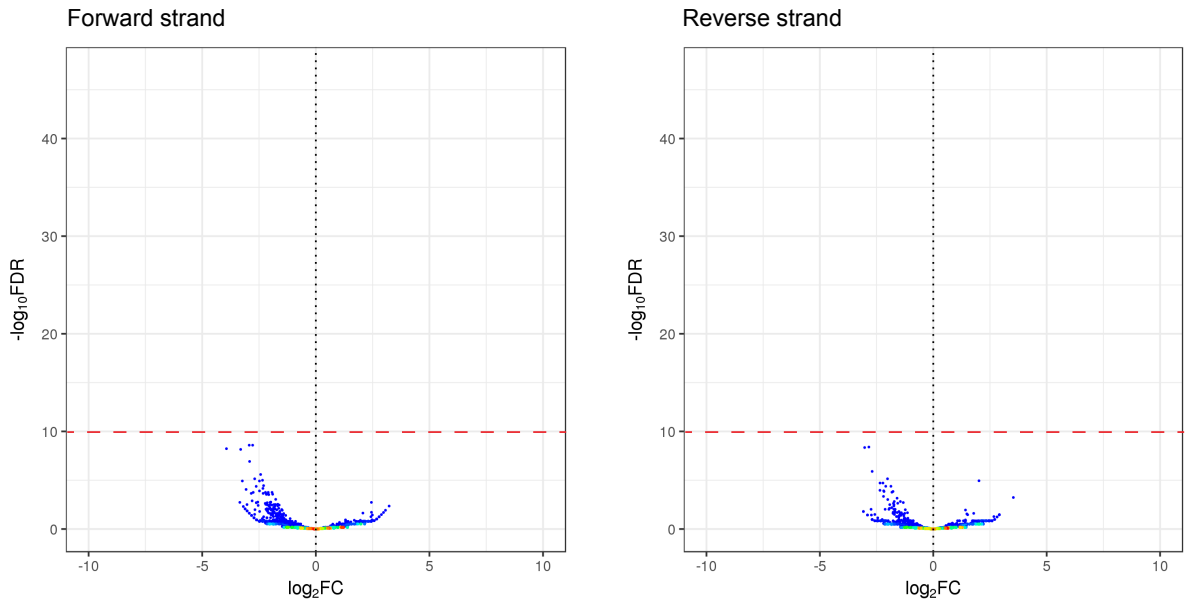


Figure 3.19: Volcano plot representing genomic loci with differential coverage in UNG-snAP-seq and input libraries. Reads were split between those aligning to the forward and reverse strands based on the reference *L. major* genome; the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (UNG-snAP-seq vs. input libraries). Analysis was carried out on two replicates of each library preparation strategy in parallel.

5-hoU, another substrate of SMUG1, is not as well studied as other thymine modifications in trypanosomatid genomes. This modification has not been reported to occur in *L. major* and furthermore, reports have suggested that 5-hoU is a product of oxidative deamination from cytosine[239,240]. Such sites are therefore largely expected to align to cytosine in the reference genome. As over 98% of SMUG1-snAP-seq sites correspond to thymine loci, 5-hoU is not expected to contribute to a large proportion of the detected sites.

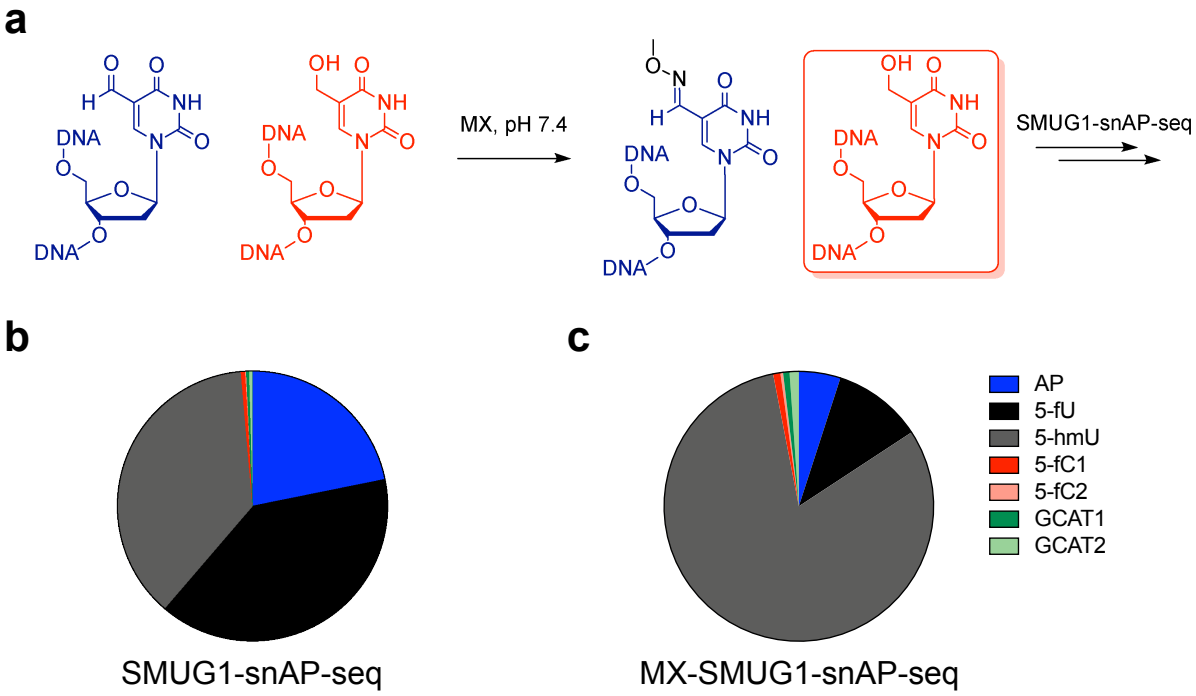


Figure 3.20: Methoxyamine (MX) blocking of reactive aldehyde modifications prior to SMUG1-snAP-seq. **a**) Treatment of DNA with methoxyamine prior to SMUG1-snAP-seq results in reaction at 5-fU to generate an oxime adduct, leading to an increased specificity towards 5-hmU during subsequent SMUG1-snAP-seq. **b**) Sequencing reads aligning to each synthetic spike-in model after SMUG1-snAP-seq. **c**) Sequencing reads aligning to each synthetic spike-in model after methoxyamine treatment followed by SMUG1-snAP-seq (MX-SMUG1-snAP-seq). Only reads aligning to the forward strand are shown. The number of reads were normalised against the number of reads aligning to the corresponding reverse strand in the input library.

As discussed in section 3.2.1, levels of 5-fU in the *L. major* genome are below that of 5-hmU. In addition, when pulldown of 5-fU was directly carried out in *L. major* DNA without prior oxidation of 5-hmU, no peaks were detected, suggesting that any 5-fU present does not cluster strongly within this genome[70]. Analysis of the spike-ins during SMUG1-snAP-seq shows that if 5-fU accumulates at sufficient levels, the method will detect these sites in addition to 5-hmU. To further assess the significance of 5-fU to SMUG1-snAP-seq results, DNA samples were treated with methoxyamine (MX) prior to SMUG1 treatment. This

nucleophilic compound reacts with 5-fU sites in DNA to form an oxime that is no longer a known substrate of SMUG1. To confirm that SMUG1 activity is reduced at this oxime adduct, synthetic spike-ins were treated MX prior to SMUG1-snAP-seq (MX-SMUG1-snAP-seq). Analysis of the read counts after sequencing demonstrated that specificity for 5-hmU is indeed improved, whilst the relative recovery of 5-fU is reduced (**Figure 3.20**).

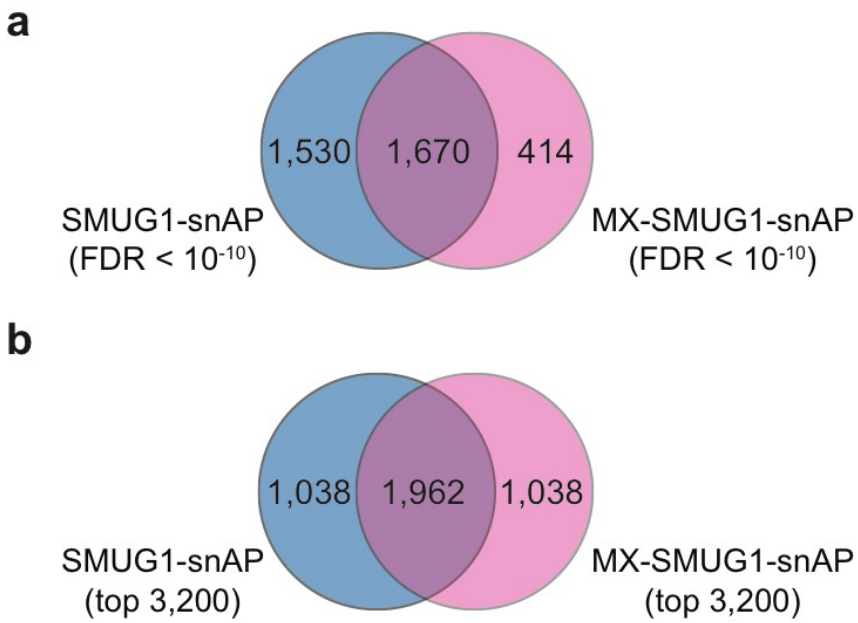


Figure 3.21: Overlap of sites detected by SMUG1-snAP-seq and MX-SMUG1-snAP-seq in *L. major* DNA. **a**) Overlap of sites detected by the two sequencing methods at a threshold of FDR < 10^{-10}. **b**) overlap of the 3,200 sites detected by the two sequencing methods with the lowest FDR values. For the calling of individual sites, two replicates of each enrichment library were processed in parallel with two input libraries.

MX-SMUG1-snAP-seq was carried out on *L. major* DNA to investigate whether 5-fU was contributing to SMUG1-snAP-seq sites. Using the same analysis pipeline and a threshold of FDR < 10^{-10}, a total of 2,084 sites were detected. 52% of the 3,200 high-confidence SMUG1-snAP-seq sites overlap with the sites detected after MX blocking (**Figure 3.21a**). This would suggest that almost half of the 3,200 sites are sensitive to methoxyamine reactivity and therefore may be due to 5-fU. However, the number of sites that remain significant beyond a given FDR threshold is heavily dependent on a number of factors including the depth of sequencing. Therefore, it was difficult to conclude with confidence that the 48% of sites that did not pass the FDR threshold were truly due to 5-fU. In an alternative analysis, sites of enrichment were ranked by FDR values, and the top 3,200 sites detected

by MX-SMUG1-snAP-seq were compared to the 3,200 high-confidence sites obtained by SMUG1-snAP-seq. This analysis was expected to be less sensitive towards some factors that lead to subtle variability between libraries, as no direct FDR threshold is used. Here, the overlap increased slightly to 61%. Given that technical variation will also be present between these two sets of data, there is a substantial degree of overlap between the MX-treated and untreated samples. These results indicate that a majority of sites are unaffected by treatment with methoxyamine, however, the enrichment of some sites is weakened. This may suggest that some of the signal detected here by SMUG1-snAP-seq are due to the oxidised derivative 5-fU, rather than 5-hmU.

Overall, application of SMUG1-snAP-seq to the *L. major* genome has demonstrated that this method is capable of detecting glycosylase-generated abasic sites in a genomic context at single-nucleotide resolution. The results also reveal that in this genome, a strong site-specific accumulation of thymine modification occurs. A preference was found for these sites in strand-switch regions, which has previously also been found for base J. SMUG1-sensitive sites were also found to be closely clustered within broad stretches and enriched in TpT and TpG motifs, suggesting that there is a preference for 5-hmU accumulation in these sequence contexts. Whilst the methodology used is highly dependent on the activity and inherent biases of the SMUG1 enzyme, control experiments have shown that the majority of sites detected by SMUG1-snAP-seq is not likely to be due to other known substrates of this enzyme or strongly influenced by any sequence preferences.

3.2.6 5-Hydroxymethyluracil in the *Trypanosoma brucei* genome

The work described in this section was carried out in collaboration with Dr Fumiko Kawasaki, Balasubramanian group and Professor Mark Carrington, Department of Biochemistry, University of Cambridge. Sequencing data were generated by Dr Kawasaki where indicated. T. brucei samples, including knockout strains, were generated by the Carrington group.

In *T. brucei*, up to 50% of base J is located outside of telomeric repeats[231]. Whilst not essential to *T. brucei* survival, base J has been associated with transcriptional regulation in a small number of genes largely found to be internal within polycistronic gene clusters. Interestingly, whilst the combined deletion of both JBP enzymes results in the complete loss of base J, a reduced level of 5-hmU remains[72]. To further study the role of 5-hmU in this genome, the location of these sites was mapped by sequencing.

Application of SMUG1-snAP-seq in the bloodstream form of *T. brucei*, where 5-hmU levels are highest, identified a total of 103 high-confidence single-nucleotide sites. This is much lower than that detected in *L. major*, despite a similar global level of 5-hmU in the two genomes, possibly suggesting that the accumulation of these sites is less strong at the single-nucleotide level in *T. brucei*. Interestingly, the overlap of these single-nucleotide sites with data generated by 5-hmU DIP-seq was relatively low, with an additional 2,355 peaks that were uniquely detected by DIP-seq (**Figure 3.22a**). Similarly, DIP-seq data overlapped poorly with peaks obtained by an independent method of 5-hmU chemical pulldown[70] in this genome (10.6%, data generated by Dr Kawasaki). Whilst each of these three methods has been demonstrated using synthetic DNA spike-ins to efficiently enrich for DNA containing 5-hmU (section 3.2.1 and ref.[70]), it is possible that the two chemical methods, SMUG1-snAP-seq and chemical pulldown, are less suitable for analysing 5-hmU particularly when these modifications are very closely spaced together. During the chemical pulldown of 5-hmU, hydrazide adducts are introduced to 5-hmU sites after oxidation, which possibly hinders PCR amplification if the adducts are in close proximity at high density. During snAP-seq it is possible that closely clustered sites lead to an excessive fragmentation of DNA during alkaline-cleavage, resulting in difficulties in sequencing such regions. In contrast, antibody pulldown is not generally hindered by a high local density of modifications and does not introduce additional chemical modifications to the DNA. Further experiments on 5-hmU in this organism were therefore carried out using DIP-seq, to explore the peaks that are detected by this method.

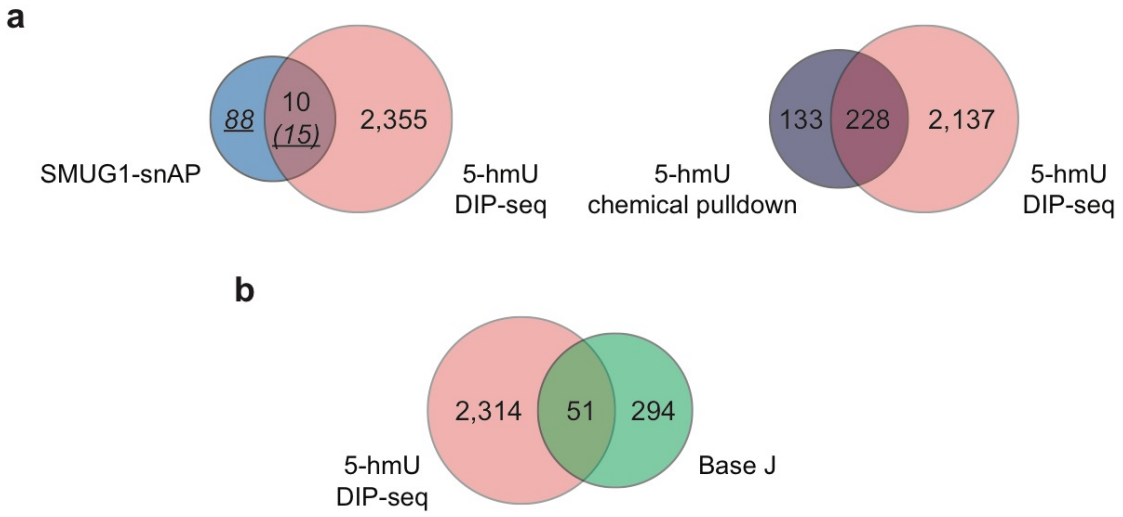


Figure 3.22: Overlap of loci enriched in thymine modifications in the *T. brucei* genome. **a**) Overlaps between 5-hmU loci detected by SMUG1-snAP-seq, chemical pulldown and DIP-seq. Single-nucleotide sites are underlined. **b**) Overlap between 5-hmU DIP-seq peaks and base J peaks. 5-hmU chemical pulldown, DIP-seq and base J datasets were generated by Dr Kawasaki.

DIP-seq experiments investigating the distribution of 5-hmU in the bloodstream form of *T. brucei* have indicated that the majority of enriched loci do not also correspond to an enrichment of base J, in contrast to that observed for 5-hmU in the *L. major* genome (**Figure 3.22b,** experiments by Dr Kawasaki). Two thymine hydroxylases, JBP1 and JBP2, have been identified in *T. brucei*. To investigate the possible source of these 5-hmU DIP-seq sites, *T. brucei* strains in which JBP1 or JBP2 was removed were generated by CRISPR-mediated knockout (experiments by the Carrington group). 5-hmU DIP-seq was carried out in these knockout samples in addition to wild-type cells and high-confidence peaks were defined as DIP-seq peaks that appear in at least two out of three technical replicates. Overlap of the high-confidence peaks detected in the knockout strains with those detected in wild-type *T. brucei* revealed that knockout of either one of the JBP enzymes does not lead to a strong depletion in 5-hmU DIP-seq peaks (**Table 3.1**). There was some variation between the two biological replicates generated for the knockout samples; however, on average around 90% of 5-hmU DIP-seq peaks remain after each knockout. To ensure that the DIP-seq peaks were not due to non-specific binding of the IgG antibody, a control pulldown library was generated using a control IgG antibody that has no known recognition specificity. A total of 89 peaks were detected in this control library using DNA from wild-type *T. brucei*, corresponding to a 1.5% overlap with 5-hmU DIP-seq peaks. This suggests that the non-specific recovery of DNA during DIP-seq was low, and the signal obtained is largely due to binding of the 5-hmU antibody. The binding specificity of this antibody to thymine modifications has previously been investigated, where it was shown that the antibody has a strong affinity to synthetic 5-hmU DNA relative to DNA containing thymine or other modifications[70,232]. This offers some support that the antibody is selective for the 5-hmU base. However, it remains unknown whether the antibody exhibits subtle binding preferences for specific sequences or motifs within DNA in the absence of 5-hmU modifications, which may influence the accuracy of the DIP-seq data.

Sample (technical triplicate)	***5-hmU DIP-seq peaks (% overlap with WT)***
Wild-type	100% (2,337)
JBP1 knockout (bio rep 1)	93.0%
JBP1 knockout (bio rep 2)	83.4%
JBP2 knockout (bio rep 1)	78.0%
JBP2 knockout (bio rep 2)	95.5%
Wild-type IgG control	1.5%

Table 3.1: Overlap of high-confidence 5-hmU DIP-seq detected in wild-type *T. brucei*, with DIP-seq peaks detected in JBP1 and JBP2 knockouts and an antibody control (IgG). Consensus peaks that appear in at least two out of three technical replicates were used for each sample described.

The peak-calling approach used here relies on detecting signal in a DIP-seq library when compared to an input library at a fixed probability threshold ($p < 10^{-10}$). Therefore, modest changes in the magnitude of signal within peaks may not be detected by only assessing the total number of peaks at this threshold. To investigate whether there were any such differences in the 5-hmU DIP-seq signal, the average read counts around the high-confidence wild-type peaks were compared between the knockout and wild-type samples. This analysis revealed that whilst there was some biological variation, overall neither JBP knockout resulted in a consistent reduction in average 5-hmU DIP-seq signal compared to wild-type cells (**Figure 3.23**). In agreement with the peak-based analysis, these results suggest that removal of either JBP1 or JBP2 alone in *T. brucei* is not sufficient to reduce the signal detected by 5-hmU DIP-seq.

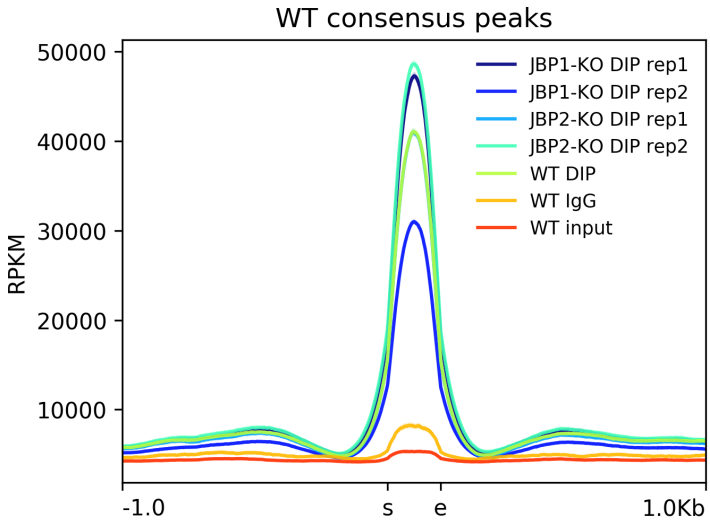


Figure 3.23: Normalised coverage in RPKM at high-confidence 5-hmU DIP-seq peaks identified in wild-type *T. brucei* (2,337, $p < 10^{-10}$, peaks that appear in at least two out of three technical replicates). The sequencing coverage from combined technical replicates of 5-hmU DIP-seq using JBP1 and JBP2 knockout samples, as well as 5-hmU DIP-seq, input and IgG control DIP using wild-type samples are shown. The two biological replicates of the knockout strains are shown separately.

Whilst it has previously been found that the majority of 5-hmU DIP-seq peaks do not overlap with base J loci in *T. brucei*, the small number of loci that are common to the two modifications showed interesting changes upon JBP knockout. At the slightly more relaxed peak-calling threshold of $q < 0.05$, the 6,149 total DIP-seq peaks were separated into 112 that overlap with base J, and 6,037 that were unique to 5-hmU. As for the total peaks (**Figure 3.23**), no consistent change in DIP-seq signal was observed between the biological replicates for either the JBP1 or JBP2 knockout in regions uniquely enriched by 5-hmU DIP-seq. In

contrast, a partial reduction in DIP-seq signal was found for the JBP1 knockout strain at the 112 DIP-seq peaks in which base J could also be detected (**Figure 3.24**). A previous study investigating base J in the absence of JBP enzymes in *T. brucei* found that JBP1 was more closely associated with the generation of base J loci internal to chromosomes, whilst JBP2 was involved in the biosynthesis of telomeric base J[233]. As telomeric regions are difficult to align during NGS due to their repetitive nature, the sequencing data analysed here are largely internal to chromosomes. Therefore, the reduction of 5-hmU DIP-seq signal upon JBP1 knockout at the 112 enriched loci associated with base J (1.8% of total 5-hmU DIP-seq peaks) supports the observation that JBP1 is essential during base J biosynthesis in regions internal to chromosomes. In contrast, knockout of either one of the JBP enzymes was not found to be sufficient to deplete the 5-hmU DIP-seq signal in loci independent of base J.

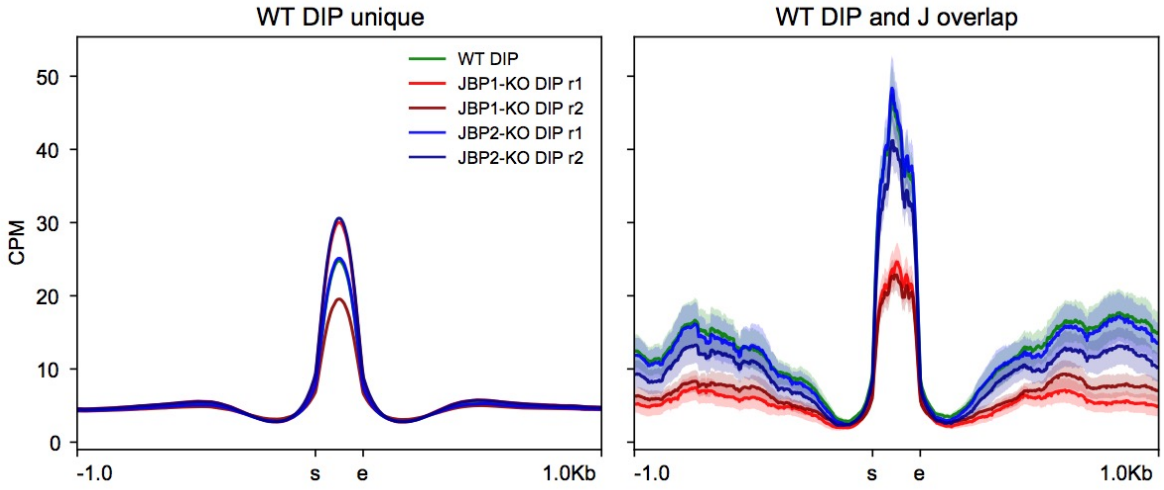


Figure 3.24: Normalised coverage in RPKM at 5-hmU DIP-seq peaks identified in wild-type *T. brucei* (6,149, $q < 0.05$, peaks that appear in at least two out of three technical replicates). The total 5-hmU DIP-seq peaks were separated into those that overlap with base J loci (right) and those unique to DIP-seq (left). The sequencing coverage from three combined technical replicates of 5-hmU DIP-seq using JBP1 and JBP2 knockout samples and wild-type *T. brucei* are shown. The two biological replicates of the knockout strains are shown separately.

The *T. brucei* knockout strains explored here lack only one of the two JBP enzymes and therefore the ability to consistently detect the majority of 5-hmU DIP-seq peaks may suggest that there is a degree of overlap in function of these two enzymes such that the loss of one can be compensated largely by the other. The sequence differences between JBP1 and JBP2 has been suggested to indicate differing functions in these two enzymes; whilst a thymine dioxygenase domain is common to both proteins, JBP1 further contains a J-binding

domain, and JBP2 contains a chromatin-remodelling domain[241]. Despite these differences, it has previously been shown that whilst base J is absent by dot-blot quantification in a *T. brucei* strain that is null in both JBP enzymes, base J can be detected again once either JBP1 or JBP2 is expressed in these cells. This suggests that both enzymes may have the ability of *de novo* base J synthesis, and can function in the absence of the other[233]. Furthermore, the biosynthesis of base J is dependent on both thymine oxidation and subsequent glucosylation of the 5-hmU intermediate; therefore, it remains unclear whether changes in base J levels reflect differences in oxidation activity, or further interactions with the JGT enzyme that functionalises 5-hmU with glucose. The data generated here by 5-hmU DIP-seq suggests that at the 5-hmU level, knockout of either JBP1 or JBP2 alone is not sufficient to deplete the majority of the 5-hmU DIP-seq signal, whilst the small proportion of enriched loci that overlap with base J appear to be dependent on JBP1 activity (**Figure 3.25**). This indicates that JBP1 is essential only in the biosynthetic pathway for 5-hmU sites that are further associated with base J. It has been reported that global levels of base J are only partially reduced when either one of the JBP enzymes are deleted in isolation, whilst the double knockout no longer contains detectable levels of base J[232]. The deletion of both JBP enzymes in combination was not explored here, but may be valuable in further investigating the oxidation activities of these enzymes at the genomic level. Interestingly, in contrast to base J, a total absence of 5-hmU has not been found in the double JBP1/2 knockout, where a three-fold reduction was observed relative to wild-type *T. brucei*[72]. This therefore suggests that alternative enzymatic pathways independent of either JBP enzymes, or possibly non-enzymatic mechanisms such as oxidative damage, may be responsible for generating a minority of 5-hmU sites in this organism.

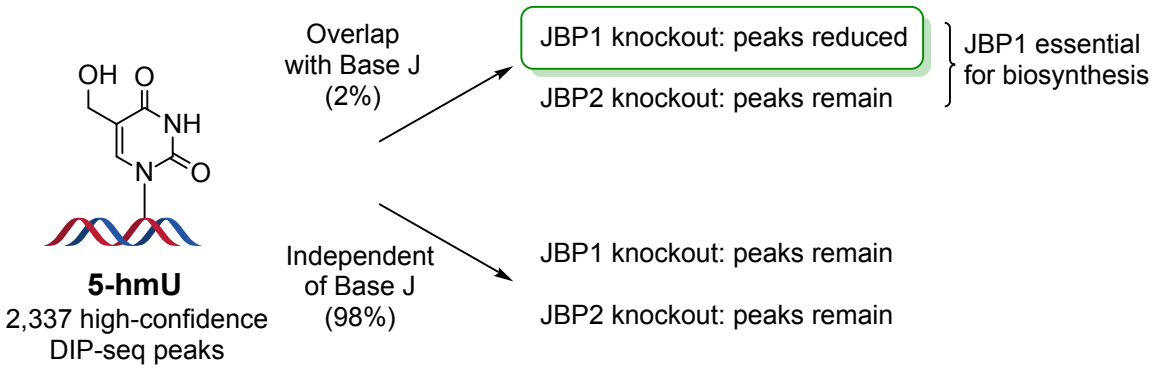


Figure 3.25: Summary of key findings from studies on the distribution of 5-hmU by DIP-seq analysis in the *T. brucei* genome. 5-hmU DIP-seq peaks identified in regions independent of base J loci are unaffected by either JBP1 or JBP2 knockout, whilst a reduction in 5-hmU DIP-seq signal is detected upon JBP1 knockout in loci that coincide with base J.

3.3 Conclusions and future directions

Using synthetic DNA models that contain modifications at known positions, it was shown in this chapter that SMUG1-snAP-seq can be used to enrich for 5-hmU DNA in addition to DNA containing other SMUG1 substrates and endogenous abasic sites. These results also emphasised the need to carry out extensive control experiments when utilising this glycosylase-mediated application of snAP-seq, in order to distinguish a substrate of interest from other possible glycosylase substrates and endogenous abasic sites. Application of SMUG1-snAP-seq to *L. major* DNA identified thousands of high-confidence sites at single-nucleotide resolution. Control libraries prepared on DNA in the absence of SMUG1 treatment showed that these sites were not due to endogenous AP sites and can therefore be considered SMUG1-sensitive. Out of the possible substrates of SMUG1, the contribution of uracil to these sites was found to be unlikely due to the lack of peaks when SMUG1 was replaced with the uracil-specific glycosylase UNG. Chemical discrimination between 5-hmU and 5-fU, a further substrate of SMUG1, suggested that some of the signal detected by SMUG1-snAP-seq may be due to 5-fU. Therefore, the SMUG1-snAP-seq sites can be concluded to be largely detecting oxidised thymine derivatives, the majority of which likely corresponds to 5-hmU.

Comparison of the genomic location of SMUG1-snAP-seq sites in *L. major* DNA with 5-hmU peaks obtained by chemical pulldown showed a large overlap between the two datasets. The consistent results obtained by these two orthogonal approaches provides validation for SMUG1-snAP-seq, and more generally the snAP-seq approach, and also provides proof-of-concept that this method can be applied genome-wide. In addition, a large overlap was also seen in the *L. major* genome between SMUG1-sensitive sites and base J pulldown peaks, further confirming the relationship between these two modifications as suggested in the biosynthetic pathway. Motif analysis of the obtained sites suggested that 5-hmU preferentially occurs within the TpG dinucleotide. Whilst it was also found that the method has some intrinsic bias for the TpG motif, the degree of enrichment of this motif in the genomic data was statistically significant when compared to the technical contribution estimated using synthetic DNA models. An enrichment in the TpT dinucleotide was also found, suggesting that 5-hmU accumulation is also preferred in these sequence contexts. Together, these results provide further insight into the accumulation of thymine modifications in the *L. major* genome at high resolution.

The distribution of 5-hmU was also studied in the *T. brucei* genome, using three independent mapping methods. Compared to the two chemical-based enrichment methods, 5-hmU DIP-seq was found to detect a larger number of peaks in this particular genome. These peaks were not detectable by a control IgG antibody with no known binding specificity. A small proportion of 5-hmU sites (2.8%) did not overlap with base J loci in the *L. major* genome, whilst for *T. brucei* 98% of peaks detected by 5-hmU DIP-seq did not coincide with base J, suggesting that these 5-hmU sites do not then get glucosylated. To further investigate how these modifications are generated, knockout of each of the two known thymine oxidases, JBP1 and JBP2, was carried out. Neither knockout strains resulted in a strong loss of 5-hmU DIP-seq signal, suggesting that JBP1 or JBP2 alone is not solely responsible for the generation of a substantial amount of the 5-hmU signal detected. However, JBP1 knockout was found to affect the 5-hmU signal in loci where 5-hmU and base J co-occur, suggesting that JBP1 is essential for the generation of 5-hmU sites associated with base J. As it is possible that the two JBP enzymes have a degree of functional overlap, future work should focus on investigating 5-hmU in cell lines where both JBP enzymes are deleted in combination. This can then reveal the extent to which the DIP-seq signal is truly due to the JBP enzymes. Any persistent 5-hmU in the absence of both JBP enzymes is then suggestive of an alternative formation pathway, either from currently unidentified enzymes in the *T. brucei* organism, or through oxidative damage. As DIP-seq was used here to study the genomic location of 5-hmU sites, future work should also focus on further investigating the binding specificity of the 5-hmU antibody. Whilst both input and IgG control libraries were used here for comparison against the 5-hmU DIP-seq libraries, validation of the detected peaks by an orthogonal method or further controls will ensure the accuracy of the data.

The work in this chapter has also shown that uracil may be mapped using UNG as part of UNG-snAP-seq. Whilst the SMUG1 enzyme has multiple substrates that must be distinguished when interpreting sequencing results, UNG is a highly specific glycosylase. The application of UNG-snAP-seq to investigate the role of uracil during DNA demethylation in embryonic stem cells is described in Chapter 5.

Chapter 4

Mapping abasic sites in the human genome

4.1 Background

The loss of a DNA nucleobase by hydrolysis leads to the formation of an abasic (AP) site. Amongst the many possible products of DNA damage, abasic sites are one of the most abundant that can arise in the event of both endogenous and exogenous damage. When AP sites are not repaired and persist within DNA, these lesions can cause polymerase stalling and mutations, leading to genomic instability[107,242]. Abasic sites are chemically unstable and single-strand breaks can form on the 3' side of the AP sugar by β-elimination at the aldehyde moiety.

Despite the severe biological consequences of AP sites, there is still a lack of understanding regarding their distribution within genomic DNA. Global levels of AP sites and the β-elimination products have been measured in mESCs by LC-MS/MS at 0.9 and 1.7 per million dN[47], respectively, with similar levels reported in other cell types including HEK293T cells. These levels are comparable to rare base modifications, such as 5-carboxycytosine (0.6 per million dN). Whilst low in abundance, the mapping of 5-caC by affinity enrichment has revealed a non-random distribution across genomic DNA such that sites of accumulation can be detected as peaks. DNA base lesions that have commonly been considered to be by-products of DNA damage such 8-oxoG and 5-fU have also been shown by genome-wide mapping to accumulate within peaks in mammalian DNA[180,243]. The accumulation of 8-oxoG has been found to differ amongst genomic features. Mammalian genomes consist of both genic and intergenic regions, where the expression of a gene is regulated by a promoter located to the 5'- end of the gene body. Protein coding genes may consist of multiple coding exons and non-coding introns, and also contain a untranslated region (UTR) on both the 5'- and 3'- end. 8-oxoG has since been proposed to act as an epigenetic marker capable of controlling gene expression in specific biological contexts[180,243]. Furthermore, DNA damage in the form of double-strand breaks has also been shown to accumulate within genomic loci[105]. Together, these observations suggest that DNA damage products are not equally distributed within genomic DNA, with some regions more prone to accumulating damage than others.

The hypothesis in this chapter is that abasic sites, whilst relatively low in abundance on a global level, may similarly be distributed non-randomly across the genome. The overall aim was therefore to sequence the genomic location of abasic sites. The focus was on human cells, and the HeLa cell line was chosen for study as it is well characterised and also relatively easy to manipulate.

At the outset of this project, no method existed for the sequencing of abasic sites. The detection of aldehyde-reactive lesions in isolated DNA fibres using ARP as a probe has suggested that the distribution of genomic abasic sites can be non-random and clustered; however, the exact locations and sequence contexts remain unexplored[126,127]. Furthermore, these experiments were carried out using ARP, which is able to react with different sources of aldehydes in DNA. In Chapter 2, this has been shown to include 5-formylpyrimidines in addition to abasic sites and therefore the true significance of abasic sites within loci detected by ARP remains unclear. During the course of this project, a method to map abasic sites by NGS, AP-seq, was reported[130]. This method also utilises ARP to mark genomic abasic sites with a biotin moiety. DNA is then sonicated into small fragments and biotinylated fragments are isolated by streptavidin pulldown and subsequently sequenced. As in the case of the DNA fibre analysis, the ARP compound used is also able to react with and therefore detect 5-formyluracil and 5-formylcytosine bases (section 2.2.1). Given that higher levels of formylpyrimidines than abasic sites have been measured in some cell types[27,47,50], it is essential that methods used to study these modifications are able to clearly distinguish between the different aldehydes in DNA. As demonstrated in Chapter 2, snAP-seq can selectively detect abasic sites in the presence of formylpyrimidines and is thus suitable for studying abasic sites in a genomic context.

As part of the base excision repair (BER) repair pathway, the repair of abasic sites in mammalian systems is largely initiated by APE1. It is estimated that up to 95% of endonuclease activity at DNA abasic sites is accounted for by APE1, whilst APE2 also has minor activity at these sites[134,136]. The product of APE1 is a single-strand break, in which the DNA backbone has been hydrolysed at the 5'- phosphate of the abasic site (**Figure 4.1a**). BER is then divided into two possible sub-pathways[94]; during short-patch BER, DNA polymerase β removes the deoxyribose unit and fills in the single-nucleotide gap, before the nick is sealed by ligase activity, whilst during long-patch BER, DNA polymerase β continues synthesis for 2-13 nucleotides beyond the initial AP site before the DNA overhang is removed by FEN1, followed again by ligation. Preference for either of these pathways *in vivo* is not

fully understood, with the type of glycosylase used to generate the initial abasic site and cell cycle stage amongst the factors that can influence the choice of mechanism[244].

Cellular repair of abasic sites by alternative pathways such as the nucleotide excision repair (NER) pathway has also been implicated in yeast[137,139], however, in mammalian cells BER remains the main route of abasic site repair. Knockdown of APE1 protein by treatment with siRNA, or inhibition of endonuclease activity by small molecules has been shown to elevate global levels of abasic sites[47,245], suggesting that the reduction of cellular APE1 activity is a convenient way to control the abundance of abasic sites. In this chapter, the distribution of abasic sites was explored in cells depleted in APE1, in addition to BER competent cells to assess whether regions of the genome are more susceptible to abasic site formation before repair *via* the BER pathway.

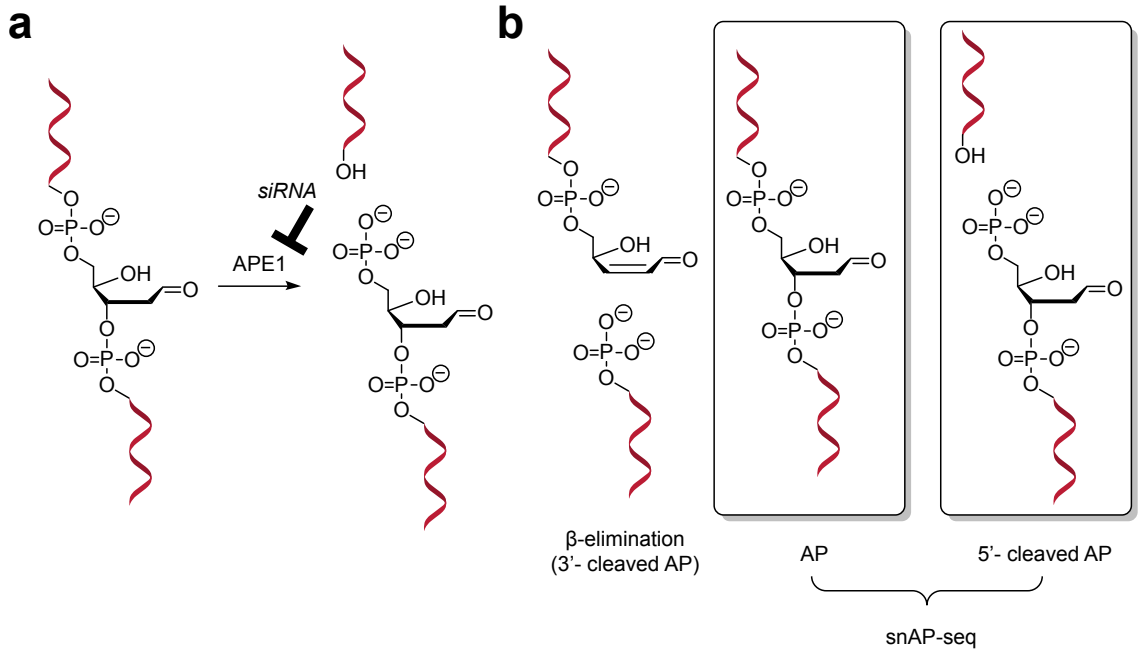


Figure 4.1: DNA strand-breaks at abasic sites. **a**) APE1 generates a strand-break on the 5'- side of an abasic site by enzymatic hydrolysis of the phosphate backbone. BER is initiated, ultimately resulting in the repair of the abasic site. Inhibition of APE1 is therefore expected to halt BER and elevate levels of unprocessed abasic sites. **b**) Abasic sites in DNA can exist in one of at least three forms: 3'-cleaved by β-elimination (left), intact (centre) and 5'-cleaved (right). Only the intact and 5'-cleaved states can be captured by snAP-seq, whilst β-eliminated AP sites cannot be detected by snAP-seq.

Abasic sites can exist in at least three different forms in a DNA strand; intact without association with an adjacent strand break, or with a single-strand break on either the 5'- or the 3'- end (**Figure 4.1b**). Both types of strand-breaks can be enzymatically generated by AP

endonuclease and lyase activity, respectively. Cleavage at the 3'- end can also occur spontaneously under physiological conditions, where rates may be further accelerated in the event of crosslinks with proteins such as histone tails[121]. Using ARP as an abasic site probe, assays on extracted genomic DNA from rat tissue suggested that the majority of abasic sites are 5'- cleaved whilst a smaller proportion are intact or 3'- cleaved[246]. This set of experiments relied on the observation that whilst intact abasic sites and those associated with one single-strand break are reactive towards ARP, simultaneous cleavage at both 5'- and 3'- sides of an abasic site results in the complete excision of the deoxyribose unit from DNA, with the released strands being no longer reactive towards ARP. Therefore, by carrying out the ARP assay before and after selectively inducing DNA cleavage on either side of abasic sites, the original status of the abasic site can be deduced. The larger amount of 5'- cleaved sites was consistent across a range of tissues in rat, although the overall level of abasic sites was highest in brain tissue. As these findings were based on the ARP assay, results should be treated with caution. Contrasting finding were reported more recently in a study by Rahimoff *et al*.[47], where it was shown that using LC-MS/MS quantification, it is possible to distinguish between 3'- cleaved and intact abasic sites due to differences in saturation of the sugar ring released upon nuclease digestion of DNA. The method detects 5'- cleaved abasic sites as intact, as both the 5'- hydrolysis product and intact abasic sites yield 2-deoxyribose upon DNA digestion and dephosphorylation. A higher level of 3'- cleaved abasic sites was detected than the combined 5'- cleaved and intact species, suggesting that a large proportion of genomic abasic sites had undergone β-elimination. A limitation of the library preparation design of snAP-seq is that 3'- cleaved abasic sites cannot be detected. The snAP-seq method relies on the pulldown of DNA sequences containing abasic sites, followed by controlled fragmentation to selectively recover the DNA fragments on the 3'- side to mark abasic sites at base-resolution. In the event that β-elimination has occurred prior to this process, the 3'- end fragment is already lost and can no longer be captured (**Figure 4.1b**).

The chemical instability of abasic sites under basic conditions, as well as the potential formation of AP artefacts during DNA processing were important considerations in this study. The genomic distribution of abasic sites can be influenced by both events, which must be avoided to accurately study endogenous abasic sites. Control experiments were carried out to assess the effects of key steps during DNA processing that may affect abasic sites. Optimised conditions were then used to sequence the location of abasic sites in both APE1 deficient and control cells, to investigate the distribution of this type of DNA damage.

Oxidative DNA base lesions such as 8-oxoG have been associated with the control of gene expression and it has been suggested that in addition to their role as DNA damage products, these modifications can be considered epigenetic in conditions of oxidative stress. The distribution of 8-oxoG has been found to differ amongst genomic features and also chromatin structure in multiple genomes[243,247], indicating that the susceptibility of formation and maintenance of these lesions can differ across the genome. The AP signal obtained by snAP-seq was also assessed relative to genomic features, chromatin state and mRNA expression levels to further investigate the biological consequences of persistent AP sites.

4.2 Results and discussion

This project was carried out in collaboration with Dr Sergio Martínez Cuesta, Balasubramanian group, who performed the bioinformatical analysis of sequencing data generated in this chapter.

4.2.1 Knockdown of APE1 protein

The overall aim of this project was to investigate the distribution of abasic sites in human DNA. As the levels of endogenous abasic sites is low in wild-type mammalian cells[47] and may be challenging to detect, global elevation of AP levels by depletion of APE1 protein was carried out. Furthermore, it was rationalised that the expected increase in abasic sites upon inhibition or knockdown of APE1 protein would be a good way to verify the accuracy of snAP-seq results. The investigation of abasic sites in APE1 depleted cells may also reveal insights into the susceptibility of AP site formation before removal by BER.

A number of small molecules that inhibit the endonuclease activity of APE1 are known, including lucanthone and CRT0044867[248,249]. Whilst an increase of global AP site levels has been observed upon treatment of cells with these molecules, lucanthone for example, also interacts with DNA through intercalation[250]. To avoid potential off-target effects of small molecules that may affect genomic DNA in unexpected ways, the direct knockdown of APE1 protein by siRNA control was explored instead. Four individual siRNAs designed to target APE1 mRNA were obtained (ON-TARGETplus, Dharmacon). Cellular delivery was achieved by transfection, and for initial optimisation the extent of knockdown was first followed on the

mRNA level by quantification using RT-qPCR. Transfection of each siRNA sequence, along with a control pool of non-targeting siRNA (Dharmacon) in HeLa cells was assessed over 96 h. RT-qPCR data obtained using primers specific to APE1 mRNA was normalised to that of a housekeeping gene, β-tubulin, where levels were not expected to vary between the siRNA treatments (**Table 7.1**). All mRNA levels were analysed by comparison against cells that were treated with transfection reagent alone and in the absence of siRNA. Normalisation was carried out by setting the APE1/β-tubulin ratio in transfection reagent-only cells to 100%.

The RT-qPCR results show that 48 h of transfection was sufficient for around 95% reduction of APE1 mRNA using three of the four siRNAs tested (**Figure 4.2**). These levels remained largely constant up to 96 h transfection. There was no strong reduction in expression for the non-targeting control pool, confirming that the observed differences were not likely due to non-specific effects of introducing siRNAs. One of the targeting sequences, siRNA #4, performed much less effectively, with around 50% of APE1 mRNA remaining after 96 h. It is possible that some siRNA sequences are not as efficient or specific as others, and this effect may be cell line dependent. Another cell line was therefore tested to investigate whether these results were consistent. In U2OS cells, an osteosarcoma line, similar results were observed. Between 90% and 95% knockdown was found for the first three siRNAs, which decreased to 60% for siRNA #4. Whilst it is possible to combine multiple siRNA sequences to further increase knockdown efficacy, three out of the four siRNAs offered over 90% knockdown of APE1 mRNA when used alone and are likely to be adequate for use.

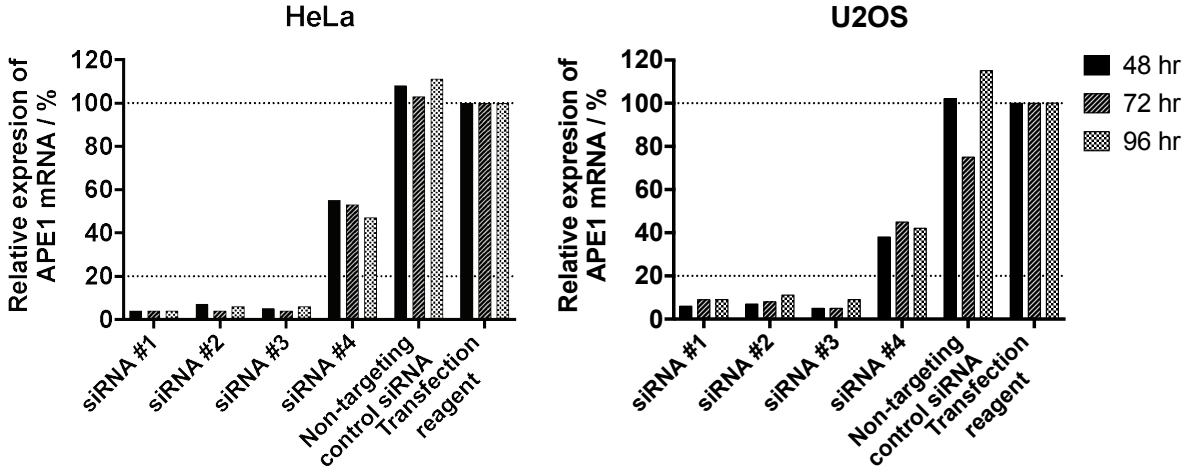


Figure 4.2: Relative expression of APE1 mRNA after treatment with siRNAs. siRNA #1-4 target APE1 mRNA, and a non-targeting siRNA pool was used as a negative control. APE1 mRNA was quantified by RT-qPCR and normalised against β-tubulin mRNA. All values are expressed as a percentage of expression relative to cells that were treated with transfection reagent only. Data from a single replicate is shown.

In addition to depletion of APE1 mRNA, it was also important to ensure that effects were ultimately observed on the protein level. Western blots were therefore carried out to verify changes in protein expression. Under the same transfection conditions as used for RT-qPCR analysis, whole cell extracts were collected from treated cells. For western blots, anti-tubulin antibody was used in addition to anti-APE1, as a loading control for the normalisation of results. The APE1/tubulin ratios obtained by quantification of western blot band intensities were normalised to control cells that were treated with transfection reagent only. The results show that in contrast to mRNA, the depletion of protein is slightly slower, and levels continue to decrease between 48-96 h transfection time (**Figure 4.3**). The most efficient knockdown was at the longest timepoint. siRNA #4 treatment was again less effective than #1-3, resulting in less than 50% protein knockdown. Between the remaining siRNAs, #3 performed slightly better. This siRNA was therefore chosen for further sequencing experiments (**Figure 4.3c**). The 96 h treatment was selected as this offered the greatest depletion of APE1 protein and in addition, the longer timepoint may allow for a greater accumulation of AP sites.

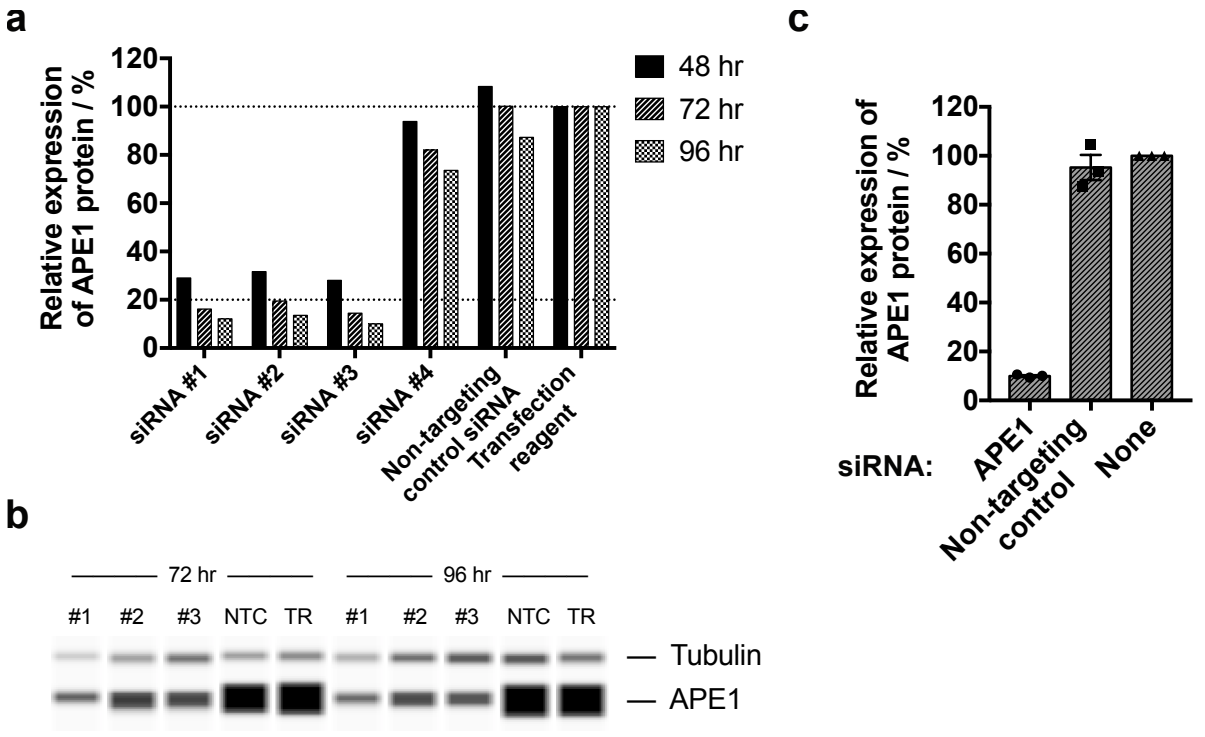


Figure 4.3: Relative levels of APE1 protein measured by western blot. Whole cell extracts from HeLa cells were used. Anti-APE1 antibody was used along with anti-tubulin as a loading control. For data analysis, the measured APE1/tubulin ratios were normalised to that in the transfection reagent only samples and expressed as a percentage of remaining APE1 protein. **a**) Normalised expression after transfection of different siRNA sequences. Data from a single replicate is shown. **b**) Western blot showing tubulin and APE1 bands. **c**) Normalised expression after transfection of siRNA #3 and controls. Mean and S.E.M of three replicates are shown.

The ultimate goal of using APE1 knockdown was to elevate genomic AP levels and therefore the observation of changes in global AP levels was also important. ARP-based assays have previously suggested an increase in reactive aldehydes upon depletion of APE1[245]; however, results may be misleading due to cross-reactivity using this approach. LC-MS/MS quantification remains the gold standard for such analysis. Using this technique, it was reported that a 50% knockdown of APE1 mRNA in mESC resulted in a 30% increase in AP site levels[47]. Although carried out in a different cell line using a differing siRNA system, these results offer some support that AP levels were expected to increase with depletion of APE1.

4.2.2 Effect of DNA extraction on abasic sites

In order to accurately detect abasic sites in genomic DNA, it was important to ensure the integrity of abasic sites prior to chemical tagging, and that the introduction of additional abasic sites by depurination is avoided. The two key steps preceding the chemical reaction with HIPS probe **20** during snAP-seq are DNA extraction and sonication.

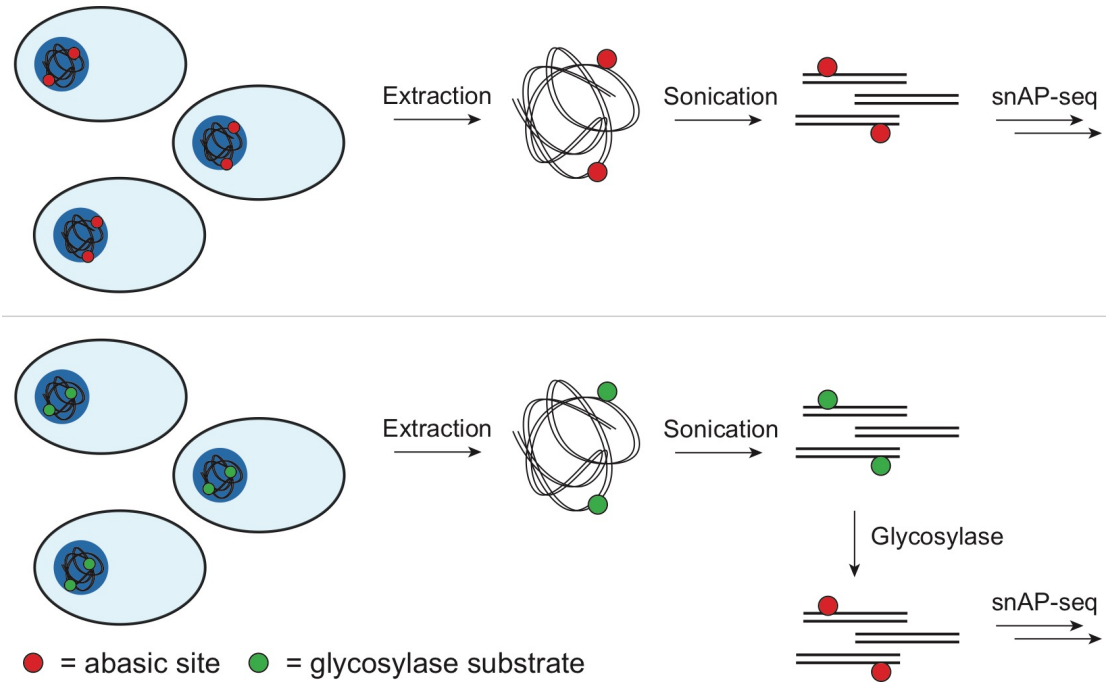


Figure 4.4: Schematic to illustrate the steps required to prepare DNA from cells during snAP-seq of endogenous AP sites (top) and glycosylase-mediated snAP-seq of a base modification such as SMUG1-snAP-seq (bottom).

As demonstrated in section 2.2.7, abasic sites introduced as artefacts after treatment with **20** are not expected to be enriched by snAP-seq and can therefore be tolerated. For glycosylase-coupled snAP-seq, it is possible to delay the glycosylase treatment to protect abasic sites. For example, in the case of SMUG1-snAP-seq in the *L. major* genome, 5-hmU is less labile than abasic sites and therefore it was beneficial to carry this modification through DNA extraction and sonication, and treat DNA with SMUG1 immediately prior to reaction with **20** to minimise the time during which free AP sites were exposed (**Figure 4.4**). In contrast, for the study of endogenous AP sites that are present at the start of the process, a similar protection is not possible.

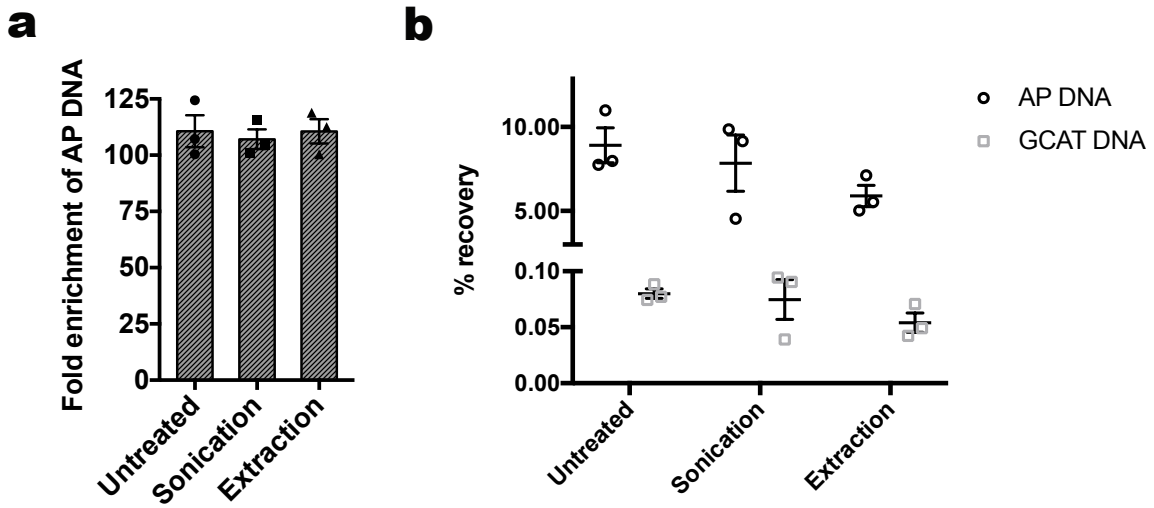


Figure 4.5: Enrichment of AP DNA relative to GCAT DNA. Pooled DNA samples were used directly, sonicated, or subjected to DNA extraction conditions. All samples were then treated with **20** and biotinylated, followed by enrichment on streptavidin beads and eluted by alkaline-cleavage (100 mM NaOH, 70 °C, 15 min). Recovered DNA was quantified by qPCR and expressed as a fold-change relative to input samples. Mean and S.E.M. of three replicates are shown. **a**) Relative enrichment of AP/GCAT DNA. **b**) Recovery of AP and GCAT DNA expressed as a percentage of the input.

To assess the effects of DNA extraction and sonication on abasic sites, the synthetic DNA sequences designed for spike-ins were used. Equal amounts of AP DNA and GCAT DNA were pooled together and subjected either to sonication using the same settings as applied for genomic DNA, or a mock extraction using a commercially available kit (Zymo Research). DNA was then treated with **20** followed by CuAAC catalysed biotinylation and subjected to streptavidin pulldown. DNA recovery and relative enrichment was assessed by qPCR. The extent of enrichment of AP DNA remained comparable to a control sample that

had not undergone extraction or sonication treatments, at over 100-fold relative to GCAT DNA (**Figure 4.5**). No significant change was detected between samples (p > 0.9999), suggesting that these steps do not introduce detectable artefacts. Furthermore, the absolute recovery of both sequences did not change significantly between treatments (Kruskal-Wallis test, p = 0.2964 (AP) and p = 0.3607 (GCAT)). The loss of abasic sites, by elimination for example, would be expected to reduce the recovery of AP DNA, whereas the introduction of artefacts may lead to increased recovery of GCAT, neither of which were observed.

The synthetic spike-in DNA may not be fully representative of genomic DNA during sonication or extraction due to the large differences in DNA length. Therefore, further control experiments were carried out using high molecular weight *L. major* DNA. In Chapter 3, the effects of DNA processing steps were not a key consideration, as all SMUG1-snAP-seq results were compared to snAP-seq without SMUG1 treatment. As these two sets of libraries were subject to the same DNA processing steps, any artefacts were expected to equally affect both conditions and be removeable by subtraction. In contrast, endogenous AP sites were the focus of interest in this chapter, and a similar control is not available in this case.

To assess whether abasic sites are stable during sonication, a set of libraries were prepared in which SMUG1 treatment was carried out on high molecular weight *L. major* DNA, which was then sonicated. The enzymatically generated abasic sites undergo sonication under these conditions, whereas in the libraries prepared in Chapter 3, SMUG1 treatment was carried out after sonication and therefore abasic sites were effectively protected from this step. Over 80% of the 3,200 high-confidence sites detected in Chapter 3 remain detectable at the same FDR threshold in these samples, suggesting that abasic sites are largely preserved during sonication (**Figure 4.6**). In addition, libraries were prepared in which high molecular weight SMUG1-treated *L. major* DNA was subjected to an additional mock re-extraction then subjected to sonication and SMUG1-snAP-seq. The SMUG1-generated AP sites in this case undergo both DNA extraction and sonication and 67% of the 3,200 control sites remained detectable. This extraction protocol utilised a commercially available kit (Zymo Research) that provides a chemical lysis system at room temperature. Interestingly, another DNA extraction kit tested gave drastically different results, with only 1% of the 3,200 control sites detected (DNeasy, Qiagen). This kit was based on proteinase K digestion to lyse cells, which requires the incubation of cells at elevated temperatures to allow efficient enzymatic digestion. The incompatibility of these conditions with AP enrichment suggests that degradation of the AP sites may be occurring at high temperatures. These results highlight the need to maintain mild conditions during DNA extraction that do not affect AP sites. Overall,

the combination of DNA extraction using the Zymo kit with sonication was found to maintain AP reactivity without the significant introduction of artefacts; these conditions were therefore selected for use on HeLa DNA.

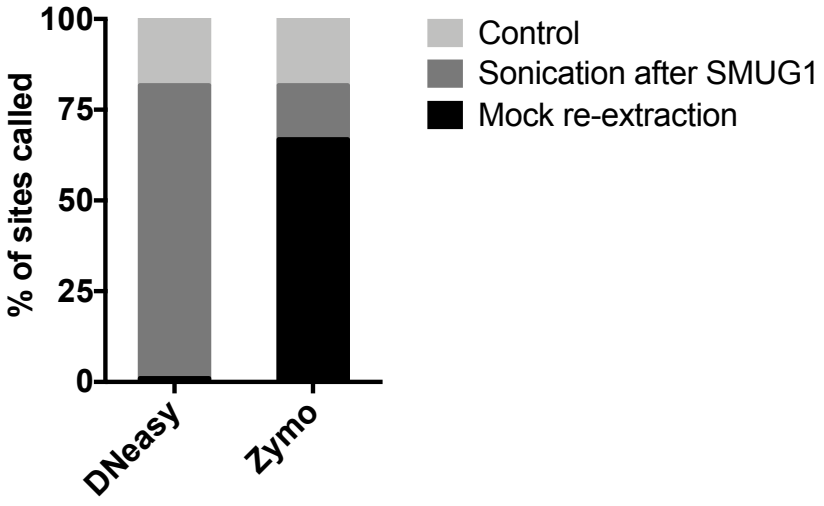


Figure 4.6: SMUG1-snAP-seq sites detected in *L. major* DNA with and without mock DNA re-extraction and sonication after generating AP sites using SMUG1. An FDR threshold of 10^{-10} was used for all analyses; the detected sites are represented as a percentage overlap with the 3,200 sites detected in the control sample without additional treatment.

4.2.3 Mapping abasic sites in HeLa cells

To investigate the distribution of endogenous abasic sites in genomic DNA, snAP-seq was carried out on DNA from APE1 deficient cells that had been treated with siRNA #3, in addition to BER competent cells that had been treated with non-targeting control siRNA. Four replicates across two independent biological samples were prepared for each cell type, along with the corresponding input libraries. To assess whether abasic sites were accumulating in these samples at the single-nucleotide level, comparative assessment of read counts was performed across the genome between snAP-seq and input libraries as previously carried out for the SMUG1-snAP-seq data (Chapter 3). The requirement for calling high-confidence enrichment sites was to set an FDR threshold at which positive $\log_2$ fold-change (snAP-seq vs. input) was favoured over negative $\log_2$ fold-change, as the site-specific accumulation of reads in the input sample without enrichment detected as negative $\log_2$ fold-change most likely corresponds to noise in the data. Inspection of the volcano plots obtained for sites in both control and APE1 siRNA treated cells revealed a largely even distribution of sites with

positive and negative log_2 fold-change, indicating that no such threshold could be set (**Figure 4.7**). Therefore, no single nucleotides of enrichment could be called with confidence.

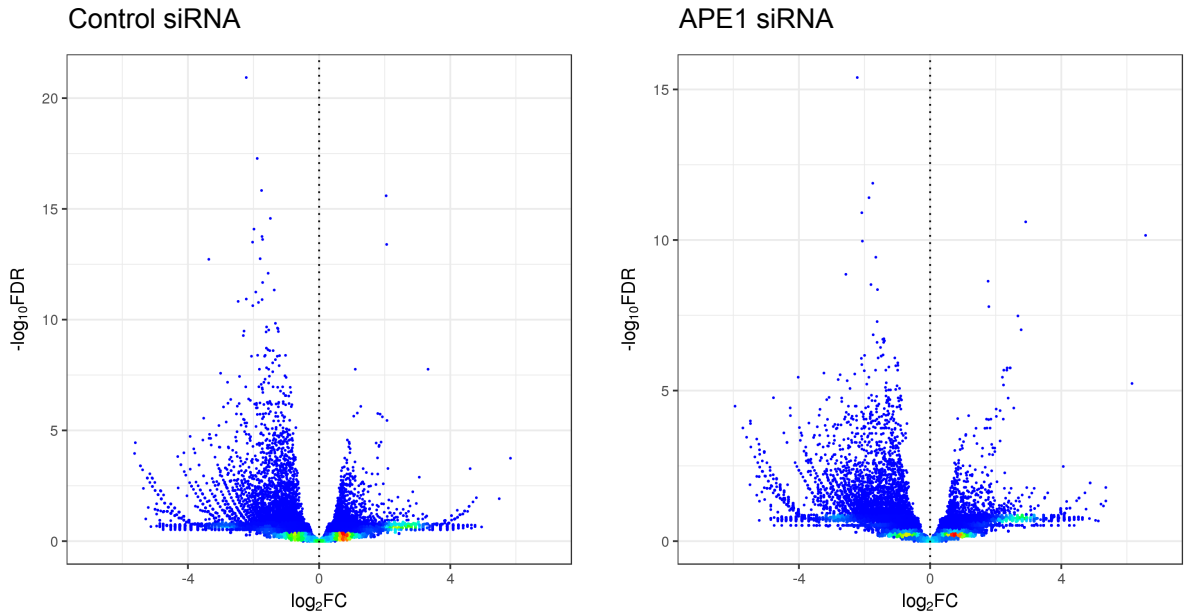


Figure 4.7: Volcano plot representing genomic loci with differential coverage in snAP-seq and input libraries. Reads were split between those aligning to forward and reverse strands based on the reference human genome; the number of reads beginning one nucleotide 3'- to each position is represented as a log_2 fold-change in coverage (snAP-seq vs. input libraries). Analysis was carried out on four replicates of each condition in parallel. Only analysis of the forward strand is shown.

Mitochondria have been associated with ROS and oxidative damage[251]; it may then be expected that oxidative DNA damage accumulates more in mitochondrial than nuclear DNA. Furthermore, the depth of sequencing achieved was much higher in mitochondrial DNA, due to the higher copy number than for chromosomal DNA. As the sensitivity of snAP-seq is likely to depend on the sequencing depth, sites may also be more easily detectable in the mitochondria. Therefore, reads aligning to mitochondrial DNA were analysed separately. Similar results were obtained to those for all reads, showing a largely even distribution in the volcano plots (**Figure 4.8**). Despite the increased depth (over 100X, compared to around 1.5X for nuclear), no snAP-seq sites were detected in mitochondrial DNA suggesting that AP sites do not accumulate at the single-nucleotide level here either.

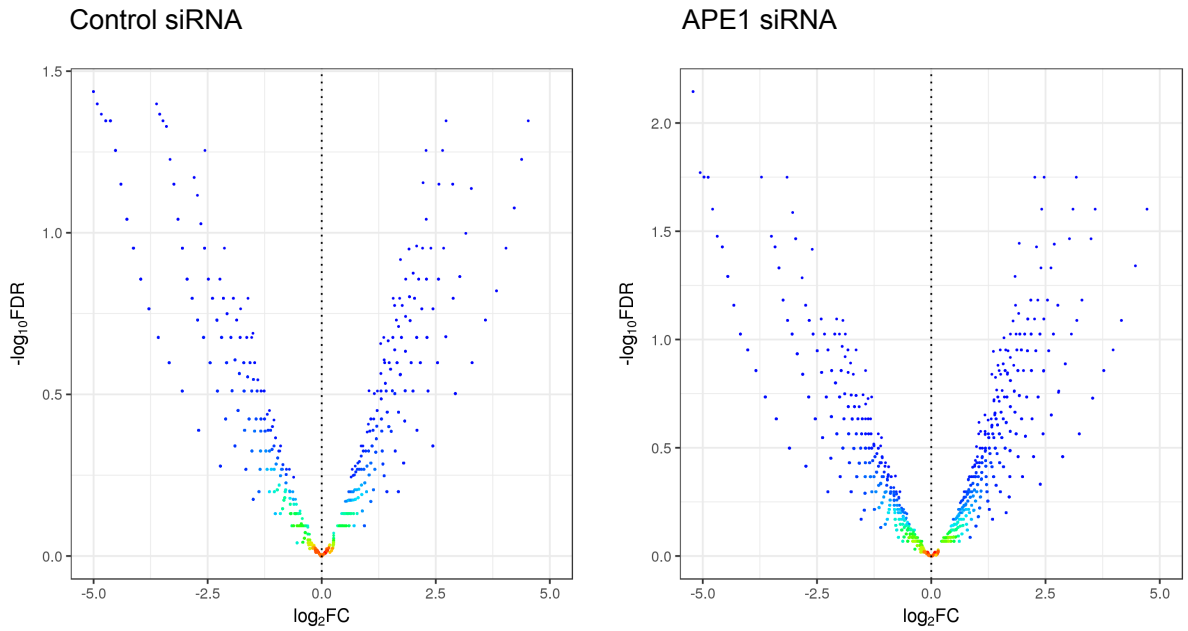


Figure 4.8: Volcano plot representing genomic loci with differential coverage in snAP-seq and input libraries in mitochondrial DNA only. Reads were split between those aligning to the forward and reverse strands based on the reference human genome; the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (snAP-seq vs. input libraries). Analysis was carried out on four replicates of each condition in parallel. Only analysis of the forward strand is shown.

The lack of individual sites that were significantly enriched after snAP-seq suggests that at the single-nucleotide levels, there may be a level of stochasticity of abasic sites across the population of cells. Typically, a snAP-seq library was generated using up to half a million HeLa cells. In order for a genomic position to be successfully enriched by snAP-seq, a substantial proportion or possibly all cells need to contain an abasic site at the same specific genomic location. The results suggest that this may not be occurring sufficiently for detection by an enrichment approach. Alternatively, knockdown of APE1 alone may not be sufficient to observe a substantial accumulation of AP sites at a given location. APE2 is also present in mammals and may become more active in the absence of APE1 activity[136], whilst the repair of AP sites by the NER pathway has also been demonstrated[139]. Methods capable of detecting DNA double-strand breaks at nucleotide-resolution have found that this type of damage accumulates within broadened peaks rather than individual nucleotides[105]. Therefore, in an alternative analysis, peak-calling was used on the snAP-seq data to investigate whether regions of the genome, rather than single nucleotides, were enriched in AP sites. Peak-calling was performed using MACS2 software[252], which is frequently used for data from ChIP-seq or enrichment sequencing[128]. This type of analysis detects enrichment

within a library in windows of the genome rather than single nucleotides, and can be used with or without an input sample for comparison to identify peaks in short stretches of the genome (**Figure 4.9a**). Regions of high coverage in the input libraries can cause problems during the analysis of enrichment-based sequencing methods including ChIP-seq data, and an available blacklist was used to filter out many of these locations[253]. As a cancer-derived cell line that has been used for many decades, the HeLa genome is heavily mutated and many regions present problems during the alignment of reads[254]. A custom genomic blacklist was therefore also generated by using MACS2 peak-calling on the input libraries alone. The regions identified were subtracted from peak-calling analysis, as they represent regions that have a higher baseline of coverage. This step was designed to further increase confidence in the peak-calling process.

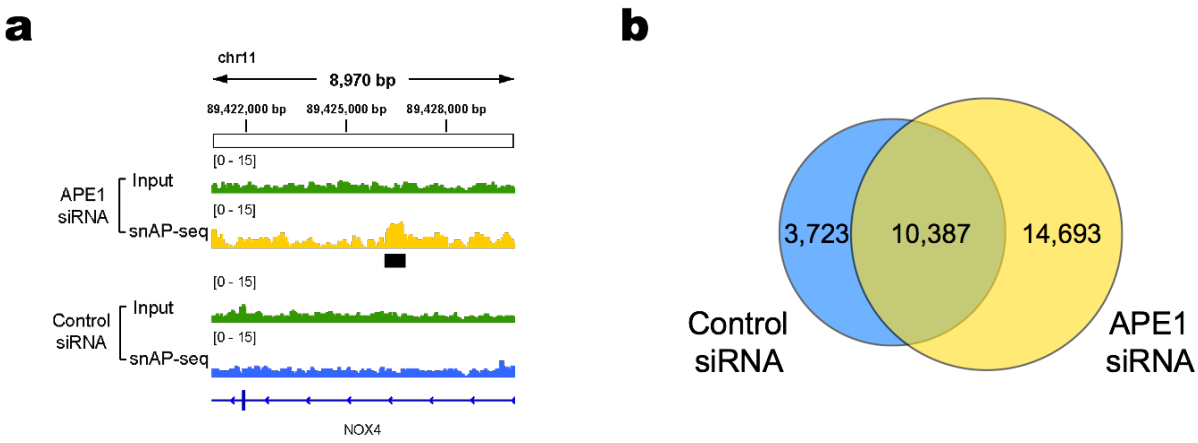


Figure 4.9: Peak-calling in HeLa snAP-seq data. **a**) Representative genome browser (IGV) view of peaks called in snAP-seq data. Black bars represents high-confidence peaks detected by MACS2 peak-calling. **b**) Overlap of high-confidence peaks determined by MACS2 software in cells treated with control and APE1 siRNA.

A stringent threshold of $p < 10^{-5}$ was set for calling peaks in each individual library using the corresponding input library as a control and only peaks that appear in at least three of four replicates were considered high-confidence. For control-siRNA treated cells, 14,110 high-confidence peaks were detected. 10,387 (74%) of these were also detected in cells treated with APE1 siRNA, where a total of 25,080 peaks were found (**Figure 4.9b**). The increase in number of peaks detected in APE1 deficient cells was consistent with expectations and is supported by reports that the number of AP sites is globally elevated when BER is compromised. These results suggest that whilst on the single-nucleotide level there was no strong accumulation of abasic sites detected, the distribution of abasic sites is

partially non-random and clusters weakly within genomic windows across a population of cells. The ability to detect AP peaks in BER competent cells indicates that some AP damage may be resistant to repair and can be better tolerated in particular regions of the genome. The additional peaks that appear upon knockdown of APE1 are therefore likely to be those associated with APE1 repair, which also cluster non-randomly across the genome.

Between the individual replicates libraries, pairwise overlap ratios between replicates range between 22% and 79% showing that results are moderately consistent. Variations between replicates may be further due to the stochastic nature of DNA damage.

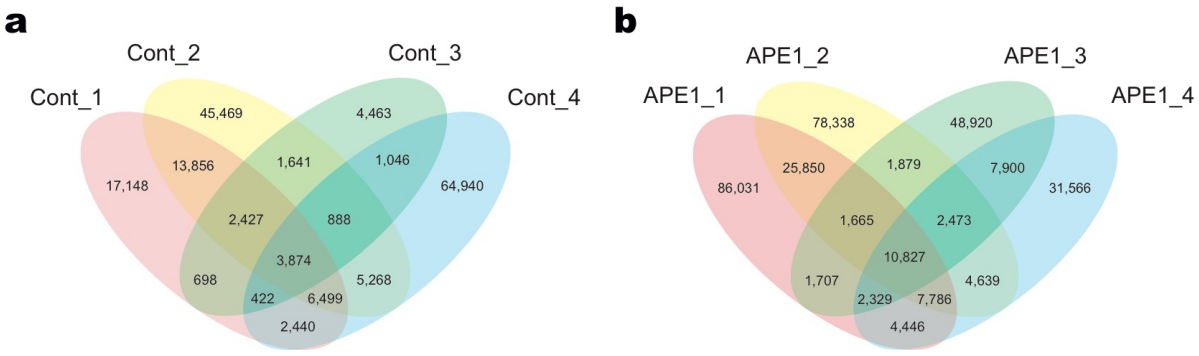


Figure 4.10: Overlap of snAP-seq peaks ($p < 10^{-5}$) between individual replicates in **a**) cells treated with control siRNA, and **b**) cells treated with APE1 siRNA.

4.2.4 Limit of detection of snAP-seq

The ability to detect individual nucleotides by enrichment using snAP-seq is ultimately determined by the limit of detection of the method. Rather than the global levels of a modification taken as an average across the genome in a pool of cells, enrichment-based sequencing methods are more likely to depend on the degree to which a modification is present at a given genomic location across the population of cells. Therefore, even for modifications at high global levels, if their distribution is stochastic between individual copies of the genome, they remain challenging to detect by enrichment. The spike-ins that have been included in each sequencing library show a strong enrichment of AP DNA relative to unmodified GCAT DNA (~200-fold), however, these values are based on a fully AP-modified DNA. In contrast, it is unlikely that across the population of cells from which the genomic DNA was extracted, the position of abasic sites will be fully consistent.

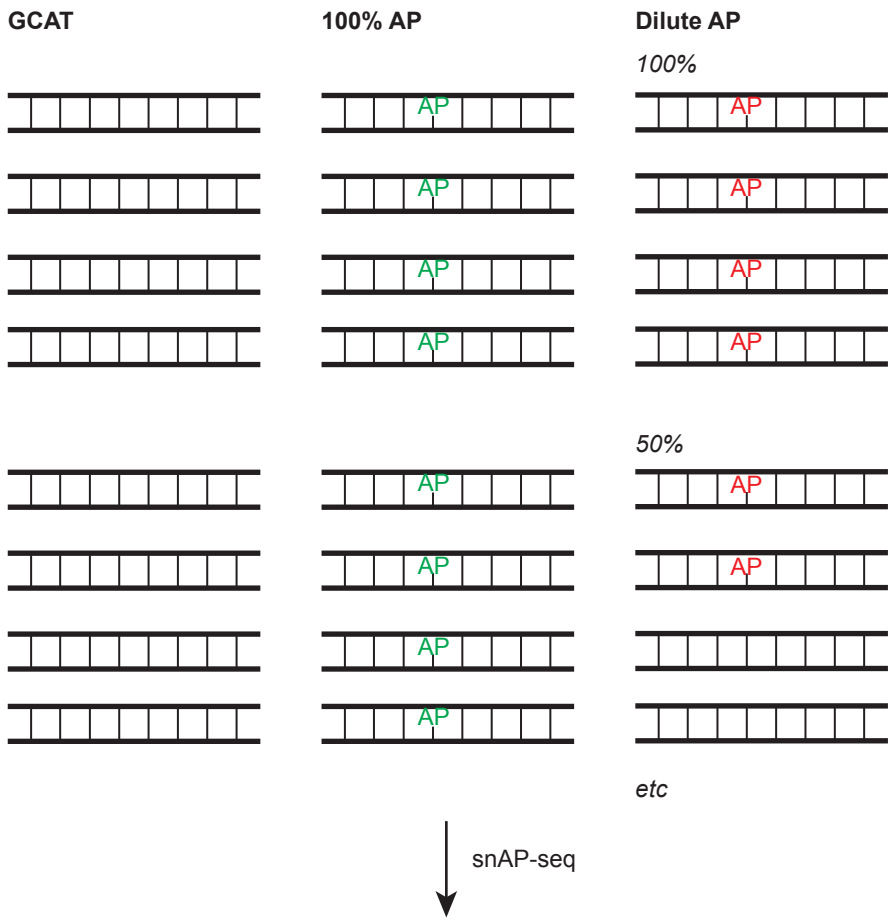


Figure 4.11: Diagram illustrating the DNA input for AP DNA dilution to investigate the limit of detection of snAP-seq. The total amounts of the three DNA models are kept consistently equal. One AP DNA is kept constant at 100% AP content, whilst the other is diluted using the same sequence where the AP site is replaced with T. Each library with varying AP/T ratio is then sequenced by snAP-seq.

To experimentally estimate the limit of detection of snAP-seq in terms of the percentage modification required at a given position within DNA to still observe an enrichment, a series of libraries were prepared in which an AP containing sequence (AP DNAx) was diluted with DNA of the same sequence in which the AP site was replaced with a thymine (**Figure 4.11**). GCAT DNA was also added as negative control to assess enrichments, at the same amount as total AP+T DNA. To avoid a scenario at low AP percentage where a library cannot be successfully generated due to lack of recovered material, a further AP model (AP DNA1) was included in each library at 100% modification. Six libraries with AP/T ratios varying from 100% down to 1% were prepared using snAP-seq and the same DNA samples were also used to prepare the six corresponding input libraries.

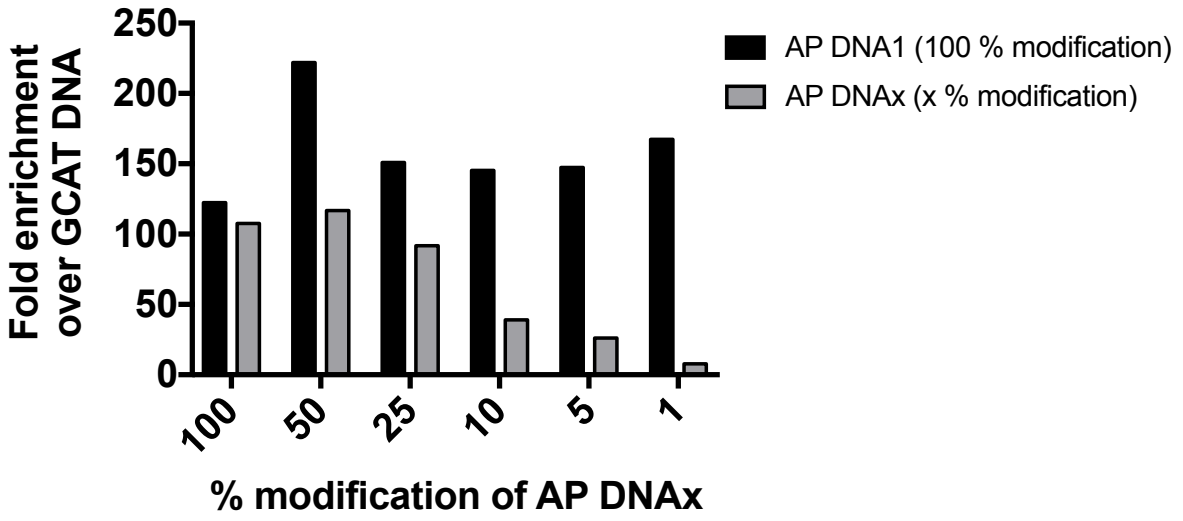


Figure 4.12: Enrichment of AP DNAs after snAP-seq. Sequencing reads aligning to the forward strand of each model were normalised against reads aligning to the reverse strand in the corresponding input library. Normalised read counts were then represented as an AP/GCAT ratio to show relative enrichment. Data from a single replicate are shown.

Sequencing reads from the six diluted snAP-seq libraries that aligned to each of the three model DNA sequences were analysed to determine enrichments. In each case, only reads aligning to the forward strand that bears a modification were considered. These were normalised against the number of reads for the unmodified, reverse strand in the corresponding input library to ensure equal overall representation of each sequence. The results show that the fold enrichment of AP DNA1, which was kept at 100% AP content throughout, fluctuates between the six libraries (**Figure 4.12**). This effect is most likely a result of technical variation. For AP DNAx, even at 1% AP/T an 8-fold enrichment relative to an equal input of GCAT could still be observed. The enrichments calculated here must be combined with the sequencing depth of genomic libraries to estimate a limit of detection. The sequencing depth of the HeLa libraries varied slightly between samples, with the lowest being 1.26X. With read lengths of 75 nucleotides, this equated to each nucleotide being covered on average 0.0168 times. Therefore, to observe a pile-up of at least two reads starting at the same position in a single replicate library, roughly 100-fold enrichment is required. From the spike-in dilutions, this would equate to somewhere between 25% and 50% modification. As four replicate libraries for each siRNA treatment were used, a further increase in the pile-up of combined reads may be expected at these percentages. It should be noted that enrichment-based sequencing approaches like snAP-seq are not quantitative, and the enrichment efficiency can depend on a number of factors including sequencing depth and

vary significantly between libraries and replicates. Therefore, these values provide only a rough estimate of the limit of detection.

Figure 4.13: Overlap of SMUG1-snAP-seq sites detected using all reads across two *L. major* libraries at a threshold of FDR < 10^{-10} (section 3.2.2), and a subset of reads obtained by sampling to represent the depth achieved for the HeLa libraries (1.26X) at the threshold of FDR < 0.05. Only sites in the forward strand of the reference *L. major* genome were analysed; results on the opposing strand are expected to be largely similar.

The relationship between sequencing depth and ability to detect enriched peaks in sequencing data has been long known. For ChIP-seq, it has been found that the number of peaks saturates with increased sequencing depth, with the maximum estimated to be reached around 40-50 million reads per library for a number of ChIP targets[255]. As snAP-seq is single-nucleotide resolution, the coverage required is expected to be higher as each read of 75 nucleotides is ultimately assigned to represent a single nucleotide position. To further investigate the limit of detection in a genomic context, the *L. major* libraries generated in Chapter 3 were used. For these libraries, a high depth of sequencing was used due to the small size of the *L. major* genome (33 Mb), which is around 100 times smaller than the human genome; sequencing at an average of 10X depth for these libraries was easily achievable at relatively low costs. Using the raw sequencing reads, the number of reads was sampled to a similar level as that achieved for the human libraries (1.26X). Data processing was then carried out in the same way on this subset of reads. Overall, the FDR values associated with sites decreases with sequencing depth. Therefore, at the same highly stringent FDR threshold of 10^{-10}, only 18 sites remain in the sampled dataset. However, the volcano plot remained strongly skewed towards positive $\log_2$ fold-change (SMUG1-snAP-seq vs. input), with 1,821 single-nucleotide sites detected in the forward strand with FDR < 0.05. Of the total high-confidence sites detected in the original dataset, 77% were also detected after sampling (**Figure 4.13**), suggesting that the majority of sites remain above the limit of detection at this lower sequencing depth. The extent to which 5-hmU occurs across populations of *L. major*

organisms at each specific genomic loci is not known, but these can be no higher than 100%. Therefore, these results suggest that at the depth to which the HeLa libraries were sequenced, it was expected that as a minimum, AP sites with quantitative representation are detectable.

4.2.5 Genomic analysis of snAP-seq sites

The lack of single-nucleotides at which AP sites accumulate suggests that on average across a population of cells, the expected ~1 per million dN sites[47] are largely not concentrated at single sites. However, the ability to detect peaks of enrichment suggests that instead, there are windows of the genome that are more susceptible to damage. The enrichment of these peaks relative to input is modest, at around 3-fold on average, but remains statistically significant at a relatively stringent *p*-value threshold of $p < 0.00005$. To further investigate the location of these enriched peaks, the distribution amongst genomic features was analysed. High-confidence peaks detected for cells treated with control siRNA, or APE1 siRNA were used, as well as peaks that were common to the two datasets. These sets of peaks were compared to sets of peaks of the same size shuffled at random across the reference genome. This allowed the fold enrichment to be calculated amongst the selected genomic features, which comprised promoters, defined as 1 kilobase (kb) upstream of transcriptional start sites, 5'- and 3'- UTRs, exons, introns and intergenic regions.

For cells treated with control siRNA, peaks were weakly enriched in intergenic regions and depleted in all other regions analysed (**Figure 4.14**). In contrast, peaks detected in cells treated with APE1 siRNA were enriched in regulatory regions including promoters, as well as 5'- and 3'- UTRs and exons ($q < 0.05$). This suggests that APE1 may be involved in removing DNA damage from these regulatory and transcribed regions. This trend was also in line with findings for the DNA damage marker, 8-oxoG, where levels were found to increase in promoters and 5'- and 3'-UTRs upon knockout of the repair enzyme OGG1 in mouse embryonic fibroblasts[243]. This effect was suggested to be due to an increased exposure of these regulatory elements to oxidative damage. Peaks in common to the two siRNA treatments were overrepresented in promoters, exons and intergenic regions whilst being depleted in 5'-UTRs and introns; however, the majority of peaks remained within intergenic regions (62%).

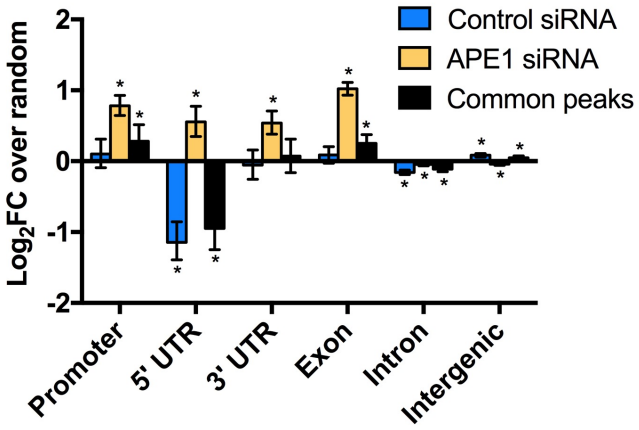


Figure 4.14: Relative enrichment of snAP-seq peaks in different genomic regions, expressed as $\log_2$ fold-change when compared to randomised sets of peaks obtained through simulation (N=10,000). The total high-confidence peaks detected in control peaks and APE1 peaks were analysed, as well as consensus peaks common to both datasets. Error bars represent 95% confidence intervals, *q < 0.05.

The location of peaks was also analysed in relation to chromatin accessibility and histone modifications. Chromatin accessibility datasets, generated by DNase-seq, FAIRE-seq and ATAC-seq also using HeLa DNA were obtained from the ENCODE or GEO database (**Table 7.2**). For FAIRE-seq and ATAC-seq, a small enrichment was seen in accessible regions of the genome for control cells, whilst no significant change from randomly shuffled peaks was seen for DNase-accessible regions (q < 0.05). A larger enrichment within accessible regions from all three datasets was seen for snAP-seq peaks in APE1 deficient cells (**Figure 4.15a**). This suggests that open-chromatin may be more susceptible to damage, particularly in the absence of BER. Accessible chromatin regions have previously been suggested to be more prone to the accumulation of a range of DNA damage products[256,257], with condensed chromatin suggested to be protected from damage.

Analysis of the distribution of peaks in relation to histone modifications revealed an enrichment of snAP-seq peaks in APE1 deficient cells within regions associated with activating histone markers (H3K27ac and H3K4me3) that was not present for control cells. Activating histone marks are largely associated with open-chromatin, and therefore these results are consistent with the enrichments found in accessible chromatin regions. Interestingly, an enrichment was also seen for AP peaks in both cell types in regions associated with repressive histone marks (H3K27me3 and H3K9me3), despite the general association of repressive histone modifications with closed-chromatin (**Figure 4.15b**). It has

been suggested that whilst regions of closed-chromatin are less accessible to DNA damaging agents, they can also be less accessible to DNA repair. For example, the repressive histone mark H3K9me3 has been associated with elevated genomic instability and mutation rates[258]. The relationship between DNA damage and chromatin landscape is therefore complex and the persistence of DNA damage products is dependent on multiple factors including the accessibility of both damage agents and repair machinery. The relative increase in AP peaks around activating histone markers when APE1 levels are reduced compared to control cells may suggest that this enzyme is most active in open and more accessible regions of the genome. Overall, the non-random distribution of AP peaks found across different chromatin states indicates that the accumulation of DNA damage can vary within the chromatin structure.

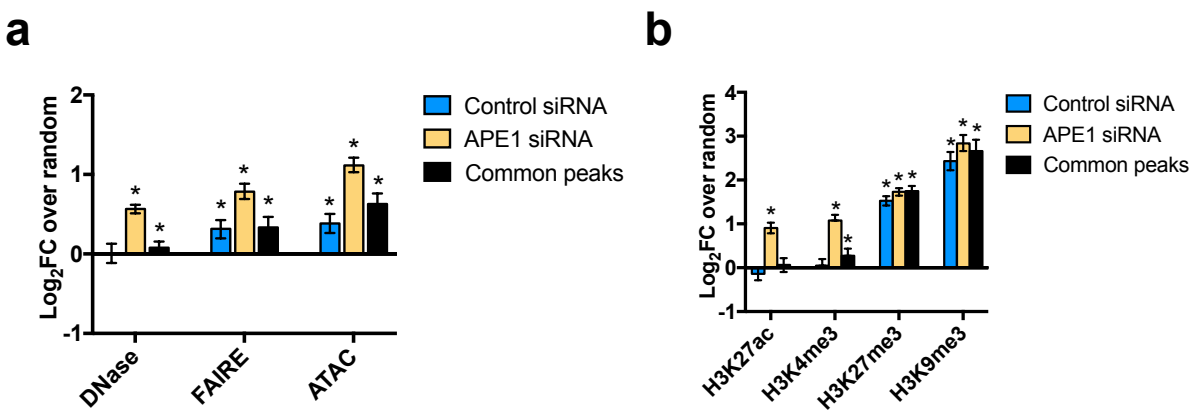


Figure 4.15: Distribution of snAP-seq HeLa peaks relative to **a**) chromatin accessibility and **b**) histone modifications expressed as $\log_2$ fold-change when compared to randomised sets of peaks obtained through simulation (N = 10,000). Error bars represent 95% confidence intervals. *q < 0.05.

Although no single-nucleotides of enrichment were called with significance from the snAP-seq data, it was possible to analyse the entire sequencing library to investigate whether there was a preference for the nucleotide immediately 5'- to the start of sequencing reads (position 0) globally across all reads. This assumes that despite the weak local accumulation of AP sites, there was a global enrichment of AP sites. This is supported by the enrichment observed within the spike-in DNA. For both control- and APE1 siRNA-treated cells, an enrichment in the purines, adenine and guanine, was observed that is supported by reports on the increased rate of depurination compared to depyrimidination observed during *in vitro* studies on DNA[112]. The variations observed in base identity at position 0 for input libraries is due to inherent bias during adapter ligation. A bias for cytosine, and depletion in purines has

previously been demonstrated in libraries that were similarly generated using dA-tailing overhangs to assist the ligation of adapters[259].

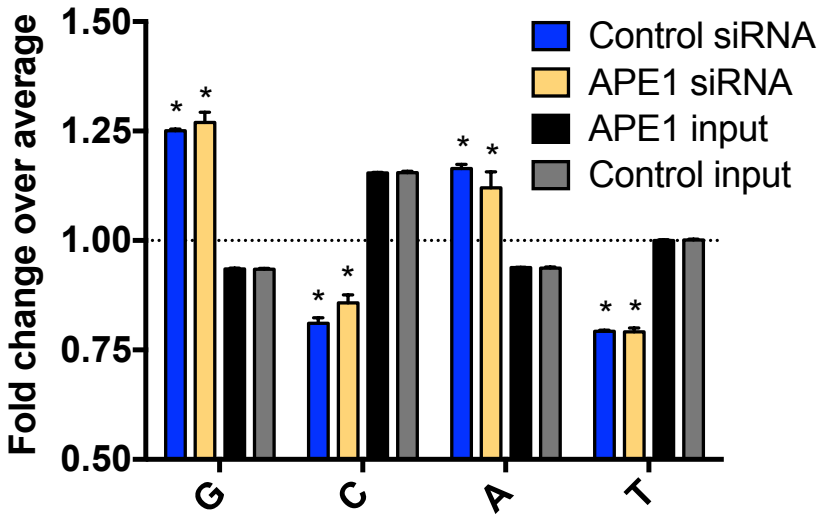


Figure 4.16: Relative enrichment of bases at the '0' position, directly 5'- to total read start sites for all genomic locations in each library. Mean and S.E.M of four replicate libraries are shown, *$p < 0.05$ (two-way ANOVA, Sidak's multiple comparisons test).

4.2.6 Association of snAP-seq peaks with gene expression

Out of the 14,110 snAP-seq peaks found in control cells, over 5,000 occur in close proximity to coding regions. Given that AP sites are roadblocks for both DNA and RNA polymerases, it is possible that they may influence transcription. To further investigate the potential biological effect persistent AP sites may have, the association between high-confidence snAP-seq peaks and gene expression was explored. RNA-seq data generated using HeLa cells were obtained from the ENCODE database. As this dataset corresponds to wild-type HeLa cells, only control siRNA snAP-seq data were used for analysis. The 14,110 snAP-seq peaks were separated into those that were located within a gene body and those that were not, and the corresponding gene expression was compared. Extended gene bodies were used here, which was defined as the gene body plus a 1 kb flanking region in either direction. There was no significant change in the expression of mRNA levels between genes that contained AP peaks and those that did not ($p < 0.05$) (**Figure 4.17**).

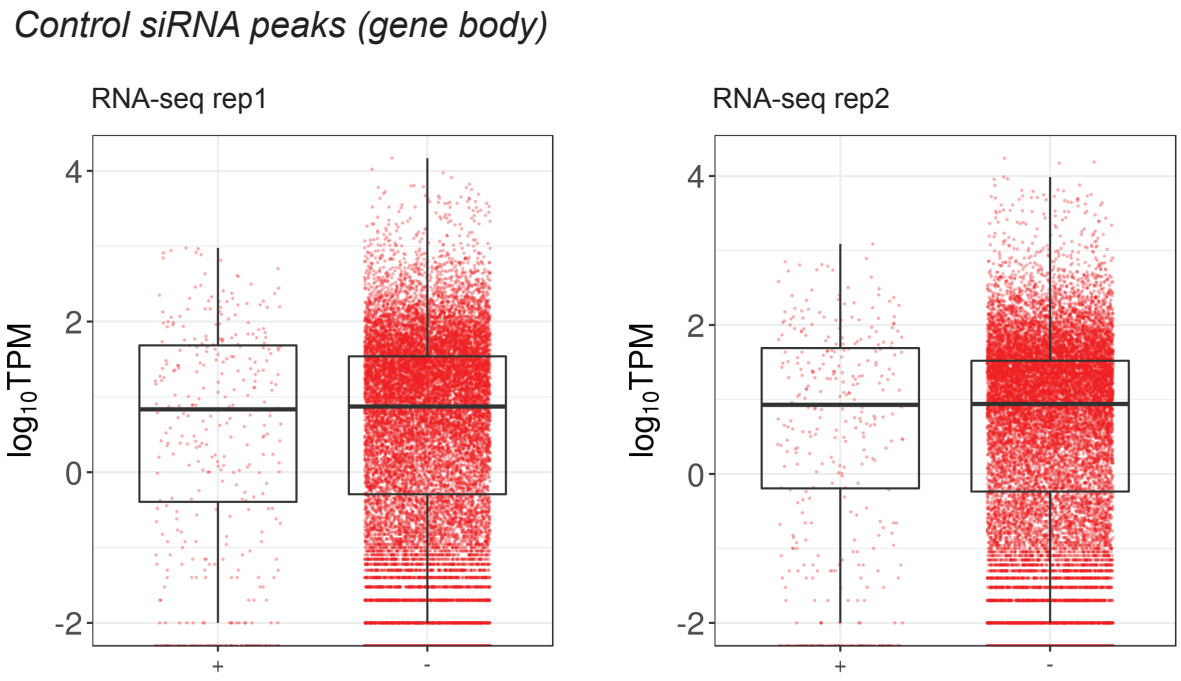


Figure 4.17: Gene expression shown as $\log_{10}$(transcripts per million reads), for genes that contain snAP-seq peaks within the extended gene body in control siRNA-treated cells (+), and genes that do not (-). Two replicates of RNA-seq were obtained from ENCODE and analysed separately. No significant change was observed for either replicate ($p < 0.05$). Box plots show first, second (median) and third quartiles, with whiskers representing 1.5 × the interquartile range.

The expression of genes in mammalian systems is also highly sensitive to changes specifically in promoter regions, such as the presence of epigenetic modifications including 5-mC[260]. mRNA expression was therefore also analysed for peaks that occur in promoters, defined here as regions 1 kb upstream of transcription start sites. Genes that contained snAP-seq peaks in their promoters were found to have on average a lower level of expression compared to those that do not (**Figure 4.18**), which was consistent across two RNA-seq replicates ($p < 0.05$). These results suggest that the accumulation of AP sites may be associated with lower levels of transcription, however, the overall number of genes identified that were associated with AP peaks in promoter regions was relatively small (88). Oxidative DNA damage has previously been associated with transcriptional downregulation[261]. Abasic sites generated during base excision repair of 8-oxoG have also been shown to lead to gene inactivation. However, there is also contradicting evidence in specific examples that suggest under hypoxic conditions, for example, 8-oxoG and BER can be associated with transcriptional upregulation[95]. The snAP-seq results here suggest that globally, there is a weak association between AP damage and genes that are lowly transcribed.

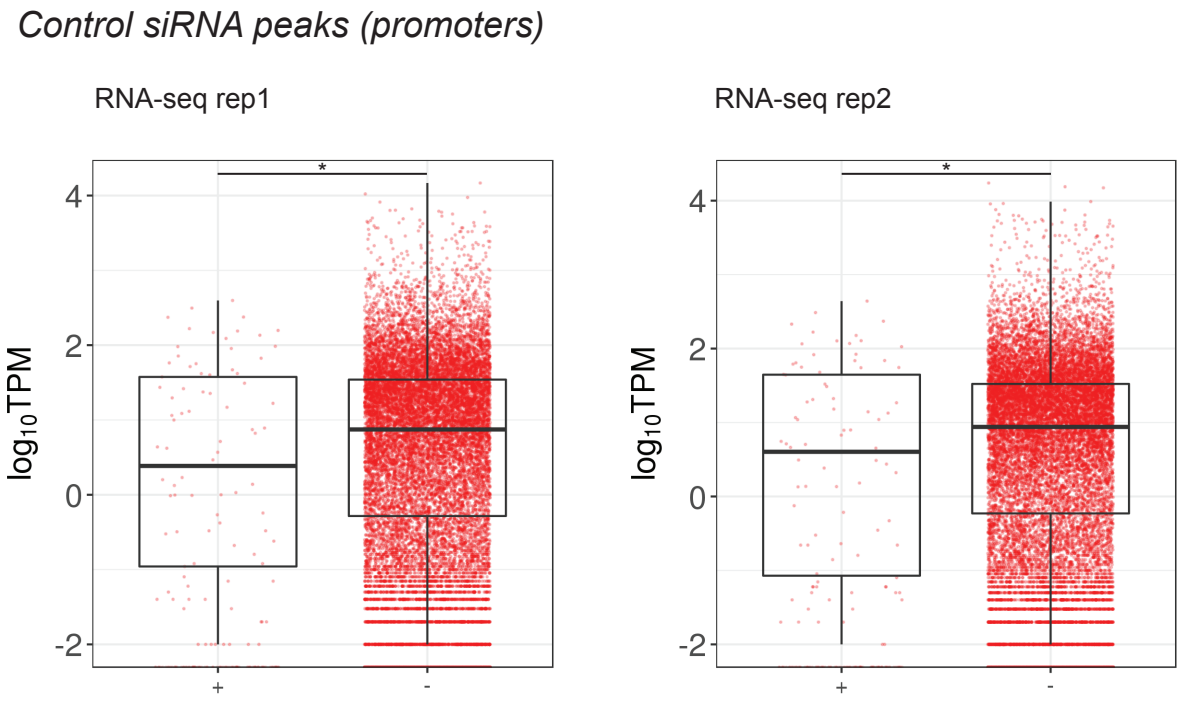


Figure 4.18: Gene expression shown as $\log_{10}$(transcripts per million reads), for genes that contain snAP-seq peaks within promoter regions (1 kb upstream of transcription start sites) in control siRNA-treated cells (+), and genes that do not (-). Two replicates of RNA-seq were obtained from ENCODE and analysed separately. *p < 0.05. Box plots show first, second (median) and third quartiles, with whiskers representing 1.5 × the interquartile range.

An existing RNA dataset in the form of microarray analysis was also used to analyse gene expression[262]. These data have been generated using HeLa cells with and without APE1 knockdown and were therefore also suitable for analysing the APE1 knockdown snAP-seq peaks. For control cells, it was found that in a similar trend to the RNA-seq data, a small decrease was seen in normalised expression for genes that contained snAP-seq peaks within promoters as well as extended gene bodies compared to genes without snAP-seq peaks; however, this decrease was no longer statistically significant (p < 0.05) (**Figure 4.19**). It should be noted that although referred to here as genes, each datapoint within the microarray data corresponds to a probe targeting a mRNA; in some cases, multiple probes may correspond to the same gene. For APE1 knockdown, a small increase in expression levels was seen instead for genes with peaks in extended gene bodies, as well as in promoters compared to genes that lacked snAP-seq peaks (p < 0.05). This suggests that in the absence of efficient DNA repair, abasic sites may be associated with more highly transcribed genes, in contrast to when BER is fully active. Although polymerases generally stall at AP sites, it is possible that when BER is compromised, highly transcribed genes become more prone to the accumulation of DNA damage. In mammalian systems, accessible regions of the genome

such as those undergoing replication have been shown to accumulate aldehyde-reactive damage, with highly transcribed regions suggested to display a similar trend[127]. In yeast, an increase in genomic instability and mutations was observed around highly transcribed genes when BER was disrupted[263]. These observations support the hypothesis that highly transcribed genes may be more susceptible to DNA damage specifically in the event of aberrant DNA repair.

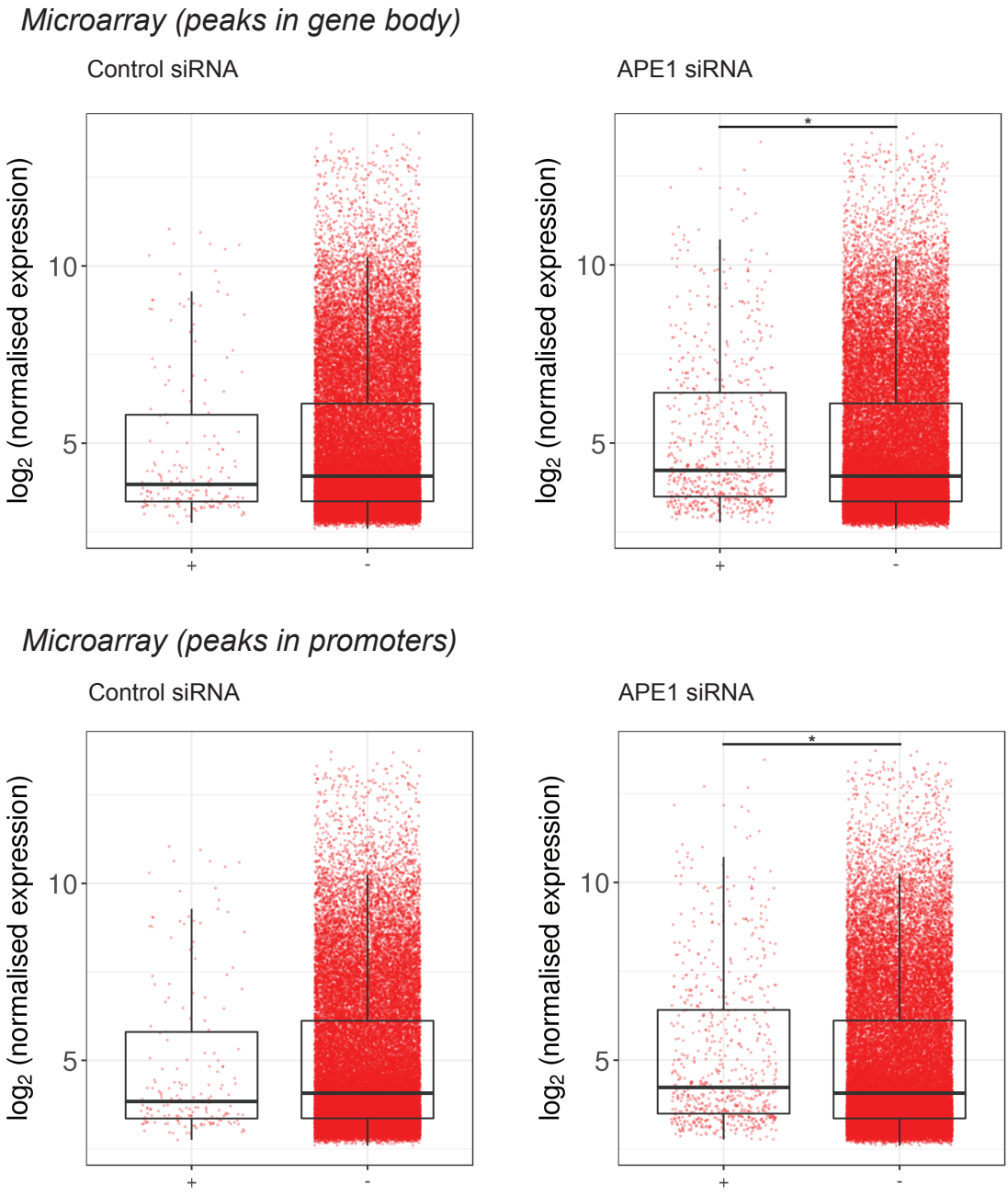


Figure 4.19: Normalised gene expression from microarray data for genes that contain snAP-seq peaks in the extended gene bodies (top) or promoters (bottom) (+), and those without (-), for control and APE1 deficient HeLa cells. *$p < 0.05$. Box plots show first, second (median) and third quartiles, with whiskers representing 1.5 × the interquartile range.

Finally, to assess how the direct loss or gain of an AP site may be influencing gene expression, differential expression analysis was carried out by comparing the relative expression of genes with and without APE1 knockdown. snAP-seq peaks were separated into those that were unique to the control siRNA treatment and therefore lost upon APE1 knockdown, unique to APE1 siRNA treatment and therefore gained upon APE1 knockdown, or those common to both treatments. Genes were divided into those that contained a snAP-seq peak of each class in their promoters and those that did not, and the fold-change in normalised expression (APE1 knockdown vs. control) was compared. The knockdown of APE1 did not lead to significant changes in transcription levels for any of three classes of peaks (**Figure 4.20**) ($p < 0.05$). These findings were consistent for peaks that fall within extended gene bodies also. Although genes with snAP-seq peaks upon APE1 knockdown were found to be on average slightly more highly transcribed than those without AP damage, locations that gain an AP peak after knockdown of APE1 do not appear to be directly associated with either transcription upregulation or downregulation.

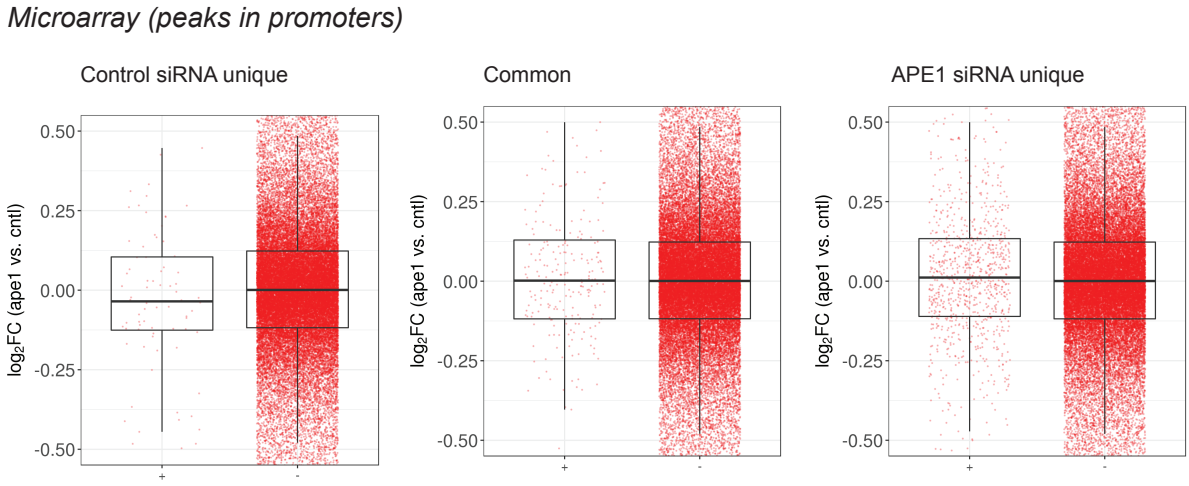


Figure 4.20: Fold-change of normalised gene expression in APE1 knockdown HeLa cells compared to control cells. Genes were separated into those that do (+) and do not (-) contain snAP-seq peaks that are unique to control or APE1 siRNA treatment, and those that are common to both cell types. No significant change is observed for any of the three conditions ($p < 0.05$). Box plots show first, second (median) and third quartiles, with whiskers representing 1.5 × the interquartile range.

Whilst a small but significant decrease in mRNA expression levels obtained by RNA-seq was seen in control cells for genes in which snAP-seq peaks occur within the promoter regions compared to genes that lack snAP-seq peaks, the number of genes that satisfy these criteria was small. Furthermore, this decrease was no longer statistically significant when

using an RNA dataset obtained instead by microarray analysis. In contrast, a small increase was seen in expression for genes associated with snAP-seq peaks with APE1 knockdown compared to genes lacking AP peaks. This may suggest that highly transcribed genes are more susceptible to DNA damage in the absence of efficient damage repair, however, the global effect of this was also relatively small. When gene expression was compared with and without APE1 knockdown, there was no direct association found between AP peaks and changes in gene expression. Overall, there is no clear correlation between the expression of genes as measured by steady-state mRNA analysis and the location of AP accumulation in peaks. On a global level, AP sites do not appear to strongly influence gene expression.

4.3 Conclusions and future directions

In this chapter, the distribution of endogenous abasic sites was studied in the human genome. HeLa cells were chosen for this study, where APE1 deficient cells, generated by siRNA-mediated knockdown were used alongside control cells that were treated with non-targeting siRNA. Controls were carried out using both synthetic and genomic DNA to ensure that the steps required for isolating and preparing DNA for snAP-seq do not alter the AP landscape. The data from two biological replicates of snAP-seq indicated that AP sites do not accumulate strongly at the single-nucleotide level for either control or APE1 depleted cells. An increase in global AP levels was expected upon knockdown of APE1, which has previously been demonstrated using different quantification techniques[47,245]. However, it is possible that the degree of knockdown achieved here (~90%) was insufficient to observe a significant site-specific accumulation of damage. Therefore, future work could focus on the complete removal of APE1 by knockout, or further knockdown of alternative repair enzymes including APE2.

Although AP accumulation at specific single-nucleotide sites was not detected, peaks were detectable in snAP-seq libraries that were not present in input libraries. This suggests that a degree of stochasticity may be present for AP sites, such that they cluster within small windows across a population. Peaks were identified in both control and APE1 deficient cells, where a high degree of overlap was seen between the two cell types, likely due to persistent AP sites that appear to be resistant to repair by BER at steady-state. An additional 14,693 peaks were found upon knockdown of APE1 protein, consistent with expectations of further AP accumulation when BER is disrupted. Within the genome, these peaks were not distributed randomly. Whilst the majority of peaks in cells treated with control siRNA were

intergenic, an enrichment was seen in regulatory regions including promoters, UTRs and exons for cells treated with APE1 siRNA. Peaks were also analysed in relation to the expression of mRNA. There was some indication that peaks detected in control cells were associated with lowly expressed genes, whilst an association of AP peaks with genes with higher transcriptional levels was detected in APE1 depleted cells. However, globally both of these effects were relatively small. Further experiments could focus on the analysis of nascent RNA, to more directly explore the influence of AP sites on RNA expression that may not be detectable using steady-state measurements.

The limit of detection of snAP-seq is likely to be dependent both on the extent to which a modification occurs at a specific genomic location across a population of cells, as well as the depth of sequencing achieved. For the depth at which libraries were sequenced here, it was estimated that between 50-100% modification may be required at a given site in order to be detected. In future work it may therefore be worthwhile to either increase the efficiency of enrichments by further optimisation of the method, or to sequence at greater depth.

Future work in this area could also focus on the exploration of AP sites in other cell lines or tissues. An association of AP damage with cancer has been suggested[133], whilst functional mutants of APE1 have been detected in the human population and associated with elevated risks of cancer[264]. HeLa cells originate from tumour samples; however, the heterogeneity of these cells may be problematic for enrichment sequencing. It may also be interesting to explore differences between cancer and non-cancer cell lines, to study the significance of AP sites and DNA damage during the development and progression of cancer.

Chapter 5

Mapping uracil in genomic DNA

5.1 Background

Uracil, one of the four canonical nucleobases in RNA, is largely absent in DNA and replaced instead with thymine. Within genomic DNA, uracil is a rare base modification that can be derived from the deamination of cytosine. Under aqueous conditions, cytosine deamination may occur spontaneously as a product of DNA damage, whilst this reaction can also be catalysed by the apolipoprotein B mRNA editing enzyme catalytic polypeptide (APOBEC) family of enzymes. Within this family, AID is a key enzyme that is active during the process of antibody diversification within B cells in mammals, involving the somatic hypermutation and class switch recombination pathways[77,78,265]. A number of other APOBEC enzymes are also involved in the inhibition of retroviruses, by catalysing the mutagenesis of cDNA from retroviruses upon infection of the host before assimilation into host DNA[266], as well as reducing the activity of endogenous retroelements within mammalian genomes[267].

Due to the change in Watson-Crick base-pairing upon deamination of cytosine, uracil can be considered a mutagenic base. Propagation of the C to T transition mutation outside of controlled pathways such as antibody diversification is prevented by a number of mechanisms. Four distinct glycosylases that recognise uracil and initiate base excision repair are known in mammals; UNG, SMUG1, TDG and MBD4. UNG is highly specific for uracil and is not active on other nucleobases with the exception of the unnatural modification 5-fluorouracil[238]. Two isoforms of UNG exist in mammals; UNG1 is predominantly localised in the mitochondria whilst UNG2 is nuclear[268]. The UNG enzyme is active on dsDNA as well as ssDNA and therefore does not require uracil to be part of a mismatch. The other uracil glycosylases have a range of pyrimidine substrates in addition to uracil[49,226]. On the triphosphate level, dUTPase dephosphorylates deoxyuridine triphosphate (dUTP) to deoxyuridine monophosphate (dUMP), thus removing uracil from the triphosphate pool to prevent misincorporation during DNA synthesis[269]. Deleting or mutating many of these enzymes involved in removing genomic uracil has been associated with an increased risk in cancers including lymphomas[80,270].

Measurements of the global levels of genomic uracil in mammalian systems vary by up to 4 orders of magnitude, from less than 1 to over 1000 per million dN. The large variability in measurements is partially due to the technical challenges in quantifying this base. For LC-MS/MS measurements, DNA must first be digested enzymatically into individual nucleosides before mass spectrometry can be carried out. During this process, uracil has been identified as an artefact arising from cytosine deamination. The rate of deamination is faster in free nucleotides and ssDNA compared to dsDNA[271] and the process of DNA digestion is particularly susceptible to uracil formation, thereby inflating measurements. Accuracy can be increased by carefully controlling the conditions under which the enzymatic steps are carried out to minimise the rate of deamination. Alternatively, the UNG enzyme may be used to recognise uracil in quantification methods that do not rely on the digestion of DNA into nucleosides. For example, abasic sites generated by treatment of genomic DNA with UNG have been measured and compared to measurements carried out in the absence of UNG treatment to calculate basal uracil levels[272]. Whilst the direct measurement of abasic sites using ARP can be unreliable due to the cross-reactivity of this probe in DNA, the subtraction of signal with and without UNG treatment provides a control for this drawback. The uracil nucleobase released from genomic DNA upon UNG treatment has also been measured using LC-MS/MS[273]. As DNA is kept in the more stable, double-stranded helical structure, the rate of artefactual deamination is expected to be greatly reduced compared to nuclease digestion. However, potential drawbacks include the possibility of incomplete excision activity of UNG that may be masking uracil levels. At the lower end of the range, uracil levels have been reported at 0.2 per million dN in mammalian genomes[82].

In addition to the role of uracil as a key intermediate during antibody diversification, there is emerging evidence that this base may also be involved in regulating the levels of 5-methylcytosine in eukaryotic systems. Whilst the installation of 5-mC, a key epigenetic modification that can control gene expression in a number of organisms, is understood to be mediated by the DNMT family of enzymes, the mechanism for its removal remains unclear. A number of pathways have been proposed for this demethylation process, the most well studied being the active demethylation pathway involving the oxidation of 5-mC by the TET enzymes followed by base excision by TDG to initiate DNA repair by the BER pathway[44,46,196] (**Figure 5.1a**). During embryonic development, a wave of demethylation occurs shortly after fertilisation in which global methylation levels in the paternal DNA are reduced genome-wide. This demethylation of the paternal genome has been observed to occur in two distinct phases, however, the exact mechanisms involved are not well understood[51]. TET3-mediated removal

of 5-mC through oxidation was found to only occur during the second phase of demethylation that is coincident with DNA replication. In contrast, TET3 oxidation of 5-mC was not found to be involved in the first phase. Interestingly, AID was also implicated during this second phase of demethylation in a TET3-independent pathway. Although AID can act directly on 5-mC to generate thymine, deamination activity is 5-10-fold lower on 5-mC compared to canonical cytosine[274]. As such, it was proposed that in the event that AID catalyses the deamination of cytosine into uracil nearby 5-mC sites, repair of the resultant uracil by long-patch BER may result in overall demethylation (**Figure 5.1b**). During long-patch BER, DNA synthesis is able to extend 2-13 nucleotides downstream of the initial deamination site. Therefore, when AID deamination occurs within this short window upstream of 5-mC, restoration of unmethylated cytosine is possible without direct enzymatic activity on 5-mC. This processive demethylation pathway appears to be distinct from active demethylation. Knockout of AID leads to decreased levels of 5-mC demethylation during the second phase of demethylation whilst 5-hmC levels remain unchanged, suggestive of a 5-hmC independent pathway. The mechanism of demethylation during the first phase remains unclear.

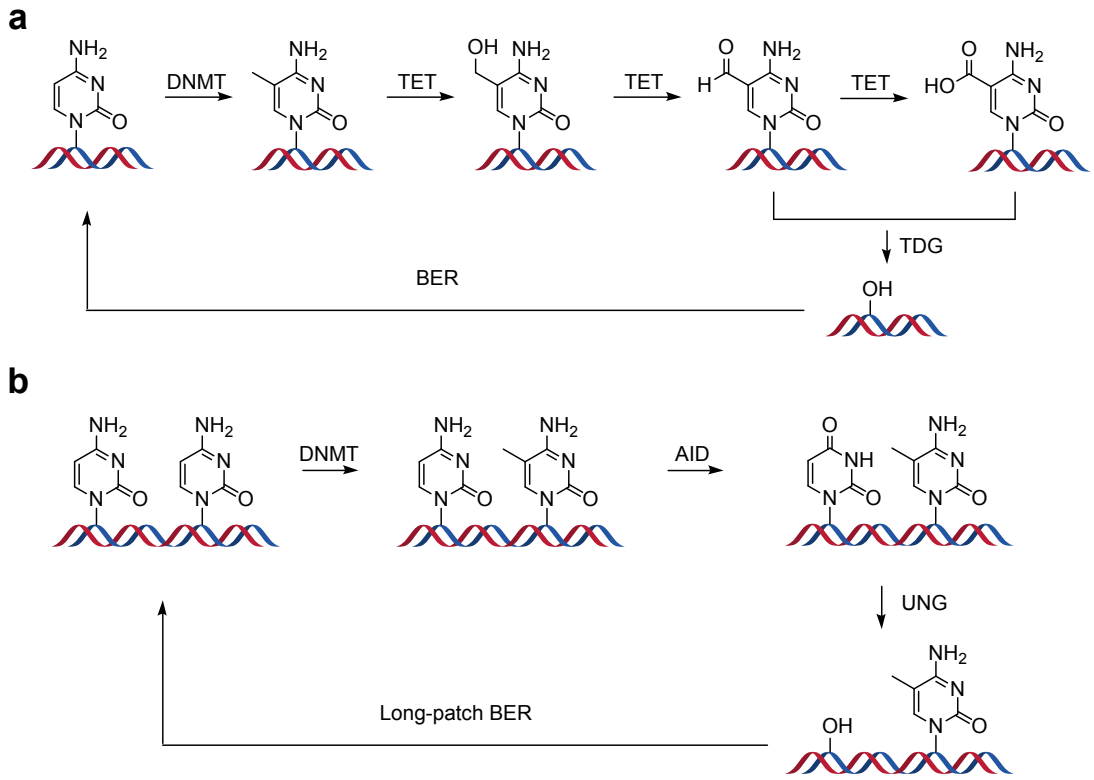


Figure 5.1: Proposed demethylation pathways for the removal of 5-methylcytosine. **a**) The active demethylation pathway, involving the successive oxidation of 5-mC followed by TDG excision activity to initiate DNA repair by the BER pathway[196]. **b**) The processive demethylation pathway, dependent on the deamination of cytosine by AID followed by the repair of uracil by the long-patch BER pathway[51,52].

In vitro studies using isolated genomic DNA and xenopus egg extracts have demonstrated that DNA demethylation can be induced by AID activity, and this effect was reduced in the presence of the UNG inhibitor UGI[52]. This process was not specific to the demethylation of 5-mC and a similar result was seen for methylation sites at 6-mA, suggesting that the involvement of AID is likely to be at canonical bases and not directly targeted at 5-mC. Furthermore, expression of an AID-GAL4 fusion protein that was targeted to GAL4 binding sites in mice induced DNA demethylation of the paternal genome in zygotes around the expected binding site, which was not seen when a catalytically inactive mutant of AID was used instead. Together, these results provide support for the hypothesis of processive demethylation, specifically in the context of paternal demethylation during embryonic development. This pathway has the potential to be more efficient than active demethylation, requiring only a single deamination reaction to initiate multiple downstream demethylation events. Therefore, in scenarios where methylation levels drop substantially on a global scale, this more efficient pathway may complement active demethylation.

The AID enzyme is active largely on ssDNA. Whilst some other deaminases in the APOBEC family can also be involved in the deamination of cytidine in RNA, this activity has not been demonstrated for AID. Human AID targets the consensus sequence WRC, where W is A or T and R is a purine[275]. During somatic hypermutation, mutations are introduced in hotspots around the RGYW motif into the variable regions of Ig antibodies[276]. The reverse complement of this motif is WRCY, which falls within the AID sequence target. Specifically, the AGCT sequence is an example of this motif, occurring in high density around Ig loci. AGCT is also enriched in the switch regions. The excision of uracil generated at this palindromic sequence has been associated with strand breaks at the resultant abasic sites, where end-joining of the double-strand breaks generated this way then allows class-switch recombination to occur[277]. Although the motif favoured by AID is concentrated around Ig loci, AID activity at non-Ig loci is also known. The aberrant expression of AID has been associated with deamination outside of Ig regions, which can lead to an increased risk of lymphoma[278]. As such, the activity of AID in pathways independent of antibody diversification is possible.

Whilst processive demethylation is an interesting hypothesis that may be important during embryonic development in addition to active demethylation, the significance of this pathway has not yet been fully determined genome-wide. If demethylation does progress significantly through this pathway, uracil should be generated and levels may be expected to be elevated above basal levels. In particular, cytosine deamination sites must be proximal to 5-mC sites and therefore the ability to detect uracil at high resolution is particularly useful in

investigating this pathway. As such, the main objective of this chapter was to sequence uracil at single-nucleotide resolution using UNG-snAP-seq to further investigate the role of processive demethylation during embryonic development. Mouse embryonic stem cells were used for this purpose. Whilst zygotes are a more sophisticated model, the amount of DNA obtainable from zygotes is limited and is not suitable for use with UNG-snAP-seq. The role of key components during the proposed pathway, namely AID and UNG, were explored by sequencing uracil when the cellular levels of these enzymes were altered.

5.2 Results and discussion

This project was carried out in collaboration with Dr Sergio Martínez Cuesta, Balasubramanian group, who performed the bioinformatical analysis of sequencing data generated in this chapter.

5.2.1 Mapping uracil in the *E. coli* genome

Before investigating the genomic distribution of uracil in the mouse genome, uracil was first explored in the *Escherichia coli* strain CJ236 where global uracil levels are much higher. The CJ236 strain contains key mutations in two genes involved in the removal of uracil from DNA; *ung* and *dut*. Ung is the primary uracil glycosylase in this organism and excises uracil from genomic DNA, whilst *dut* codes for dUTPase, an enzyme that dephosphorylates dUTP to prevent the misincorporation of uracil during DNA replication. Deactivation of these two enzymes leads to an elevation in global uracil content, however, like many measurements of genomic uracil, reports vary depending on the method of detection used. In CJ236, this ranges between 3,000-8,000 per million dN[272]. In comparison, in the JM105 strain where both Ung and dUTPase enzymes are active, basal uracil levels were around 1 per million dN. When the *ung* gene was deactivated in the JM105 strain, uracil levels increased to around 20-30 per million dN. The high levels of uracil in CJ236 are due to the combined effects of increased uracil misincorporation during DNA replication and an inability to remove these sites. As proof of concept for UNG-snAP-seq, the distribution of uracil was first studied in this *E. coli* strain where uracil levels are high and should be readily detectable[279]. This also allowed optimisation of the data analysis strategy before exploration of the lower uracil levels in mESCs.

CJ236 *E. coli* was grown in the recommended media (LB broth) and harvested without further manipulation. Total DNA was extracted from cells, and two biological replicates of UNG-snAP-seq were carried out. The UNG treatment was kept relatively short at 2 h to reduce the risk of further DNA damage and deamination. This treatment was sufficient to quantitatively excise uracil from a short ssODN (section 2.2.1). Sequencing reads were aligned to a reference *E. coli* genome (K-12 MG1655) and the read counts at individual genomic loci were compared between the UNG-snAP-seq and input libraries. Sites detected corresponding to positive enrichments (positive $\log_2$ fold-change, UNG-snAP-seq vs. input) were possible sites of uracil accumulation, whilst negative enrichments (negative $\log_2$ fold-change) where an accumulation was seen preferentially in the input library enrichment were considered to be noise. Comparison of the FDR values associated with the detected sites revealed that it was not possible to set a threshold at which positive enrichments were seen without negative enrichments. However, at FDR < 0.05, a strong skew towards positive $\log_2$ fold-change was observable and 38,455 UNG-snAP-seq sites were called with positive enrichment, compared to 534 sites (1.4%) with negative enrichments (**Figure 5.2**).

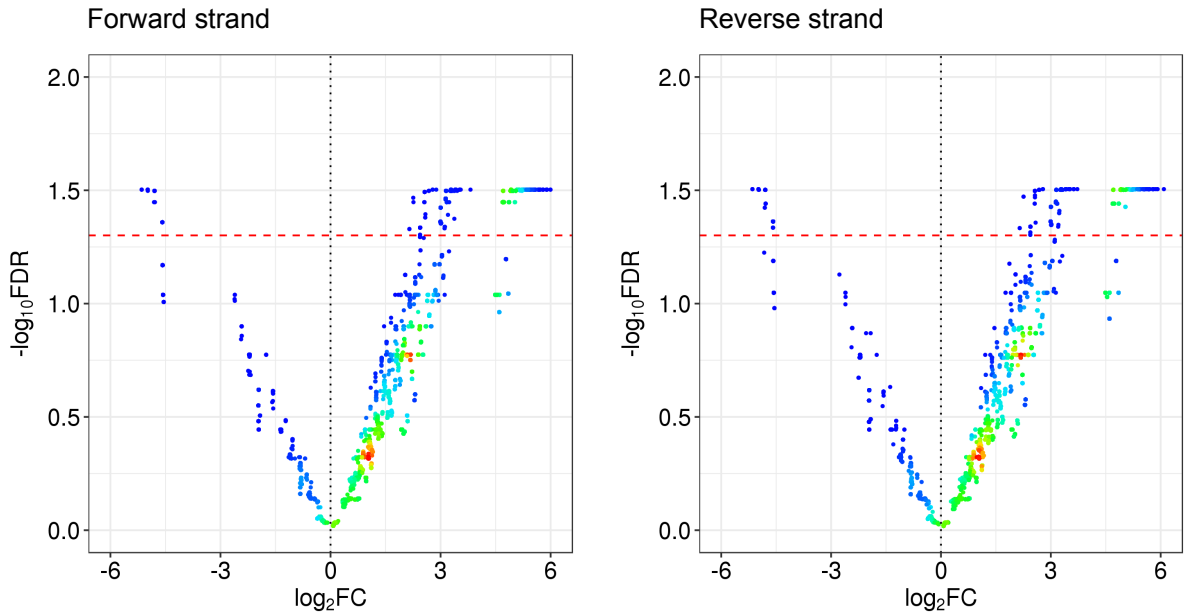


Figure 5.2: Volcano plot representing genomic loci with differential coverage in UNG-snAP-seq and input libraries and corresponding FDR values in *E. coli* strain CJ236. Reads were split between those aligning to the forward and reverse strands based on the reference *E. coli* genome (K-12); the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (UNG-snAP-seq vs. input libraries). Analysis was carried out on two biological replicates in parallel.

When enrichments were ranked instead by *p*-value, it was possible to use a similar approach to that taken in Chapter 3 to set a *p*-value at which no false positives in the form of negative enrichments were detectable. The site with the smallest *p*-value associated with negative enrichment was at $p = 0.000628$ (**Figure 5.3**). Beyond this threshold, 5,505 sites can still be detected with positive enrichment. The sites detected at this stringent threshold were considered high-confidence UNG-snAP-seq sites, whilst the FDR < 0.05 threshold was more relaxed and likely to capture more weakly enriched loci.

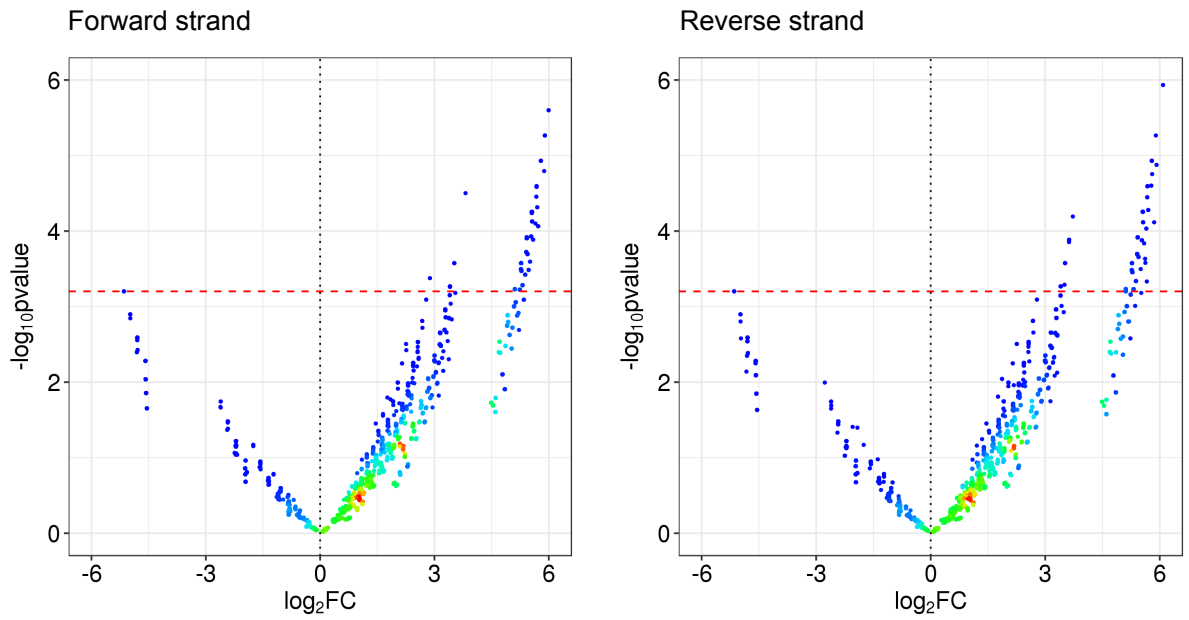


Figure 5.3: Volcano plot representing genomic loci with differential coverage in UNG-snAP-seq and input libraries and corresponding *p*-values in *E. coli* strain CJ236. Reads were split between those aligning to the forward and reverse strands based on the reference *E. coli* genome (K-12 MG1655); the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (UNG-snAP-seq vs. input libraries). Analysis was carried out on two biological replicates in parallel.

Given that the global level of uracil in CJ236 is up to 20 times higher than 5-hmU in *L. major*, the reduction in glycosylase-generated snAP-seq signal is somewhat surprising. These results highlight the fact that the strength of signal in snAP-seq is dependent not on the global levels of AP sites, but rather the localised accumulation across a population at a given genomic position. Despite the high abundance of uracil in this *E. coli* strain, the misincorporation of uracil appears to occur throughout the genome. Accumulation is not entirely random, as suggested by the ability to detect UNG-snAP-seq sites, however, the extent of site-specific accumulation is much weaker than that observed for 5-hmU in *L. major*.

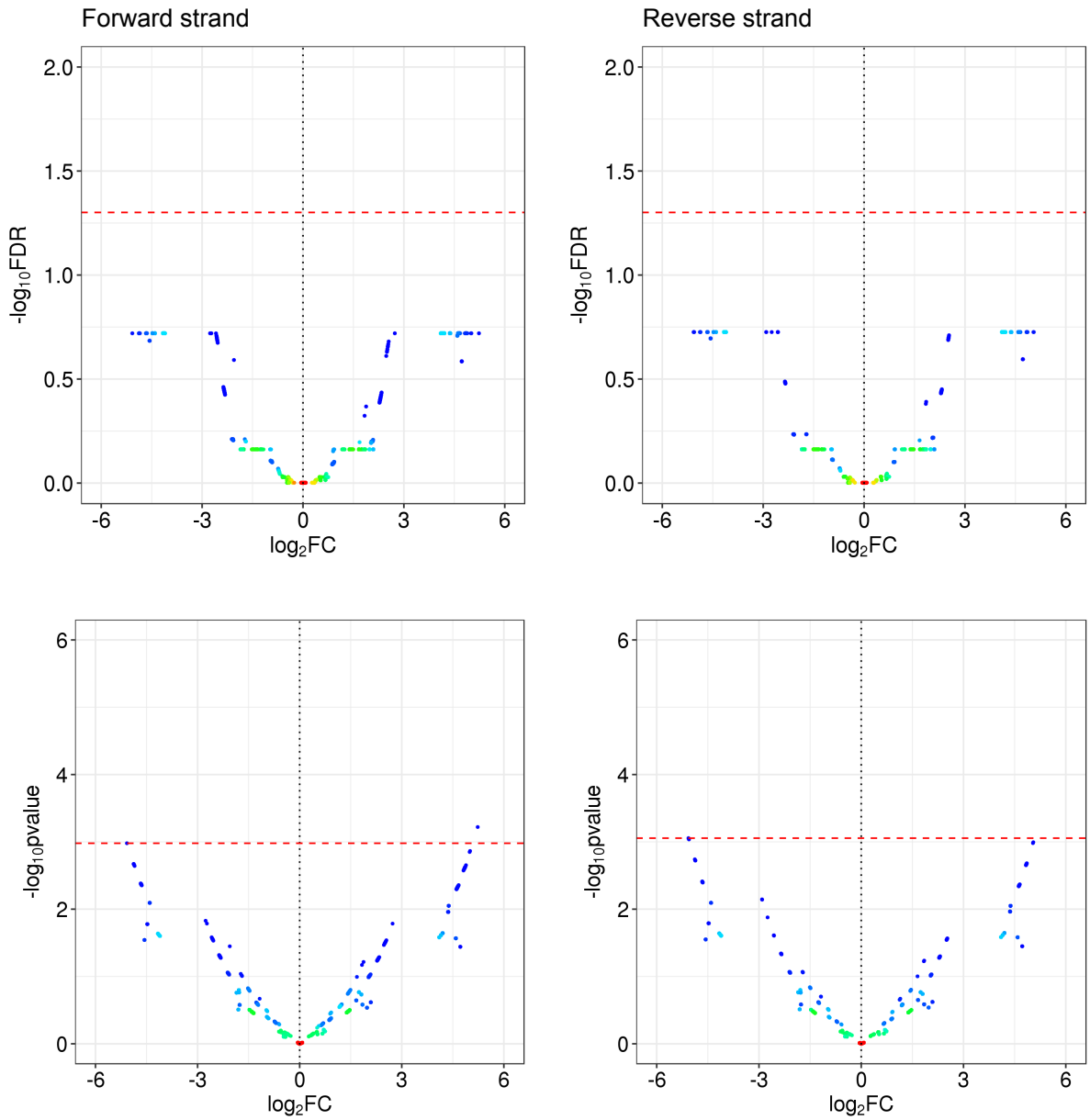


Figure 5.4: Volcano plot representing genomic loci with differential coverage in snAP-seq and input libraries and corresponding FDR or *p*-values in *E. coli* strain CJ236. Reads were split between those aligning to the forward and reverse strands based on the reference *E. coli* genome (K-12 MG1655); the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (snAP-seq vs. input libraries). Analysis was carried out on two biological replicates in parallel.

In contrast to a number of other glycosylases that have multiple possible substrates, UNG has a high specificity for uracil. The only known exception is the ability of this enzyme to also excise 5-fluorouracil, however, this unnatural nucleobase is not expected to occur significantly in genomic DNA without the introduction of fluorine labels. The use of UNG to

study uracil in genomic DNA has also been demonstrated previously with high selectivity[272,273], and therefore the signal generated by UNG-snAP-seq is expected to be specific for uracil. A further possible source of signal in UNG-snAP-seq data is from endogenous abasic sites. To assess the contribution of non-UNG derived abasic sites to the sites detected here, snAP-seq without UNG treatment was carried out. At FDR < 0.05, no sites were detected after snAP-seq, whilst a single site was detected at p < 0.000628 (**Figure 5.4**). These results suggest that the UNG-snAP-seq sites detected at both thresholds are strongly dependent on UNG excision activity.

5.2.2 Genomic analysis of UNG-snAP-seq sites in the *E. coli* genome

The sequence identity of sites called at either the more relaxed FDR threshold (< 0.05), or the high-confidence sites using the stringent p-value threshold (< 0.000628) revealed a strong enrichment for T (**Figure 5.5**). In CJ236, uracil is largely expected to be derived from misincorporation of uracil and should therefore correspond to thymine in the reference genome due to the similarities in Watson-Crick base pairing between these two bases.

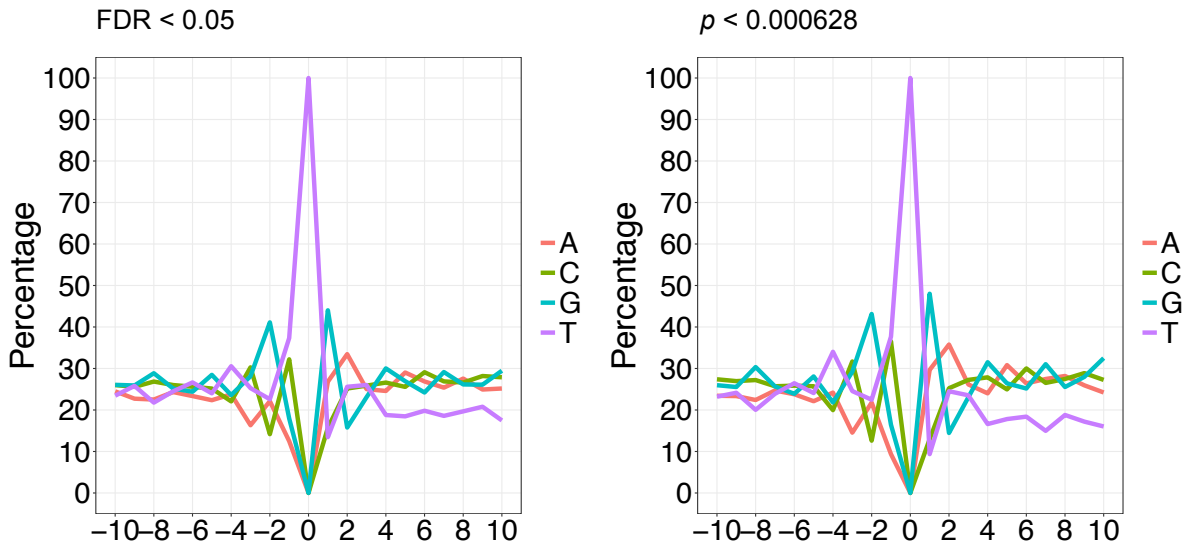


Figure 5.5: Base composition of nucleotides around identified UNG-snAP-seq sites in *E. coli* strain CJ236. Sites called at the stated thresholds were centred at position 0, and the primary sequence of bases in the reference genome of flanking sites are shown.

Without focusing on sites, analysis of the '0' position of total reads, one nucleotide before aligned start positions also revealed a strong enrichment in T globally across the UNG-snAP-seq libraries, which was much larger than that observed for the snAP-seq or input libraries (**Figure 5.6**). This further suggests that uracil enrichment was successful using HIPS pulldown and that uracil is abundant in these samples. These findings are in contrast to that observed in the SMUG1-snAP-seq of *L. major*, where although the site-specific accumulation of 5-hmU was strong, a global enrichment in T could not be detected (section 3.2.3).

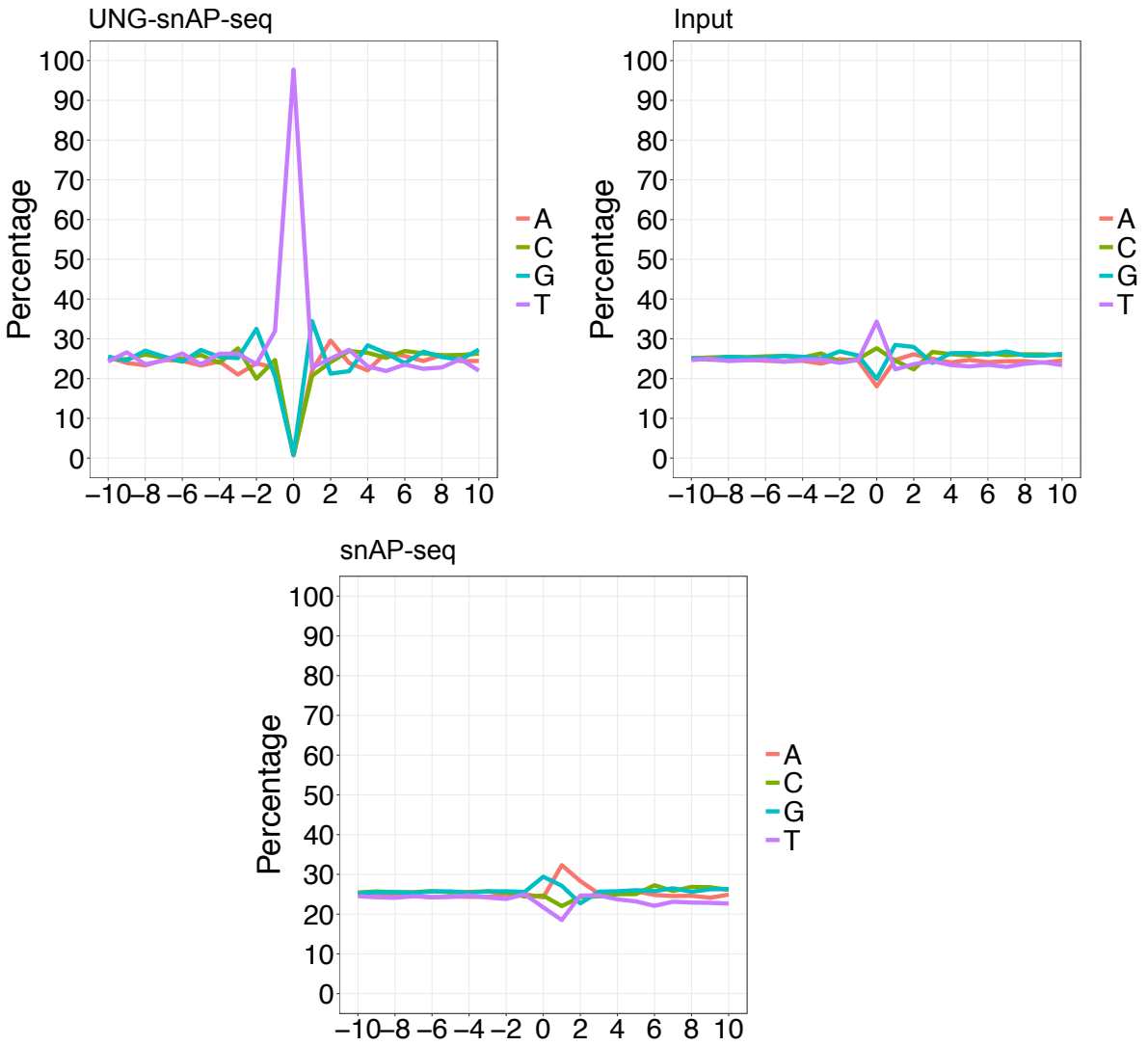


Figure 5.6: Base composition of total aligned reads after UNG-snAP-seq, snAP-seq and input library preparation using DNA from *E. coli* strain CJ236. Reads begin at position 1, and therefore captured sites are expected at position 0. Only reads aligning to the forward strand of the reference genome are shown.

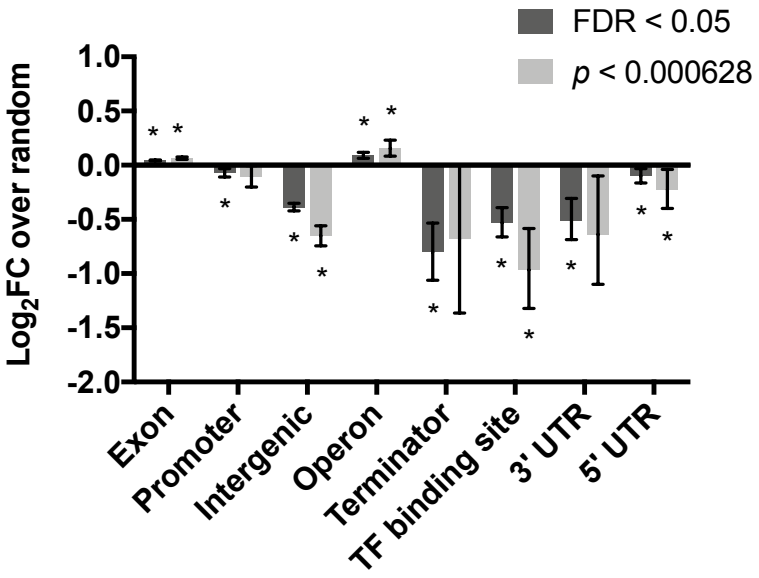


Figure 5.7: Relative enrichment of UNG-snAP-seq sites within selected genomic features. The number of sites within each region was compared to that of shuffled sets of sites of the same size. Analysis for sites called at the more relaxed FDR threshold of 0.05 and more stringent *p*-value threshold of 0.00628, are shown. Error bars represent 95% confidence intervals. *q < 0.05.

Although the UNG-snAP-seq signal was relatively weak in *E. coli*, the ability to detect thousands of sites even at the stringent *p*-value threshold suggests that the genomic accumulation of uracil is non-random. To assess the significance of detected sites in a genomic context, the relative enrichment of sites within selected genomic features was compared to sites that had been shuffled at random. This was carried out for uracil detected at the FDR < 0.05 threshold, as well as the stringent *p*-value threshold. The analysis showed that uracil was modestly depleted in a number of non-coding regions and regulatory parts of the genome, including intergenic, introns, transcription factor binding sites and UTRs (**Figure 5.7**). In contrast, a weak enrichment in exons and operons was found (q < 0.05). The overall result was similar for the sets of sites at different thresholds. Unlike mammalian systems, introns are largely depleted in the *E. coli* genome and many genes consist of a single coding sequence, whilst clusters of genes controlled by the same promoter are grouped in operons. The distribution of UNG-snAP-seq peaks in this genome suggests that transcribed regions are more susceptible to uracil accumulation, although the effect is very small. Whilst this enrichment analysis normalises for differences in the size of genomic features by random shuffling, the GC content is not taken into account. However, exons are on average more GC rich and correspondingly AT poor compared to non-coding regions[280]; therefore the enrichment of uracil within exons is not likely to be explained by higher thymine content.

To additionally investigate the sequence context of the detected UNG-snAP-seq sites, motif analysis was carried out. When the proportion of bases around the detected sites were averaged, the most abundant base flanking the UNG-snAP-seq sites were T and G at the 5'- and 3'- sides, respectively (**Figure 5.8**). Motif analysis using the DREME software[237] identified GBTGB, where B is any base excluding A, as the most enriched 5-mer (**Figure 5.9**). The identified motifs suggest there may be a preference for uracil to occur in T-rich and G-rich sequences. Together, the findings here from motif analysis further suggest that in the absence of Ung and dUTPase proteins, the relatively high levels of uracil that accumulate are not distributed entirely at random and some subtle sequence preferences exist. The enrichment of motifs around identified sites is dependent on the excision activity of UNG. Reports have suggested that any sequence bias of UNG is minimised when used in excess relative to the DNA substrate[281], which was used here. UNG isolated from *E. coli*, which was used to generate these libraries, was previously suggested to display reduced excision reactivity when two uracil bases are either adjacent to each other or separated by a single base[282], whilst the local sequence context in terms of other bases was not found to be as influential on activity. It is therefore possible that closely clustered uracil sites are underrepresented in the UNG-snAP-seq data.

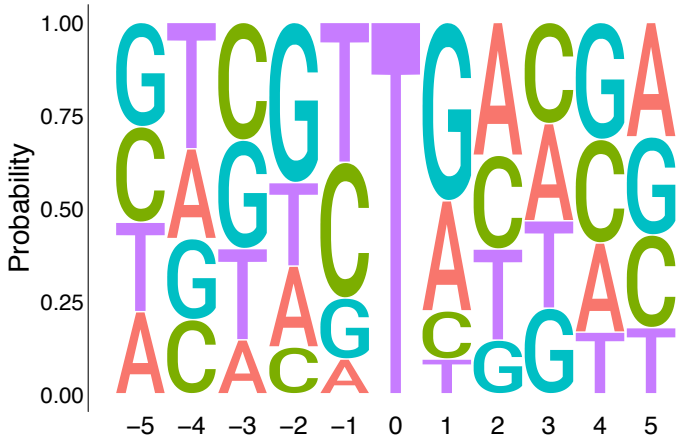


Figure 5.8: Sequence logo plot of nucleotides 5 bases upstream and downstream of high-confidence UNG-snAP sites ($p < 0.000628$) in *E. coli* CJ236 (base '0'). Only sites aligning to the forwards strand of the reference genome are shown.

UNG-snAP-seq was carried out in this section on DNA from the uracil-tolerant *E. coli* strain CJ236. Despite the high levels of total uracil in this strain accounting for up to 0.8% of nucleobases, the signal at individual nucleotides within the genome was weak. This was likely

explained by a lack of strong accumulation of uracil at specific loci across the population of cells and a large proportion of uracil was randomly distributed instead. Analysis of the total reads after UNG-snAP-seq, however, confirms a strong global enrichment for uracil. Together, these results highlight that the limit of detection of snAP-seq experiments depends strongly on the local and not global accumulation of modifications. Furthermore, different ways of bioinformatically identifying UNG-snAP-seq sites were tested, where it was found that ranking sites by corresponding *p*-values allowed a more stringent method of calling enriched sites.

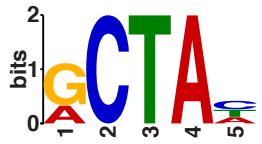
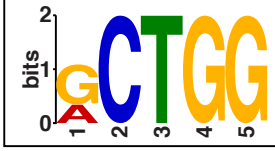
Motif	E-value
	1.9e-106
	4.9e-045
	1.5e-041
	1.8e-018
	2.0e-014

Figure 5.9: Enriched motifs around high-confidence UNG-snAP-seq sites ($p < 0.000628$) obtained by DREME[237]. The detected uracil is centred at position 3 within each motif. Only sites aligning to the forward strand of the reference genome are shown.

5.2.3 Detection of UNG-snAP-seq sites in mESC DNA

The work described in this section was carried out in collaboration with Dr Fátima Santos and Dr Poppy Gould from the Reik Group, Babraham Institute, University of Cambridge. Cell lines were generated and cultured by Dr Gould and Dr Santos. The project was designed in collaboration.

Shortly after the fertilisation of oocytes, a wave of 5-mC demethylation occurs in the paternal genome, resulting in an asymmetry between the maternal and paternal DNA. This demethylation has been observed to occur in two distinct phases; one prior to, and one coincident with, DNA replication during the first cell cycle after fertilisation. Evidence for the involvement of both an active demethylation and processive demethylation pathways in the second phase has been reported[51,52] (**Figure 5.1b**). Whilst the processive demethylation pathway has been demonstrated to lead to demethylation of both 5-mC and 6-mA in a non-specific manner *in vitro*, as well as around targeted sites *in vivo*, the extent to which this pathway occurs endogenously during embryonic development remains unknown. On the DNA level, a key intermediate within this mechanism is uracil. In order for long-patch BER to be able to repair uracil and replace 5-mC with unmethylated cytosine as part of a single process, the position of uracil must be within roughly 13 nucleotides of 5-mC, on the same strand of DNA. Therefore, single-nucleotide resolution mapping of uracil by UNG-snAP-seq was carried out to investigate whether demethylation by this pathway could be detected.

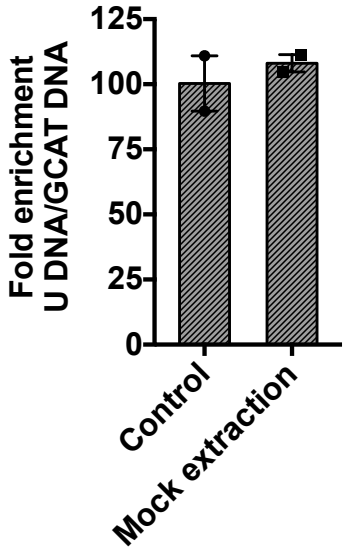


Figure 5.10: Enrichment of uracil DNA relative to GCAT DNA. The combined DNA was subjected to DNA extraction (DNeasy, Qiagen) and then treated with UNG, followed by HIPS and biotinylation reactions. Purified DNA was subjected to streptavidin pulldown and the extent of enrichment was followed by qPCR, by comparison of DNA recovery relative to input DNA. As a control, DNA that had not undergone extraction was also enriched by the same protocol. Mean and S.E.M of two replicates are shown.

The deamination of cytosine in dsDNA has been reported to be reduced compared to ssDNA[271]. Therefore, extraction of DNA in the double-stranded form prior to UNG-snAP-seq was not expected to introduce substantial amounts of uracil artefacts to genomic samples.

Furthermore, LC-MS/MS experiments have shown that global uracil levels are largely unaffected by a variety of DNA extraction methods (measurements by Dr Gould). To further verify these findings in the context of UNG-snAP-seq, the effect of DNA extraction on the efficiency of HIPS enrichment was investigated. Using synthetic ODNs previously designed as spike-ins, uracil DNA and unmodified GCAT DNA were subjected to a mock DNA extraction. The purified DNA was treated with UNG, followed by HIPS and biotinylation reactions. The tagged DNA was enriched using streptavidin (see Chapter 2) and the overall extent of enrichment for uracil DNA was compared to a sample that had not undergone DNA extraction. No significant change was observed between the two samples (p = 0.67, Mann-Whitney test), suggesting that the extraction method does not significantly influence the enrichment of uracil (**Figure 5.10**). Together with the findings using mass spectrometry, the effects of DNA extraction were concluded to not substantially impact uracil levels.

The hypothesised processive demethylation pathway is initiated by AID-mediated deamination of cytosine residues to generate uracil, followed by BER initiated by UNG, the main uracil glycosylase in mammals. To elevate the levels of uracil generated by this pathway and enhance detection, both overexpression of AID and knockout of UNG were carried out. Although earlier work on this pathway was carried out in mouse zygotes, the amount of DNA extractable from this system was impractical for UNG-snAP-seq. Therefore, stable cell lines were generated using mouse embryonic stem cells where much more DNA can be obtained (experiments by Dr Gould). In wild-type mESCs, global uracil levels have been measured by LC-MS/MS at around 3.1 per million dN. An AID-GFP fusion protein was overexpressed in mESCs (AID-OE), in which global uracil levels have been confirmed to increase to 6.6 per million dN (measurements by Dr Gould). A UNG knockout obtained by CRISPR was generated in which the AID-GFP fusion was also expressed (AID-OE/UNG-KO). Knockout of UNG was found to increase uracil levels by around 70% compared to wild-type mESCs, and the combined effects of overexpressing AID and removal of UNG was expected to further elevate the global levels of uracil in this cell line.

UNG-snAP-seq was applied first to the AID-OE cells, where deamination activity is elevated. A single replicate was prepared with an input library in parallel. Sequencing to a maximum practical depth was prioritised over replicates at this stage, in order to generate preliminary data to assess whether the uracil signal was observable in this cell type. The sequencing data was aligned to a reference mouse genome (mm10) and individual sites of enrichment were called by comparing the relative number of reads directly adjacent to each genomic position in the enriched and input libraries. Although individual sites of positive

enrichment (positive log_2 fold-change, UNG-snAP-seq vs. input) could be found, it was not possible to set a threshold at which these were favoured over negative enrichments (negative log_2 fold-change), thus indicating that enriched UNG-snAP-seq sites could not be called with confidence (**Figure 5.11**).

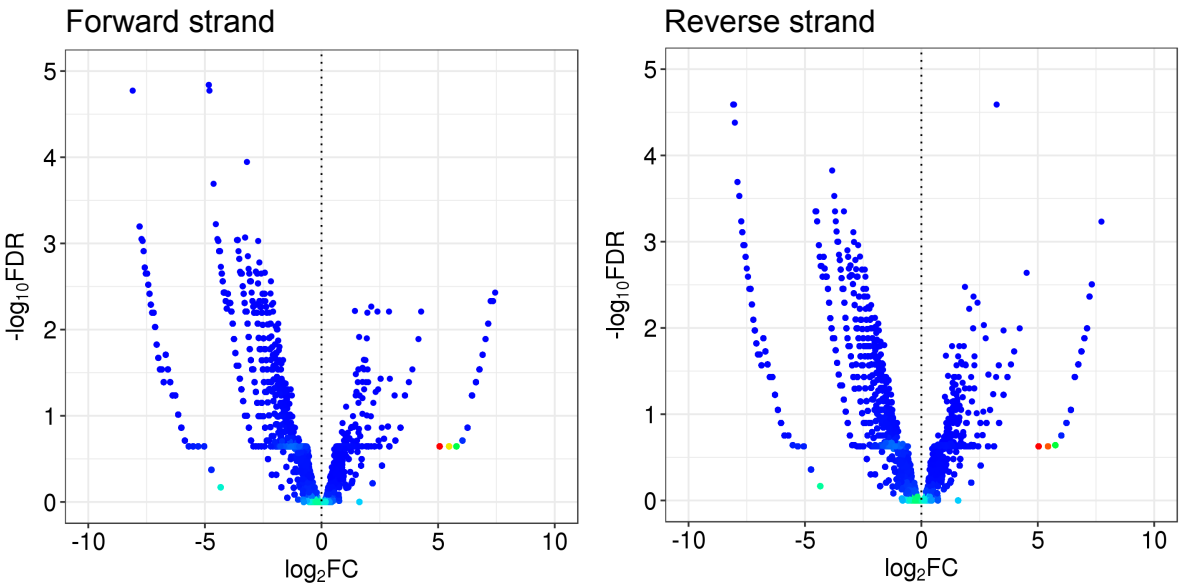


Figure 5.11: Volcano plot representing genomic loci with differential coverage in UNG-snAP-seq and input libraries and corresponding FDR values in AID-OE mESCs. Reads were split between those aligning to the forward and reverse strands based on the reference genome; the number of reads beginning one nucleotide 3'- to each position is represented as a log_2 fold-change in coverage (UNG-snAP-seq vs. input).

Visual inspection of the sites with high read accumulation specifically in the input library revealed that a large proportion of these were in mitochondrial DNA (Chr M). The higher copy number of mitochondrial DNA relative to nuclear DNA results in a much higher coverage. Therefore, small relative increases in coverage in mitochondrial DNA are large in absolute read counts and are subsequently associated with very small FDR values. As mitochondrial DNA is not involved in the processive demethylation pathway, site-calling was carried out again where mitochondrial reads were discarded after alignment. This was expected to reduce noise in the input libraries, and possibly enhance the detectability of signal within the nuclear genome in the UNG-snAP-seq library. However, the volcano plot after the removal of mitochondrial reads did not show a relative increase in sites with positive log_2 fold-change (**Figure 5.12**). Instead, more sites were called with negative enrichment. Further inspection of these additional sites revealed that they were located at loci of high-coverage in the input. This is a similar scenario to the high-coverage mitochondrial DNA and detected largely due to biases in the statistical tests used for calling sites. In an attempt to further remove these

areas of high input coverage, MACS2 peak-calling was carried out on the input library alone without using a control, to identify the regions with highest coverage. A custom blacklist for the mouse genome was then made using these peaks. Calling sites after exclusion of the blacklisted regions still did not shift the volcano plot in the expected direction, and further additional sites in the input library were called, largely in locations with the next highest coverage. From the *E. coli* data, it was found that ranking sites by *p*-value instead of FDR gave better sensitivity. This was not found to change the outcome here and it was concluded that sites of enrichment at the single-nucleotide level could not be confidently detected in the UNG-snAP-seq data that was significant above the noise observed in the input library.

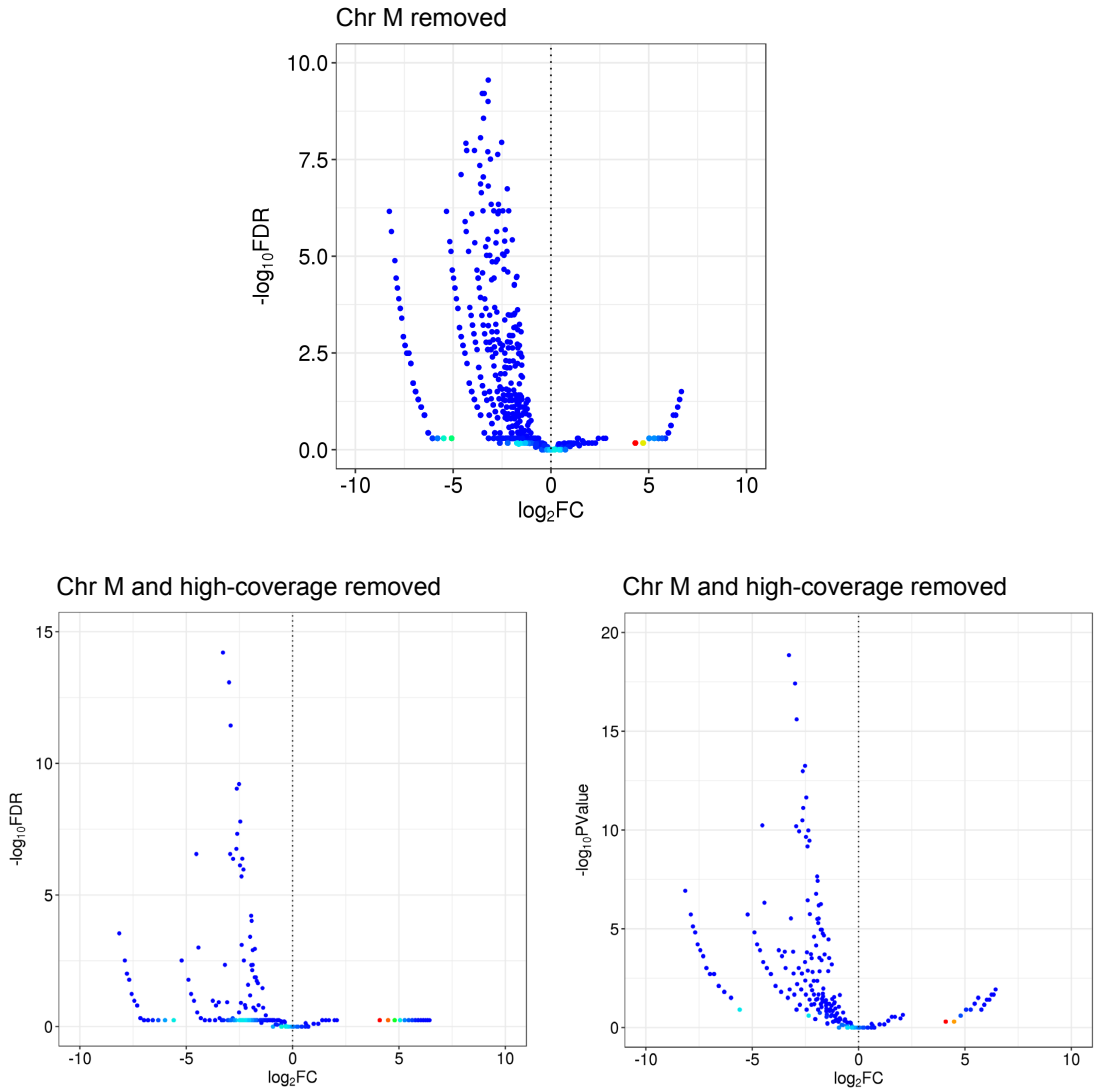


Figure 5.12: Volcano plot representing genomic loci with differential coverage in UNG-snAP-seq and input libraries and corresponding FDR or *p*-values in AID-OE mESCs. Reads were split between those aligning to the forward and reverse strands based on the reference genome; the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (UNG-snAP-seq vs. input). Only reads aligning to the forward strand are shown.

It was rationalised that the lack of uracil accumulation at single nucleotides in the AID-OE cells may be due to efficient removal of the potentially mutagenic uracil product and therefore UNG knockout cells, also expressing the AID-GFP fusion protein (AID-OE/UNG-KO) were investigated. UNG-snAP-seq on these cells gave similar results to the AID-OE cells (**Figure 5.13**). A comparable number of positive and negative enrichment sites were detected across the genome, which remained consistent when mitochondrial, or high-coverage areas were removed. Ranking of sites by *p*-value also gave a similar outcome.

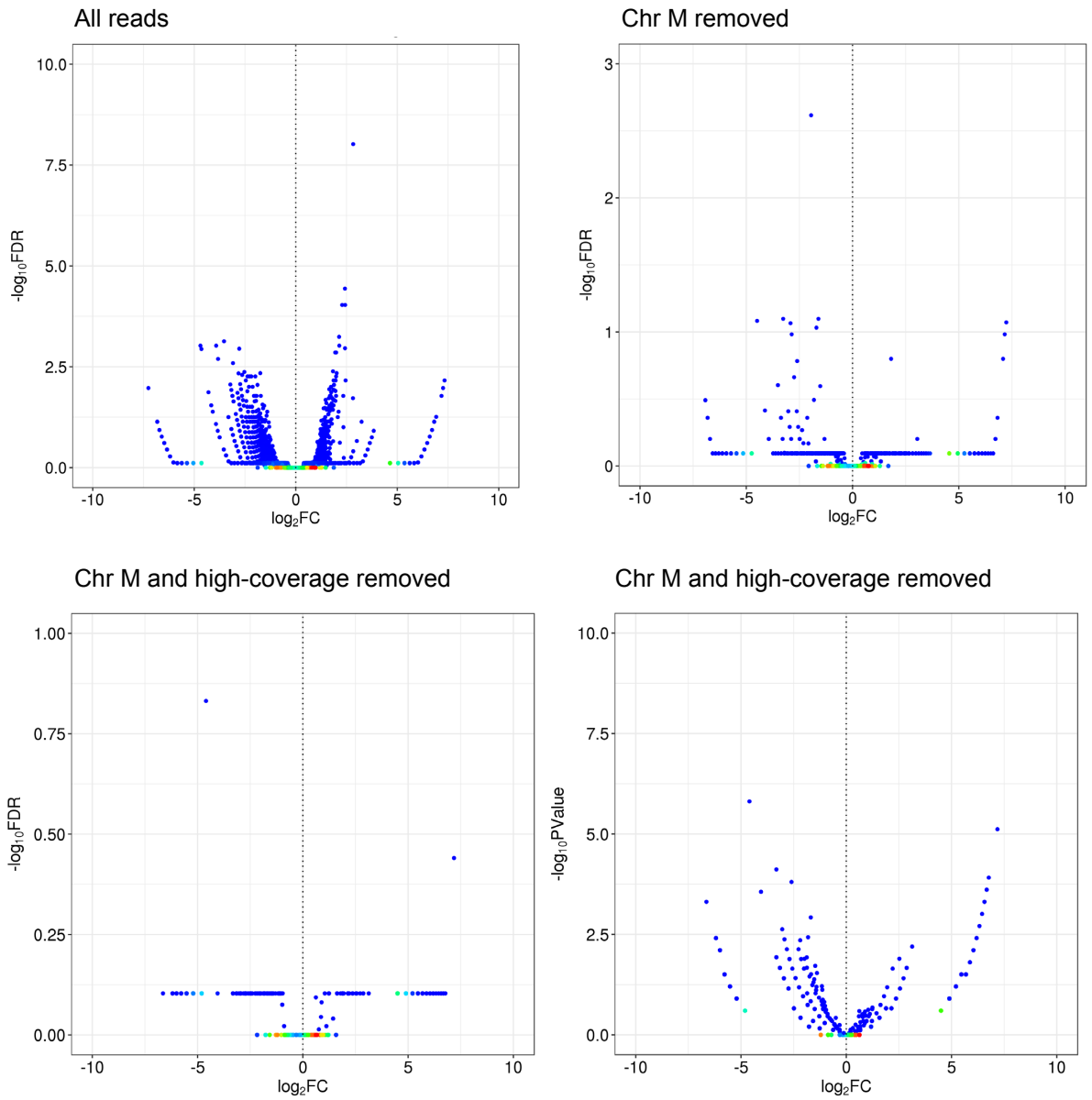


Figure 5.13: Volcano plot representing genomic loci with differential coverage in UNG-snAP-seq and input libraries and corresponding FDR or *p*-values in AID-OE/UNG-KO mESCs. Reads were split between those aligning to the forward and reverse strands based on the reference mouse genome; the number of reads beginning one nucleotide 3'- to each position is represented as a $\log_2$ fold-change in coverage (UNG-snAP-seq vs. input). Only reads aligning to the forward strand are shown.

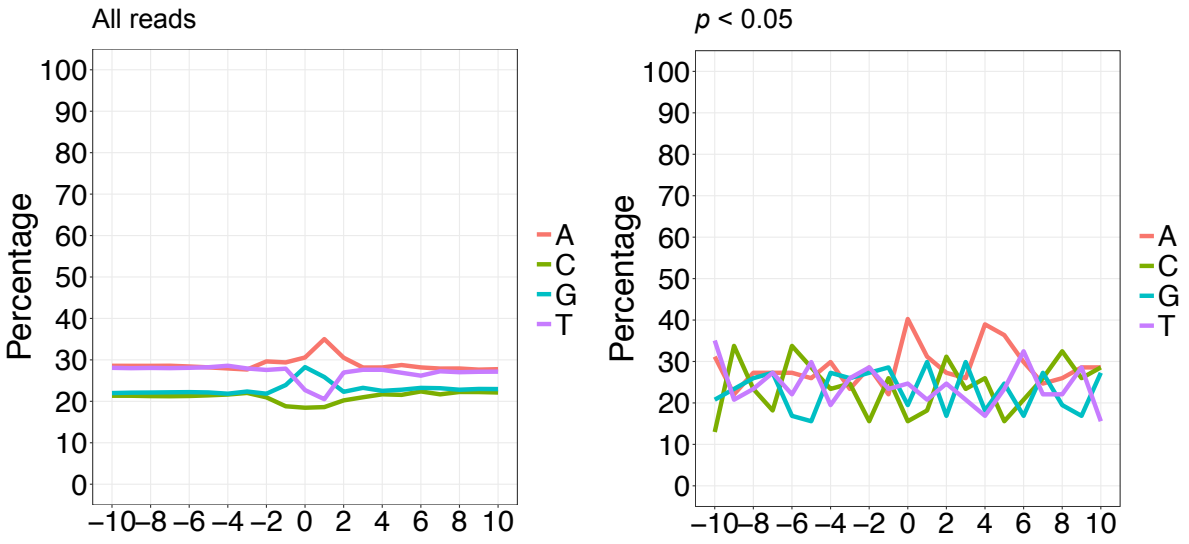


Figure 5.14: Analysis of the base composition of sequencing reads in UNG-snAP-seq libraries generated using DNA from AID-OE mESCs. Total reads were analysed (left), where reads begin at position 1 and the expected UNG-AP site is located at position 0. A set of peaks with $p < 0.05$ were also analysed (right) where the identified sites were centred at position 0. Only reads or sites aligning to the forward strand of the reference genome are shown.

Although the volcano plots indicate that sites detected with positive enrichment were not statistically significant when compared to the accumulation of reads also seen in the input library throughout the genome, the base identity of total reads in the enriched library was analysed to investigate whether a global enrichment for cytosine deamination could be observed. The sites of positive enrichment detected with $p < 0.05$ was also investigated, as a set of potential UNG-snAP-seq sites. For total reads across the entire library, a small enrichment in the purines was seen for the nucleotide preceding read start sites, whilst cytosine was depleted relative to flanking regions (**Figure 5.14**). For the potential sites of enrichment ($p < 0.05$), a weak depletion in cytosine was seen once again and adenine was the most abundant base. These results suggest that overall, cytosine deamination had not occurred at a frequency above the limit of detection of UNG-snAP-seq. The ability to observe a global enrichment in cytosine from averaging total reads is dependent on the overall abundance of deamination events. In Chapter 3, it was found that a low global abundance of SMUG1-senstive sites (~32 5-hmU per million dN[70]) was insufficient to observe a global enrichment in thymine, the 5-hmU precursor, for total reads in SMUG1-snAP-seq libraries. However, the direct base identity of the vast majority of high-confidence sites aligned to thymine suggesting that site-specific accumulation was strong. In the UNG-snAP-seq data generated in *E. coli* (~3,000-8,000 uracil per million dN[272]), both the total reads and the

identified sites were highly enriched in thymine, suggesting a higher global abundance of uracil in addition to site-specific accumulation (section 5.2.2). In contrast, application of this method to the mouse samples shows that uracil derived from cytosine deamination is not strongly accumulating on either the global or nucleotide level, even with overexpression of AID protein.

Analysis of the AID-OE/UNG-KO revealed similar results (**Figure 5.15**). The number of sites of enrichment with $p < 0.05$ was very low in these samples, at 67 and 58 in the forward and reverse strands of the reference genome, respectively. In the total reads, an enrichment only in the purines was observed, whilst the sites ($p < 0.05$) were weakly depleted in cytosine and guanine was the most prominent base.

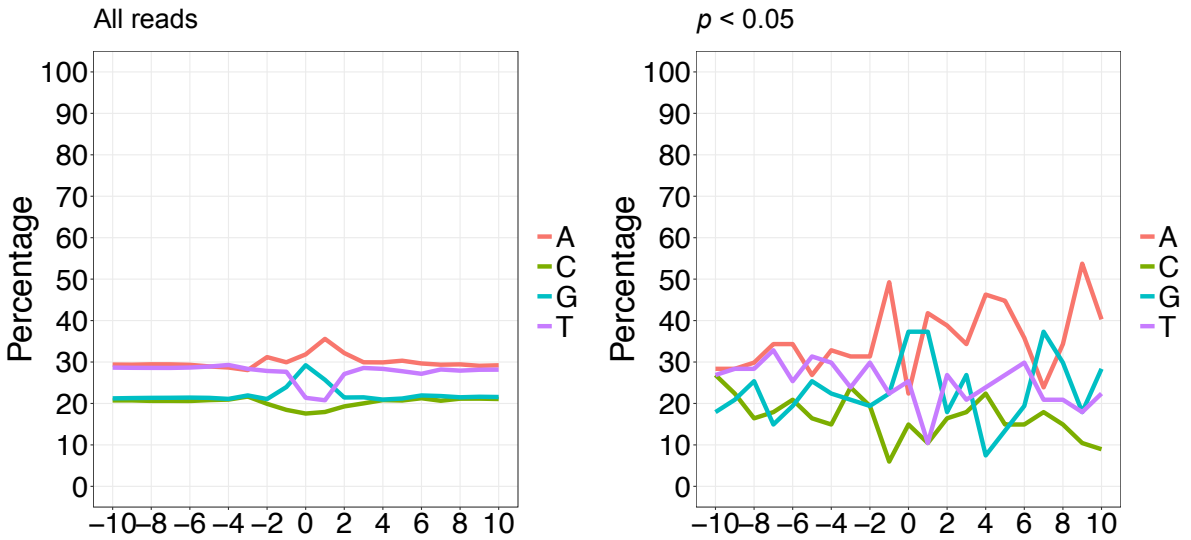


Figure 5.15: Analysis of the base composition of sequencing reads in UNG-snAP-seq libraries generated using DNA from AID-OE/UNG-KO mESCs. Total reads were analysed (left), where reads begin at position 1 and the expected UNG-AP site is located at position 0. A set of peaks with $p < 0.05$ was also analysed (right) where the identified sites were centred at position 0. Only reads or sites aligning to the forward strand of the reference genome are shown.

The work in this section was carried out to explore the involvement of a processive demethylation pathway during embryonic development. Using embryonic stem cells, the overexpression of AID, which has been shown to increase global uracil levels, was not associated with detectable single-nucleotide sites of uracil accumulation using UNG-snAP-seq. Further knockout of UNG in mESCs also did not enhance uracil accumulation above the limit of detection. As such, no direct support of the proposed mechanism was found.

5.2.4 Limit of detection of UNG-snAP-seq

The limit of detection of UNG-snAP-seq is expected to be highly dependent on the depth of sequencing. The lack of detectable UNG-sensitive sites at the single-nucleotide level in the mESC data can only be used to conclude that uracil is not accumulating at levels above this limit of detection. To attempt to quantify this limit, a similar method to that described in section 4.2.4 was carried out. A bioinformatical approach was taken, using the SMUG1-snAP-seq data generated from *L. major* DNA to estimate the sensitivity of the sequencing used here. Samples were taken from the *L. major* sequencing reads to obtain a comparable depth to that achieved with the mouse libraries. Although a similar depth was aimed for in each of the UNG-snAP-seq and input libraries, the total number of final aligned reads varied slightly, with an average depth of 1.4X achieved for the AID-OE library, and 2.0X for the AID-OE/UNG-KO library. Sites in the sampled *L. major* data were called using a single replicate library compared to the corresponding input library, to maintain the same experimental design as used for UNG-snAP-seq. At FDR < 0.05, 20% of the high-confidence sites prior to sampling remained detectable at 1.4X, and 63% of sites were detected at 2.0X. At $p < 0.05$, this increases to 77% and 94%. The sensitivity of detection at the single-nucleotide level is expected to depend on the proportion of cells that consistently contain a modification at the same given genomic location. Whilst the extent to which this occurs at each of the 3,200 SMUG1-snAP-seq sites in *L. major* is not known, these cannot be higher than 100%. Together, these findings suggest that sites in which uracil occurs in high-abundance, potentially for up to 100% of the population, should remain largely detectable for the experimental design used here. Particularly when *p*-values are used to assess the statistical significance of sites, the majority of high-abundance sites are expected to remain above the limit of detection.

The lack of sites detected by UNG-snAP-seq in the mESCs explored in this section suggests that despite removal of UNG from the BER pathway and elevating cellular AID activity, a significant accumulation of uracil at the single-nucleotide level does not occur in genomic DNA. It is possible that alternative glycosylases, such as SMUG1, that are also active at uracil[225] are able to quickly remove a large proportion of uracil generated by AID when UNG is removed. *In vitro* studies on processive demethylation have shown that inhibition of UNG by UGI during AID-mediated demethylation only partially reduces demethylation, suggestive of alternative repair mechanisms in the absence of UNG activity such as the involvement of alternative glycosylases or possibly MMR[52]. Whilst evidence in support of the proposed processive demethylation pathway could not be found here in the

distribution of genomic uracil, the results from UNG-snAP-seq can only be interpreted relative to the limit of detection of the method. Using SMUG1-snAP-seq data on the *L. major* genome, it can be concluded that in the event that specific genomic loci are quantitatively deaminated, these are most likely detectable. However, rare deamination events that are not consistent within the population of cells remain difficult to detect. *In vitro* studies have shown that AID activity is not targeted around 5-mC loci when present at high levels and is equally able to lead to demethylation of 5-mC and 6-mA in an unbiased manner[52]. This further suggests that deamination activity in the event of overexpression of AID may be largely not site-specific. As for all enrichment-based methods, UNG-snAP-seq is only able to detect sites that are substantially higher in signal relative to background that is consistent across a population, and rare events at low levels are challenging for such approaches. Together with the relatively low depth of sequencing that is practical to achieve in large mammalian genomes, the current UNG-snAP-seq design may not be suitable for the detection of low abundance AID deamination sites. To this end, it may be more beneficial to select regions of the genome in which the demethylation events of interest are most prominent and target sequencing reads to these areas, than to repeatedly sequence the entire genome at significantly greater depth. These regions could for example be identified through bisulfite sequencing carried out on DNA from wild-type and AID-OE cells to determine sites where 5-mC is consistently lost. A targeted uracil enrichment around these identified regions may then reveal interesting results that are not easily detected during genome-wide sequencing. A preliminary experimental design towards targeted AP site sequencing is discussed in the following section.

5.2.5 Towards a targeted design of snAP-seq

To focus sequencing reads at selected regions of the genome without the need to cover the entire genome, a targeted version of snAP-seq was designed. Changes to the protocol were only in the library preparation strategy whilst chemical tagging steps remain unchanged (**Figure 5.16**). In this targeted design, the P7 adapter is not introduced by ligation. Instead, tagged and purified DNA is first treated with phosphatase, to deactivate free DNA ends as in standard snAP-seq. DNA is then enriched on streptavidin beads, released by cleavage and repurified on beads to remove non-specific DNA release. Following this, a primer with a partial overhang is introduced. The priming region is complementary to the sequence of interest and thus serves to target libraries towards these locations, whilst the overhang contains the complement of the P7 adapter. Between the primer region and the P7 adapter

is a random hexamer. This is necessary as during standard snAP-seq, the 5'- ends of multiple inserts are expected to align to the same position, and therefore random variation in the 3'- end is required to allow the identification of PCR duplicates as those that also align at the same position from the 3'- end. In this targeted approach, the location of both 5'- and 3'- ends are potentially fixed and therefore an alternative way of identifying PCR duplicates is required. Using randomised hexamers, any reads that contain the same hexamer sequence can be classified as a duplicate and discarded. Polymerase extension of this custom primer then generates dsDNA, to which the P5 adapter can be introduced by ligation. PCR amplification of this product generates a sequencing library.

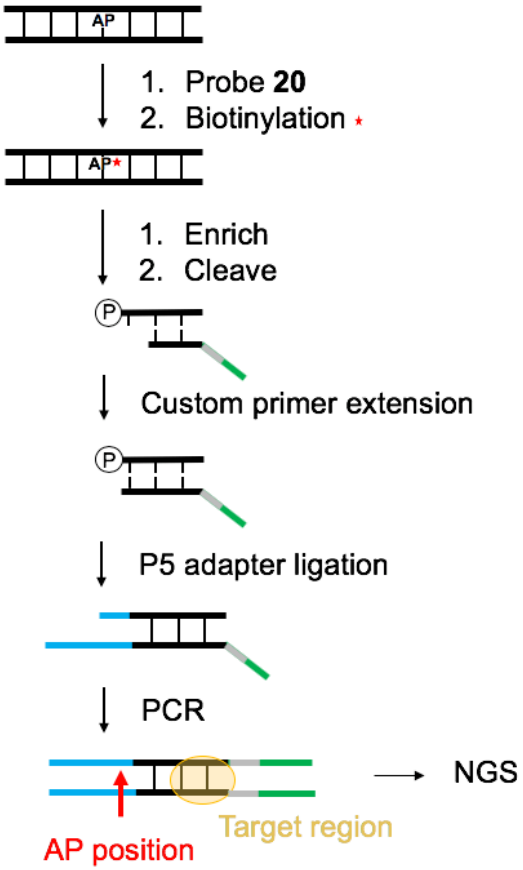


Figure 5.16: Workflow of targeted snAP-seq. DNA is treated with HIPS probe **20** followed by biotinylation. DNA is enriched using streptavidin beads and recovered by cleavage. The P7 adapter (green) is introduced by primer extension as part of an overhang, where the priming region is complementary to the target. This primer also contains a random hexamer (grey) to allow the identification of PCR duplicates as reads that have the same hexamer sequence. Polymerase extension generates dsDNA, to which the P5 adapter (blue) is introduced. The position directly 5'- to the start of the insert corresponds to the AP sites, whilst the 3'- end of the insert corresponds to the target region.

To test the feasibility of this design, the spike-in DNA models were used. Primers targeting the 3'- end of AP DNA2 and GCAT DNA2 were designed. The inclusion of a primer specific for a GCAT control sequence was to ensure that the targeting step does not falsely enrich for non-AP DNA. A further AP and GCAT sequence were included in libraries that were not targeted, along with 5-fU and 5-fC sequences as additional controls. The described protocol was then applied. Upon sequencing, it was found that the large majority of reads (96%) did not align to any of the known sequences. This unexpected outcome suggests that non-specific priming, or error-prone DNA synthesis may have occurred during the protocol leading to unrecognised sequencing reads, which may be due to shortcomings in primer design. However, out of the small proportion of reads that could be aligned to any of the model DNA sequences, those aligning to AP DNA2 accounted for a large majority (97.8%) (**Figure 5.17**). The number of reads aligning to the targeted negative control, GCAT DNA2, for which a primer was also included, was slightly higher than for the other negative controls. However, a 65-fold AP2/GCAT2 enrichment was still observed overall, suggesting that the primer targeting does not interfere strongly with enrichment.

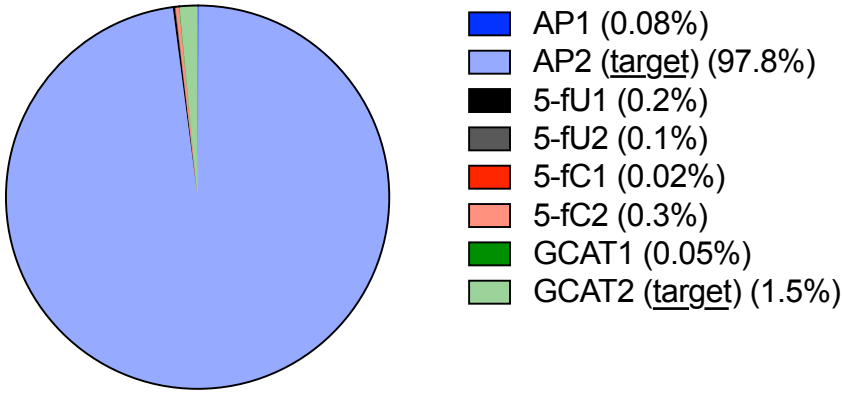


Figure 5.17: Normalised number of sequencing reads aligned to DNA sequences after targeted snAP-seq. Primers targeting the 3'- end of AP DNA2, and GCAT DNA2 were used. The number of reads aligning only to the forward strand of each sequence is shown, with normalisation to the number of reads aligning to the corresponding reverse strand in an input library.

It remains unclear whether the alignment problem observed for spike-in DNA using the targeted snAP-seq protocol was a sequence specific artefact, and whether further application to genomic DNA would reveal a similar issue. The complementary region between the primer and target sequences was limited to 20 nucleotides in this library; it is possible that optimisation of this by varying the length and melting temperature, along with the choice of

polymerase and primer extension conditions may be able to improve polymerase fidelity and specificity. These preliminary results provide some support that this method of sequence targeting is possible and compatible with snAP-seq, however, further optimisation is required to overcome technical challenges.

5.3 Conclusions and future directions

In a further extension of snAP-seq, the distribution of genomic uracil was explored through use of the UNG enzyme. Uracil is a putative intermediate in the processive demethylation pathway, in which 5-mC is removed and replaced by unmethylation cytosine as part of the long-patch BER pathway. This demethylation pathway has been implicated during embryonic development, where the precise mechanism or mechanisms that mediate the wave of demethylation observed for the paternal DNA is still largely unknown. Within the proposed pathway, the location of uracil generated by AID deamination activity must be no more than 13 nucleotides away from 5-mC sites. Therefore, the ability to map uracil, particularly at high-resolution, is of value in further exploring this possible pathway.

Using embryonic stem cells from mouse, the importance of two key enzymes in the pathway, AID and UNG was examined. In particular, the precise location of uracil sites was of interest. Uracil levels in wild-type mESCs are relatively low, at around 4 per million dN. UNG-snAP-seq was first carried out on an *E. coli* strain, CJ236, in which uracil levels are much higher and these libraries were expected to aid in further optimisation of the methodology or data analysis. Despite the high global levels of uracil in this organism, at up to 8,000 per million dN, accumulation at individual sites within the genome remained modest. It was found that when using FDR thresholds to call sites, it was not possible to completely avoid false positives in the input library. Ranking of sites by *p*-value instead allowed for more stringent site-calling, where a total of 5,505 high-confidence sites were detected. These sites were found to be weakly enriched in transcribed regions in the genome, and depleted in others, possibly suggesting that a further level of control in regulating uracil levels may be in place even when two key components currently known to regulate uracil were removed.

The results from UNG-snAP-seq in *E. coli* show that detecting individual nucleotides of enrichment in this type of data can be challenging even when uracil levels are high. With this in mind, UNG-snAP-seq was carried out on two mouse cell lines derived from mESCs; AID-OE, where an AID-GFP fusion protein was expressed in addition to endogenous AID protein,

and AID-OE/UNG-KO, where in addition to AID-OE, UNG activity was removed. No clear signal could be detected by UNG-snAP-seq in either of these cell types, suggesting that uracil levels remained low at the single-nucleotide level across the genome. Thresholding using *p*-value, which proved to be more suitable in analysing *E. coli* data, was also used instead of FDR, however, this did not affect the overall outcome. Further analysis of total reads in the library, as well as a potential set of sites found at the $p < 0.05$ threshold showed no clear enrichment in cytosine, which would be expected for AID-mediated deamination events. Therefore, it was concluded that uracil does not accumulate strongly in either of these cell types.

The limit of detection of UNG-snAP-seq must be considered when drawing conclusions from the sequencing data. This in turn is highly dependent on the depth of sequencing, which is a limiting factor particularly when studying large genomes such as the mouse genome. Using *L. major* sequencing data where the signal from SMUG1-mediated snAP-seq was very strong, it was estimated bioinformatically that for the sequencing depth achieved with the mouse samples, quantitative deamination events at a given genomic location is likely to be detectable. These results therefore suggest that at this limit of detection, deamination events mediated by AID as part of the processive demethylation pathway are not significantly occurring. To further explore this pathway at greater sensitivity, future work should focus on applying UNG-snAP-seq in a targeted approach. A library preparation strategy to target genomic regions through complementary priming in combination with snAP-seq was designed and demonstrated on synthetic DNA. A large proportion of sequencing reads obtained by this method were not alignable to the input sequences, however, the small number of reads that were alignable showed some success in this approach. Further optimisation in improving priming specificity may overcome the problems with read alignment. This targeted UNG-snAP-seq may then be used to detect uracil at a small number of selected genomic loci in which processive demethylation is proposed to be most active, to further explore whether this pathway occurs at significant levels during embryonic development.

Chapter 6

Conclusions

The main focus of this thesis has been on the sequencing of abasic sites in DNA at high resolution. The chemical reactivity of abasic sites was explored in Chapter 2, to develop a functionalised nucleophilic probe that allows the affinity enrichment of DNA fragments containing abasic sites in the presence of background DNA. A key requirement for the methodology was to allow abasic sites to be distinguished from other potentially reactive sites in DNA, such as naturally occurring formylpyrimidine bases. The developed strategy, snAP-seq, was demonstrated to detect abasic sites with high selectivity in the presence of both unmodified DNA and formylpyrimidine base modifications. Due to a selective chemical cleavage of the DNA backbone at chemically tagged abasic sites, this approach was also shown to reveal the location of abasic sites at single-nucleotide resolution.

In addition to exploring endogenous abasic sites, methodology aimed at studying abasic sites is also useful in the investigation of any base modification that is a substrate of a glycosylase enzyme. Glycosylases excise base modifications from DNA to generate an abasic site with high specificity, and thus the *in vitro* treatment of isolated DNA can be used to convert modifications of interest into an abasic site intermediate. The extension of the snAP-seq methodology by combination with glycosylase treatment was explored in Chapters 3 and 5. The study of thymine modifications as part of SMUG1-snAP-seq in the *Leishmania major* genome in Chapter 3 was useful in further validating the accuracy of the overall snAP-seq method, as the data generated was found to overlap well with a previously reported low-resolution map of 5-hmU generated in the same genome. Furthermore, the single-nucleotide resolution data generated here provided insights into the distribution and sequence context preference of 5-hmU accumulation. 5-hmU was also studied in the *Trypanosoma brucei* genome, where it has previously been found that despite the role of 5-hmU as a precursor to the hypermodification, base J, a large proportion of 5-hmU sites does not corelate with base J. Genome-wide mapping of 5-hmU in *T. brucei* cell lines in which the JBP1 or JBP2 enzyme was removed by gene knockout revealed that the signal detected by 5-hmU DIP-seq is not strongly altered in the absence of either one of these oxidases. These findings suggest that the two JBP enzymes may have overlapping functions, or that the detected 5-hmU is largely

generated by alternative pathways. However, a dependence on JBP1 was found for the small number of 5-hmU loci that overlap with base J loci, suggesting that JBP1 is essential during the biosynthesis of these sites.

In Chapter 4, endogenous abasic sites were studied in the human genome by using DNA isolated from HeLa cells. To further explore the formation of abasic sites, cells in which the levels of the repair protein, APE1, were reduced by siRNA knockdown were studied in addition to control cells. The ability to detect abasic sites with high selectivity was particularly important in this study to avoid potential cross-reactivity at formylpyrimidine sites. Such cross-reactivity has been a major drawback to many other studies on endogenous abasic sites, which snAP-seq has been demonstrated to overcome in Chapter 2. The selective detection of abasic sites revealed that this type of DNA damage does not cluster significantly at the single-nucleotide level. Instead, windows of accumulation could be detected throughout the genome as peaks. These were found to be non-randomly distributed, and an enrichment in regulatory and coding regions was found when APE1 protein was depleted.

Finally, in Chapter 5 snAP-seq was combined with UNG treatment to study the distribution of uracil during embryonic development. C to U deamination has been proposed to be a key event during the demethylation pathway of 5-mC during epigenetic programming shortly after the process of fertilisation, and therefore UNG-snAP-seq was used to further explore this pathway in mouse embryonic stem cells. The cellular levels of two key enzymes involved in the pathway, UNG and AID, were altered from endogenous levels to enhance global levels of uracil; however, no high-confidence uracil signal could be detected in either cell types. It was concluded that uracil was not accumulating significantly at the single-nucleotide level in these samples. As this conclusion can only be drawn in relation to the sensitivity of the method, an estimate of the limit of detection was also explored. Future work on this area could focus on mapping uracil in a targeted approach, in regions where the proposed mechanism is likely to be most active. A preliminary design for a targeted snAP-seq strategy was investigated and demonstrated with some success using synthetic DNA to enable further exploration of this pathway at higher sensitivity.

Overall, a versatile sequencing method has been developed that can be used to study both endogenous abasic sites and a range of base modifications in DNA. The methodology was validated using synthetic DNA, and further applied to a number of genomes to study DNA damage as well as the role of pyrimidine modifications in different biological contexts.

Chapter 7

Materials and methods

7.1 General experimental details

7.1.1 Organic synthesis

All solvents and reagents were used as supplied from commercial sources (Sigma Aldrich unless stated otherwise). LC-MS was performed on an amaZon ESI-MS (Bruker) connected to a Dionex UltiMate 3000 UHPLC system (Thermo Fisher Scientific). LC-MS data was analysed using Bruker Compass DataAnalysis 4.2. Flash chromatography was carried out using a CombiFlash Rf system (Teledyne Isco) with puriFlash columns (Interchim). NMR spectra were recorded in Chloroform-d unless stated otherwise on a Bruker 400 MHz Avance III HD spectrometer or a 500 MHz DCH cryoprobe spectrometer. Collected NMR spectra were processed using MestReNova software. Accurate mass spectra were recorded on a Waters LCT Premier (ESI) spectrometer.

7.1.2 Oligonucleotides

All oligonucleotides were purchased from Invitrogen or Sigma Aldrich with HPLC (up to 30 nucleotides) or PAGE (30-105 nucleotides) purification unless stated otherwise. Oligonucleotide stocks were prepared in ultra-pure water at high concentration and diluted with buffer prior to individual experiments. Base modifications in short ODNs (up to 20 bp) were incorporated by solid-phase synthesis. ODNs containing uracil were also obtained through solid-phase synthesis, which were then treated with UNG enzyme to generate the corresponding AP ODNs (see section 7.2.2 for detailed protocol). Long ODNs (100-105 bp) containing base modifications other than uracil or AP sites were obtained by primer extension using modified dNTPs and an unmodified template DNA strand (see section 7.2.2 for detailed protocols). Primer extensions were also used to obtain dsDNA models from ssDNA templates. Sequences and further details of all ODNS used are given in **Table 7.3**.

7.1.3 LC-MS analysis of short ODNs

LC-MS was performed using an XTerra MS C18 column (2.5 µM, 2.1 x 50 mm). The separation of ODNs was carried out using 5-30% solvent A (100 mM 1,1,1,3,3,3-hexafluoro-2-propanol, 10 mM triethylamine) in solvent B (methanol) over 25 min, at a flow rate of 0.2 mL/min. The area underneath peaks detected by UV (260 nm) was integrated, and the % conversion of reactions was calculated by comparing the peak area between starting material and product peaks. The identity of each species detected by UV was confirmed by ESI-MS.

7.1.4 Enzymatic oligonucleotide reactions

Enzymatic and PCR reactions using oligonucleotides and genomic DNA were performed in a T100 Thermal cycler (Bio-Rad) or peqSTAR 96X Universal Gradient (Peqlab). qPCRs were performed using a CFX96 Real-Time System (BioRad), and data was processed using CFX software manager 3.1 (BioRad).

7.1.5 Sequencing

Illumina sequencing was carried out on either an MiSeq or NextSeq 500 machine (Illumina). All consumables were purchased from Illumina and used according to the manufacturer's instructions. Paired-end sequencing using 150-cycle kits were used unless stated otherwise. Sequencing libraries were quantified using a universal library quantification kit (KAPA Biosystems) according to the manufacturer's instructions.

7.1.6 Quantification and visualisation of DNA

DNA concentration was measured with a broad range dsDNA detection kit using a Qubit 2.0 fluorometer (Thermo Fisher Scientific) as well as a Nanodrop one system (Thermo Fisher Scientific). For visualisation of DNA length, samples were analysed by Tapestation (Agilent), using either a D1000 or high sensitivity D1000 screentape.

7.1.7 DNA sonication

Genomic or synthetic DNA was sheared by sonication using a Covaris M220 system, to an average fragment length of 450 bp or as stated. Fragmentation was confirmed by Tapestation analysis.

7.1.8 Cell culture

HeLa and U2OS cells were cultured in Dulbecco's modified eagle medium (DMEM, Gibco, 41965039) supplemented with 10% heat-inactivated fetal bovine serum (Thermo Fisher Scientific). All cell lines were maintained at 37 °C with 5% carbon dioxide. Cells were authenticated by short tandem repeat (STR) genotyping and routinely tested for mycoplasma (Microprobe mycoplasma detection kit, R&D systems). For harvesting, cells were washed once in cold PBS and harvested by scraping into cold PBS (500 µL) followed by centrifugation at 300 × g. Cell pellets were flash frozen at -80 °C and stored until further use.

7.1.9 Data analysis

Sequencing data processing: The quality of raw sequencing reads was evaluated using FastQC v0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality bases were filtered and Illumina TruSeq adapters were trimmed from the 3' ends of reads using cutadapt v1.12[283]. Reads smaller than 15 bp after adapter removal and base quality trimming were discarded. Trimmed reads containing the P7 adapter sequence (GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T) at the 5' end were also discarded.

Alignment of reads: bwa v0.7.15-r1140[284] was used to prepare reference sequences (spike-ins and genomes), which were then used to align trimmed sequencing reads using bwa mem. References genomes were obtained from publicly available sources as specified. The resulting alignments were cleaned, merged, sorted and indexed using samtools v1.3.1[285]**.** Duplicate reads in genomic libraries were marked using sambamba v0.6.5[286]. The filtering of alignments involved the removal of unaligned reads, secondary/alternative alignments, PCR duplicates, alignments with a quality score of less than 10, reads overlapping blacklisted regions (see Chapter 4 and 5 methods) and spike-ins. Clean alignments were further split

into alignments of R1 reads mapping to the forward and to the reverse strands respectively. These were then converted to tdf format using igvtools v2.3.91[287] for visualisation. The bamCoverage function available in deeptools v2.4.2-5-f439d22[288] was used to perform sequencing coverage calculation normalised by RPKM at single-nucleotide resolution (after merging replicates).

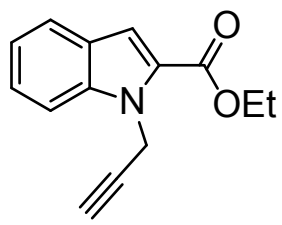
Calling single-nucleotide AP sites: bedtools v2.27.0 genomecov[289] was used to obtain the sequencing coverage of 5' positions for R1 reads aligning to the forward and reverse strands respectively. Genome-wide modelling and comparative assessment of counts from enriched and input libraries were performed using negative binomial generalized linear models as implemented in edgeR v3.16.5[290]. This analysis was carried out on all available replicates along with corresponding input libraries in parallel. High-confidence sites were obtained when selecting single nucleotides with positive $\log_2$ fold-change (snAP-seq vs. input) at specified FDR or *p*-value thresholds. Volcano plots were generated using ggplot2 v2.2.1 (https://cran.r-project.org/web/packages/ggplot2/citation.html).

Statistical analysis: statistical tests were performed in the R programming language unless stated otherwise. For detailed methods, see data analysis section for each chapter. For peak-calling, the statistical testing of candidate peaks was performed within the MACS2 software and thresholds were set at $p < 10^{-10}$ or $q < 0.05$ for the *Trypanosoma brucei* libraries, and $p < 10^{-5}$ for the human libraries as indicated.

7.2 Chapter 2 methods

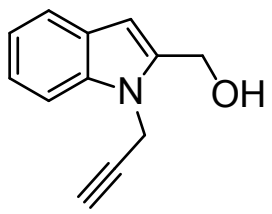
7.2.1 Synthesis of probes

Ethyl 1-propargylindole-2-carboxylate (**22**)



Sodium hydride (60 w% dispersion in oil, 0.315 g, 7.95 mmol) was dissolved in dry dimethylformamide (38 mL). A solution of ethyl indole-2-carboxylate **21** (1.50 g, 7.95 mmol) in dimethylformamide (4.5 mL) was then added dropwise at 0 °C. After stirring for 30 min at 0 °C, propargyl bromide (80 w% in toluene, 1.35 mL, 11.9 mmol, 1.5 equiv) was added dropwise and the resulting brown solution stirred for a further 4 h at 0 °C. Ammonium chloride (sat. solution, 35 mL) was added to quench the reaction, and the mixture poured into brine (35 mL). The mixture was extracted with ethyl acetate (3 × 75 mL), and the combined organic fractions was washed with brine (25 mL), dried over sodium sulfate and concentrated *in vacuo*. The crude product was purified by column chromatography (0-40% ethyl acetate in hexane) to give a white solid (1.39 g, 77%). **^{1}H NMR** (400 MHz, chloroform-d) δ 7.72 (d, *J* = 7.9 Hz, 1H), 7.53 (d, *J* = 8.4 Hz, 1H), 7.42 (dd, *J* = 8.4, 7.0 Hz, 1H), 7.38 (s, 1H), 7.21 (dd, *J* = 7.9, 7.0 Hz, 1H), 5.48 (d, *J* = 2.5 Hz, 2H), 4.43 (q, *J* = 7.1 Hz, 2H), 2.28 (t, *J* = 2.5 Hz, 1H), 1.45 (t, *J* = 7.1 Hz, 3H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 162.0, 138.9, 126.9, 126.2, 125.4, 122.8, 121.1, 111.4, 110.5, 78.7, 72.0, 60.8, 33.9, 14.3 ppm. **HRMS** (ESI-TOF) calcd for $C_{14}H_{14}NO_2$ $[M+H]^+$: 228.1025; found: 228.1022.

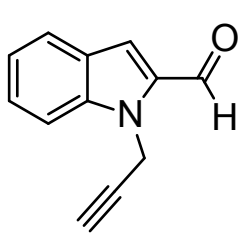
[1-(2-Propynyl)-1*H*-indol-2-yl]methanol (**23**)



Lithium aluminium hydride (1.0 M solution, 6.6 mL, 6.6 mmol, 1.2 equiv) was added dropwise to a solution of ethyl 1-propargylindole-2-carboxylate **22** (1.25 g, 5.5 mmol) in diethyl ether (20 mL) at 0 °C. After stirring at room temperature for 2 h, the mixture was added to a solution of ethyl acetate (70 mL) and water (35 mL), and the layers separated. The organic layer was washed with water (2 × 30 mL), aqueous sodium hydroxide (1 M, 20 mL), and brine (20 mL) and dried over sodium sulfate and then concentrated *in vacuo*. The crude product was purified by column chromatography (0-40% ethyl acetate in hexane) to give a white solid (0.79 g, 78%). **^{1}H NMR** (400 MHz, chloroform-d) δ 7.62 (d, *J* = 7.9, Hz, 1H), 7.45 (d, *J* = 8.2 Hz, 1H), 7.30 (t, *J* = 8.2 Hz, 1H), 7.16 (t, *J* = 7.9 Hz, 1H), 6.51

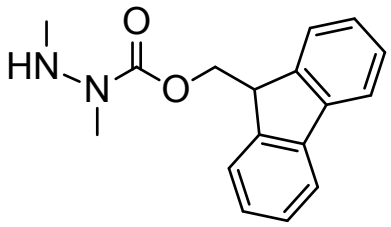
(s, 1H), 5.04 (d, *J* = 2.5 Hz, 2H), 4.90 (d, *J* = 6.0 Hz, 2H), 2.31 (t, *J* = 2.5 Hz, 1H), 1.72 (t, *J* = 6.0 Hz, 1H), 1.59 (s, 1H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 137.7, 137.3, 127.5, 122.5, 121.1, 120.2, 109.4, 102.7, 78.5, 72.4, 57.5, 32.9 ppm. **HRMS** (ESI-TOF) calcd for $C_{12}H_{12}NO$ $[M+H]^+$: 186.0919; found: 186.0921.

1-(Prop-2-yn-1-yl)-1*H*-indole-2-carbaldehyde (**24**)



Dess-Martin periodinane (1.87 g, 4.4 mmol) was dissolved in a mixture of pyridine (1 mL) and dichloromethane (8 mL). After stirring for 5 minutes at room temperature, the solution was transferred to a solution of [1-(2-propynyl)-1H-indol-2-yl]methanol **23** (0.75 g, 4.0 mmol) in dichloromethane (4 mL) and the solution stirred for a further 3 h. The reaction was then quenched by the addition of sodium thiosulphate (10% aqueous solution, 4 mL) and sodium hydrogen carbonate (sat. aqueous solution, 4 mL). The aqueous layer was extracted with dichloromethane (3 × 30 mL) and the combined organic phases dried over sodium sulfate and concentrated *in vacuo*. The crude product was purified by column chromatography (0-25% ethyl acetate in hexane) to give a white solid (0.63 g, 85%). **^{1}H NMR** (400 MHz, chloroform-d) δ 9.92 (s, 1H), 7.79 (d, *J* = 8.1 Hz, 1H), 7.57 (d, *J* = 8.5 Hz, 1H), 7.50 (dd, *J* = 8.5, 6.9 Hz, 1H), 7.32(s, 1H), 7.25 (dd, *J* = 8.1, 6.9 Hz, 1H), 5.49 (d, *J* = 2.5 Hz, 2H), 2.29 (t, *J* = 2.5 Hz, 1H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 182.7, 140.1, 134.5, 127.4, 126.7, 123.6, 121.5, 118.7, 110.8, 78.2, 72.5, 33.9 ppm. **HRMS** (ESI-TOF) calcd for $C_{12}H_{10}NO$ $[M+H]^+$: 184.0762; found: 184.0758.

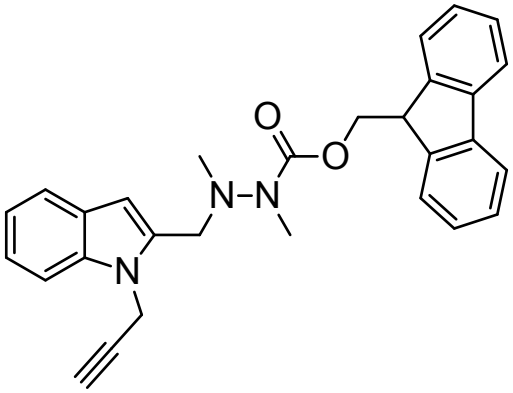
(9H-Fluoren-9-yl)Methyl 1,2-Dimethylhydrazinecarboxylate (**26**)



N,*N*'-dimethylhydrazine dihydrochloride **25** (3.990 g, 30 mmol, 2 equiv) was dissolved in acetonitrile (60 mL). Triethylamine (18 mL, 43 mmol) was added, and the solution centrifuged for 5 min. To the supernatant, a solution of Fmoc-chloride (3.870 g, 15 mmol) in acetonitrile (30 mL) was added dropwise over 2.5 h at -18 °C under argon. The solution was diluted with ethyl acetate (60 mL), washed with water (50 mL) and brine (50 mL), dried over sodium sulfate and concentrated *in vacuo*. The crude product was purified by column chromatography (0-50% ethyl acetate in hexane) to give a yellow oil (2.104 g, 50%). **^{1}H NMR** (400 MHz, chloroform-d) δ 7.80 (d, *J* = 7.6 Hz, 2H), 7.61 (d, *J* = 7.5 Hz, 2H), 7.43 (m, 2H), 7.34 (m, 2H), 4.49 (br s, 2H), 4.28 (m, 1H), 3.07 (s, 3H), 2.57 (br s, 3H) ppm. **^{13}C**

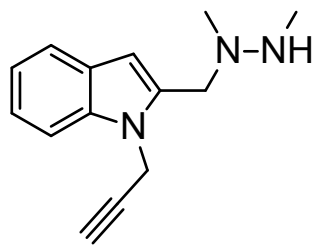
NMR (101 MHz, chloroform-d) δ 143.9, 141.4, 127.7, 127.1, 124.9, 120.0, 67.5, 47.3, 36.3, 35.7 ppm. **HRMS** (ESI-TOF) calcd for $C_{17}H_{19}N_2O_2$ $[M+H]^+$: 283.1447; found: 283.1447.

(9H-Fluoren-9-yl)methyl-1-methyl-2-((1-(prop-2-yn-1-yl)-1*H*-indol-2-yl)methyl)hydrazine-1-carboxylate (**27**)



1-(Prop-2-yn-1-yl)-1*H*-indole-2-carbaldehyde **24** (380 mg, 2.1 mmol) and (9H-fluoren-9-yl)Methyl 1,2-dimethylhydrazinemarboxylate (600 mg, 2.1 mmol) was dissolved in 1,2-dichloroethane (10 mL). Sodium triacetoxyborohydride (614 mg, 2.9 mmol, 1.4 equiv) was then added, and the resulting suspension stirred for 3 h at room temperature under Argon before quenching with sodium hydrogen carbonate (sat. aqueous solution, 5 mL). The organic layer was separated and the aqueous layer extracted with dichloromethane (5 × 5 mL). The pooled extracts were dried over sodium sulfate and concentrated *in vacuo*. The crude product was purified by column chromatography (0-40% ethyl acetate in hexane) to give a yellow oil (350 mg, 37%) **^{1}H NMR** (400 MHz, chloroform-d) δ 7.80 (d, J = 7.5 Hz, 2H), 7.56 (d, J = 7.8 Hz, 3H), 7.75–7.53 (m, 3H), 7.51–7.18 (m, 5H), 7.13 (dd, J = 8.0, 7.0 Hz, 1H), 6.43 (s, 0.45H), 6.15 (s, 0.55H), 5.29–4.21 (m, 6H), 3.70 (s, 1H), 2.80 (m, 4H), 2.19 (m, 3H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 157.0, 155.2*, 144.3*, 143.9, 141.5, 137.2, 134.8*, 134.3, 127.7, 127.1, 125.0, 124.7, 122.0, 120.7, 120.0, 119.8, 109.2, 103.8, 78.9, 72.0, 66.8, , 51.6, 51.0*, 47.3, 40.2*, 39.5, 34.3*, 32.7, 32.2*, 30.4, 29.7* ppm (* = weaker peaks, likely due to different rotamers). **HRMS** (ESI-TOF) calcd for $C_{29}H_{28}N_3O_2$ $[M+H]^+$: 450.2182; found: 450.2175.

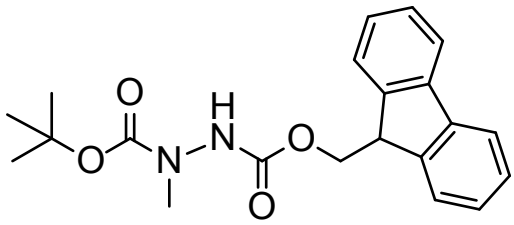
2-((2-Methylhydrazinyl)methyl)-1-(prop-2-yn-1-yl)-1*H*-indole (**19**)



(9H-Fluoren-9-yl)methyl-1-methyl-2-((1-(prop-2-yn-1-yl)-1H-indol-2-yl)methyl)hydrazine-1-carboxylate **27** (150 mg, 0.33 mmol) was dissolved in a mixture of piperidine (0.65 mL, 6.7 mmol, 20 equiv) and *N*,*N*-dimethylformamide (2.5 mL) and the resultant solution was stirred at room temperature for 30 min. The reaction mixture was then diluted with ethyl acetate (15 mL), washed with water (5 mL) and brine (3 × 5 mL), dried over sodium sulfate and concentrated *in vacuo*. The crude product

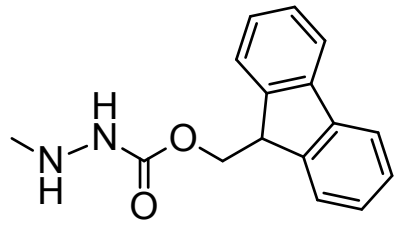
was purified by column chromatography (0-8% methanol in dichloromethane) to give a pale yellow oil (61 mg, 81%). **^{1}H NMR** (400 MHz, chloroform-d) δ 7.59 (d, *J* = 7.8 Hz, 1H), 7.45 (d, *J* = 8.2 Hz, 1H), 7.26 (t, *J* = 8.2 Hz, 1H), 7.14 (t, *J* = 7.8 Hz, 1H), 6.43 (s, 1H), 5.18 (d, *J* = 2.5 Hz, 2H), 3.95 (s, 2H), 2.62 (s, 3H), 2.44 (s, 3H), 2.27 (t, *J* = 2.5 Hz, 1H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 137.2, 135.3, 127.7, 121.8, 120.5, 119.9, 109.3, 103.5, 79.0, 72.0, 56.0, 43.4, 35.5, 32.9 ppm. **HRMS** (ESI-TOF) calcd for $C_{14}H_{18}N_3$ $[M+H]^+$: 228.1501; found: 228.1499.

1-((9*H*-fluoren-9-yl)methyl) 2-(*tert*-butyl) hydrazine-1,2-dicarboxylate (**29**)



To a solution of 1-Boc-1-methylhydrazine **28** (1 mL, 0.985 g, 6.74 mmol) in tetrahydrofuran (5 mL) and water (5 mL) was added sodium hydrogen carbonate (1.13 g, 13.5 mmol, 2 equiv) with rapid stirring. A solution of Fmoc-chloride (1.74 g, 6.74 mmol) in tetrahydrofuran (5 mL) was then added dropwise, and the reaction mixture stirred at room temperature for a further 1 h. Ether (10 mL) was added, the organic layer was washed with brine (15 mL), dried over sodium sulfate and concentrated *in vacuo* to give a yellow oil (1.64 g, 4.46 mmol, 66%). **^{1}H NMR** (400 MHz, chloroform-d) δ 7.78 (d, *J* = 7.5 Hz, 2H), 7.62 (d, *J* = 7.4 Hz, 2H), 7.43 (t, *J* = 7.4 Hz, 2H), 7.33 (t, *J* = 7.5, 2H), 4.49 (d, *J* = 7.0 Hz, 2H), 4.27 (t, *J* = 7.0 Hz, 1H), 3.16 (br s, 3H), 1.49 (s, 9H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 143.6, 141.3, 127.8, 127.1, 125.1, 120.0, 68.0, 65.9, 47.1, 28.2, 25.6 ppm. **HRMS** (ESI-TOF) calcd for $C_{21}H_{25}N_2O_4$ $[M+H]^+$: 369.1809; found: 369.1819.

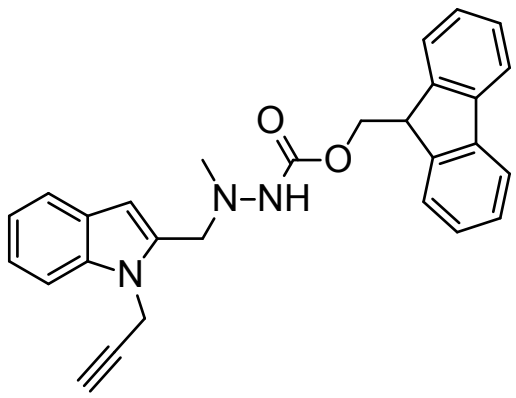
(9*H*-fluoren-9-yl)methyl 2-methylhydrazine-1-carboxylate (**30**)



1-((9*H*-fluoren-9-yl)methyl) 2-(*tert*-butyl) hydrazine-1,2-dicarboxylate **29** (1.2 g, 3.26 mmol) was dissolved in dichloromethane (6 mL) and trifluoroacetic acid (2 mL) and stirred at room temperature for 2 h. The solvent was then removed *in vacuo*, and the resulting oil was dissolved in ethyl acetate (15 mL). Saturated sodium hydrogen carbonate (10 mL) was added, and the precipitate formed was collected, re-dissolved in dichloromethane (30 mL), washed with sodium hydrogen carbonate (sat. aqueous solution 15 mL), dried over sodium sulfate and concentrated *in vacuo* to give a white solid (0.651 g, 2.43 mmol, 75%). **^{1}H NMR** (400 MHz,

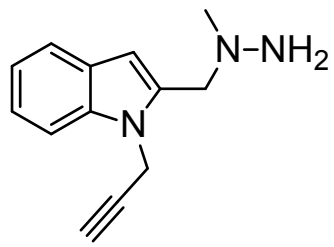
chloroform-d) δ 7.86 – 7.72 (m, 2H), 7.61 (d, *J* = 7.4 Hz, 2H), 7.51 – 7.39 (m, 2H), 7.34 (td, *J* = 7.4, 1.2 Hz, 2H), 4.48 (d, *J* = 6.7 Hz, 2H), 4.26 (t, *J* = 6.7 Hz, 1H), 2.67 (s, 3H) ppm. **^{13}C NMR** (101 MHz, chloroform-d) δ 157.1, 143.7, 141.3, 127.8, 127.1, 125.0, 120.0, 67.0, 47.2, 39.3 ppm. **HRMS** (ESI-TOF) calcd for $C_{16}H_{17}N_2O_2$ [M+H]$^+$: 269.1285; found: 269.1380.

(9*H*-fluoren-9-yl)methyl 2-methyl-2-((1-(prop-2-yn-1-yl)-1*H*-indol-2-yl)methyl)hydrazine-1-carboxylate (**31**)



1-(Prop-2-yn-1-yl)-1*H*-indole-2-carbaldehyde **24** (200 mg, 1.1 mmol) and (9*H*-fluoren-9-yl)methyl 2-methylhydrazine-1-carboxylate **30** (351 mg, 1.68 mmol, 1.5 equiv) was dissolved in dichloromethane (15 mL). Sodium triacetoxyborohydride (462 mg, 2.2 mmol, 2 equiv) was then added, and the resulting suspension stirred for 16 h at room temperature under argon before quenching with sodium hydrogen carbonate (sat. aqueous solution, 10 mL). The organic layer was separated and the aqueous layer extracted with dichloromethane (5 × 10 mL). The pooled extracts were dried over sodium sulfate and concentrated *in vacuo*. The crude product was purified by column chromatography (0-40% ethyl acetate in hexane) to give a colourless oil (359 mg, 76%).**^{1}H NMR** (500 MHz, chloroform-d) δ 7.77 (t, *J* = 7.0 Hz, 2H), 7.64 – 7.46 (m, 3H), 7.46 – 7.34 (m, 3H), 7.34 – 7.26 (m, 2H), 7.26 – 7.21 (m, 1H), 7.11 (dd, *J* = 7.9, 7.0, 1H), 6.39 (s, 1H), 5.88 (s, 1H), 5.20 (br s, 2H), 4.43 (br s, 2H), 4.15 (br s, 2H), 2.67 (s, 3H), 2.22 (br s, 1H) ppm.**^{13}C NMR** (126 MHz, chloroform-d) δ 155.1, 143.7, 141.3, 137.3, 133.9, 127.7, 127.5, 127.1, 125.1, 125.0, 122.1, 120.7, 120.0, 109.4, 104.2, 79.0, 72.0, 66.6, 47.2, 44.3, 32.9, 28.2 ppm. **HRMS** (ESI-TOF) calcd for $C_{28}H_{26}N_3O_2$ [M+H]$^+$: 436.2020; found: 436.2031.

2-((1-methylhydrazinyl)methyl)-1-(prop-2-yn-1-yl)-1*H*-indole (**20**)



(9*H*-Fluoren-9-yl)methyl 2-methyl-2-((1-(prop-2-yn-1-yl)-1*H*-indol-2-yl)methyl)hydrazine-1-carboxylate **31** (180 mg, 0.41 mmol) was dissolved in a solution of piperidine (0.65 mL, 6.7 mmol, 16 equiv) in *N,N*-dimethylformamide (2.5 mL) and stirred at room temperature for 30 min. The reaction mixture was diluted with ethyl acetate (20 mL), washed with brine (4 × 8 mL), dried

over sodium sulphate and concentrated *in vacuo*. The crude product was purified by column chromatography (0-20% methanol in dichloromethane) to give a pale yellow oil (47 mg, 54%). Probe **20** was stored neat at -80 °C until further use. **^{1}H NMR** (400 MHz, chloroform-d) δ 7.58 (d, *J* = 7.9 Hz, 1H), 7.42 (d, *J* = 8.2 Hz, 1H), 7.24 (m, 1H), 7.12 (m, 1H), 6.42 (s, 1H), 5.12 (d, *J* = 2.5 Hz, 2H), 3.84 (s, 2H), 2.53 (s, 3H), 2.25 (t, *J* = 2.5 Hz, 1H) ppm. **^{13}C NMR** (126 MHz, chloroform-d) δ 137.3, 135.0, 127.7, 122.0, 120.6, 120.0, 109.4, 104.0, 77.5, 72.1, 60.1, 48.6, 32.9 ppm. **HRMS** (ESI-TOF) calcd for $C_{13}H_{16}N_3$ $[M+H]^+$: 214.1339; found: 214.1342.

7.2.2 Generation of ODNs

Generation of AP sites from U-ODNs: Uracil containing ODNs (up to 2.5 µg) were incubated with UNG (10 U, NEB) and UNG Buffer (2.5 µL) in 25 µL reactions at 37 °C for 2 h. Reactions were cleaned up with either a mini quick-spin column (Roche) or a DNA clean and concentrator-5 kit (Zymo research) according to the manufacturer's instructions.

Generation of double-stranded AP DNA, 5-fU DNA, 5-fC DNA, 5-hmU DNA: Reaction volumes (20 µL) contained template ODN (2.5 µg), primer (10 µM), either dGTP, dCTP, dATP, dfUTP (200 µM) for 5-fU DNA, dGTP, dfCTP, dATP, dTTP (200 µM) for 5-fC DNA, dGTP, dCTP, dATP, dhmUTP (200 µM) for 5-hmU DNA, or dNTPs (200 µM), for AP DNA, DreamTaq Buffer (2 µL) and DreamTaq polymerase (0.4 µL, 2 U). Samples were heated to 95 °C for 30 s, then annealed at the stated temperatures (**Table 7.3**) for 60 s then held at 72 °C for 10 min, before purification with a GeneJET PCR purification kit according to the manufacturer's protocol.

Generation of double-stranded GCAT DNA: Reaction volumes (20 µL) contained template ODN (100 ng), forward and reverse primers (10 µM), dNTPs (200 µM), DreamTaq Buffer (2 µL) and DreamTaq polymerase (0.4 µL, 2 U). Samples were heated to 95 °C for 1 min, followed by 30 cycles of denaturation at 95 °C for 30s, annealing for 30s (**Table 7.3**), and extension at 72 °C for 20s. The final extension was held at 72 °C for 10 min, before purification with a GeneJET PCR purification kit according to the manufacturer's protocol.

7.2.3 Chemical reactions on DNA

Screening of probes: Purified AP ODNs were treated with probes at the concentration stated, in aqueous buffers as detailed. All reactions were incubated at room temperature for 2 h unless stated otherwise and purified using a mini-quick spin oligo column (Roche) according to the manufacturer's instructions. Purified DNA was then analysed by LC-MS.

CuAAC biotinylation: Purified DNA after reaction with **20** was incubated with CuBr (250 μM), THPTA (1.25 mM) and biotin-PEG3-azide (500 μM) at 37 °C for 2 h. Samples were purified using either a mini-quick oligo column (Roche) or a pre-washed Amicon Ultra-0.5 mL 10K centrifugal filter (500 μL water) and washed on the filter with water (450 μL) and Tris-HCl buffer (450 μL, 10 mM, pH 7.4) and eluted in Tris-HCl buffer (50 μL).

Sodium hydroxide cleavage: Sodium hydroxide (100 mM final concentration) was added to purified ODNs in 50 μL reactions and incubated at 70 °C for 15 min. Reactions were immediately quenched with either Tris-HCl (pH 7.0, 5 μL, 1 M) or hydrochloric acid (5 μL, 1 M) and purified with a mini quick-spin column (Roche).

Methoxyamine displacement: Methoxyamine hydrochloride (10 mM) was added to purified ODNs in 50 μL reactions in sodium phosphate buffer (40 mM, pH 6.0) and incubated for 2 h at room temperature. Reactions were purified with a mini quick-spin column (Roche).

7.2.4 Assessment of enrichment by qPCR

DNA enrichment: Enrichments were based on a reported protocol with modifications[128]. MagneSphere streptavidin magnetic beads (50 μg, Promega), were washed with 1 × binding buffer (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20) (3 × 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). Input DNA (1 ng/ODN) and poly dI:dC (2 μg, Thermo Fisher Scientific) were mixed and made up to a final volume of 50 μL, and then added to the magnetic beads, before incubation for 15 minutes at room temperature with gentle rotation. Beads were washed with 1 × binding buffer (6 × 500 μL), then incubated with NaOH (100 μL, 100 mM) at room temperature for 10 min. The beads were washed again with NaOH (100 μL, 100 mM) followed by 1 × binding buffer (3 × 500 μL) then eluted in NaOH (50 μL, 100 mM) at 70 °C for 15 min

and quenched immediately with Tris-HCl (25 µL, 500 mM, pH 7.0). A fresh sample of pre-washed streptavidin beads (75 µg) was incubated with poly dI:dC (2 µg) and resuspended in 2 × binding buffer (75 µL), to which the neutralised DNA eluent was added. The sample was incubated at room temperature for a further 15 min, before separating from the beads. The recovered DNA was purified using a ssDNA/RNA clean and concentrator (Zymo Research) according to the manufacturer's instructions with the exception that the IIC column step was omitted, and eluted in water (25 µL).

qPCR quantification of enrichment: Reaction volumes (10 µL) contained enriched DNA (1 µL), Brilliant III ultra-Fast SYBR green qPCR master mix (5 µL, Agilent Technologies), and the corresponding forward and reverse primers (1 µM each). The mixture was subject to qPCR according to the protocol outlined by the manufacturer. The extent of DNA amplification was compared to that of input samples. Primers were designed 3'- to modifications so that any possible strand cleavage would not affect amplification.

7.2.5 snAP-seq

Reaction with **20** and biotinylation: Purified DNA was treated with **20** (10 mM) in sodium phosphate buffer (40 mM, pH 7.4) at room temperature for 2 h, then purified using a mini-quick spin oligo column (Roche) according to the manufacturer's instructions. The eluted DNA was incubated with CuBr (250 µM), THPTA (1.25 mM) and biotin-PEG3-azide (500 µM) at 37 °C for 2 h. Samples were purified using a pre-washed Amicon Ultra-0.5 mL 10K centrifugal filter (500 µL water) and washed on the filter with water (450 µL) and Tris-HCl buffer (450 µL, 10 mM, pH 7.4) and eluted in Tris-HCl buffer (50 µL).

Custom P7 and P5 adapter generation: Oligos were purchased from ATDBio with double HPLC purification. Top and bottom oligos (15 µM each in 10 mM Tris-HCl pH 7.4, 1 mM EDTA, 50 mM NaCl) were annealed by heating to 95 °C for 2 min, cooled at 0.1 °C/s to 70 °C and held for 5 min, then cooled at 0.1 °C/s to 20 °C. Annealed adapters were stored at -20 °C until further use.

P7 adapter ligation: Sequencing adapters were ligated onto labelled DNA using a NEBNext Ultra II DNA library preparation kit according to the manufacturer's instructions, with the exception that the adapter was replaced with custom P7 adapter (2.5 µL). Ligated DNA was

then treated with shrimp alkaline phosphatase (NEB, 3 U) in CutSmart Buffer (NEB) for 30 min at 37 °C, before purification with AMPure XP beads (1.4 × volume) and eluted in Tris-HCl (10 mM pH 7.4, 48 µL).

Streptavidin pulldown: Magnesphere streptavidin beads (Promega, 50 µL) were pre-washed three times with 1 × binding buffer (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20) and resuspended in 50 µL of 2 × binding buffer (10 mM Tris pH 7.5, 1 mM EDTA, 2 M NaCl, 0.1% Tween 20). Poly dI:dC (2 µg, Thermo Fisher Scientific) was added to ligated DNA samples and incubated with resuspended streptavidin beads at room temperature for 15 min. Beads were washed with 1 × binding buffer (6 × 500 µL), then incubated with NaOH (100 µL, 100 mM) at room temperature for 10 min. The beads were washed again with NaOH (100 µL, 100 mM) followed by 1 × binding buffer (3 × 500 µL). DNA was eluted in NaOH (50 µL, 100 mM) at 70 °C for 15 min and quenched immediately with Tris-HCl (25 µL, 500 mM, pH 7.0). A fresh sample of pre-washed streptavidin beads (75 µL) was incubated with poly dI:dC (2 µg) and resuspended in 2 × binding buffer (75 µL), to which the neutralised DNA eluent was added. The sample was incubated at room temperature for 15 min, then the supernatant was separated and purified using a ssDNA clean & concentrator (Zymo Research) according to the manufacturer's guidelines with the exception that the IIC column step was omitted.

Primer extension: Reaction (30 µL) containing purified ssDNA, dNTPs (200 µM), P7 primer (1 µM) and NEBuffer 2 was heated to 95 °C for 1 min, annealed at 65 °C for 30 s and held at 37 °C for 30 min, at which point Klenow fragment (3'->5'- exo-, NEB, 2 U) was added. The synthesised dsDNA was purified using a DNA clean and concentrator-5 kit (Zymo Research) according to the manufacturer's instructions and eluted in Tris-HCl (10 mM, pH 7.4).
P5 adapter ligation: DNA (22.5 µL), Blunt/TA ligase master mix (25 µL, NEB) and custom P5 adapter (2.5 µL) were incubated at 20 °C for 30 min. Libraries were purified with AMPure XP beads (1.5 × volume) and eluted in Tris-HCl (10 mM, pH 7.4) before amplification using a Q5 hot start high-fidelity master mix (NEB) with library amplification primers (10 µM each). Libraries were quantified using a KAPA library quantification kit and sequenced on either an Illumina MiSeq, or NextSeq machine.

Input library preparation: DNA was treated in the same way as above with the NEBNext kit, with the exception that the custom P7 adapter was substituted by TruSeq Nano indexed adapters (Illumina, 2.5 µL). Samples were purified twice using AMPure XP beads (1.0 × volume) to remove excess adapters and then amplified as in snAP-seq.

7.2.6 Data analysis

The distribution of reads aligning to reference oligonucleotide sequences and patterns of alignment start sites were obtained using samtools view and standard Unix tools. For each library, the total number of reads aligned to the forward strand each ODN was first normalised to the corresponding input library. Only the reverse strand was used for normalisation due to a large underrepresentation of AP site-containing strands (forward) in input libraries. See **Table 7.3** for definition of forward and reverse strands.

qPCR enrichment: the relative recovery of each dsDNA ODN was calculated from the ΔC_t value when compared to the amplification of a corresponding input sample prepared without enrichment. The recovery of AP DNA was expressed as a fold enrichment relative to GCAT DNA, 5-fC DNA or 5-fU DNA.

7.3 Chapter 3 methods

7.3.1 DNA purification

Leishmania major DNA (product 30012D, lot 62762024) was purchased from ATCC. DNA was purified before use with a DNeasy blood and tissue kit (Qiagen) according to the manufacturer's instructions, with the exception that the protein digestion step was omitted. The purified DNA was eluted in Tris-HCl (50 µL, 10 mM, pH 7.4).

Trypanosoma brucei DNA was obtained from the Carrington Group and re-purified before use with a DNeasy blood and tissue kit (Qiagen) according to the manufacturer's instructions. The purified DNA was eluted in Tris-HCl (50 µL, 10 mM, pH 7.4).

7.3.2 Enzymatic reactions

SMUG1 excision: DNA was incubated with hSMUG1 (25 U, NEB) in NEBuffer 1 supplemented with BSA (100 µg/mL) at 37 °C for 18 h, and purified using either a oligo clean and concentrator (Zymo Research), DNA clean & concentrator kit (Zymo Research), or AMPure XP beads (2.0 ×), depending on the size of DNA used. Short ssODNs were then analysed by LC-MS.

7.3.3 Sequencing methods

SMUG1-snAP-seq of *Leishmania major* DNA: Spike-in DNA (50 pg) was added to purified genomic DNA (400 ng) and the combined DNA was sonicated to an average of 450 bp using a Covaris M220 system. Samples were treated with hSMUG1 (25 U, NEB) in NEBuffer 1 supplemented with BSA (100 µg/mL) at 37 °C for 18 h before purification using AMPure XP beads (2.0 ×), or a DNA clean & concentrator kit (Zymo Research). Samples were then subjected to snAP-seq, see section 7.2.5 for detailed protocol.

snAP-seq of *Leishmania major* DNA: as a control without SMUG1 treatment, standard snAP-seq was carried out as described above in section 7.2.5 and amplified using the same PCR conditions as that for SMUG1-snAP-seq libraries.

UNG-snAP-seq of *Leishmania major* DNA: All steps were carried out as described above for SMUG1-snAP-seq, with the exception that SMUG1 enzyme was replaced by UNG (10 U, NEB) and BSA was omitted.

Methoxyamine blocking SMUG1-snAP-seq: sonicated DNA was first incubated with methoxyamine hydrochloride (10 mM) in sodium phosphate buffer (40 mM) for 2 h at room temperature and purified using AMPure XP beads (2.0 ×). Samples were then subjected to the standard SMUG1-AP-seq protocol detailed above.

5-hmU DIP-seq: Spike-in DNA (20 pg) was added to *T. brucei* DNA (200 ng per replicate) and sonicated to an average length of 200 bp using a Covaris M220 system. Adapters (TruSeq Nano, Illumina) were introduced onto DNA fragments using a NEBNext Ultra II library preparation kit (NEB) according to the manufacturer's instructions, where the TruSeq adapters were used in place of NEB adapters. DNA was purified using AMPure XP beads (0.75 ×) and eluted in ultra-pure water. DNA was made up in PBS with 0.1% (v/v) tween-20 to a total volume of 200 µL, and heat denatured by incubation at 95 °C for 10 min. Samples were immediately placed on ice. For antibody binding, the denatured DNA was incubated with 5-hmU antibody (ab19735, Abcam, 10 µL) and rabbit anti-goat IgG antibody (ab6697, Abcam, 5 µL) along with salmon sperm DNA (1 µL, 10 mg/ml, Thermo Fisher Scientific) overnight at 4 °C with rotation. For immunoprecipitation, Dynabeads protein G (25 µL, Life Technologies) were pre-washed with citrate-phosphate buffer (200 µL, 0.5 ×, Alfa Aesar) followed by PBS with 0.1% tween 20 (2 × 200 µL). DNA samples were then added to beads

and incubated at 4 °C with rotation for 2 h. The mixture was then placed on a magnetic rack, and beads were washed with PBS with 0.1% tween 20 (5 × 200 µL) with mild vortexing during each wash. For elution, beads were suspended in elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, and 0.5% SDS) and proteinase K (20 µg/µL, 1.75 µL) was added. The mixture was incubated at 50 °C with vigorous shaking (1300 rpm) overnight. The supernatant was collected and purified using a DNA clean and concentrator-5 kit (Zymo Research) according to the manufacturer's instructions and eluted in ultra-pure water (27 µL). Libraries were amplified using TruSeq PCR Master Mix and PCR Primer Cocktail (Illumina) according to the manufacturer's guidelines, for a total of 12 cycles. DNA libraries were purified using AMPure XP beads (0.8 ×).

DIP-seq input libraries: For input libraries, adapter ligation was carried out as described for DIP-seq libraries. Samples were diluted around 200-fold, then subjected to the same PCR conditions as for DIP-seq.

IgG DIP-seq: IgG antibody was used as a control for DIP-seq libraries. All steps were carried out as described for 5-hmU DIP-seq, with the exception that 5-hmU antibody was replaced with IgG control antibody (ab37373, Abcam). Due to the reduced amount of DNA recovered, PCR amplification was carried out for 3 addition cycles, for a total of 15 cycles.

All DIP-seq libraries, including inputs and controls were sequenced by single-end sequencing, using a 150-cycle kit (Illumina).

7.3.4 Data analysis

Alignment of reads: Trimmed reads were aligned to the reference *L. major* genome (Sanger Institute) using bwa mem. Aligned reads were then filtered with removal of unaligned reads, secondary/alternative alignments, PCR duplicates, alignments with a quality score of less than 10 and spike-ins. Clean alignments were further split into alignments of R1 reads mapping to the forward and to the reverse strands respectively.

Calling high-confidence SMUG1-snAP-seq sites: Single-nucleotides with positive $\log_2$ fold-change (SMUG1-snAP-seq vs. input) at an FDR threshold smaller than 10^{-10} across two technical replicates analysed in parallel were considered high-confidence sites. See section 7.1.9 for details.

Overlap with 5-hmU and base J datasets: intersections of detected sites with 5-hmU and base J peaks obtained in Kawasaki *et al*. 2017[70] were performed with bedtools intersect.

Base composition profiles: Fasta files containing sequences flanking either the high-confidence SMUG1-snAP-seq sites, or total reads in sequencing libraries by 10 bp were generated using bedtools getfasta. Base composition plots were then produced with ggplot2.

Coverage profiles: coverage profile plots were obtained using the computeMatrix and plotProfile functions available in deeptools, for regions 1 kb on either side of high-confidence SMUG1-snAP-seq in SMUG1-snAP-seq, snAP-seq and input libraries.

Motif analysis: Fasta files containing sequences flanking the SMUG1-snAP-seq sites by 5 bp were generated using bedtools getfasta. As a control, similar fasta files were obtained for all T bases genome-wide. The ggseqlogo library in R was used to generate sequence logo representations. The dreme tool as available in meme v4.11.2[237] was used to extract sequence motifs.

Synthetic 5-hmU *N*-oligo analysis: trimmed reads were deduplicated based on the two randomised N10 stretches flanking the 5-hmU site. Fasta files and tables of nucleotide counts were generated with in-house Python v2.7.12 scripts. The ratio of nucleotide counts between the flanking regions of high-confidence SMUG1-snAP-seq sites and the flanking regions of a randomised set of control thymine positions in the *L. major* genome was calculated. The same ratio was also calculated between the enriched and control libraries generated using the 5-hmU *N*-oligo. For statistical testing, Fisher's Exact and Pearson's Chi-squared tests were used to compare the nucleotide counts and calculated ratios between the genomic and synthetic libraries, as implemented in the R v3.3.2 programming language. This analysis confirmed that the enrichment in the TG motif within SMUG1-snAP-seq data was significant compared to that of the synthetic libraries ($p < 0.05$). Sequence logo representations were generated with the ggseqlogo v0.1 R library (https://cran.rstudio.com/web/packages/ggseqlogo/index.html).

DIP-seq: macs2 v2.1.1.20160309[252] callpeak was used to obtain regions of enriched signal using input libraries as control, with options –nomodel and either $p < 10^{-10}$, or $q < 0.05$ as indicated. High-confidence peaks were defined as those that appear in at least two out of three technical replicates.

7.4 Chapter 4 methods

7.4.1 siRNA knockdown

siRNAs were purchased from Dharmacon (ON-TARGETplus human APEX1 siRNA, product code J-010237-08-0002 and ON-TARGETplus non-targeting pool, product code D-001810-10-05) and resuspended in water to give 20 μM stocks. For transfections, HeLa cells were seeded in 6-well plates at 100,000-200,000 cells/well and grown overnight. On the second day, cells were transfected with 10 nM of siRNA or control siRNA using Lipofectamine RNAiMAX according to the manufacturer's instructions. After 48-96 h cells were washed once in cold PBS and harvested by scraping into cold PBS (500 μL) followed by centrifugation at 300 × g. Cell pellets were flash frozen at –80 °C for further use.

For analysis of mRNA expression, cell pellets were collected as above and total RNA was extracted using a RNeasy mini kit (Qiagen) according to the manufacturer's instructions. Isolated RNA was quantified by UV spectrophotometry (Nanodrop One, Thermo Fisher Scientific), then used as the template for cDNA synthesis using a High capacity cDNA reverse transcription kit (Applied Biosystems) according to the manufacturer's instructions. cDNA samples were diluted 1:4 and 1 μL was quantified in 10 μL qPCR reactions containing Brilliant III Ultra-Fast SYBR® green qPCR master mix (Agilent Technologies) and the relevant primer sets (1 μM each) (**Table 7.1**). The extent of amplification (Cq value) for APE1 mRNA was normalised against that of β-Tubulin and compared to siRNA-free cells.

Name	***Primer sequence (5'-3')***	***PCR product size (bp)***
APE1	Forward: TGG AAT GTG GAT GGG CTT CGA GCC Reverse: AAG GAG CTG ACC AGT ATT GAT GA	169
β-Tubulin	Forward: TTG GCC AGA TCT TTA GAC CAG ACA AC Reverse: CCG TAC CAC ATC CAG GAC AGA ATC	122

Table 7.1: Primer sequences used for quantification of mRNA levels by RT-qPCR.

For analysis of protein expression, cell pellets were collected as above and lysed in 100-150 µL cold RIPA buffer (Thermo Fisher Scientific) supplemented with Halt protease inhibitor cocktail (Thermo Fisher Scientific) and shaken on ice for 15 min. The lysate was centrifuged at maximum speed for 15 min and the supernatant was collected. The total protein concentration was quantified using a BCA protein assay kit (Thermo Fisher Scientific) according to the manufacturer's instructions. A 0.2 mg/mL sample of each lysate was analysed by Simple Western blot using a 12-230 kDa Wes separation module (Protein Simple). APE1 protein was detected by anti-APE1 antibody (Abcam ab194, 1:500 dilution), with anti-β-Tubulin (Cell Signaling #86298, 1:50 dilution) used as a loading control.

7.4.2 DNA extraction

Frozen cell pellets were thawed at room temperature, then DNA was extracted using a Quick-DNA kit (Zymo Research) according to the manufacturer's guidelines, where the genomic lysis buffer was supplemented with TEMPO (20 mM) immediately before each extraction to reduce DNA damage formation[291]. Purified DNA was eluted in Tris-HCl (10 mM, pH 7.4) and quantified using a Qubit 2.0 fluorometer as well as a Nanodrop.

For comparison, DNA was also extracted using a DNeasy blood & tissues kit (Qiagen) according to the manufacturer's instructions. RNase A (400 µg) was added according to the manufacturer's guidelines to remove RNA. DNA was eluted in Tris-HCl (10 mM, pH 7.4).

7.4.3 Sequencing methods

snAP-seq: the standard snAP-seq protocol was carried out as described above, using up to 2.5 µg purified genomic DNA per replicate during the chemical reaction with **20**, supplemented with spike-in DNA (50 pg). The total amount of purified DNA subjected to P7 adapter ligation did not exceed 1 µg for library, in line with the manufacturer's guidelines.

Extraction and sonication controls on *L. major* DNA: For mock AP sonication, SMUG1 treatment was carried out on full-length genomic *L. major* DNA and purified with AMPure XP beads (2.0 ×), after which DNA was sonicated then processed as above for snAP-seq. For re-extraction, SMUG1 treatment was carried out on full-length *L. major* DNA, which was then purified using AMPure XP beads (2.0 ×) and re-extracted using either the DNeasy blood and

tissue kit (Qiagen), or the Quick-DNA mini kit (Zymo Research). Both extractions were carried out according to the manufacturer's instructions, with the exception that proteinase K digestion was carried out for 30 min at 37 °C for the DNeasy kit. All lysis buffers were supplemented with TEMPO (20 mM) to reduce additional DNA damage.

7.4.4 Data analysis

Alignment of reads: Trimmed reads were aligned to the reference human genome (UCSC hg38) using bwa mem. Aligned reads were then filtered with removal of unaligned reads, secondary/alternative alignments, PCR duplicates, alignments with a quality score of less than 10, reads overlapping blacklisted regions (http://mitra.stanford.edu/kundaje/akundaje /release/blacklists/hg38-human/) and spike-ins. Clean alignments were further split into alignments of R1 reads mapping to the forward and to the reverse strands respectively.

Calling single-nucleotides: Genome-wide modelling and comparative assessment of counts from enriched and input libraries was carried out as described in section 7.1.9 using four replicates in parallel, and assessment of the volcano plots generated from this analysis revealed that no threshold could be set in which positive $\log_2$ fold-change (snAP-seq vs. input) could be favoured without the detection of sites with negative $\log_2$ fold-change. Therefore, no high-confidence sites could be called from this data.

Peak-calling: macs2 v2.1.1.20160309[252] callpeak was used to obtain regions of enriched signal of snAP-seq using input libraries as control, with options $p < 0.00001$, and --nomodel. To further correct for regions with naturally high coverage of reads, macs2 callpeak with option --nomodel was also used in the input libraries only. Overlaps between the enriched and naturally high regions were obtained using bedtools intersect and subtracted from the regions of enriched signal. High-confidence peaks were defined as consensus peak regions obtained between three out of four (HeLa) replicates. For visualisation of enrichment within a given genomic region, raw read counts within the selected region were normalised by dividing by the total number of reads in the region × 1,000.

Testing genomic associations: hg38 gene feature annotations (promoters, 5'UTR, 3'UTR, exons, introns and intergenic regions) were extracted from the UCSC's genes.gtf file using the library GenomicFeatures[292] in R and searching for regular expressions using Python (https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder). Computing

the significance of overlap between the genomic annotations and snAP-seq peaks was performed with the Genomic Association Tester (GAT)[293]. Genomic associations were also tested against available datasets of chromatin accessibility and histone modifications (**Table 7.2**). The Storey's *q*-value for the relative fold enrichment of each genomic feature was obtained from the Genomic Association Tester, where a threshold of $q < 0.05$ was used for statistical significance.

Genomic assay	*Accession code*
DNase-seq	ENCFF950NDW
FAIRE-seq	ENCFF001UYM
ATAC-seq	GSM2830381
H3K27ac	ENCFF392EDT
H3K4me3	ENCFF862LUQ
H3K27me3	ENCFF512TQI
H3K9me3	ENCFF712ATO

Table 7.2: Encode and GEO datasets used to analyse HeLa snAP-seq peaks.

Nucleotide analysis: To analyse the nucleotide at position '0' corresponding to reads in sequencing libraries, the frequency of each base occurring at the nucleotide directly 5'- to sequencing read start sites after alignment was normalised by the frequency of each nucleobase as an average across the genome, and the fold enrichment was plotted using Prism. A two-way ANOVA (Sidak's multiple comparisons test) was performed to compare the fold enrichment of each nucleobase between different libraries.

qPCR enrichment: the fold enrichment of AP DNA relative to GCAT DNA was calculated as described in section 7.2.6. The enrichment between the different conditions was compared using a one-way ANOVA, and the percentage recovery of each model sequence was compared using a Kruskal-Wallis test.

RNA expression analysis:

RNA-seq data generated using HeLa cells was obtained from ENCODE (accession code ENCSR000CPR). Raw reads were trimmed and quality checked, before aligning to the reference human genome (UCSC hg38) using rsem[294]. To investigate the relationship between AP peaks and gene expression levels, the co-ordinates of high-confidence snAP-seq peaks detected in cells treated with control siRNA were intersected with extended gene

bodies, defined here as the gene body ± 1 kilobase using bedtools. The two RNA-seq replicates were analysed separately, and the normalised expression level (TPM, transcripts per million reads) was compared between genes that contained at least one snAP-seq peak in the extended gene body (positive), and genes that did not contain any snAP-seq peak (negative). In all cases, the number of negative genes was larger than the number of positive genes. Therefore, to keep the sizes of the two datasets comparable, samples were taken from the negative genes to reflect the number of positive genes, where such sampling was repeated a total of five times. Pairwise testing for differences in the normalised gene expression was carried out for each of the five samples that did not contain snAP-seq peaks against the genes containing snAP-seq peaks (Wilcoxon test), and the combined samples were also tested against positive genes using a Kruskal Wallis test. For visualisation purposes only, normalised gene expression for genes containing at least one snAP-seq peak was plotted alongside that for all other genes. Data analysis and visualisation was performed using R programming language unless stated otherwise.

snAP-seq peaks which fall within promoter regions, defined as 1 kilobase upstream of transcription start sites were analysed using the same pipeline as described above.

Microarray data, generated in control and APE1 siRNA knockdown cells were obtained from Vascotto *et al.*[262]. Analysis was carried out in the same pipeline as for RNA-seq, with the exception that normalised counts for each RNA probe were used instead of TPMs for each gene. Analysis was carried out separately for snAP-seq peaks which were detected in control cells and those detected in cells treated with APE1 siRNA, where the corresponding microarray dataset was used. To investigate the changes in expression between APE1 knockdown and control cells, snAP-seq peaks detected in cells treated with control or APE1 siRNA were intersected using bedtools and separated into those that were unique to control cells, unique to APE1 siRNA cells, or common to both. RNA probes corresponding to genes were intersected with each of the three classes of snAP-seq peaks and separated into those that contained at least one snAP-seq peak of each class, and those that do not. This analysis was performed for genes that fall within extended gene bodies, as well as promoters. The change in normalised expression, defined as the $\log_2$ fold change in normalised counts (APE1 knockdown vs. control) for each RNA probe was compared for genes containing snAP-seq peaks and those that did not, where sampling was performed as described above to ensure equal sample sizes during testing.

7.5 Chapter 5 methods

7.5.1 *E. coli* culture and DNA extraction

E. coli strain CJ236 (NEB) stock was streaked on LB agar containing chloramphenicol (15 µg/ml) and grown overnight. A single colony was selected and streaked onto fresh LB agar. A sample was selected, and culture overnight in LB broth at 37 °C. After incubation, stocks were made by diluting cultures 1:1 in glycerol, and frozen for further use.

For harvesting of *E. coli* cells, samples of CJ236 taken from glycerol stocks were grown overnight in LB broth. Aliquots were collected by centrifugation and washed in PBS. DNA was extracted using a Quick-DNA fungal/bacterial kit (Zymo Research) according to the manufacturer's instructions, with vortexing of BashingBeads for 10 min. DNA was eluted in Tris-HCl (50 µL, 10 mM, pH 7.4).

7.5.2 mESC DNA extraction

mESCs were cultured by the Reik Group (Dr Fátima Santos). Cells were grown in 10 cm dishes, and flash frozen after harvesting. Cell pellets were thawed at room temperature, and DNA was extracted using a DNeasy blood & tissues kit (Qiagen) according to the manufacturer's instructions. RNase A (400 µg) was added according to the manufacturer's guidelines to remove RNA. DNA was eluted in Tris-HCl (100 µL, 10 mM, pH 7.4).

7.5.3 Sequencing methods

UNG-snAP-seq: Spike-in DNA (50 pg) was added to purified genomic DNA (up to 2.5 µg per replicate) and sonicated to an average length of 450 bp using a Covaris M220 system. DNA was treated with UNG (2 µg, NEB) in UNG buffer (1 ×) in a total reaction volume of 25 µL for 2 h at 37 °C. The DNA was purified using a DNA clean and concentrator-5 kit (Zymo Research) according to the manufacturer's instructions and eluted in Tris-HCl (10 mM, pH 7.4). Samples were then subjected to snAP-seq as detailed above.

Targeted snAP-seq: DNA was treated with shrimp alkaline phosphatase (NEB, 3 U) in CutSmart Buffer (NEB) for 30 min at 37 °C, then purified using AMPure XP beads (1.4 ×). Samples were then subjected to streptavidin pulldown as described in standard snAP-seq, with re-purification in a second streptavidin incubation to remove non-specific release of DNA. The recovered DNA was purified using a ssDNA/RNA clean and concentrator kit (Zymo Research) according to the manufacturer's instructions and eluted in water (25 µL). The ssDNA recovered was then annealed to one or more P7 target primer (1 µM each) by heating to 95 °C for 1 min, annealing at 65 °C for 30 s and held at 37 °C for 30 min, in the presence of dNTPs (200 µM), and NEBuffer 2. Klenow fragment (3'→5'- exo-, NEB, 2 U) was added once the reaction mixture reached 37 °C. The synthesised dsDNA was purified using a DNA clean and concentrator-5 kit (Zymo Research) according to the manufacturer's instructions and eluted in Tris-HCl (10 mM, pH 7.4). P5 target adapter was introduced by ligation where DNA (22.5 µL), Blunt/TA ligase master mix (25 µL, NEB) and P5 target adapter (2.5 µL) were incubated at 20 °C for 30 min. Libraries were purified with AMPure XP beads (1.5 × volume) and eluted in Tris-HCl (10 mM, pH 7.4) before amplification using a Q5 hot start high-fidelity master mix (NEB) with library amplification primers (10 µM each). Libraries were quantified using a KAPA library quantification kit and sequenced on either an Illumina MiSeq, or NextSeq machine.

7.5.4 Data analysis

Alignment of reads: Trimmed reads were aligned to the reference *E. coli* genome (K-12 MG1655, Ensembl bacteria) or GRCm38 using bwa mem for *E. coli* and mESC libraries, respectively. Aligned reads were filtered with removal of unaligned reads, secondary/alternative alignments, PCR duplicates, alignments with a quality score of less than 10, reads overlapping blacklisted regions and spike-ins. Clean alignments were further split into alignments of R1 reads mapping to the forward and to the reverse strands respectively.

Calling single-nucleotides: Genome-wide modelling and comparative assessment of counts from enriched and input libraries was carried out as described in section 7.1.9 using either two replicates in parallel (*E. coli* CJ236), or a single replicate with corresponding input (mESCs).

For *E. coli* data, a set of sites were obtained using a threshold of FDR < 0.05 and $\log_2$ fold-change > 0 (UNG-snAP-seq vs. input). A further set of high-confidence sites were obtained by setting the *p*-value threshold below the minimum *p*-value of sites with negative $\log_2$fold-change (UNG-snAP-seq vs. input), at $p < 0.000628$.

For mESC data, assessment of the volcano plots generated from this analysis revealed that no threshold could be set in which positive $\log_2$ fold-change (UNG-snAP-seq vs. input) could be favoured without the detection of sites with negative $\log_2$ fold-change. Therefore, no high-confidence sites could be called from this data. A potential set of sites, with $p < 0.05$ and $\log_2$ fold-change (UNG-snAP-seq vs. input) > 0 were selected for further analysis.

Testing genomic associations: K-12 MG1655 gene feature annotations (exons, promoters, intergenic, operons, terminators, transcription factor binding sites, 5'UTR and 3'UTR) were extracted from the reference genome .gtf file using the library GenomicFeatures[292] in R, or obtained from the RegulonDB database. Computing the significance of overlap between the genomic annotations and UNG-snAP-seq peaks was performed with the Genomic Association Tester (GAT)[293]. The Storey's *q*-value for the relative fold enrichment within each genomic feature was obtained from the Genomic Association Tester, where a threshold of $q < 0.05$ was used for statistical significance.

Base composition profiles: Fasta files containing sequences flanking either total reads in sequencing libraries, or the potential set of peaks with $p < 0.05$, by 10 bp were generated using bedtools getfasta. Base composition plots were then produced using ggplot2 for visualisation.

7.6 Oligonucleotides

ODN	*Sequence 5'-3'*
U-ODN1	AGC GAC A**U**A TCT TGT
AP-ODN1	AGC GAC A-**AP**-A TCT TGT
fU-ODN2	ATC GCA **5-fU**GT A
fC-ODN3	ATC G**5-fC**G CGT A
fC-ODN4	TAA TTA TC**5-fC** GGA CTC ATA AG
U-ODN5	AAC ACG A**U**A GGA AGC
AP-ODN5	AAC ACG A-**AP**-A GGA AGC
hmU-ODN	p-ATC GCA **5-hmU**GT A
Base J-ODN	GAA C**J**G **J**C**J** GAG
AP DNA1 template (forward)	CAC ACC GCC AGC CAC AGC AAC GAA CG-**AP** GCA GCG CCC CTC ACG CCA CAG AAC ATC GCA TTT ACG ACG ATT GAT GTA CTA AAT AGT GGG TGG TCG GTT CGC G
AP DNA1/ U DNA primer	CGC GAA CCG ACC ACC CAC TA (T_{anneal} = 60 °C)
AP DNA1/ U DNA (reverse)	CGC GAA CCG ACC ACC CAC TAT TTA GTA CAT CAA TCG TCG TAA ATG CGA TGT TCT GTG GCG TGA GGG GCG CTG CAC GTT CGT TGC TGT GGC TGG CGG TGT G
U DNA template (forward)	CAC ACC GCC AGC CAC AGC AAC GAA CG**U** GCA GCG CCC CTC ACG CCA CAG AAC ATC GCA TTT ACG ACG ATT GAT GTA CTA AAT AGT GGG TGG TCG GTT CGC G
5-fU DNA1 template (reverse)	CTG TTC GCT GGT CTG CTT GTG CTC TTC TGT CTT GCT TGG TTC TCT TGG TTC GTG TGT CTT GGC TTT CCT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A
5-fU DNA1 primer	TAG CCA TAC TGC CTC GTC CG (T_{anneal} = 68 °C)
5-fU DNA1 (forward)	TAG CCA TAC TGC CTC GTC CGG ACA CG**5-fU** CAG GAG GAA AGC CAA GAC ACA CGA ACC AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G
5-fC DNA1 template (reverse)	CCT CAC TCA CCT CCA CCC TCT CAC TAC CTC ACT CTT CCT CCT AAC CCT CTC CAA CCA CCT CTC CAC CCT CCT AGA TCT CTA CCT GAC TGA GCG TGT GCG A
5-fC DNA1 primer	TCG CAC ACG CTC AGT CAG GT (T_{anneal} = 68 °C)
5-fC DNA1 (forward)	TCG CAC ACG CTC AGT CAG GTA GAG AT**5-fC** TAG GAG GGT GGA GAG GTG GTT GGA GAG GGT TAG GAG GAA GAG TGA GGT AGT GAG AGG GTG GAG GTG AGT GAG G

GCAT DNA1 template (forward)	GGC CAC CAC CCG CAC ATA CTC TGG TAC GAT TAC GAA CAC AGC CCG ACA CCA CCT CTA ATG AAC GTC GCT TAT AGT GAT TAA CGC CCC GTA GAC ACC ATG G
GCAT DNA1 primer	Forward: GGC CAC CAC CCG CAC ATA CT Reverse: CCA TGG TGT CTA CGG GGC GT (T_{anneal} = 60 °C)
GCAT DNA1 (reverse)	CCA TGG TGT CTA CGG GGC GTT AAT CAC TAT AAG CGA CGT TCA TTA GAG GTG GTG TCG GGC TGT GTT CGT AAT CGT ACC AGA GTA TGT GCG GGT GGT GGC C
5-hmU DNA template (reverse)	CTG TTC GCT GGT CTG CTT GTG CTC TTC TGT CTT GCT TGG TTC TCT TGG TTC GTG TGT CTT GGC TTT CCT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A
5-hmU DNA primer	TAG CCA TAC TGC CTC GTC CG (T_{anneal} = 65 °C)
5-hmU DNA (forward)	TAG CCA TAC TGC CTC GTC CGG ACA CG**5-hmU** CAG GAG GAA AGC CAA GAC ACA CGA ACC AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G
AP DNA2 template (reverse)	GCG TCA GGG AGT GCG AAT TCC CTC CTA GAC TCC AAT GAG CCT AAC GAT TCG GTA GCC AGC GAG ACG CCG CTG CAG GAT TCT AGG CAG CTA GGT ATG CGT AGG TTC
AP DNA2 primer	GCG TCA GGG AGT GCG AAT TC(T_{anneal} = 65 °C)
AP DNA2 (forward)	GAA CCT ACG CAT ACC TAG CTG CCT AGA ATC CTG CAG CGG CGT CTC GCT GGC TAC CGA ATC GT-**AP** AGG CTC ATT GGA GTC TAG GAG GGA ATT CGC ACT CCC TGA CGC
GCAT DNA2 template (forward)	GGC CAC CAC CTG CAC ATA CTC CAT CTG TAG AGA CGG TGG ACA CGT CCG AGT ACG CTG GCA TAC CCG TAG TAC TCT GCC TAA TGA GGA AGC TAC TTC CAC CCA CGG
GCAT DNA2 primer	Forward: CCG TGG GTG GAA GTA GCT TC Reverse: GGC CAC CAC CTG CAC ATA CT (T_{anneal} = 60 °C)
GCAT DNA2 (reverse)	CCG TGG GTG GAA GTA GCT TCC TCA TTA GGC AGA GTA CTA CGG GTA TGC CAG CGT ACT CGG ACG TGT CCA CCG TCT CTA CAG ATG GAG TAT GTG CAG GTG GTG GCC
5-fU DNA2 (forward)	TAG CCA TAC TGC CTC GTC CGG ACA CG**5-fU** CAG GAG GAA AGC CAA GAC ACA CGA ACC AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G
5-fU DNA2 primer	TAG CCA TAC TGC CTC GTC CG (T_{anneal} = 65 °C)
5-fC DNA2 template (reverse)	CCT CAC TCA CCT CCA CCC TCT CAC TAC CTC ACT CTT CCT CCT AAC CCT CTC CAA CCA CCT CTC CAC CCT CCT AGA TCT CTA CCT GAC TGA GCG TGT GCG A
5-fC DNA2 primer	TCG CAC ACG CTC AGT CAG GT (T_{anneal} = 65 °C)

5-fC DNA2 (forward)	TCG CAC ACG CTC AGT CAG GTA GAG AT**5-fC** TAG GAG GGT GGA GAG GTG GTT GGA GAG GGT TAG GAG GAA GAG TGA GGT AGT GAG AGG GTG GAG GTG AGT GAG G
AP DNAx template (reverse)	CTG TTC GCT GGT CTG CTT GTG CTC TTC TGT CTT GCT TGG TTC TCT TGG TTC GTG TGT CTT GGC TTT CCT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A
AP DNAx primer	TAG CCA TAC TGC CTC GTC CG (T_{anneal} = 65 °C)
AP DNAx (forward)	TAG CCA TAC TGC CTC GTC CGG ACA CG-**AP** CAG GAG GAA AGC CAA GAC ACA CGA ACC AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G
P7 adapter (top) (also P7 primer)	Me-GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T Where Me indicates a 5'-OMe modification
P7 adapter (bottom)	p-GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC-NNNNNN-ATC TCG TAT GCC GTC TTC TGC TTG-spacerC3 where NNNNNN indicates index sequence (TruSeq Nano) and spacerC3 represents a 3'-C3 spacer modification
P5 adapter (top)	GAA TGA TAC GGC GAC CAC CGA GAT CTA CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CT
P5 adapter (bottom)	p-GAT CGG AAG AGC G
P7 target primer	GAC TGG AGT TCA GAC GTG TGC TCT TCC GAT CT-N'N'N'N'N'N'-COMP' where N'N'N'N'N'N' indicates the reverse complement of index sequence s(TruSeq nano) and COMP' indicates the reverse complement of the target sequence.
P5 target adapter	Top: CGC TCT TCC GAT CddT Bottom: p-GAT CGG AAG AGC GTC GTG TAG GGA AAG AGT GTA GAT CTC GGT GGT CGC CGT ATC ATT C where ddT indicates a dideoxythymine modification
Library amplification primers	Forward: AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GA Reverse: CAA GCA GAA GAC GGC ATA CGA GAT
AP (U) DNA qPCR primers	Forward: GCC CCT CAC GCC ACA GAA CA Reverse: CGC GAA CCG ACC ACC CAC TA
GCAT DNA qPCR primers	Forward: GGC CAC CAC CCG CAC ATA CT Reverse: CCA TGG TGT CTA CGG GGC GT
5-fU DNA qPCR primers	Forward: AAG CAA GGA AGC AAG GCA GA

	Reverse: TGC TTG GCT GCG TGG TCT CG
5-fC DNA qPCR primers	Forward: GTA TGG AGG AAT GAG TGT GG Reverse: TAA CTA CCT ATC TAC CAT TC
5-hmU *N*-oligo template (forward)	GTC TAC CTG AAC GCC GCT GTN NNN NNN NNN **5-hmU**NN NNN NNN NNG TAG TAG TCG ACT AGA CG TCC AAC CAA CGG AAG GGT ATT CGG ACG AGG CAG TAT GGC TA where N indicates randomised bases
5-hmU *N*-oligo primer	TAG CCA TAC TGC CTC GTC CG (T_{anneal} = 65 °C)

Table 7.3: Sequences of all synthetic ODNs used. AP DNA and GCAT DNA templates were purchased from IDT with PAGE purification. 5-hmU, 5-fC and 5-fC DNA template were purchased from Biomers with PAGE purification. 5-hmU *N*-oligo was purchased from ATDBio with HPLC purification. 5-fU-ODN**2** was synthesised using a protected 5-formyldeoxyuridine phosphoramidite[129]. 5-fC-ODN**3** was purchased from Eurogentec with HPLC purification. 5-fC-ODN4 was purchased from ATDBio with HPLC purification. Custom adapters were purchased with double HPLC purification from ATDBio.

ODN	*Calculated MW*	*Calculated ion*	*ESI-MS found*
U-ODN1	4554	M^{-3} = 1517	M^{-3} = 1516.34
AP-ODN1	4459	M^{-3} = 1485	M^{-3} = 1485.07
AP-ODN1 + HIPS 1	4654	M^{-3} = 1550	M^{-3} = 1550.04
AP-ODN1 + HIPS 1 + Biotin PEG3 azide	5099 5095 (oxidised)	M^{-3} = 1699 M^{-3} = 1697	- M^{-3} = 1696.99
AP-ODN1 β-δ-elimination products	2165 2194	M^{-2} = 1082 M^{-2} = 1096	M^{-2} = 1082.09 M^{-2} = 1096.07
AP-ODN1 + ARP	4772	M^{-3} = 1590	M^{-3} = 1589.66
fU-ODN2	3041	M^{-2} = 1520	M^{-2} = 1519.08
fU-ODN2 + HIPS 1	3236	M^{-2} = 1617	M^{-2} = 1616.74
fU-ODN2 + HIPS 1 + Biotin PEG3 azide	3681 3677 (oxidised)	M^{-2} = 1840 M^{-2} = 1837	- M^{-2} = 1836.92
fU-ODN2 +ARP	3354	M^{-2} = 1676	M^{-2} = 1675.84
fC-ODN3	3056	M^{-2} = 1527	M^{-2} = 1526.69
fC-ODN3 + HIPS 1	3251	M^{-2} = 1625	M^{-2} = 1624.23
fC-ODN3 + HIPS 1 + Biotin PEG3 azide	3696 3692 (oxidised)	M^{-2} = 1847 M^{-2} = 1845	- M^{-2} = 1844.52
fC-ODN3 + ARP	3369	M^{-2} = 1684	M^{-2} = 1683.45
hmU-ODN	3122	M^{-2} = 1560	M^{-2} = 1560.27
hmU-ODN AP	2998	M^{-2} = 1498	M^{-2} = 1498.33
hmU-ODN + HIPS 1	3193	M^{-2} = 1596	M^{-2} = 1595.89
hmU-ODN + HIPS 1 + Biotin PEG3 azide	3638 3634 (oxidised)	M^{-2} = 1819 M^{-2} = 1816	- M^{-2} = 1816.02
hmU-ODN β-δ-elimination product	1935	M^{-1} = 1934	M^{-1} = 1934.17
BaseJ-ODN	4220	M^{-3} = 1406	M^{-2} = 1405.44 (**1**) 1405.45 (**1**-CuAAC) 1405.63 (SMUG1)

Table 7.4: Calculated and detected masses of short ssODNs (≤ 20 nucleotides). ODNs were resolved by LC-MS, and mass values given were detected within peaks observed at 260 nm UV.

References

1. Miescher, F. Ueber die chemische Zusammensetzung der Eiterzellen. *Med. Chem. Untersuchungen* **4,** 441–460 (1871).
2. Kossel, A. Über die chemische zusammensetzung der zelle. *Arch. Für Physiol.* 181–186 (1891).
3. Levene, P. A. & Mandel, J. A. On the carbohydrate group in the nucleoproteid of the spleen. *J. Exp. Med.* **8,** 178–179 (1906).
4. Levene, P. A. The structure of yeast nucleic acid IV. Ammonia hydrolysis. *J. Biol. Chem.* **40,** 415–424 (1919).
5. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the chemical nature of the substance including transformation of pneumococcal types. *J. Exp. Med.* **79,** 137–158 (1944).
6. Chargaff, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia* **6,** 201–209 (1950).
7. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
8. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171,** 740 (1953).
9. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular structure of nucleic acids: molecular structure of deoxypentose nucleic acids. *Nature* **171,** 738 (1953).
10. DeVoe, H. & Tinoco, I. The stability of helical polynucleotides: Base contributions. *J. Mol. Biol.* **4,** 500–517 (1962).

11. Poater, J., Swart, M., Matthias Bickelhaupt, F. & Guerra, C. F. B-DNA structure and stability: the role of hydrogen bonding, π–π stacking interactions, twist-angle, and solvation. *Org. Biomol. Chem.* **12,** 4691–4700 (2014).
12. Wang, J. C. Helical repeat of DNA in solution. *Proc. Natl. Acad. Sci.* **76,** 200–203 (1979).
13. Franklin, R. E. & Gosling, R. G. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallogr.* **6,** 673–677 (1953).
14. Mitsui, Y., Langridge, R., Shortle, B. E., Cantor, C. R., Grant, R. C., Kodama, M. & Wells, R. D. Physical and enzymatic studies on poly d(I-C)-poly d(I-C), an unusual double-helical DNA. *Nature* **228,** 1166–1169 (1970).
15. Ghosh, A. & Bansal, M. A glossary of DNA structures from A to Z. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 620–626 (2003).
16. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13,** 770–780 (2012).
17. Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* **311,** 532 (1984).
18. Black, J. C. & Whetstine, J. R. Chromatin landscape. *Epigenetics* **6,** 13–19 (2011).
19. Cao, B., Chen, C., DeMott, M. S., Cheng, Q., Clark, T. A., Xiong, X., Zheng, X., Butty, V., Levine, S. S., Yuan, G., Boitano, M., Luong, K., Song, Y., Zhou, X., Deng, Z., Turner, S. W., Korlach, J., You, D., Wang, L., Chen, S. & Dedon, P. C. Genomic mapping of phosphorothioates reveals partial modification of short consensus sequences. *Nat. Commun.* **5,** 3951 (2014).
20. Zhao, Y. & Garcia, B. A. Comprehensive catalog of currently documented histone modifications. *Cold Spring Harb. Perspect. Biol.* **7,** a025064 (2015).
21. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
22. Bird, A. Perceptions of epigenetics. *Nature* **447,** 396–398 (2007).

23. Johnson, T. B. & Coghill, R. D. Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in Tuberculinic acid, the nucleic acid of the Tubercle Bacillus 1. *J. Am. Chem. Soc.* **47,** 2838–2844 (1925).

24. Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J. Biol. Chem.* **175,** 315–332 (1948).

25. Wyatt, G. R. Occurrence of 5-methylcytosine in nucleic acids. *Nature* **166,** 237–238 (1950).

26. Le, T., Kim, K.-P., Fan, G. & Faull, K. F. A sensitive mass-spectrometry method for simultaneous quantification of DNA methylation and hydroxymethylation levels in biological samples. *Anal. Biochem.* **412,** 203–209 (2011).

27. Pfaffeneder, T., Hackner, B., Truß, M., Münzel, M., Müller, M., Deiml, C. A., Hagemeier, C. & Carell, T. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem.* **123,** 7146–7150 (2011).

28. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321,** 209 (1986).

29. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* **20,** 116–117 (1998).

30. Schulz, W. A., Steinhoff, C. & Florl, A. R. Methylation of endogenous human retroelements in health and disease. *Curr. Top. Microbiol. Immunol.* **310,** 211–250 (2006).

31. Goto, T. & Monk, M. Regulation of X-chromosome inactivation in development in mice and humans. *Microbiol. Mol. Biol. Rev.* **62,** 362–378 (1998).

32. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.* **103,** 1412–1417 (2006).

33. Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., Bibel, M. & Schübeler, D. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30,** 755–766 (2008).

34. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).

35. 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).

36. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1,** 239–259 (2009).

37. Hon, G. C., Hawkins, R. D., Caballero, O. L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L. E., Camargo, A. A., Stevenson, B. J., Ecker, J. R., Bafna, V., Strausberg, R. L., Simpson, A. J. & Ren, B. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22,** 246–258 (2012).

38. Raiber, E.-A., Beraldi, D., Martínez Cuesta, S., McInroy, G. R., Kingsbury, Z., Becq, J., James, T., Lopes, M., Allinson, K., Field, S., Humphray, S., Santarius, T., Watts, C., Bentley, D. & Balasubramanian, S. Base resolution maps reveal the importance of 5-hydroxymethylcytosine in a human glioblastoma. *Npj Genomic Med.* **2,** 6 (2017).

39. Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21,** 5427–5440 (2002).

40. Gruenbaum, Y., Cedar, H. & Razin, A. Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature* **295,** 620 (1982).

41. Bashtrykov, P., Jankevicius, G., Smarandache, A., Jurkowska, R. Z., Ragozin, S. & Jeltsch, A. Specificity of Dnmt1 for Methylation of Hemimethylated CpG Sites Resides in Its Catalytic Domain. *Chem. Biol.* **19,** 572–578 (2012).

42. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. *et al*. The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

43. Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286,** 35334–35338 (2011).

44. He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C. & Xu, G.-L. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).

45. Dawlaty, M. M., Breiling, A., Le, T., Barrasa, M. I., Raddatz, G., Gao, Q., Powell, B. E., Cheng, A. W., Faull, K. F., Lyko, F. & Jaenisch, R. Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell* **29,** 102–111 (2014).

46. Weber, A. R., Krawczyk, C., Robertson, A. B., Kuśnierczyk, A., Vågbø, C. B., Schuermann, D., Klungland, A. & Schär, P. Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism. *Nat. Commun.* **7,** 10806 (2016).

47. Rahimoff, R., Kosmatchev, O., Kirchner, A., Pfaffeneder, T., Spada, F., Brantl, V., Müller, M. & Carell, T. 5-Formyl- and 5-carboxydeoxycytidines do not cause accumulation of harmful tepair intermediates in stem cells. *J. Am. Chem. Soc.* **139,** 10359–10364 (2017).

48. Matsubara, M., Masaoka, A., Tanaka, T., Miyano, T., Kato, N., Terato, H., Ohyama, Y., Iwai, S. & Ide, H. Mammalian 5-formyluracil−DNA glycosylase. 1. Identification and characterization of a novel activity that releases 5-formyluracil from DNA. *Biochemistry* **42,** 4993–5002 (2003).

49. Masaoka, A., Matsubara, M., Hasegawa, R., Tanaka, T., Kurisu, S., Terato, H., Ohyama, Y., Karino, N., Matsuda, A. & Ide, H. 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).

50. Pfaffeneder, T., Spada, F., Wagner, M., Brandmayr, C., Laube, S. K., Eisen, D., Truss, M., Steinbacher, J., Hackner, B., Kotljarova, O., Schuermann, D., Michalakis, S.,

Kosmatchev, O., Schiesser, S., Steigenberger, B., Raddaoui, N., Kashiwazaki, G., Müller, U., Spruijt, C. G., Vermeulen, M., Leonhardt, H., Schär, P., Müller, M. & Carell, T. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat. Chem. Biol.* **10,** 574–581 (2014).

51. Santos, F., Peat, J., Burgess, H., Rada, C., Reik, W. & Dean, W. Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. *Epigenetics Chromatin* **6,** 39 (2013).
52. Franchini, D.-M., Chan, C.-F., Morgan, H., Incorvaia, E., Rangam, G., Dean, W., Santos, F., Reik, W. & Petersen-Mahrt, S. K. Processive DNA demethylation via DNA deaminase-induced lesion resolution. *PLOS ONE* **9,** e97754 (2014).
53. Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. & Sperens, B. DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. *J. Biol. Chem.* **252,** 3286–3294 (1977).
54. Friedberg, E. C., Ganesan, A. K. & Minton, K. N-glycosidase activity in extracts of Bacillus subtilis and its inhibition after infection with bacteriophage PBS2. *J. Virol.* **16,** 315–321 (1975).
55. Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A. & Balasubramanian, S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6,** 1049–1055 (2014).
56. Bachman, M., Uribe-Lewis, S., Yang, X., Burgess, H. E., Iurlaro, M., Reik, W., Murrell, A. & Balasubramanian, S. 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11,** 555–557 (2015).
57. Pfeifer, G. P., Xiong, W., Hahn, M. A. & Jin, S.-G. The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res.* **356,** 631–641 (2014).
58. Ji, S., Shao, H., Han, Q., Seiler, C. L. & Tretyakova, N. Y. Reversible DNA-protein cross-linking at epigenetic DNA marks. *Angew. Chem.* **56,** 14130–14134 (2017).

59. Raiber, E.-A., Portella, G., Martínez Cuesta, S., Hardisty, R., Murat, P., Li, Z., Iurlaro, M., Dean, W., Spindel, J., Beraldi, D., Liu, Z., Dawson, M. A., Reik, W. & Balasubramanian, S. 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat. Chem.* **10,** 1258 (2018).

60. Warren, R. A. J. Modified bases in bacteriophage DNAs. *Annu. Rev. Microbiol.* **34,** 137–158 (1980).

61. Weigele, P. & Raleigh, E. A. Biosynthesis and function of modified bases in bacteria and their viruses. *Chem. Rev.* **116,** 12655–12687 (2016).

62. Kropinski, A. M., Bose, R. J. & Warren, R. A. 5-(4-Aminobutylaminomethyl)uracil, an unusual pyrimidine from the deoxyribonucleic acid of bacteriophage phiW-14. *Biochemistry* **12,** 151–157 (1973).

63. Brandon, C., Gallop, P. M., Marmur, J., Hayashi, H. & Nakanishi, K. Structure of a new pyrimidine from Bacillus subtilis phage SP-15 nucleic acid. *Nature. New Biol.* **239,** 70–71 (1972).

64. Ehrlich, M. & Ehrlich, K. C. A novel, highly modified, bacteriophage DNA in which thymine is partly replaced by a phosphoglucuronate moiety covalently bound to 5-(4',5'-dihydroxypentyl)uracil. *J. Biol. Chem.* **256,** 9966–9972 (1981).

65. Miller, P. B., Scraba, D. G., Leyritz-Wills, M., Maltman, K. L. & Warren, R. A. Formation and possible functions of alpha-putrescinylthymine in bacteriophage phi W-14 DNA: analysis of bacteriophage mutants with decreased levels of alpha-putrescinylthymine in their DNAs. *J. Virol.* **47,** 399–405 (1983).

66. Burza, S., Croft, S. L. & Boelaert, M. Leishmaniasis. *Lancet Lond. Engl.* **392,** 951–970 (2018).

67. Büscher, P., Cecchi, G., Jamonneau, V. & Priotto, G. Human African trypanosomiasis. *Lancet Lond. Engl.* **390,** 2397–2409 (2017).

68. Wheeler, R. J., Gluenz, E. & Gull, K. The cell cycle of Leishmania: morphogenetic events and their implications for parasite biology. *Mol. Microbiol.* **79,** 647–662 (2011).

69. Borst, P. & Sabatini, R. Base J: discovery, biosynthesis, and possible functions. *Annu. Rev. Microbiol.* **62,** 235–251 (2008).

70. Kawasaki, F., Beraldi, D., Hardisty, R. E., McInroy, G. R., van Delft, P. & Balasubramanian, S. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biol.* **18,** 23 (2017).

71. Liu, S., Ji, D., Cliffe, L., Sabatini, R. & Wang, Y. 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).

72. 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).

73. Daniels, J.-P., Gull, K. & Wickstead, B. Cell biology of the trypanosome genome. *Microbiol Mol Biol Rev* **74,** 552–569 (2010).

74. Genest, P.-A., Ter Riet, B., Cijsouw, T., van Luenen, H. G. A. M. & Borst, P. Telomeric localization of the modified DNA base J in the genome of the protozoan parasite Leishmania. *Nucleic Acids Res.* **35,** 2116–2124 (2007).

75. van Luenen, H. G. A. M., Farris, C., Jan, S., Genest, P.-A., Tripathi, P., Velds, A., Kerkhoven, R. M., Nieuwland, M., Haydock, A., Ramasamy, G., Vainio, S., Heidebrecht, T., Perrakis, A., Pagie, L., van Steensel, B., Myler, P. J. & Borst, P. Glucosylated hydroxymethyluracil, DNA Base J, prevents transcriptional readthrough in leishmania. *Cell* **150,** 909–921 (2012).

76. Reynolds, D., Cliffe, L., Förstner, K. U., Hon, C.-C., Siegel, T. N. & Sabatini, R. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Res.* **42,** 9717–9729 (2014).

77. Stavnezer, J., Guikema, J. E. J. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26,** 261–292 (2008).

78. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76,** 1–22 (2007).

79. Okazaki, I., Hiai, H., Kakazu, N., Yamada, S., Muramatsu, M., Kinoshita, K. & Honjo, T. Constitutive expression of AID leads to tumorigenesis. *J. Exp. Med.* **197,** 1173–1181 (2003).

80. Nilsen, H., Stamp, G., Andersen, S., Hrivnak, G., Krokan, H. E., Lindahl, T. & Barnes, D. E. Gene-targeted mice lacking the Ung uracil-DNA glycosylase develop B-cell lymphomas. *Oncogene* **22,** 5381–5386 (2003).

81. Muha, V., Horváth, A., Békési, A., Pukáncsik, M., Hodoscsek, B., Merényi, G., Róna, G., Batki, J., Kiss, I., Jankovics, F., Vilmos, P., Erdélyi, M. & Vértessy, B. G. Uracil-containing DNA in drosophila: stability, stage-specific accumulation, and developmental involvement. *PLOS Genet.* **8,** e1002738 (2012).

82. Galashevskaya, A., Sarno, A., Vågbø, C. B., Aas, P. A., Hagen, L., Slupphaug, G. & Krokan, H. E. A robust, sensitive assay for genomic uracil determination by LC/MS/MS reveals lower levels than previously reported. *DNA Repair* **12,** 699–706 (2013).

83. Shu, X., Liu, M., Lu, Z., Zhu, C., Meng, H., Huang, S., Zhang, X. & Yi, C. Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA. *Nat. Chem. Biol.* **14,** 680 (2018).

84. 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).

85. Marmorstein, R. & Zhou, M.-M. Writers and readers of histone acetylation: structure, mechanism, and inhibition. *Cold Spring Harb. Perspect. Biol.* **6,** a018762 (2014).

86. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Exp. Mol. Med.* **49,** e324 (2017).

87. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).

88. Du, Z., Li, H., Wei, Q., Zhao, X., Wang, C., Zhu, Q., Yi, X., Xu, W., Liu, X. S., Jin, W. & Su, Z. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in Oryza sativa L. Japonica. *Mol. Plant* **6,** 1463–1472 (2013).

89. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–357 (2012).

90. Krokan, H. E. & Bjørås, M. Base excision repair. *Cold Spring Harb. Perspect. Biol.* **5,** (2013).

91. Jacobs, A. L. & Schär, P. DNA glycosylases: in DNA repair and beyond. *Chromosoma* **121,** 1–20 (2012).

92. Krokan, H. E., Standal, R. & Slupphaug, G. DNA glycosylases in the base excision repair of DNA. *Biochem. J.* **325,** 1–16 (1997).

93. McCullough, A. K., Dodson, M. L. & Lloyd, R. S. Initiation of base excision repair: glycosylase mechanisms and structures. *Annu. Rev. Biochem.* **68,** 255–285 (1999).

94. 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. CMLS* **66,** 981–993 (2009).

95. Pastukh, V., Ruchko, M., Gorodnya, O., Wilson, G. L. & Gillespie, M. N. Sequence-specific oxidative base modifications in hypoxia-inducible genes. *Free Radic. Biol. Med.* **43,** 1616–1626 (2007).

96. Ba, X., Bacsi, A., Luo, J., Aguilera-Aguirre, L., Zeng, X., Radak, Z., Brasier, A. R. & Boldogh, I. 8-Oxoguanine DNA glycosylase-1 augments proinflammatory gene expression by facilitating the recruitment of site-specific Ttranscription factors. *J. Immunol.* **192,** 2384–2394 (2014).

97. Schärer, O. D. Nucleotide excision repair in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **5,** (2013).

98. Hsieh, P. & Yamane, K. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mech. Ageing Dev.* **129,** 391–407 (2008).

99. Sparks, J. L., Chon, H., Cerritelli, S. M., Kunkel, T. A., Johansson, E., Crouch, R. J. & Burgers, P. M. RNase H2-initiated ribonucleotide excision repair. *Mol. Cell* **47,** 980–986 (2012).

100. Caldecott, K. W. Single-strand break repair and genetic disease. *Nat. Rev. Genet.* **9,** 619–631 (2008).

101. Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harb. Perspect. Biol.* **5,** a012740 (2013).

102. Weterings, E. & van Gent, D. C. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA Repair* **3,** 1425–1435 (2004).

103. de Lange, T. Shelterin-mediated telomere protection. *Annu. Rev. Genet.* **52,** 223–247 (2018).

104. Vilenchik, M. M. & Knudson, A. G. Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. *Proc. Natl. Acad. Sci.* **100,** 12871–12876 (2003).

105. Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E. Y., Tannahill, D. & Balasubramanian, S. DSBCapture: in situ capture and sequencing of DNA breaks. *Nat. Methods* **13,** 855–857 (2016).

106. Hogg, M., Wallace, S. S. & Doublié, S. Crystallographic snapshots of a replicative DNA polymerase encountering an abasic site. *EMBO J.* **23,** 1483–1493 (2004).

107. Boiteux, S. & Guillet, M. Abasic sites in DNA: repair and biological consequences in Saccharomyces cerevisiae. *DNA Repair* **3,** 1–12 (2004).

108. Haracska, L., Washington, M. T., Prakash, S. & Prakash, L. Inefficient bypass of an abasic site by DNA polymerase η. *J. Biol. Chem.* **276,** 6861–6866 (2001).

109. Strauss, B. S. The 'A rule' of mutagen specificity: A consequence of DNA polymerase bypass of non-instructional lesions? *BioEssays* **13,** 79–84 (1991).

110. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochemistry* **11,** 3610–3618 (1972).

111. Hevesi, L., Wolfson-Davidson, E., Nagy, J. B., Nagy, O. B. & Bruylants, A. Contribution to the mechanism of the acid-catalyzed hydrolysis of purine nucleosides. *J. Am. Chem. Soc.* **94,** 4715–4720 (1972).

112. Lindahl, T. & Karlstrom, O. Heat-induced depyrimidination of deoxyribonucleic acid in neutral solution. *Biochemistry* **12,** 5151–5154 (1973).

113. Brozmanová, J., Dudás, A. & Henriques, J. A. Repair of oxidative DNA damage--an important factor reducing cancer risk. *Neoplasma* **48,** 85–93 (2001).

114. Osborne, M. & Merrifield, K. Depurination of benzo[a]pyrene-diolepoxide treated DNA. *Chem. Biol. Interact.* **53,** 183–195 (1985).

115. Kim, J., Gil, J. M. & Greenberg, M. M. Synthesis and characterization of oligonucleotides containing the C4′-oxidized abasic site produced by bleomycin and other DNA damaging agents. *Angew. Chem.* **42,** 5882–5885 (2003).

116. Morland, I., Luna, L., Gustad, E., Seeberg, E. & Bjørås, M. Product inhibition and magnesium modulate the dual reaction mode of hOgg1. *DNA Repair* **4,** 381–387 (2005).

117. Sugiyama, H., Fujiwara, T., Ura, A., Tashiro, T., Yamamoto, K., Kawanishi, S. & Saito, I. 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).

118. Zheng, Y. & Sheppard, T. L. Half-life and DNA strand scission products of 2-deoxyribonolactone oxidative DNA damage lesions. *Chem. Res. Toxicol.* **17,** 197–207 (2004).

119. Price, N. E., Johnson, K. M., Wang, J., Fekry, M. I., Wang, Y. & Gates, K. S. Interstrand DNA–DNA cross-link formation between adenine residues and abasic sites in duplex DNA. *J. Am. Chem. Soc.* **136,** 3483–3490 (2014).

120. Rashidian, M., Dozier, J. K., Lenevich, S. & Distefano, M. D. Selective labeling of polypeptides using protein farnesyltransferase via rapid oxime ligation. *Chem. Commun.* **46,** 8998–9000 (2010).

121. Sczepanski, J. T., Wong, R. S., McKnight, J. N., Bowman, G. D. & Greenberg, M. M. Rapid DNA–protein cross-linking and strand scission by an abasic site in a nucleosome core particle. *Proc. Natl. Acad. Sci.* **107,** 22475–22480 (2010).

122. Hinz, J. M., Mao, P., McNeill, D. R. & Wilson, D. M. Reduced nuclease activity of apurinic/apyrimidinic endonuclease (APE1) variants on nucleosomes: identification of acess residues. *J. Biol. Chem.* **290,** 21067–21075 (2015).

123. Kurisu, S., Miya, T., Terato, H., Masaoka, A., Ohyama, Y., Kubo, K. & Ide, H. Quantitation of DNA damage by an aldehyde reactive probe (ARP). *Nucleic Acids Res. Suppl. 2001* 45–46 (2001).

124. Kubo, K., Ide, H., Wallace, S. S. & Kow, Y. W. A novel sensitive and specific assay for abasic sites, the most commonly produced DNA lesion. *Biochemistry* **31,** 3703–3708 (1992).

125. Ide, H., Akamatsu, K., Kimura, Y., Michiue, K., Makino, K., Asaeda, A., Takamori, Y. & Kubo, K. Synthesis and damage specificity of a novel probe for the detection of abasic sites in DNA. *Biochemistry* **32,** 8276–8283 (1993).

126. Chastain, P. D., Nakamura, J., Swenberg, J. & Kaufman, D. Nonrandom AP site distribution in highly proliferative cells. *FASEB J.* **20,** 2612–2614 (2006).

127. Chastain, P. D., Nakamura, J., Rao, S., Chu, H., Ibrahim, J. G., Swenberg, J. A. & Kaufman, D. G. Abasic sites preferentially form at regions undergoing DNA replication. *FASEB J.* **24,** 3674–3680 (2010).

128. Raiber, E.-A., Beraldi, D., Ficz, G., Burgess, H. E., Branco, M. R., Murat, P., Oxley, D., Booth, M. J., Reik, W. & Balasubramanian, S. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13,** R69 (2012).

129. 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).

130. Poetsch, A. R., Boulton, S. J. & Luscombe, N. M. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.* **19,** 215 (2018).

131. Hänsel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell Biol.* **18,** 279–284 (2017).

132. 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.* **114,** 2604–2609 (2017).

133. Kidane, D., Murphy, D. L. & Sweasy, J. B. Accumulation of abasic sites induces genomic instability in normal human gastric epithelial cells during Helicobacter pylori infection. *Oncogenesis* **3,** e128 (2014).

134. Masani, S., Han, L. & Yu, K. Apurinic/Apyrimidinic endonuclease 1 is the essential nuclease during immunoglobulin class switch recombination. *Mol. Cell. Biol.* **33,** 1468–1473 (2013).

135. Mundle, S. T., Delaney, J. C., Essigmann, J. M. & Strauss, P. R. Enzymatic mechanism of human apurinic/apyrimidinic endonuclease against a THF AP site model substrate. *Biochemistry* **48,** 19–26 (2009).

136. Burkovics, P., Szukacsov, V., Unk, I. & Haracska, L. Human Ape2 protein has a 3′–5′ exonuclease activity that acts preferentially on mismatched base pairs. *Nucleic Acids Res.* **34,** 2508–2515 (2006).

137. Torres-Ramos, C. A., Johnson, R. E., Prakash, L. & Prakash, S. Evidence for the involvement of nucleotide excision repair in the removal of abasic sites in yeast. *Mol. Cell. Biol.* **20,** 3522–3528 (2000).

138. Kim, N. & Jinks-Robertson, S. Abasic sites in the transcribed strand of yeast DNA are removed by transcription-coupled nucleotide excision repair. *Mol. Cell. Biol.* **30,** 3206–3215 (2010).

139. Sakurai, E., Susuki, M., Kanamitsu, K., Kawano, S. & Ikeda, S. Global genome nucleotide excision repair proteins Rhp7p and Rhp41p are involved in abasic site repair of Schizosaccharomyces pombe. *Adv. Biosci. Biotechnol.* **06,** 265–274 (2015).

140. Barnes, T., Kim, W.-C., Mantha, A. K., Kim, S.-E., Izumi, T., Mitra, S. & Lee, C. H. Identification of apurinic/apyrimidinic endonuclease 1 (APE1) as the endoribonuclease that cleaves c-myc mRNA. *Nucleic Acids Res.* **37,** 3946–3958 (2009).

141. Wong, D., DeMott, M. S. & Demple, B. Modulation of the 3'-->5'-exonuclease activity of human apurinic endonuclease (Ape1) by its 5'-incised Abasic DNA product. *J. Biol. Chem.* **278,** 36242–36249 (2003).

142. Tell, G., Quadrifoglio, F., Tiribelli, C. & Kelley, M. R. The many functions of APE1/Ref-1: not only a DNA repair enzyme. *Antioxid. Redox Signal.* **11,** 601–619 (2009).

143. Thakur, S., Sarkar, B., Cholia, R. P., Gautam, N., Dhiman, M. & Mantha, A. K. APE1/Ref-1 as an emerging therapeutic target for various human diseases: phytochemical modulation of its functions. *Exp. Mol. Med.* **46,** e106 (2014).

144. Luo, M. & Kelley, M. R. Inhibition of the human apurinic/apyrimidinic endonuclease (Ape1) repair activity and sensitization of breast cancer cells to DNA alkylating agents with lucanthone. *Anticancer Res.* **24,** 2127–2134 (2004).

145. Xanthoudakis, S., Smeyne, R. J., Wallace, J. D. & Curran, T. The redox/DNA repair protein, Ref-1, is essential for early embryonic development in mice. *Proc. Natl. Acad. Sci.* **93,** 8919–8923 (1996).

146. Mohni, K. N., Wessel, S. R., Zhao, R., Wojciechowski, A. C., Luzwick, J. W., Layden, H., Eichman, B. F., Thompson, P. S., Mehta, K. P. M. & Cortez, D. HMCES maintains genome integrity by shielding abasic sites in single-strand DNA. *Cell* **176,** 144-153.e13 (2019).

147. Spruijt, C. G., Gnerlich, F., Smits, A. H., Pfaffeneder, T., Jansen, P. W. T. C., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F., Eberl, H. C., Mensinga, A., Brinkman, A. B., Lephikov, K., Müller, U., Walter, J., Boelens, R., van Ingen, H., Leonhardt, H., Carell,

T. & Vermeulen, M. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152,** 1146–1159 (2013).

148. Kremer, D., Metzger, S., Kolb-Bachofen, V. & Kremer, D. Quantitative measurement of genome-wide DNA methylation by a reliable and cost-efficient enzyme-linked immunosorbent assay technique. *Anal. Biochem.* **422,** 74–78 (2012).
149. Kow, Y. W. & Dare, A. Detection of abasic sites and oxidative DNA base damage using an ELISA-like assay. *Methods* **22,** 164–169 (2000).
150. Koc, H. & Swenberg, J. A. Applications of mass spectrometry for quantitation of DNA adducts. *J. Chromatogr. B* **778,** 323–343 (2002).
151. Tretyakova, N., Villalta, P. W. & Kotapati, S. Mass spectrometry of structurally modified DNA. *Chem. Rev.* **113,** 2395–2436 (2013).
152. Traube, F. R., Schiffers, S., Iwan, K., Kellner, S., Spada, F., Müller, M. & Carell, T. Isotope-dilution mass spectrometry for exact quantification of noncanonical DNA nucleosides. *Nat. Protoc.* **14,** 283–312 (2019).
153. Crain, P. F. Preparation and enzymatic hydrolysis of DNA and RNA for mass spectrometry. *Methods Enzymol.* **193,** 782–790 (1990).
154. Quinlivan, E. P. & Gregory, J. F. DNA digestion to deoxyribonucleoside: A simplified one-step procedure. *Anal. Biochem.* **373,** 383–385 (2008).
155. Wheeler, O. H. The Girard reagents. *J. Chem. Educ.* **45,** 435 (1968).
156. 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).
157. Jiang, H.-P., Liu, T., Guo, N., Yu, L., Yuan, B.-F. & Feng, Y.-Q. Determination of formylated DNA and RNA by chemical labeling combined with mass spectrometry analysis. *Anal. Chim. Acta* **981,** 1–10 (2017).

158. Roberts, K. P., Sobrino, J. A., Payton, J., Mason, L. B. & Turesky, R. J. Determination of apurinic/apyrimidinic lesions in DNA with high-performance liquid chromatography and tandem mass spectrometry. *Chem. Res. Toxicol.* **19,** 300–309 (2006).

159. Li, J., Leung, E. M. K., Choi, M. M. F. & Chan, W. Combination of pentafluorophenylhydrazine derivatization and isotope dilution LC-MS/MS techniques for the quantification of apurinic/apyrimidinic sites in cellular DNA. *Anal. Bioanal. Chem.* **405,** 4059–4066 (2013).

160. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74,** 5463–5467 (1977).

161. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74,** 560–564 (1977).

162. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).

163. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).

164. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** (2012).

165. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

166. Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev,

E., McKernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).

167. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
168. McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., Vega, F. M. D. L. & Blanchard, A. P. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19,** 1527–1541 (2009).
169. Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., Marma, M. S., Meng, Q., Cao, H., Li, X., Shi, S., Yu, L., Kalachikov, S., Russo, J. J., Turro, N. J. & Ju, J. Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci.* **105,** 9145–9150 (2008).

170. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** (2010).

171. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40,** e3 (2012).

172. Thu, K. L., Vucic, E. A., Kennett, J. Y., Heryet, C., Brown, C. J., Lam, W. L. & Wilson, I. M. Methylated DNA immunoprecipitation. *J. Vis. Exp. JoVE* **23,** e935 (2009).

173. Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A. J., Zheng, L., Zhang, H., Huang, S., Min, J., Nicholson, T., Chen, T., Xu, G., Shi, Y., Zhang, K. & Shi, Y. G. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol. Cell* **42,** 451–464 (2011).

174. Koziol, M. J., Bradshaw, C. R., Allen, G. E., Costa, A. S. H., Frezza, C. & Gurdon, J. B. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat. Struct. Mol. Biol.* **23,** 24–30 (2016).

175. Xiao, Y., Yu, F., Pang, L., Zhao, H., Liu, L., Zhang, G., Liu, T., Zhang, H., Fan, H., Zhang, Y., Pang, B. & Li, X. MeSiC: A model-based method for estimating 5 mC levels at single-CpG resolution from MeDIP-seq. *Sci. Rep.* **5,** 14699 (2015).

176. Lentini, A., Lagerwall, C., Vikingsson, S., Mjoseng, H. K., Douvlataniotis, K., Vogt, H., Green, H., Meehan, R. R., Benson, M. & Nestor, C. E. A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nat. Methods* **15,** 499 (2018).

177. Xia, B., Han, D., Lu, X., Sun, Z., Zhou, A., Yin, Q., Zeng, H., Liu, M., Jiang, X., Xie, W., He, C. & Yi, C. Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12,** 1047–1050 (2015).

178. 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).

179. Liu, C., Wang, Y., Zhang, X., Wu, F., Yang, W., Zou, G., Yao, Q., Wang, J., Chen, Y., Wang, S. & Zhou, X. Enrichment and fluorogenic labelling of 5-formyluracil in DNA. *Chem. Sci.* **8,** 4505–4510 (2017).

180. Wang, Y., Liu, C., Wu, F., Zhang, X., Liu, S., Chen, Z., Zeng, W., Yang, W., Zhang, X., Zhou, Y., Weng, X., Wu, Z. & Zhou, X. Highly selective 5-formyluracil labeling and genome-wide mapping using (2-benzimidazolyl)acetonitrile probe. *iScience* **9,** 423–432 (2018).

181. Green, N. M. Avidin. *Adv. Protein Chem.* **29,** 85–133 (1975).

182. Green, N. M. in *Methods Enzymol.* **184,** 51–67 (1990).

183. Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. & Paul, C. L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.* **89,** 1827–1831 (1992).

184. Clark, S. J., Harrison, J., Paul, C. L. & Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* **22,** 2990–2997 (1994).

185. Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S. & Jaenisch, R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33,** 5868–5877 (2005).

186. Waalwijk, C. & Flavell, R. A. MspI, an isoschizomer of hpaII which cleaves both unmethylated and methylated hpaII sites. *Nucleic Acids Res.* **5,** 3231–3236 (1978).

187. Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W. & Balasubramanian, S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336,** 934–937 (2012).

188. 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).

189. Kawasaki, F., Martínez Cuesta, S., Beraldi, D., Mahtey, A., Hardisty, R. E., Carrington, M. & Balasubramanian, S. Sequencing 5-hydroxymethyluracil at single-base resolution. *Angew. Chem.* **57,** 9694–9696 (2018).

190. Chambers, V. S., Marsico, G., Boutell, J. M., Di Antonio, M., Smith, G. P. & Balasubramanian, S. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33,** 877–881 (2015).

191. Kwok, C. K. & Balasubramanian, S. Targeted detection of G-quadruplexes in cellular RNAs. *Angew. Chem.* **54,** 6751–6754 (2015).

192. Shu, X., Xiong, X., Song, J., He, C. & Yi, C. Base-resolution analysis of cisplatin–DNA adducts at the genome scale. *Angew. Chem.* **55,** 14246–14249 (2016).

193. Dai, Q., Moshitch-Moshkovitz, S., Han, D., Kol, N., Amariglio, N., Rechavi, G., Dominissini, D. & He, C. Nm-seq maps 2′-O-methylation sites in human mRNA with base precision. *Nat. Methods* **14,** 695–698 (2017).

194. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. & Turner, S. Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

195. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34,** 518–524 (2016).

196. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502,** 472–479 (2013).

197. Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. & Turner, S. W. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7,** 461–465 (2010).

198. Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M. & Paten, B. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14,** 411–413 (2017).

199. Kumar, R., El-Sagheer, A., Tumpane, J., Lincoln, P., Wilhelmsson, L. M. & Brown, T. Template-directed oligonucleotide strand ligation, covalent intramolecular DNA circularization and catenation using click chemistry. *J. Am. Chem. Soc.* **129,** 6859–6864 (2007).

200. An, R., Jia, Y., Wan, B., Zhang, Y., Dong, P., Li, J. & Liang, X. Non-enzymatic depurination of nucleic acids: factors and mechanisms. *PLOS ONE* **9,** e115950 (2014).

201. Heyn, P., Stenzel, U., Briggs, A. W., Kircher, M., Hofreiter, M. & Meyer, M. Road blocks on paleogenomes—polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Res.* **38,** e161 (2010).

202. Lim, S., Song, I., Guengerich, F. P. & Choi, J.-Y. Effects of N(2)-alkylguanine, O(6)-alkylguanine, and abasic lesions on DNA binding and bypass synthesis by the euryarchaeal B-family DNA polymerase vent (exo(-)). *Chem. Res. Toxicol.* **25,** 1699–1707 (2012).

203. McInroy, G. R., Raiber, E.-A. & Balasubramanian, S. Chemical biology of genomic DNA: minimizing PCR bias. *Chem. Commun.* **50,** 12047–12049 (2014).

204. Edwards, J. O. & Pearson, R. G. The factors determining nucleophilic reactivities. *J. Am. Chem. Soc.* **84,** 16–24 (1962).

205. Fina, N. J. & Edwards, J. O. The alpha effect. A review. *Int. J. Chem. Kinet.* **5,** 1–26 (1973).

206. Dirksen, A., Hackeng, T. M. & Dawson, P. E. Nucleophilic catalysis of oxime ligation. *Angew. Chem.* **45,** 7581–7584 (2006).

207. Kalia, J. & Raines, R. T. Hydrolytic stability of hydrazones and oximes. *Angew. Chem.* **47,** 7523–7526 (2008).

208. Lindahl, T. & Andersson, A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry* **11,** 3618–3623 (1972).

209. Küpfer, P. A. & Leumann, C. J. The chemical stability of abasic RNA compared to abasic DNA. *Nucleic Acids Res.* **35,** 58–68 (2007).

210. Rodrigues, F., Canac, Y. & Lubineau, A. A convenient, one-step, synthesis of β-C-glycosidic ketones in aqueous media. *Chem. Commun.* **20,** 2049–2050 (2000).

211. Riemann, I., Fessner, W.-D., Papadopoulos, M. A. & Knorst, M. C-Glycosides by aqueous condensation of β-dicarbonyl compounds with unprotected sugars. *Aust. J. Chem.* **55,** 147–154 (2002).

212. Price, N. P. J., Bowman, M. J., Le Gall, S., Berhow, M. A., Kendra, D. F. & Lerouge, P. Functionalized C-glycoside ketohydrazones: carbohydrate derivatives that retain the ring integrity of the terminal reducing gugar. *Anal. Chem.* **82,** 2893–2899 (2010).

213. Guo, P., Xu, X., Qiu, X., Zhou, Y., Yan, S., Wang, C., Lu, C., Ma, W., Weng, X., Zhang, X. & Zhou, X. Synthesis and spectroscopic properties of fluorescent 5-benzimidazolyl-2′-deoxyuridines 5-fdU probes obtained from o-phenylenediamine derivatives. *Org. Biomol. Chem.* **11,** 1610–1613 (2013).

214. Pictet, A. & Spengler, Theod. Über die Bildung von Isochinolin-derivaten durch Einwirkung von Methylal auf Phenyl-äthylamin, Phenyl-alanin und Tyrosin. *Berichte Dtsch. Chem. Ges.* **44,** 2030–2036 (1911).

215. Maresh, J. J., Giddings, L.-A., Friedrich, A., Loris, E. A., Panjikar, S., Trout, B. L., Stöckigt, J., Peters, B. & O'Connor, S. E. Strictosidine synthase: mechanism of a Pictet−Spengler catalyzing enzyme. *J. Am. Chem. Soc.* **130,** 710–723 (2008).

216. Kawate, T., Yamada, H., Soe, T. & Nakagawa, M. Enantioselective asymmetric Pictet-Spengler reaction catalyzed by diisopinocampheylchloroborane. *Tetrahedron Asymmetry* **7,** 1249–1252 (1996).

217. Agarwal, P., Kudirka, R., Albers, A. E., Barfield, R. M., de Hart, G. W., Drake, P. M., Jones, L. C. & Rabuka, D. Hydrazino-Pictet-Spengler ligation as a biocompatible method for the generation of stable protein conjugates. *Bioconjug. Chem.* **24,** 846–851 (2013).

218. Bailey, P. Direct proof the involvement of a spiro intermediate in the Pictet-Spengler reaction. *J. Chem. Res.* **8,** 202–203 (1987).

219. Vranken, C., Deen, J., Dirix, L., Stakenborg, T., Dehaen, W., Leen, V., Hofkens, J. & Neely, R. K. Super-resolution optical DNA mapping via DNA methyltransferase-directed click chemistry. *Nucleic Acids Res.* **42,** e50–e50 (2014).

220. Suvorov, N. N., Ovchinnikova, Zh. D., Peresleni, E. M. & Sheinker, Yu. N. Azacarbazoles. *Chem. Heterocycl. Compd.* **1,** 631–636 (1966).

221. Holmberg, A., Blomstergren, A., Nord, O., Lukacs, M., Lundeberg, J. & Uhlén, M. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis* **26,** 501–510 (2005).

222. Gansauge, M.-T., Gerber, T., Glocke, I., Korlević, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A. & Meyer, M. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* **45,** e79–e79 (2017).

223. Marchand, V., Ayadi, L., Ernst, F. G. M., Hertler, J., Bourguignon-Igel, V., Galvanin, A., Kotter, A., Helm, M., Lafontaine, D. L. J. & Motorin, Y. AlkAniline-Seq: profiling of m7G and m3C RNA modifications at single nucleotide resolution. *Angew. Chem.* **57,** 16785–16790 (2018).

224. Cervantes-Cervantes, M. P., Calderón-Salinas, J. V., Albores, A. & Muñoz-Sánchez, J. L. Copper increases the damage to DNA and proteins caused by reactive oxygen species. *Biol. Trace Elem. Res.* **103,** 229–248 (2005).

225. Alexeeva, M., Moen, M. N., Grøsvik, K., Tesfahun, A. N., Xu, X. M., Muruzábal-Lecumberri, I., Olsen, K. M., Rasmussen, A., Ruoff, P., Kirpekar, F., Klungland, A. &

Bjelland, S. Excision of uracil from DNA by hSMUG1 includes strand incision and processing. *Nucleic Acids Res.* **47,** 779–793 (2019).

226. Sjolund, A., Senejani, A. & Sweasy, J. MBD4 and TDG: multifaceted DNA glycosylases eith ever expanding biological roles. *Mutat. Res.* **743,** 12–25 (2013).
227. Boiteux, S. & Radicella, J. P. The human OGG1 gene: structure, functions, and its implication in the process of carcinogenesis. *Arch. Biochem. Biophys.* **377,** 1–8 (2000).
228. Kimber, S. T., Brown, T. & Fox, K. R. A mutant of uracil DNA glycosylase that distinguishes between cytosine and 5-methylcytosine. *PLOS ONE* **9,** e95394 (2014).
229. Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G. *et al*. The genome of the kinetoplastid parasite, Leishmania major. *Science* **309,** 436–442 (2005).
230. Booth, M. J., Ost, T. W. B., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W. & Balasubramanian, S. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protoc.* **8,** 1841–1851 (2013).
231. van Leeuwen, F., Wijsman, E. R., Kuyl-Yeheskiely, E., van der Marel, G. A., van Boom, J. H. & Borst, P. The telomeric GGGTTA repeats of Trypanosoma brucei contain the hypermodified base J in both strands. *Nucleic Acids Res.* **24,** 2476–2482 (1996).
232. Cliffe, L. J., Kieft, R., Southern, T., Birkeland, S. R., Marshall, M., Sweeney, K. & Sabatini, R. 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).
233. 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).
234. Ulbert, S., Eide, L., Seeberg, E. & Borst, P. Base J, found in nuclear DNA of Trypanosoma brucei, is not a target for DNA glycosylases. *DNA Repair* **3,** 145–154 (2004).

235. Genest, P.-A., Baugh, L., Taipale, A., Zhao, W., Jan, S., van Luenen, H. G. A. M., Korlach, J., Clark, T., Luong, K., Boitano, M., Turner, S., Myler, P. J. & Borst, P. Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing. *Nucleic Acids Res.* **43,** 2102–2115 (2015).

236. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33,** 2908–2916 (2005).

237. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinforma. Oxf. Engl.* **27,** 1653–1659 (2011).

238. Schormann, N., Ricciardi, R. & Chattopadhyay, D. Uracil-DNA glycosylases—structural and functional perspectives on an essential family of DNA repair enzymes. *Protein Sci. Publ. Protein Soc.* **23,** 1667–1685 (2014).

239. Wagner, J. R., Hu, C. C. & Ames, B. N. Endogenous oxidative damage of deoxycytidine in DNA. *Proc. Natl. Acad. Sci.* **89,** 3380–3384 (1992).

240. Thiviyanathan, V., Somasunderam, A., Volk, D. E. & Gorenstein, D. G. 5-Hydroxyuracil can form stable base pairs with all four bases in a DNA duplex. *Chem. Commun.* **3,** 400–402 (2005).

241. DiPaolo, C., Kieft, R., Cross, M. & Sabatini, R. Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. *Mol. Cell* **17,** 441–451 (2005).

242. Loeb, L. A. & Preston, B. D. Mutagenesis by apurinic/apyrimidinic sites. *Annu. Rev. Genet.* **20,** 201–230 (1986).

243. 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).

244. Fortini, P. & Dogliotti, E. Base damage and single-strand break repair: Mechanisms and functional significance of short- and long-patch repair subpathways. *DNA Repair* **6,** 398–409 (2007).

245. Roychoudhury, S., Nath, S., Song, H., Hegde, M. L., Bellot, L. J., Mantha, A. K., Sengupta, S., Ray, S., Natarajan, A. & Bhakat, K. K. Human apurinic/apyrimidinic endonuclease (APE1) Is acetylated at DNA damage sites in chromatin, and acetylation modulates its DNA repair activity. *Mol. Cell. Biol.* **37,** (2017).

246. Nakamura, J. & Swenberg, J. A. Endogenous apurinic/apyrimidinic sites in genomic DNA of mammalian tissues. *Cancer Res.* **59,** 2522–2526 (1999).

247. Wu, J., McKeague, M. & Sturla, S. J. Nucleotide-resolution genome-wide mapping of oxidative DNA damage by Click-Code-Seq. *J. Am. Chem. Soc.* **140,** 9783–9787 (2018).

248. Mendez, F., Goldman, J. D. & Bases, R. E. Abasic sites in DNA of HeLa cells induced by lucanthone. *Cancer Invest.* **20,** 983–991 (2002).

249. Madhusudan, S., Smart, F., Shrimpton, P., Parsons, J. L., Gardiner, L., Houlbrook, S., Talbot, D. C., Hammonds, T., Freemont, P. A., Sternberg, M. J. E., Dianov, G. L. & Hickson, I. D. Isolation of a small molecule inhibitor of DNA base excision repair. *Nucleic Acids Res.* **33,** 4711–4724 (2005).

250. Bailly, C. & Waring, M. J. Preferential intercalation at AT sequences in DNA by lucanthone, hycanthone, and indazole analogs. A footprinting study. *Biochemistry* **32,** 5985–5993 (1993).

251. Ott, M., Gogvadze, V., Orrenius, S. & Zhivotovsky, B. Mitochondria, oxidative stress and cell death. *Apoptosis Int. J. Program. Cell Death* **12,** 913–922 (2007).

252. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

253. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.* **5,** (2014).

254. Mittelman, D. & Wilson, J. H. The fractured genome of HeLa cells. *Genome Biol.* **14,** 111 (2013).

255. Jung, Y. L., Luquette, L. J., Ho, J. W. K., Ferrari, F., Tolstorukov, M., Minoda, A., Issner, R., Epstein, C. B., Karpen, G. H., Kuroda, M. I. & Park, P. J. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.* **42,** e74–e74 (2014).

256. Takata, H., Hanafusa, T., Mori, T., Shimura, M., Iida, Y., Ishikawa, K., Yoshikawa, K., Yoshikawa, Y. & Maeshima, K. Chromatin compaction protects genomic DNA from radiation damage. *PloS One* **8,** e75622 (2013).

257. Xue, G., Chu, X.-Y. & Zhang, H.-Y. Oxidative DNA damage is associated more with genome accessibility than spatial positioning in the nucleus. *J. Biomol. Struct. Dyn.* **37,** 1857–1862 (2019).

258. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488,** 504–507 (2012).

259. Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., Alberdi, M. T., Prado, J. L., Prieto, A., Willerslev, E. & Orlando, L. Ligation bias in Illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLOS ONE* **8,** e78575 (2013).

260. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38,** 23–38 (2013).

261. 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).

262. Vascotto, C., Cesaratto, L., Zeef, L. A. H., Deganuto, M., D'Ambrosio, C., Scaloni, A., Romanello, M., Damante, G., Taglialatela, G., Delneri, D., Kelley, M. R., Mitra, S., Quadrifoglio, F. & Tell, G. Genome-wide analysis and proteomic studies reveal APE1/Ref-1 multifunctional role in mammalian cells. *Proteomics* **9,** 1058–1074 (2009).

263. Morey, N. J., Greene, C. N. & Jinks-Robertson, S. Genetic analysis of transcription-associated mutation in Saccharomyces cerevisiae. *Genetics* **154,** 109–120 (2000).

264. Hadi, M. Z., Coleman, M. A., Fidelis, K. & Mohrenweiser, H. W. Functional characterization of Ape1 variants identified in the human population. *Nucleic Acids Res.* **28,** 3871–3879 (2000).

265. Kumar, R., DiMenna, L. J., Chaudhuri, J. & Evans, T. Biological function of activation-induced cytidine deaminase (AID). *Biomed. J.* **37,** 269–283 (2014).

266. Harris, R. S. & Liddament, M. T. Retroviral restriction by APOBEC proteins. *Nat. Rev. Immunol.* **4,** 868 (2004).

267. Arias, J. F., Koyama, T., Kinomoto, M. & Tokunaga, K. Retroelements versus APOBEC3 family members: No great escape from the magnificent seven. *Front. Microbiol.* **3,** (2012).

268. Otterlei, M., Haug, T., Nagelhus, T. A., Slupphaug, G., Lindmo, T. & Krokan, H. E. Nuclear and mitochondrial splice forms of human uracil-DNA glycosylase contain a complex nuclear localisation signal and a strong classical mitochondrial localisation signal, respectively. *Nucleic Acids Res.* **26,** 4611–4617 (1998).

269. Harris, J. M., McIntosh, E. M. & Muscat, G. E. Structure/function analysis of a dUTPase: catalytic mechanism of a potential chemotherapeutic target. *J. Mol. Biol.* **288,** 275–287 (1999).

270. Tricarico, R., Cortellino, S., Riccio, A., Jagmohan-Changur, S., van der Klift, H., Wijnen, J., Turner, D., Ventura, A., Rovella, V., Percesepe, A., Lucci-Cordisco, E., Radice, P., Bertario, L., Pedroni, M., de Leon, M. P., Mancuso, P., Devarajan, K., Cai, K. Q., Klein-Szanto, A. J. P., Neri, G., Møller, P., Viel, A., Genuardi, M., Fodde, R. & Bellacosa, A. Involvement of MBD4 inactivation in mismatch repair-deficient tumorigenesis. *Oncotarget* **6,** 42892–42904 (2015).

271. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13,** 3405–3410 (1974).

272. Lari, S.-U., Chen, C.-Y., Vertéssy, B. G., Morré, J. & Bennett, S. E. Quantitative determination of uracil residues in Escherichia coli DNA: Contribution of ung, dug, and dut genes to uracil avoidance. *DNA Repair* **5,** 1407–1420 (2006).

273. Ren, J., Ulvik, A., Refsum, H. & Ueland, P. M. Uracil in human DNA from subjects with normal and impaired folate status as determined by high-performance liquid chromatography−tandem mass spectrometry. *Anal. Chem.* **74,** 295–299 (2002).

274. Morgan, H. D., Dean, W., Coker, H. A., Reik, W. & Petersen-Mahrt, S. K. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues. *J. Biol. Chem.* **279,** 52353–52360 (2004).

275. Bransteitter, R., Pham, P., Calabrese, P. & Goodman, M. F. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. *J. Biol. Chem.* **279,** 51612–51621 (2004).

276. Dörner, T., Foster, S. J., Farner, N. L. & Lipsky, P. E. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol.* **28,** 3384–3396 (1998).

277. Han, L., Masani, S. & Yu, K. Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. *Proc. Natl. Acad. Sci.* **108,** 11584–11589 (2011).

278. Robbiani, D. F., Bunting, S., Feldhahn, N., Bothmer, A., Camps, J., Deroubaix, S., McBride, K. M., Klein, I. A., Stone, G., Eisenreich, T. R., Ried, T., Nussenzweig, A. & Nussenzweig, M. C. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol. Cell* **36,** 631–641 (2009).

279. Bryan, D. S., Ransom, M., Adane, B., York, K. & Hesselberth, J. R. High resolution mapping of modified DNA nucleobases using excision repair enzymes. *Genome Res.* **24,** 1534–1542 (2014).

280. Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., Pupko, T. & Ast, G. Differential GC content

between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* **1,** 543–556 (2012).

281. Slupphaug, G., Eftedal, I., Kavli, B., Bharati, S., Helle, N. M., Haug, T., Levine, D. W. & Krokan, H. E. Properties of a recombinant human uracil-DNA glycosylase from the UNG gene and evidence that UNG encodes the major uracil-DNA glycosylase. *Biochemistry* **34,** 128–138 (1995).
282. Delort, A.-M., Duplaa, A.-M., Molko, D., Teoule, R., Leblanc, J.-P. & Laval, J. Excision of uracil residues in DNA: mechanism of action of *Escherichia coli* and *Micrococcus luteus* uracil-DNA glycosylases. *Nucleic Acids Res.* **13,** 319–335 (1985).
283. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
284. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv 13033997 Q-Bio* (2013).
285. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
286. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31,** 2032–2034 (2015).
287. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
288. Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. & Manke, T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44,** W160–W165 (2016).
289. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

290. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

291. Nakamura, J., La, D. K. & Swenberg, J. A. 5′-Nicked apurinic/apyrimidinic sites are resistant to β-elimination by β-polymerase and are persistent in human cultured cells after oxidative stress. *J. Biol. Chem.* **275,** 5323–5328 (2000).

292. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9,** e1003118 (2013).

293. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29,** 2046–2048 (2013).

294. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).