



(51) International Patent Classification:

C07D 401/00 (2006.01) C12Q 1/6806 (2018.01)

Published:

— with international search report (Art. 21(3))

(21) International Application Number:

PCT/EP2019/070255

(22) International Filing Date:

26 July 2019 (26.07.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

1812283.8 27 July 2018 (27.07.2018) GB

(71) Applicant: CAMBRIDGE ENTERPRISE LIMITED

[GB/GB]; The Old Schools, Trinity Lane, Cambridge Cambridgeshire CB2 1TN (GB).

(72) Inventors: BALASUBRAMANIAN, Shankar; c/o Uni-

versity of Cambridge, Department of Chemistry, Lensfield Road, Cambridge Cambridgeshire CB2 1EW (GB).

LIU, Zheng; c/o University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge Cambridgeshire CB2 1EW (GB).

VAN DELFT, Pieter; c/o University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge Cambridgeshire CB2 1EW (GB).

(74) Agent: MEWBURN ELLIS LLP; City Tower, 40 Basing-

hall Street, London Greater London EC2V 5DE (GB).

(81) Designated States (unless otherwise indicated, for every

kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every

kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: HIGH RESOLUTION DETECTION OF DNA ABASIC SITES

(57) Abstract: The present invention provides method of labelling a nucleic acid containing an abasic (AP) site. The method includes the step of reacting a nucleic acid comprising an AP site with an aldehyde probe, such that the probe covalently binds to the AP site to produce a nucleic acid labelled with the probe; optionally oxidising the nucleic acid labelled with the probe; optionally isolating the labelled nucleic acid, and selectively cleaving the nucleic acid at the AP site to produce nucleic acid fragments.

High Resolution Detection of DNA Abasic Sites

Related Application

The present application claims the benefit of and priority to GB 1812283.8 filed on 27 July 2019 (27.07.2019), the contents of which are hereby incorporated by reference in their entirety.

Field

This invention relates to the detection and mapping of abasic (AP) sites in nucleic acids.

Background

Genomic DNA is continually subject to both endogenous and exogenous forms of damage. Spontaneous hydrolysis of the N-glycosylic bond between the DNA backbone and nucleobases leads to the formation of an estimated 10,000 abasic (AP) sites per cell, per day¹. In the event of exogenous damage, these levels can be further elevated². AP sites are also generated enzymatically as an intermediate in the base excision repair (BER) pathway. Damaged DNA bases such as 8-oxoguanine or uracil from deamination of cytosine are efficiently removed by this pathway^{3,4}. An active demethylation pathway to remove the epigenetic marker 5-methylcytosine (5-mC) in DNA involving TET and TDG has also been proposed, in which abasic sites are a putative intermediate^{5,6}. AP sites are therefore an interesting and important target for study in a range of pathways. However, little is known regarding their location within genomic DNA. Chastain et al reported that the formation of AP sites from exogenous damage is clustered and non-random by DNA fiber analysis^{7,8}. However, the exact sequence context of AP sites remains largely unexplored.

Aldehyde Reactive Probe (ARP) has been widely used in dot-blot and ELISA assays to detect the aldehyde moiety revealed in the ring-open form of DNA AP sites^{9,10}. However, this approach suffers from cross-reactivity with bases including 5-formylcytosine (5-fC)¹¹ and 5-formyluracil (5-fU)^{12,13} that will confound the interpretation of any results (Figure 1).

Quantitative measurements of these formyl bases by mass spectrometry have revealed levels comparable to those estimated for AP sites in a range of tissues and cell lines^{14,15,16}. Therefore, care must be taken to distinguish between these different groups during chemical labelling.

Summary

The present inventors have developed aldehyde-reactive probes that chemically label aldehyde residues within nucleic acids and form adducts of different stabilities with abasic

(AP) sites, 5-fU and 5-fC. This allows the selective cleavage of the labelled nucleic acid strands at AP sites and may be useful for example for the high-resolution mapping of abasic (AP) sites in genomic DNA.

- 5 An aspect of the invention provides a probe for reaction with an aldehyde of a nucleic acid having an AP site, 5-fU or 5-fC. The probe may be referred to as an aldehyde probe. The probe is typically a nitrogen nucleophile, which forms a nitrogen adduct upon reaction with the aldehyde. The nitrogen adduct may be an imine, an enamine, a hydrazone, an oxime or an amine, and these groups may be present within a linear or cyclic system, or the nitrogen
10 adduct may be a heteroaromatic group having nitrogen as a ring atom, such as an aromatic ring atom.

The aldehyde probe may be a hydrazine, including a hydrazide, or a hydroxylamine probe. The hydrazine, and hydroxylamine functional groups may be connected to an alkyl, aryl,
15 cycloalkyl or heterocyclyl group, and these groups may be further substituted.

The probe may be connected to a capture tag and/or a detectable label, or the probe may have a coupling moiety, such as a functional group, for connection with to a capture tag and/or a detectable label. For example, the probe may be connected or connectable to
20 biotin.

The probe may be selected from a compound of formula A, ARP (*N*-(Aminooxyacetyl)-*N*'-biotinyl-hydrazine), truncated ARP (*N*-biotinyl-hydrazine), O-(4-nitrobenzyl)hydroxylamine, biotinamidocaproyl hydrazide, biotin-dPEG-hydrazide and alkyne hydrazide.
25

The probe may be a compound of formula A, as described in further detail below.

A further aspect of the invention provides a probe of formula A;



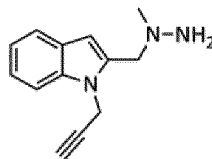
30

where -Ar is optionally substituted aryl, such as heteroaryl, such as indolyl,
-L- is alkylene, such as methylene,
-A- is -CR³R⁴-, -N(R⁵)- or -O-, such as -N(R⁵)-, where each of -R³ and -R⁴ is
independently hydrogen or alkyl, and -R⁵ is hydrogen or alkyl, such as alkyl,
35 -R¹ is hydrogen,
-R² is hydrogen or alkyl, such as hydrogen,

and salts, solvates and protected forms thereof.

Preferably, -Ar is optionally substituted indolyl, -L- is methylene, and -A- is -N(R⁵)-. For example, -Ar is indolyl substituted, such as substituted at the aromatic ring N atom, with alkynyl, such as propargyl.

Preferably, the probe has the formula B;



10 Another aspect of the invention provides a method of labelling an abasic (AP) site in a nucleic acid comprising;

reacting a nucleic acid containing an abasic (AP) site with a probe of formula A, such that the probe covalently binds to the abasic (AP) site of the nucleic acid.

15 Another aspect of the invention provides the use of an aldehyde probe, such as a probe of formula A, for labelling an abasic (AP) site in a nucleic acid.

Another aspect of the invention provides a method of isolating a nucleic acid containing an abasic (AP) site, the method comprising;

20 providing a population of nucleic acids,
contacting the population with an aldehyde probe, such as a probe of formula A, such
that the probe covalently binds to abasic (AP) sites to produce nucleic acid strands labelled
with the probe,
optionally oxidising the nucleic acid strands labelled with the probe,
25 optionally adding a capture tag or detectable label to the probe,
isolating labelled nucleic acid strands from the population of nucleic acids,
selectively cleaving the isolated nucleic acid strands at abasic sites (AP) that are
covalently bound to the probe to produce a population of nucleic acid fragments, and
isolating the nucleic acid fragments.

The isolated nucleic acid fragments correspond to the sequences 5' and 3' of abasic (AP) sites in the population of double-stranded nucleic acids.

Another aspect of the invention provides a method of mapping abasic (AP) sites in genomic nucleic acids comprising;

providing a population of double-stranded genomic nucleic acids,

contacting the population with an aldehyde probe, such as a probe of formula A, such

5 that the probe covalently binds to AP sites in nucleic acid strands within the population and labels the nucleic acid strands with the probe,

ligating a first sequencing adapter to both ends of the double-stranded genomic nucleic acids in the population, wherein the first sequencing adapter comprises a non-ligatable end,

10 isolating nucleic acid strands labelled with the probe from the population of double-stranded genomic nucleic acids,

cleaving the isolated nucleic acid strands at probe-bound AP sites to produce a population of unlabelled nucleic acid fragments,

isolating the unlabelled nucleic acid fragments,

15 annealing an extension primer to the first sequencing adapter sequence of the unlabelled nucleic acid fragments,

extending the extension primer along the unlabelled nucleic acid fragments to produce double stranded nucleic acid fragments with a non-adapted end,

20 ligating a second sequencing adapter to the non-adapted end of the double stranded nucleic acid fragments to produce a population of adapted double stranded nucleic acid fragments,

amplifying the adapted double stranded nucleic acid fragments to produce a population of amplified nucleic acid fragments, and,

sequencing the population of amplified nucleic acid fragments,

25 wherein the sequences of the amplified nucleic acid fragments correspond to the sequences 3' of AP sites in the population of double-stranded genomic nucleic acids.

The method may include the steps of optionally oxidising the nucleic acid strands labelled with the probe, and optionally adding a capture tag or detectable label to the probe.

30

Another aspect of the invention provides a kit for use in labelling AP sites; isolating nucleic acids containing AP sites; or mapping AP sites, wherein the kit comprises an aldehyde probe, such as a probe of formula A, optionally together with a reagent for cleaving abasic sites, such as a base.

35

Aspects and embodiments of the invention are described in more detail below.

Brief Description of Figures

Figure 1 shows structures of aldehyde containing moieties found in DNA.

Figure 2 shows HIPS probe **1** (the compound of formula B), *o*-phenylenediamine derivative **2**, N-biotinyl-hydrazine and ARP.

Figure 3 shows (a) Reactivity of probes with AP-ODN1. AP-ODN1 (10 μ M) was incubated with probes (1 mM) at room temperature for 2 hr. Reactions were buffered by sodium acetate (pH 5.0) or sodium phosphate (pH 6.0-7.4 at 40 mM, and followed by LC-MS. The conversion % was calculated by integration of the ligated-ODNs at 260 nm UV absorption. (b) DNA sequences of ODNs used in LC-MS studies. AP-ODN1 is obtained after treatment of ODN1 with UNG.

Figure 4 shows the stability of adducts formed between **1** and AP-ODN1, fU-ODN2 and fC-ODN3 to alkaline-cleavage (100 mM NaOH, 15 minutes). ODNs were treated with **1** at room temperature for 2 hr, apart from 5-fC which was at 37 °C for 24 hr, followed by copper-catalysed biotinylation except where labelled 'pre-click'. All cleavage reactions were carried out at 70 °C, unless labelled RT. % cleaved product was calculated by integration of UV absorption at 260 nm. Mean and S.E.M of three replicates are shown.

Figure 5 shows HIPS probe **1** conjugated ds-ODNs before and after alkaline-cleavage assay (100 mM NaOH, 15 mins at 70 °C). All samples were first treated with **1** at room temperature for 2 hr, except 5-fC which was over 24 hr at 37 °C before copper-catalysed biotinylation.

Figure 6 shows (a) Enrichment of AP ds-ODN relative to 5-fC or unmodified ds-ODNs after first round of enrichment. All samples were treated with **1** at room temperature for 2 hr followed by biotinylation. Fold-enrichment was calculated by comparison of qPCR amplification to input samples. Mean and S.E.M of three replicates are shown. (b) Recovery of modified ds-ODNs after second round of enrichment, quantified by qPCR against the input. Results from three independent replicates are shown.

Figure 7 shows a workflow of AP-seq to generate adapter ligated fragments where the first base after the P5 adapter (blue) corresponds to the position directly 3'- to captured AP sites.

Figure 8 shows the number of sequencing reads aligned to each modified or unmodified ds-ODN after treatment with **1** and biotinylation, followed by (A) AP-seq or (B) standard Illumina library preparation. (C) Number of aligned reads beginning exactly 1 base pair after site of

modification. Mean and S.E.M of two replicates are shown, with the total number of reads in each library normalized to 500,000. (D) Representative view of sequencing coverage across modified ODNs after AP-seq. Black arrows indicate the site of modification.

Figure 9 shows the overlap of sites identified by SMUG1-AP-seq with 5-hmU enriched regions.

Figure 10 shows, from left to right, the distribution of normalised ODN read counts after standard AP-seq, AP-seq after SMUG1 treatment and AP-seq after methoxyamine and SMUG1 treatment.

Figure 11 shows the distribution of normalized ODN read counts after UNG treatment followed by AP-seq.

Figure 12 shows, from left to right, the distribution of normalized ODN read counts after AP-seq with and without hOGG1 treatment.

Figure 13 shows the mapping of AP sites in HeLa DNA. (A) Enrichment of synthetic DNA before and after DNA sonication and mock re-extraction. Mean $\pm$ S.E.M. of three replicates are shown. (B) Western blot of APE1 protein after siRNA knockdown. Mean $\pm$ S.E.M. of three independent replicates are shown. (C) Overlap of AP-seq peaks called in HeLa cells treated with control or APE1 siRNA. (D) Relative enrichment of AP-seq peaks in different genomic regions expressed as $\text{Log}_2(\text{fold change})$ when compared to peaks shuffled at random. Error bars represent 95 % confidence intervals. * $q < 0.05$.

Figure 14 shows a workflow of a targeted version of AP-seq to generate adapter ligated fragments where the first base after the P5 adapter (blue) corresponds to the position directly 3'- to captured AP sites. An adapter oligonucleotide containing the P7 primer is annealed to a target sequence in the fragments, so only fragments containing the target sequence are adapted for amplification and sequencing.

Detailed Description

This invention relates to the labelling of abasic (AP; apurinic/apyrimidinic) sites and the isolation and sequencing of nucleic acids containing AP sites. Nucleic acids are reacted with an aldehyde probe, such as a probe of formula A. The probe covalently binds to AP sites and thereby labels nucleic acid strands in the nucleic acids that contain AP sites. The labelled strands are then isolated from the population of nucleic acids and may be selectively cleaved at the probe-bound AP sites. This produces a population of nucleic acid fragments

that correspond to the nucleotide sequences 5' and 3' of the AP sites in the nucleic acids. The nucleic acid fragments may be amplified, sequenced and used to map the AP sites at high resolution.

- 5 An abasic site (AP) site is a position in the backbone of a nucleic acid, such as DNA or RNA, that lacks a nucleobase i.e. the ribose of the RNA backbone or the deoxyribose group of the DNA backbone is not covalently linked to either a purine base, such as A or G, or a pyrimidine base, such as C, U or T. An AP site may be internal within the nucleotide sequence of the nucleic acid i.e. there may be nucleotides both 5' and 3' of the AP site.

10

AP sites may be endogenous i.e. a naturally occurring AP site that has not been introduced artificially through chemical, enzymatic or other treatment. Endogenous AP sites may arise from spontaneous hydrolysis of the *N*-glycosylic bond between the DNA backbone and a nucleobase or enzymatically, for example by the action of DNA glycosylases during Base
15 Excision Repair.

AP sites may be exogenous i.e. an artificial AP site introduced through chemical, enzymatic or other treatment. Suitable treatments for the introduction of abasic sites into a population of nucleic acids include chemical treatments, such as bisulfite treatment (Tanaka K et al.

- 20 (2007) *Bioorganic and Medicinal Chemistry Letters* 17, 1912), acid hydrolysis (Lindahl T et al (1972) *Biochemistry* 11, 3618; Tamm C et al. (1952) *J. Biol. Chem.* 195, 49) and dimethylsulfate treatment, and enzymatic treatments, such as Uracil-DNA glycosylase (UNG) treatment (Lindahl, T. *et al.* (1977) *J. Biol. Chem.* 252, 3286).

- 25 AP sites in a population of nucleic acids may be labelled by a probe as described herein.

A nucleic acid containing an AP site may contain one or more AP sites i.e. at least one position in the phosphate-ribose or phosphate-deoxyribose backbone of the nucleic acid lacks a nucleobase. For example, a nucleic acid may contain 1, 2, 3, 4, 5 or more AP sites.

30

The nucleic acids in the population may be single-stranded, double-stranded or a mixture of single and double-stranded nucleic acids. For example, cellular nucleic acids, such as cellular genomic DNA, may be double-stranded and cell-free nucleic acids, such as cfDNA, may be a mixture of single and double-stranded nucleic acids.

35

Preferably the nucleic acids in the population are DNA molecules, such as plasmids, synthetic DNA, viral DNA, genomic DNA preferably mammalian or human genomic DNA, and cell-free circulating DNA (cfDNA).

- 5 In other embodiments, the nucleic acids may be RNA molecules, such as genomic RNA (e.g. mammalian, plant or viral genomic RNA), mRNA, tRNA, rRNA and non-coding RNA. Genomic RNA may include mammalian, plant or viral genomic RNA.

10 The nucleic acids in the population may be 10 bases to 50 kbases in length, such as 20 to 3000 bases in length. Nucleic acids isolated from cellular sources may be greater than 1000 bases in length and may be fragmented, for example by sonication, for use as described herein. In some embodiments, the choice of the sequencing technique may determine the size of the nucleic acids in the population. For example, nucleic acids of 100-1000 bases may be compatible with Illumina sequencing.

15

In some preferred embodiments, the nucleic acids in the population may be mammalian, preferably human nucleic acids.

20 A method described herein may comprise isolating a population of nucleic acids from a sample. For example, the population of nucleic acids may be isolated from a sample of intact or disrupted cells or cellular material, such as mammalian cells, preferably human cells. Suitable samples include isolated cell and tissue samples, such as biopsies, including solid tissue or tumour biopsies. In some embodiments, the sample may be obtained from a formalin fixed paraffin embedded (FFPE) tissue sample or other stored sample of cellular
25 material.

The sample may be obtained from an individual, preferably a human individual, for example a patient having or suspected of having a disease condition, such as cancer; or a healthy or at risk individual for health monitoring or assessment; or a patient undergoing treatment to
30 assess response to a drug.

Methods of extracting and isolating genomic DNA from samples of cells are well-known in the art. For example, genomic DNA may be isolated using any convenient isolation technique, such as phenol/chloroform extraction and alcohol precipitation, caesium chloride
35 density gradient centrifugation, solid-phase anion-exchange chromatography and silica gel-based techniques.

Whole genomic DNA isolated from cells obtained from a sample may be used directly as a population of nucleic acids as described herein, after isolation or may be subjected to further preparation steps before labelling with a probe as described herein.

- 5 For example, the genomic DNA may be fragmented, for example by sonication, shearing or endonuclease digestion, to produce genomic DNA fragments. The whole or a fraction of the genomic DNA may be used as described herein. Suitable fractions of genomic DNA may be based on size or other criteria.
- 10 Suitable populations of nucleic acids may include human genomic DNA, for example from tissue samples and human cell lines, and genomic DNA from model organisms such as *C. elegans*, yeast, bacteria, such as *E. coli*, plants, such as *Arabidopsis thaliana* and mammalian models, such as mouse.
- 15 Suitable populations of nucleic acids may also include genomic DNA from cancer cells or tumours, xenografts and other cancer models, cell-free plasma DNA, and single-cell DNA.

Following fractionation, denaturation, adaptation and/or other preparation steps, the population of nucleic acids may be optionally further purified, and provided in a suitable form
20 for reaction with the probe as described herein. For example, the population of nucleic acids may be in aqueous solution in the absence of buffers before treatment as described herein.

A probe is a compound that reacts with the free aldehyde groups in a nucleic acid, for example through a Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction to form an adduct
25 (reaction 1).

A probe may react selectively with the free aldehyde group in an AP site or may also react with the free aldehyde groups in 5-formyluracil (5-fU) and/or 5-formylcytosine (5-fC) residues in the nucleic acids.

30

The probe reacts with the AP site in a nucleic acid strand to form an adduct, which is typically a cyclic adduct. The adduct formed with the AP site in the nucleic acid strand undergoes an elimination reaction (such as β - and/or β - δ - elimination) (reaction 2) in the methods described herein to cleave a phosphodiester bond in the backbone of the nucleic
35 acid strand at the AP site. The adduct formed by the probe with an AP site is more susceptible to the elimination reaction than the adduct formed by the probe with 5-fU or 5-fC residues. This allows the selective cleavage of nucleic acid strands at AP sites over

cleavage at 5-fU or 5-fC residues, for example using selective reaction conditions which favour the cleavage of nucleic acid strands at AP sites over 5-fU or 5-fC residues.

5 The probe may comprise a hydrazine group or a hydroxylamine group that is reactive with free aldehyde groups in a nucleic acid strand.

The probe may further comprise a coupling moiety, such as a functional group, which allows conjugation of the probe to another compound, such as a capture tag. Suitable coupling moieties include an alkyne and a carboxy group.

10

The probe may be a hydrazine, including a hydrazide, or a hydroxylamine probe. The hydrazine and hydroxylamine functional groups may be connected to an alkyl, aryl, cycloalkyl or heterocyclyl group, and these groups may be optionally further substituted. The aldehyde probe forms a covalent bond between a nitrogen atom of the probe and the
15 aldehyde carbon. Typically a carbon-nitrogen double bond is formed, although the final adduct may not possess this functionality, and may have a carbon-nitrogen single bond. Accordingly, the aldehyde probe may be used to prepare an adduct that is a hydrazone, an oxime or a hydrazine.

20 The formation of the product adduct may include a ring formation step, which may be the formation of a cyclic hydrazone or a cyclic amidine product, for example. The ring formation step may include the formation of an aromatic ring, or such a ring may be subsequently formed during cleavage, such as oxidative cleavage, of the labelled nucleic acid strand.

25 A hydrazine probe may react with an aldehyde to form a hydrazone product. A hydrazine probe may be a hydrazide probe. The hydrazine probe may contain an alkyl hydrazine group and may contain an alkyl hydrazide group.

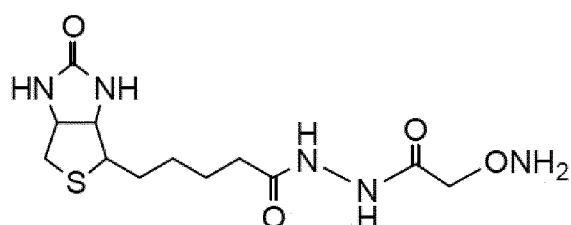
The hydrazine probe may be a hydrazide probe, such as a truncated version of the so-called
30 Aldehyde Reactive Probe (ARP), which is the Aldehyde Reactive Probe without the aminooxyacetyl group of the ARP. Thus, the truncated probe is N-biotinyl-hydrazine. Further examples of hydrazide probes for use in the methods of the invention are biotinamidocaproyl hydrazide (biotinamidoheptanoic acid hydrazide), biotin-dPEG₃-hydrazide (available from Quanta Biodesign) and alkyne hydrazide (available from Lumiprobe).

35

In the hydrazide probes mentioned above, the hydrazide group may be replaced with a hydrazine group. In this way, biotinylated hydrazine probes, for example, may be used in the methods of the invention.

- 5 A hydroxylamine probe may react with an aldehyde to form an oxime product. The hydroxylamine probe may contain an alkyl hydrazine group.

The hydroxylamine probe may be the so-called Aldehyde Reactive Probe (ARP), where the hydroxylamine functionality is connected to a biotin capture tag *via* a hydrazide-containing linker. The compound may be referred to as *N*-(aminooxyacetyl)-*N'*-biotinyl-hydrazine. ARP has the structure shown below:



Further examples of hydroxylamine probes for use in the methods of the invention are biotin-dPEG₃-oxyamine and biotin-dPEG₁₁-oxyamine (available from Quanta Biodesign).

15

A suitable probe may have the formula A above. The probe is a compound of formula A:



20

where -Ar is optionally substituted aryl, such as heteroaryl, such as indolyl,
 -L- is alkylene, such as methylene,
 -A- is -CR³R⁴-, -N(R⁵)- or -O-, such as -N(R⁵)-, where each of -R³ and -R⁴ is independently hydrogen or alkyl, and -R⁵ is hydrogen or alkyl,
 -R¹ is hydrogen,
 -R² is hydrogen or alkyl, such as hydrogen,
 25 and salts, solvates and protected forms thereof.

25

The group -R¹ is hydrogen.

30

The group -R² may be hydrogen or alkyl, such as hydrogen, methyl or ethyl, such as hydrogen or methyl, such as hydrogen.

Typically, the group -A- contains a heteroatom, such as where -A- is -N(R⁵)- or -O-. It is believed that such groups have an enhanced nucleophilicity compared with those compounds where -A- contains carbon, such as where -A- is -CR³R⁴-.

- 5 Preferably -A- is -N(R⁵)-. Here, -R⁵ may be hydrogen or alkyl, such as hydrogen, methyl or ethyl, such as hydrogen or methyl, such as methyl.

When the compound of formula A is a compound where -A- is -N(R⁵)- the compound may be referred to as a hydrazine compound.

10

Alternatively, the group -A- may be CR³R⁴- or -O-, although this is less preferred. Each of R³ and R⁴ is independently hydrogen or alkyl, such as hydrogen, methyl or ethyl. Preferably, each of R³ and R⁴ is hydrogen.

- 15 The group -L- is alkylene, which may be linear or branched. The alkylene group may be C₁₋₆ alkylene, such as C₁₋₄ alkylene, such as C₁₋₃ alkylene, such as C₁₋₂ alkylene, such as C₁ alkylene (methylene, -CH₂-).

- 20 The group -L- may be selected from -CH₂-, -CH(CH₃)- and -C(CH₃)₂-. Preferably, -L- is methylene (-CH₂-).

- 25 The group -Ar is aryl including heteroaryl and carboaryl. The aryl may be a fused ring system, and one or more of the rings in the fused system may be substituted. At least one ring within the fused ring system is an aromatic ring. The group -Ar is preferably a heteroaryl. The group -Ar preferably has a fused ring system.

30

Where -Ar contains a fused ring system, each ring may be an aromatic ring. Alternatively, a fused ring system can contain one or more non-aromatic rings, which may be fused to an aromatic ring that is present within the aryl group.

- The carboaryl may be C₆₋₁₄ carboaryl, such as phenyl or naphthyl, such as phenyl. Carboaryl groups are less preferred owing to their lower reactivity in the ring-forming reactions described herein. The inventors have also established that adducts formed from phenyl-containing compounds require strong conditions to cleave the nucleic acid.

35

The heteroaryl may be C₅₋₁₄ heteroaryl, such as C₅₋₁₀ heteroaryl, such as C₉₋₁₀ heteroaryl. The heteroaryl is preferably a nitrogen-containing heteroaryl. Thus, heteroaryl group has an

aromatic ring containing a nitrogen ring atom. Example heteroaryl groups include pyrrolyl, imidazolyl, pyrazolyl, benzoimidazolyl, and indolyl. The heteroaryl is preferably indolyl.

5 Preferably -Ar includes a nitrogen-containing aromatic ring, which is preferably a five-membered ring, such as a pyrrole ring.

10 The group -L- is connected to an aromatic ring of the aryl group, and it is typically connected to a nitrogen-containing aromatic ring. The nitrogen-containing aromatic ring may be fused to another ring, such as another aromatic ring. For example, the nitrogen-containing aromatic ring may be a pyrrole ring and this may be fused to a benzene ring, as in an indole ring system.

15 The aryl group typically bears the group -L-A-NR¹R² on a carbon aromatic ring atom. Where the aryl group contains a nitrogen aromatic ring atom, such as where the aryl group is indolyl, the aryl typically bears the group -L-A-NR¹R² on a carbon aromatic ring atom that is α (adjacent) or β to the nitrogen ring atom, such as α to the nitrogen ring atom. In an indole system, the group -L-A-NR¹R² may be a 2- or 3-substituent, such as a 2-substituent, to the indole ring.

20 The aryl group is optionally substituted, and is preferably substituted. Typically, a substituent to the aryl group is or contains a functional group, and this may be suitable for further functionalisation of the compounds of formula A. These functional groups may serve as points for the attachment of a detectable label, such as a chromophore, a fluorescent or phosphorescent label or a radiolabel. Such labels may be directly or indirectly connected to the aryl group after the probe is connected to the nucleic acid.

30 In an alternative embodiment, the compound of formula A may contain a detectable label. Such a label may be or may be part of a substituent to the aryl group. Here, the reaction of the compound of formula A with an aldehyde of the nucleic acid serves to directly label that nucleic acid.

35 Where the aryl group is a heteroaryl group, a substituent may be provided on a ring carbon or ring nitrogen atom, where appropriate. For example, where -Ar is indolyl, the nitrogen of the pyrrole ring of the indolyl group may be substituted, such as substituted with an alkynyl group, such as C₂₋₆ alkynyl, such as C₂ or C₃ alkynyl, such as C₃ alkynyl (propargyl).

For example, a substituent for the aryl group may be or contain an alkynyl or azide group, such as an alkynyl or azide group, such as an alkynyl group.

The compound of formula A may have a functional group that is connected to the aryl group either directly or *via* a linker group. The functional group may be a group selected from amine, halo, hydroxyl, thiol, carboxyl and activated carboxyl, alkenyl, alkynyl, nitro, azide and maleimide. The functional group may be selected from amine, hydroxyl, thiol, carboxyl and activated carboxyl, alkenyl, alkynyl, azide and maleimide, such as alkynyl and azide.

Preferably, an alkynyl or an azide group is connected directly to -Ar, and most preferably an alkynyl group, such as propargyl, is connected directly to -Ar.

The linker group connecting the functional group the aryl group may be selected from alkylene, heteroalkylene, cycloalkylene, heterocycloalkylene, arylene, alkylene-arylene (aralkylene), and heteroalkylene-arylene.

The functional group is a group that does not react during the reaction of the aldehyde with the probe. Additionally or alternatively, the functional group may be protected with a protecting group. The protecting group may be removed for reaction of the functional group with a labelling agent, for example after the probe is connected to the nucleic acid.

Additionally or alternatively, the aryl group may be substituted with one or more groups selected from alkyl, heteroalkyl, cycloalkyl, heterocycloalkyl, aryl, alkyl-aryl (aralkylene), and heteroalkyl-aryl.

The aryl group may be substituted with one substituent (in addition to substitution with the group -L-A-NR¹R²).

A reference to alkyl may be a reference to a C₁₋₁₂ alkyl group, which may be linear or branched. Example alkyl groups include C₁₋₆ alkyl, such as C₁₋₄ alkyl, such as C₁₋₂ alkyl, such as methyl or ethyl. A reference to alkylene may be construed accordingly.

A reference to aryl may be a reference to a C₅₋₁₄ aryl group, such as a C₅₋₁₀ aryl group, which may be carboaryl or heteroaryl. An aryl group may be phenyl, for example. A reference to arylene may be construed accordingly.

The compounds of formula A may be provided as salts, for example in a protonated form together with a suitable counter anion.

The compounds of formula A may be provided in solvated form, such as hydrated form.

5 The compounds of formula A may be provided in protected form, for example where the terminal amino group $-NR^1R^2$ is protected with an amino-protecting group. The protecting group may be removed as required for reaction of the probe with a nucleic acid.

10 Methods for the preparation of compounds of formula A are known in the art, and example methods are set out in Agarwal *et al.* The worked examples in the present case also show the preparation of a compound of formula A.

A preferred probe may have the formula B above.

15 Also provided by the present invention are the compounds of formula A.

For example, the invention provides a compound where $-R^5$ is alkyl, such as methyl.

20 For example, the invention provides a compound of formula A having an alkynyl group. Such a group is provided for useful connection to a detectable label, such as by reaction of the alkynyl group with a labelling reagent having an azide group. The alkynyl group may be a substituent to the aryl group and may be connected directly or *via* a linker.

25 The aldehyde probe reacts with free aldehyde groups within a nucleic acid to form a covalent bond which chemically labels the nucleic acid with the probe. The probe of formula A typically reacts with free aldehyde groups within a nucleic acid through a Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction.

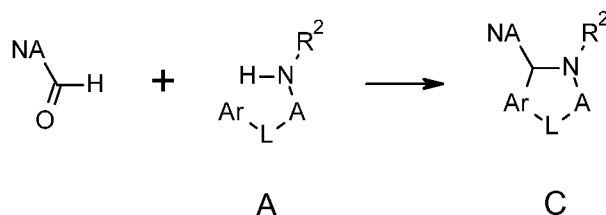
30 Free aldehyde groups may be present at AP sites, 5-formyluracils (5-fU) or 5-formylcytosines (5-fC) in the nucleic acid.

The aldehyde probe may react with AP sites, 5-fU sites and 5-fC sites or may react selectively with AP sites; AP sites and 5-fU sites; or AP sites and 5-fC sites.

35 For example, under standard conditions (such as 10 mM probe to 10 μ M DNA, room temperature for 2 hrs at pH 6.0-7.4), the aldehyde probe may react quantitatively with AP sites and 5-fU but may not react with 5-fC (i.e. may display less than < 5% reaction with 5-fC).

The compounds of formula A are reacted with an aldehyde-containing nucleic acid to generate a cyclic addition product of formula C. The reaction may be a ring-forming reaction. The reaction may be a Pictet-Spengler reaction, and more specifically a hydrazino-*iso*-Pictet-Spengler (HIPS) reaction. The probe is preferably a compound of formula A as these compounds are seen to react with aldehyde-containing nucleic acids under relatively benign conditions (such as at a pH of 6 to 7.4 at ambient temperature) and with high conversion to the adduct product (such as greater than 75%).

- 10 The ring-forming reaction is the formation of a nitrogen-containing ring, which ring is fused to a ring of the aryl group. Typically the ring is a 6-membered nitrogen-containing ring, and such is formed when -L- has a single carbon atom linking Ar- and -A- in the compounds of formula A, such as where -L- is -CH₂-.
- 15 The method of the invention may include the reaction of an aldehyde-containing nucleic acid (NA) with a compound of formula A to generate a product of formula C, as shown below:



- 20 where the aldehyde is an aldehyde group of the nucleic acid (NA), which may be an aldehyde of an abasic site, or the aldehyde of a base, such as 5-fU or 5-fC. The adduct of formula C may be subsequently functionalised with a capture tag or a detectable label to give a labelled nucleic acid D. Optionally, and preferably, the adduct C or the labelled nucleic acid D may be oxidised. The oxidation may convert the heterocycle in the adduct to
- 25 a heteroaromatic group.

- Typically, the nitrogen-containing ring is formed with a connection to an aromatic ring carbon atom. This carbon atom is typically α (adjacent) to the aromatic ring atom that is substituted with -L-A-. The nitrogen-containing ring is preferably a 6- or 7-membered ring, most
- 30 preferably a 6-membered ring. The group -L- dictates the size of the ring formed. Where -L- is a group having one carbon separating -Ar and -A-, such as where -L- is -CH₂-, a 6-membered ring will be formed. Where -L- is a group having two carbon atoms separating -Ar and -A-, such as where -L- is -CH₂CH₂-, a 7-membered ring will be formed.

Where the aryl group contains a nitrogen aromatic ring atom, such as where the aryl group is indolyl, the nitrogen-containing ring is typically formed at the ring atoms that are α and β to the nitrogen aromatic ring atom. Such a product is formed where the aryl of the compound of formula A bears the group $-L-A-NR^1R^2$ on a carbon aromatic ring atom that is α (adjacent) or β to the nitrogen ring atom. The ring forming reaction therefore forms a covalent bond at a carbon aromatic ring atom that is not substituted with the group $-L-A-NR^1R^2$.

Where the aryl group is a nitrogen-containing heteroaryl, such as a heteroaryl-containing a pyrrole ring, for example indole, the reaction may be performed under relatively benign conditions, with good conversion of the aldehyde to the cyclic addition product. Thus, the reaction may not require significant heating and may not require strongly acidic or basic conditions.

Where the aryl group is a carboaryl, or the group $-L-A-NR^1R^2$ is connected to an aromatic ring containing only carbon atoms, the reaction of the aldehyde with the probe may require heating and may require acidic reaction conditions. These are less favourable conditions for connecting a probe to a nucleic acid.

The duration and temperature of the labelling reaction are sufficient to allow the aldehyde probe to covalently bind to one or more AP sites in a nucleic acid or population. Preferably, the conditions lead to minimal nucleic acid degradation.

The reaction may be performed in an aqueous medium. The pH of the aqueous medium may be pH 4 or more, such as 5 or more, such as 6 or more. The pH of the aqueous medium may be pH 10 or less, such as 9 or less, such as 8 or less, such as 7.4 or less. For example, suitable pH may include pH 5-7.4, preferably pH 6-7.4.

The aqueous medium may contain one or more co-solvents together with water.

The reaction may be performed at ambient temperature, such as a temperature in the range 10 to 30°C. If necessary, the reaction may be performed at an elevated temperature, such as a temperature that is greater than 30°C. Here the reaction is typically performed at a temperature that is no greater than 70°C, such as no greater than 60°C, such as no greater than 50°C. For example, labelling may be performed at less than 40°C, for example 20°C to 40°C, preferably about room temperature or 37°C.

The reaction may be performed for 15 mins to 24 hours, for example under the preferred conditions mentioned above, such as pH 5-8, and 20-37°C.

- 5 The product of formula C has a heterocyclic ring formed from the aryl group together with the group L-A-NR²- and the carbon atom from the aldehyde of the nucleic acid. This heterocyclic ring is partially unsaturated, as it is fused with the aryl group. The methods of the invention may include the step of oxidising the heterocyclic ring to increase the level of unsaturation in the ring, for example such that the ring is fully unsaturated, such as the ring is aromatic.

10

For example, where the adduct contains a tetrahydropyridazine ring (for example, where where -L- is -CH₂- and -A- is -N(R⁵)- in the adduct C), this may be converted to a pyridazine ring in the oxidation reaction. Where the adduct contains a tetrahydropyridine ring (for example, where -L- is -CH₂- and -A- is -CR³R⁴- in the adduct C), this may be converted to a pyridine ring in the oxidation reaction.

15

The inventors have found that the cleavage of the labelled nucleic, as described below, is optimal when the heterocyclic group is oxidised. Thus, the methods of the invention preferably include the step of oxidising the adduct, such as the adduct C, formed from the reaction of the aldehyde with the probe of formula B. The oxidation may be undertaken using a standard oxidising agent, including an inorganic oxidising agent such as a Cu(I) or Cu(II) salt, Fe(III) salt, and a Mn (VI) salt, or an organic oxidising agent such as TEMPO (2,2,6,6-Tetramethyl-1-piperidinyloxy).

20

- 25 Advantageously, the oxidation reaction may also be combined with the steps for functionalisation of the adduct, for example to add a capture tag or a detectable label. In the preferred methods of the invention, an alkynyl group of the adduct is reacted with an azide group of a capture tag or a detectable label thereby to form a triazole connection. In this reaction, the formation of the triazole in the azide-alkyne cycloaddition is a metal-catalysed, such as a copper-catalysed, cycloaddition reaction. The reagents for use in this cycloaddition reaction to promote the formation of the triazole may also effect the oxidation of the heterocyclic ring. Thus, the functionalisation of the adduct C may be performed with the oxidation of the heterocyclic ring.

30

- 35 In other embodiments, the oxidation step is performed after the functionalisation of the adduct C. Thus, the labelled nucleic acid D may be oxidised, for example using the oxidising agents described above.

The adduct formed from the reaction may be functionalised, for example to add a capture tag or a detectable label, which may be a fluorescent label, for example. For example, the cyclic addition product C may be reacted with a functionalised capture tag or functionalised detectable label to give a labelled nucleic acid D.

Typically the functionalisation step is performed after the probe is connected to the nucleic acid. As noted above, the probe may itself include a label, and the reaction of the probe with the nucleic acid may label the nucleic acid directly.

The invention also provides a nucleic acid that is connected to a probe of the invention, and optionally where the probe is provided with a detectable label. Thus, the invention also provides the adduct of formula C and the labelled nucleic acid of formula D, together with the oxidised forms of the compounds.

In some embodiments, a capture tag may be attached to the aldehyde probe. The capture tag may facilitate the isolation of nucleic acid strands that are labelled with the probe as described below. In some embodiment, a detectable label may be attached to the probe. This label may facilitate the identification of nucleic acid strands that are labelled with the probe.

An AP site may present in one strand of a double stranded nucleic acid molecule. This labelled strand of the nucleic acid molecule may be isolated from unlabelled nucleic acids, including the complementary strand of the double stranded nucleic acid molecule.

Suitable techniques are well-known in the art. In some embodiments, the labelled nucleic acid may be isolated by contacting the population with an immobilised specific binding member that binds to labelled nucleic acid strands and isolating the labelled nucleic acid strands that are bound to the immobilised binding member. For example, nucleic acids not bound to the specific binding member may be removed by washing.

The binding member may be immobilised on a solid support. A solid support is an insoluble, non-gelatinous body which presents a surface on which the capture molecule can be immobilised for capture of the labelled nucleic acid. Examples of suitable supports include glass slides, microwells, membranes, or microbeads. The support may be in particulate or solid form, including for example a plate, a test tube, bead, a ball, filter, fabric, polymer or a membrane. Nucleic acids may, for example, be fixed to an inert polymer, a 96-well plate,

other device, apparatus or material which is used in a nucleic acid sequencing or other investigative context. The immobilisation of nucleic acids to the surface of solid supports is well-known in the art. In some embodiments, the solid support itself may be immobilised. For example, microbeads may be immobilised on a second solid surface. In preferred

embodiments, the solid support may be a magnetic bead.

Following immobilisation, the labelled nucleic acid-binding member complex may be washed, for example, to remove non-immobilised molecules from its environment, including unlabelled nucleic acids and other reagents and molecules. Suitable techniques and reagents for washing immobilised complexes are well-known in the art.

In some embodiments, the labelled nucleic acid may be tagged with a capture tag, such as biotin. The tagged nucleic acid may be isolated by contacting the population with an immobilised specific binding member that binds to the capture tag and isolating the tagged nucleic acid strands bound to the immobilised binding member. For example, untagged nucleic acids not bound to the specific binding member may be removed by washing.

The capture tag may be attached to the probe before or more preferably after the probe is reacted with the population of nucleic acids. In the worked examples of the present case, the probe ARP is used to label abasic sites in a nucleic acid. This probe contains a biotin capture tag. In other worked examples, a probe of formula B is reacted with abasic sites in a nucleic acid. The adduct is then subsequently connected to a biotin capture tag.

The capture tag or the detectable label may react with the coupling moiety of the probe to form a covalent bond that couples the capture tag or the detectable label to the probe. Any convenient chemical coupling procedure may be employed.

For example, covalent linkage of the capture tag or the detectable label to the coupling moiety of the probe may be achieved through click chemistry. For example, the coupling moiety may comprise an alkyne ($C\equiv C$) or an azide group. A coupling moiety comprising one of an alkyne or an azide group may react with a capture tag comprising the other of the alkyne or the azide group to form covalent linkage via a 1,2,3-triazole moiety. In some preferred embodiments, the coupling moiety may be an alkyne and the capture tag may comprise an azide group and the coupling moiety may react with the azide group of the capture tag through an azide-alkyne cycloaddition (AAC), for example a copper(I)-catalysed azide-alkyne cycloaddition (CuAAC).

Alternatively, covalent linkage of the capture tag or the detectable label to the coupling moiety of the probe may be achieved through sulfhydryl/maleimide or amine/activated ester reactions. For example, a capture tag comprising one of a sulfhydryl or maleimide group may react with a coupling moiety comprising the other of the sulfhydryl or maleimide group to form a 3-thiosuccinimidyl ether linkage. Alternatively, a capture tag comprising one of an amine group or an activated ester, such as N-hydroxysuccinimide ester, may react with a coupling moiety comprising the other of the amine group or the activated ester to form an amide linkage.

- 10 The capture tag may comprise any tag, molecule or group which allows the isolation of the nucleic acid to which it is attached. For example, the capture tag may be capable of binding covalently or non-covalently to a specific binding member.

Preferably, the capture tag is capable of binding non-covalently with a specific binding member to form a specific binding pair. Suitable specific binding pairs include antibody/immunogenic epitope, glutathione-S-transferase/glutathione and biotin/biotin binding protein. For example, the capture tag may be an immunogen, such as digoxigenin or a short peptide, glutathione, or preferably biotin. Other suitable capture tags are known in the art.

20 Preferably, the capture tag is biotin. Suitable specific binding members for use in binding biotin include a biotin-binding protein, such as streptavidin, avidin, anti-biotin antibody or neutravidin.

- 25 Following isolation, the labelled nucleic acid strands may be released from the immobilised specific binding member.

The detectable label may comprise any label, molecule or groups which allows for the identification, such as the localisation, of the nucleic acid to which it is attached. For example, the detectable label may be a radiolabel, a chromophore, or a fluorescent label, such as a fluorescent label.

Preferably, the detectable label is detectable by spectroscopic techniques.

- 35 A probe may contain both a capture tag and a detectable label. An example of such a probe is the biotinylated o-phenylenediamine probe described by Liu *et al.*, which contains a detectable naphthalimide group.

In some embodiments, labelled nucleic acid strands may be released and then selectively cleaved at AP sites in two separate steps. In other embodiments, labelled nucleic acid strands may be released by selective cleavage at the AP sites in a single step. The selective
5 cleavage of the immobilised strands at the AP sites generates nucleic acid fragments that are not labelled with the probe or tagged with capture tag and not bound to the immobilised specific binding member.

The labelled nucleic acid strands may be subjected to conditions that cause cleavage of a
10 phosphodiester bond in the nucleic acid backbone at the AP site for example by β -elimination and β - δ - elimination.

The reaction conditions may be selective for cleavage of the nucleic acid backbone at AP sites. AP-probe adducts in the labelled nucleic acid strand may be selectively cleaved
15 relative to 5-fU-probe or 5-fC-probe adducts. For example, AP site-probe adducts may display at least 100-fold, at least 200-fold, at least 500-fold bound or at least 1000-fold more cleavage under basic conditions than 5fU-probe or 5fC-probe adducts. The present inventors have found that AP-probe adducts formed from the compounds of formula A are
20 cleavable with excellent selectivity over the corresponding 5-fU-probe and 5-fC-probe adducts

Suitable conditions for inducing β -elimination at an AP site in a nucleic acid strand are well-known in the art and may be readily determined using standard techniques .

25 For example, the population of nucleic acid strands may be subjected to basic conditions, to cause β - and β - δ - elimination of the AP sites and produce strand-cleaved nucleic acid fragments. Basic conditions may include exposure to a base such as NaOH or piperidine at elevated temperature (i.e. higher than 20°C). Suitable conditions include 0.01 M to 1 M NaOH, preferably 0.1M, at 50-90°C, preferably 70°C.

30 Basic conditions may cause AP-probe adducts in a nucleic acid strand to undergo β - and/or β - δ -elimination reactions that cleave the labelled nucleic acid strand and generate a first fragment comprising the nucleic acid 5' of the AP site and a second fragment comprising the nucleic acid 3' of the AP site. The ends of the first and second fragments adjacent the
35 cleaved AP site may be phosphorylated.

The nucleic acid fragments generated by selective cleavage of a labelled nucleic acid strand at an AP site are single stranded.

The nucleic acid fragments may be isolated. For example, the nucleic acid fragments may be separated from labelled nucleic acids, which may include for example nucleic acid strands labelled at 5-fU or 5-fC residues that are released from the immobilised binding member under the selective cleavage conditions.

The nucleic acid fragments may be isolated by reverse selection (i.e. selection and removal of the labelled nucleic acid strands). For example, the nucleic acid fragments may be contacted with an immobilised binding member that binds to labelled DNA strands. Nucleic acid fragments that do not bind to the immobilised binding member and remain in solution may be collected and further purified.

Following the isolation, the nucleic acid fragments may be amplified and/or sequenced. The sequences of the nucleic acid fragments, in particular the second nucleic acid fragment (3' of the AP site), may be useful in locating or mapping the positions of AP sites in the original population of nucleic acids.

The methods described herein may be useful in high resolution mapping of AP sites in genomic nucleic acid (AP-seq). A method of mapping AP sites may comprise;

- providing a population of double-stranded genomic nucleic acids,
- reacting the population with an aldehyde probe, such as a probe of formula A, such that the probe covalently binds to AP sites, thereby labelling AP-site containing strands in the population of double-stranded genomic nucleic acids,
- ligating a first sequencing adapter to both ends of the double-stranded genomic nucleic acids, wherein the first sequencing adapter comprises a non-ligatable end,
- isolating nucleic acid strands labelled with the probe from the population of double-stranded genomic nucleic acids,
- cleaving the isolated nucleic acid strands at probe-bound AP sites to produce a population of unlabelled nucleic acid fragments,
- isolating the unlabelled nucleic acid fragments,
- annealing an extension primer to the first sequencing adapter sequence of the unlabelled nucleic acid fragments,
- extending the extension primer along the unlabelled nucleic acid fragments to produce double stranded fragments with a non-adapted end,

ligating a second sequencing adapter to the non-adapted end of the double stranded fragments to produce a population of adapted nucleic acid fragments,

amplifying the adapted nucleic acid fragments to produce a population of amplified nucleic acid fragments, and,

5 sequencing the population of amplified nucleic acid fragments,

wherein the sequences of the amplified nucleic acid fragments correspond to the sequences located 3' of AP sites in the population of double-stranded genomic nucleic acids.

Suitable methods for ligation of sequencing adapters are well-known in the art. For example, 10 the population of double-stranded genomic nucleic acids may contain dA overhangs (dA tails), for example following amplification or extension with a dA tailing polymerase, such as DreamTaq™ or Klenow exo-, or the double-stranded nucleic acid molecules may be blunt-ended and dA overhangs may be added to facilitate ligation of the first sequencing adapter. Suitable methods for adding dA overhangs to blunt ended nucleic acid molecules are well- 15 known in the art.

Suitable sequencing adapters for the production of adapted nucleic acids for sequencing may include a region that is complementary to the universal primers on the solid support (e.g. a flowcell or bead) and a region that is complementary to universal sequencing primers 20 (i.e. which when annealed to the adapter oligonucleotide and extended allows the sequence of the nucleic acid molecule to be read). For example, the first and second sequencing adapters may comprise a sequence that hybridises to complementary primers immobilised on the solid support (e.g. 20-30 nucleotides); a sequence that hybridises to sequencing primer (e.g. 30-40 nucleotides) and a unique index sequence (e.g. 6-10 nucleotides).

25 Suitable first and second sequencing adapters may be 56-80 nucleotides in length.

In some embodiments, one of the first and second sequencing adapters may be a P7 sequencing adapter and the other of the first and second sequencing adapters may be a P5 sequencing adapter. Suitable nucleotide sequences for sequencing adapters are well known in the art and depend on the sequencing platform to be employed. Suitable sequencing 30 platforms include Illumina (e.g. TruSeq™), LifeTech IonTorrent, Roche 454 and PacBio RS.

The first sequencing adapter is a double-stranded or partially double stranded molecule comprising a non-ligatable end and a ligatable end. The first sequencing adapter is ligated to the DNA molecules in the population at its ligatable end. The non-ligatable end of the first 35 sequencing adapter is blocked or inactivated to prevent inter- or intra-molecular ligation. Suitable techniques for blocking ligation are well known in the art. For example, the 5'

terminus at the free end may be blocked with a 5'-OMe group and/or the 3' terminus may be blocked with a single stranded spacer sequence.

After ligating the first sequencing adapter to both ends of the double-stranded genomic nucleic acids, the population of nucleic acids may be treated with alkaline phosphatase to remove terminal phosphate groups and prevent the ligation of any remaining ends.

A suitable extension primer may be complementary to the sequence of the first sequencing adaptor. The extension primer hybridises to the sequence of the first sequencing adaptor in the second DNA fragment (3' of the AP site). The extension primer is extended in a 5'-3' direction along the second DNA fragment using a polymerase. Suitable DNA polymerases are well-known in the art and include the Klenow fragment. Suitable techniques and protocols for the hybridisation of oligonucleotide primers and primer extension along a single-stranded template using polymerases are well-known in the art and reagents are available from commercial sources.

In some embodiments, the extension primer may be extended using a DNA with inherent dA-tailing ability e.g. DreamTaq™ (Thermo Fisher). This adds a dA tail to the non-adapted end of the double stranded fragments and facilitates the ligation of the second sequencing adaptor. In other embodiments, the double stranded fragments may be dA tailed in a separate step before ligation of the second sequencing adapter.

The adapted nucleic acid fragments may be amplified following ligation of the second sequencing adapter. This may facilitate further manipulation and/or sequencing. Suitable polynucleotide amplification techniques are well known in the art and include PCR. The design and use of amplification primers to amplify nucleic acid is well known in the art. Suitable amplification reactions include the polymerase chain reaction (PCR) (reviewed for instance in "PCR protocols; A Guide to Methods and Applications", Eds. Innis *et al.*, 1990, Academic Press, New York, Mullis *et al.*, Cold Spring Harbor. Symp. Quant. Biol., 51:263, (1987), Ehrlich (ed.), PCR technology, Stockton Press, NY, 1989, and Ehrlich *et al.*, Science, 252:1643-1650, (1991)).

The adapted nucleic acid fragments may be sequenced using any convenient low or high throughput sequencing technique or platform, including Sanger sequencing, Solexa-Illumina sequencing (for example, TruSeq™), ligation-based sequencing (SOLiD™), pyrosequencing; single molecule real-time sequencing (SMRT™); PacBioscience sequencing; and semiconductor array sequencing (Ion Torrent™). Preferably, sequencing is performed by

next-generation sequencing. Suitable protocols, reagents and apparatus for nucleic acid sequencing are well-known in the art and are available commercially.

The sequencing technique or platform employed will be compatible with the first and second sequencing adapters present on the adapted nucleic acid fragments.

- 5 Methods described herein may be useful, for example, in identifying and/or characterising nucleic acids in the population which contain AP sites. In particular, nucleotide sequences containing AP sites may be identified and mapped within the genome.

In some embodiments, AP sites may be mapped in a subset of nucleic acids in the population of genomic nucleic acids, for example genomic nucleic acids in the population
10 that contain a target genomic sequence. A method of mapping AP sites in genomic nucleic acid containing a target genomic sequence comprising;

- providing a population of double-stranded genomic nucleic acids,
- contacting the population with an aldehyde probe, such as a probe of formula A, such that the probe covalently binds to AP sites in nucleic acid strands within the population and
15 labels the nucleic acid strands with the probe,
- isolating nucleic acid strands labelled with the probe from the population of double-stranded genomic nucleic acids,
- cleaving the isolated nucleic acid strands at probe bound AP sites to produce a population of unlabelled nucleic acid fragments,
- 20 isolating the unlabelled nucleic acid fragments,
- annealing an adapter oligonucleotide to a target genomic sequence in the unlabelled nucleic acid fragments, wherein the adapter oligonucleotide comprises a first sequencing adapter and a targeting sequence that hybridises to the target genomic sequence,
- extending the adapter oligonucleotide along the unlabelled nucleic acid fragments
25 comprising the target sequence to produce double stranded nucleic acid fragments with a non-adapted end,
- ligating a second sequencing adapter to the non-adapted end of the double stranded nucleic acid fragments to produce a population of adapted double stranded nucleic acid fragments,
- 30 amplifying the adapted double stranded nucleic acid fragments to produce a population of amplified nucleic acid fragments, and,
- sequencing the population of amplified nucleic acid fragments,
- wherein the sequences of the amplified nucleic acid fragments correspond to the sequences 3' of AP sites in double-stranded genomic nucleic acids in the population that
35 comprise the target genomic sequence.

The targeting sequence of the adapter oligonucleotide may be complementary to the target genomic sequence. The presence of the targeting sequence in the adapter oligonucleotide causes the first sequencing adapter to be introduced only to those nucleic acid fragments
5 that contain the target genomic sequence. The first sequencing adapter of the adapter oligonucleotide does not hybridise to the target genomic sequence and may be present as an overhang or non-complementary portion.

The adapter oligonucleotide may also contain a random sequence, for example a random 3-
10 9-mer sequence, such as a hexamer, to facilitate distinguishing between PCR duplicates.

The identification and mapping of AP sites in the genome may be useful in the study of neural development and function, and cell differentiation, division and proliferation, as well as the prognosis and diagnosis of diseases, such as cancer.
15

In some embodiments, a set of sequence reads of adapted nucleic acid fragments may be determined, for example 10 or more, 100 or more or 1000 or more sequence reads may be determined.

The sequence reads may be analysed by routine bioinformatic techniques. For example, low quality sequence reads and reads arising only from sequencing adaptors may be removed and the sequence reads may be aligned with reference sequences.
20

The identified sequence reads of the adapted nucleic acid fragments may be analysed to determine the location of AP sites in the population of nucleic acids. When the population of nucleic acids are genomic DNA, the adapted nucleic acid fragments may be analysed to determine the location of AP sites in the genome. For example, a sequence read of the adapted nucleic acid fragments that terminates at a position in the sequence of a nucleic acid in the population may be indicative of the presence of an AP site at that position. In
25 some embodiments, an increased proportion of sequence reads that terminate at a position in the sequence of a nucleic acid in the population relative to other positions may be indicative of the presence of an AP site at that position.
30

A pattern or map of AP sites in the population of nucleic acids may be determined from the set of sequence reads. For a population of genomic DNA molecules obtained from a sample
35 of cells, the pattern or map of AP sites in the genome or part of the genome of the cells may be determined from the set of sequence reads.

This pattern may be indicative of the progress or status of a disease condition. For example, the AP site pattern of the target species may be useful in determining the progress of a disease condition or its prognosis or the responsiveness of a disease condition to treatment.

- 5 The AP site pattern of the target species may be also useful in monitoring the response of an individual with a disease condition to treatment.

This pattern of AP sites may be indicative of the tissue of origin of the target species of the subject-nucleic acid. For example, the pattern of AP sites may be useful in identifying a
10 diseased or cancerous tissue in an individual or diagnosing a disease condition such as cancer in an individual.

Methods of labelling and mapping AP sites as described may also be useful in mapping base mismatches and target modified bases, such as uracil, alkylpurine,
15 5-hydroxymethyluracil, 5-formyluracil, 8-oxoguanine, 5-formylcytosine, and 5-carboxycytosine.

A method of labelling and mapping modified bases may comprise;
treating a population of nucleic acids with a glycosylase specific for a target modified
20 base to replace the target modified base in the nucleic acids with an AP site.

The AP sites introduced by the glycosylase may then be mapped as described above to identify the locations of the target modified bases in the nucleic acids. For example, AP sites introduced by the glycosylase into a population of genomic DNA molecules obtained from a
25 sample of cells may be useful in mapping the positions of the target modified bases in the genome of the cells.

In some embodiments, endogenous AP sites in the population of nucleic acids may be silenced before treatment with the glycosylase.
30

Endogenous AP sites in the population of nucleic acids may be silenced by any convenient method. Suitable methods include chemical reduction or reaction with a probe such as methoxyamine. For example, the endogenous AP sites may be reacted with methoxyamine to form stable oximes, or reduced, for example using NaBH_4 , to form alcohols. 5-fU and 5-fC
35 sites may be silenced along with endogenous AP sites in the population of nucleic acids so that only synthetic AP sites generated by the glycosylase are labelled with the aldehyde

probe; optionally the step of separating DNA fragments from labelled DNA strands may be omitted in these embodiments.

In other embodiments, AP sites identified in the glycosylase-treated population of nucleic acids may be compared to AP sites in a control population that is not been treated with the glycosylase. The AP sites identified in the glycosylase untreated control population (i.e. endogenous AP sites) may be subtracted from the AP sites identified in the glycosylase treated population in order to identify those AP sites that are introduced by the glycosylase. These AP sites may then be mapped as described above to identify the locations of the target modified bases in the nucleic acids. Suitable control populations may include populations of nucleic acids from the same sample as the glycosylase-treated population.

Suitable glycosylases for different target modified bases are well-known in the art and include any glycosylase for which the inherent AP lyase activity is halted or outcompeted by reaction with the aldehyde probe. For example, uracil may be excised to leave an AP site using Uracil-DNA-glycosylase (UNG/UDG); alkylpurine may be excised to leave an AP site using AlkC or AlkD; 5-hydroxymethyluracil or 5-formyluracil may be excised to leave an AP site using single-strand selective monofunctional uracil DNA glycosylase (SMUG1); oxo-G, FapyG or 8-oxoA may be excised to leave an AP site using 8-oxoguanine DNA glycosylase 1 (OGG1) and 5-formylcytosine or 5-carboxycytosine may be excised to leave an AP site using Thymine DNA glycosylase (TDG) .

Mismatch sites in a nucleic acid may be identified by converting the mismatch sites into AP sites using a glycosylase.

A method of labelling and mapping base mismatches may comprise;

treating a population of nucleic acids with a glycosylase specific for a base mismatch to replace the base mismatch in the nucleic acids with an AP site.

In some embodiments, endogenous AP sites in the population of double stranded nucleic acids may be silenced as described above before treatment with the glycosylase.

In other embodiments, endogenous AP sites may be identified in a control population of nucleic acids and subtracted from the total AP sites identified in the glycosylase-treated population to identify the AP sites introduced by the glycosylase.

The AP sites introduced by the glycosylase may then be mapped as described above to identify the locations of the base mismatches in the nucleic acids. For example, AP sites introduced by the glycosylase into a population of genomic DNA molecules obtained from a sample of cells may be useful in mapping the positions of the base mismatches in the genome of the cells.

Suitable glycosylases for different base mismatches are well-known in the art. For example, G/T mismatches may be excised to leave an AP site using Thymine DNA glycosylase (TDG) or methyl-CpG-binding domain protein 4 (MBD4).

Another aspect of the invention provides kits for use in labelling and mapping AP sites. A kit may comprise an aldehyde probe, such as a probe of formula A.

Suitable aldehyde probes are described in detail above.

The kit may further comprise a capture tag as described above for coupling to the probe. The kit may further comprise reagents for labelling the probe-AP adduct with the capture tag, such as Cu (I) or (II), Tris(3-hydroxypropyltriazolylmethyl)amine (THPTA) ligand and sodium ascorbate.

The kit may further comprise nucleic acid isolation reagents. Suitable reagents are well-known in the art and include spin-chromatography columns.

The kit may further comprise a labelling buffer for attachment of the probe to nucleic acid containing AP sites.

The kit may further comprise a cleavage buffer for selective cleavage of the nucleic acid backbone at the positions of AP-probe adducts. A suitable cleavage buffer may be basic i.e. pH >10, >11 or >12, and may for example comprise 10mM to 1M NaOH, for example 100 mM NaOH. The kit may further comprise a base, which may be present within the cleavage buffer, for selective cleavage of the nucleic acid.

The kit may further comprise a specific binding member. The specific binding member may bind specifically to the label or capture tag of the aldehyde probe in the kit. For example, the specific binding member may bind to a biotin capture tag. Suitable members include streptavidin. The specific binding member may be immobilised or immobilisable on a solid support.

The kit may further comprise a solid support. The solid support may be coated or coatable with the specific binding member. Suitable solid supports are described above and include magnetic beads. In some preferred embodiments, the capture tag of the aldehyde probe is biotin and the solid support is streptavidin-coated magnetic beads. A magnet may be included in the kit for purification of the magnetic beads.

A kit may include one or more other reagents required for the method, such as buffer solutions, sequencing and other reagents. For example, a kit may include one or more reagents for primer extension from the target nucleic acid specific primer. Suitable reagents may include a polymerase, such as Klenow exo-, dNTPs and an appropriate buffer. The kit may also comprise reagents for DNA ligation, such as T4 ligase; reagents for end repair, such as T4 DNA Polymerase, Klenow Fragment, T4 Polynucleotide Kinase and dNTPs; and reagents for dA tailing, such as Taq DNA Polymerase and Klenow exo-.

A kit may include sequencing adapters and one or more reagents for the attachment of sequencing adapters to the ends of isolated nucleic acids, such as T4 ligase.

A kit may include one or more reagents for the amplification of a population of nucleic acids using the amplification primers. Suitable reagents may include a thermostable polymerase, for example a high discrimination polymerase, dNTPs and an appropriate buffer.

A kit may include one or more reagents for silencing endogenous AP sites and a glycosylase for converting modified bases or base mismatches into synthetic AP sites. Suitable reagents are described above.

The kit may further comprise one or more oligonucleotides for use as controls. A suitable positive control oligonucleotide may comprise at least one AP site.

A suitable negative control oligonucleotide may be devoid of AP sites. In some embodiments, the negative control oligonucleotide may comprise at least one 5-fC and/or 5-fU residue. Control oligonucleotides may be made synthetically by standard methods.

A kit for use in labelling, enrichment or detection of AP sites may include one or more articles and/or reagents for performance of the method, such as means for providing the test sample itself, including DNA and/or RNA isolation and purification reagents, sample handling

containers (such components generally being sterile), and other reagents required for the method, such as buffer solutions, sequencing and other reagents.

The kit may include instructions for use in a method of labelling AP sites as described above.

5

Other aspects and embodiments of the invention provide the aspects and embodiments described above with the term “comprising” replaced by the term “consisting of” and the aspects and embodiments described above with the term “comprising” replaced by the term “consisting essentially of”.

10

It is to be understood that the application discloses all combinations of any of the above aspects and embodiments described above with each other, unless the context demands otherwise. Similarly, the application discloses all combinations of the preferred and/or optional features either singly or together with any of the other aspects, unless the context demands otherwise.

15

Modifications of the above embodiments, further embodiments and modifications thereof will be apparent to the skilled person on reading this disclosure, and as such, these are within the scope of the present invention.

20

All documents and sequence database entries mentioned in this specification are incorporated herein by reference in their entirety for all purposes.

“and/or” where used herein is to be taken as specific disclosure of each of the two specified features or components with or without the other. For example “A and/or B” is to be taken as specific disclosure of each of (i) A, (ii) B and (iii) A and B, just as if each is set out individually herein.

25

Certain aspects and embodiments of the invention will now be illustrated by way of example and with reference to the figures described above.

30

Experiments

1. Methods

A modified P7 adapter sequence was ligated onto DNA sequences. This P7 adapter contained a 5'-OMe modification of the top strand, and a 3'-spacer on the bottom to prevent self-ligation, since the 5'-OH is blocked. DNA was then treated with alkaline phosphatase to inactivate any remaining ends that have not successfully undergone ligation. Biotinylated

35

DNA was then captured using magnetic streptavidin beads, and complementary strands that were not themselves biotinylated were washed away in 0.1 M NaOH at room temperature. The biotinylated DNA strand was eluted from the beads and was cleaved as in Fig.2 in a single step (0.1 M NaOH, 70°C). Although the adduct formed between probe **1** and 5-fU after CuAAC was stable under these conditions, 5-fU DNA was still released due to disruption of the biotin-streptavidin interaction under such harsh conditions. To separate AP site-derived fragments, which no longer contain biotin and any 5-fU-containing strands which are still biotinylated, the eluent was subjected to a second round of reverse-selection. AP site-derived fragments remained in the supernatant (since they were no longer biotinylated) whilst biotinylated 5-fU-DNA was captured and removed from the process. A single polymerase extension was then carried out on these enriched AP-containing fragments using the P7 adapter sequence for priming. As β - and/or β - δ -elimination generated fragmented DNA already bearing 5'-phosphate groups, either Klenow or another polymerase with inherent dA-tailing ability (DreamTaq™) was used for this extension. Thus, after primer extension the DNA was directly ready for adapter ligation without the need for end-prep. The P5 adapter was then introduced by ligation, followed by PCR amplification to generate the sequencing library. The 5'-OMe modification of the P7 adapter also functioned as a protecting group during the second P5 adapter ligation for any non-AP sequences that may have been carried through the workflow.

2. Results

We designed a probe (Fig. 2 #1) that bears an alkyne handle for ease of functionalization, to react with abasic sites through the Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction (Reaction 1). *o*-Phenylenediamine has been previously shown to react with the aldehyde group in 5-fU to form a stable adduct^{12,18}. A derivative of *o*-phenylenediamine compatible with AP sites, (Fig. 2 #2), was chosen along with ARP (Fig. 2 #3), and the reactivity of these probes was screened on a model oligodeoxynucleotide (ODN) containing a single AP site (Fig. 2). 10 μ M AP-ODN1 was incubated with each probe at 1 mM in buffered solutions. At pH 6 and above, **1** outperformed both **2** and ARP, and reactivity was retained well up to pH 7.4 (Fig. 3). Due to the accelerated rate of formation of AP site artefacts under acidic conditions, this is an important advantage when studying these sites. When the concentration of **1** was increased to 10 mM, at pH 7.4, quantitative conversion to ligated product was observed for both AP-ODN1 and fU-ODN2, whilst no detectable reactivity with fC-ODN3 was observed under these conditions (< 5 %). This is in agreement with previous reports of the reduced reactivity of 5-fC with amine probes¹². Sequencing by enrichment, using a biotinylated probe followed by pulldown with magnetic streptavidin beads has been demonstrated for multiple DNA modifications. We adopted this approach for AP-seq, by incorporating a biotin moiety onto

HIPS-labelled AP sites *via* the CuAAC reaction. Quantitative biotinylation with biotin-PEG3-azide was achieved for AP-ODN1 pre-treated with **1**, in the presence of CuBr and the stabilizing ligand THPTA. However, under these conditions, oxidation of the 6-membered ring formed during the HIPS reaction was also observed, presumably due to small amounts of Cu(II) in solution. The resulting oxidized adduct in AP-ODN1 was found to be sensitive to β - and β - δ - elimination when heated under basic conditions (100 mM NaOH, 15 min at 70 °C) (Reaction 2). This is similar to unfunctionalised AP sites, which are known to fragment under similar conditions¹⁹ and was observed in unlabeled AP-ODN1, whilst only very small amounts of elimination were observed for HIPS-labelled AP-ODN1 in the absence of copper.

The analogous adduct on fU-ODN2 was stable to fragmentation, as well as fC-ODN3, even when the HIPS reaction with **1** was extended to 24 hr at 37 °C to obtain quantitatively labelled fC-ODN3 (Fig. 4).

To confirm the difference in stability of biotinylated HIPS adducts in longer, double-stranded DNA more representative of that used during next-generation sequencing (NGS), model double-stranded ODNs (ds-ODNs) containing a single AP, 5-fU or 5-fC site on one strand were labelled with **1** and biotinylated. In the case of 5-fC, reaction with **1** was again extended to 24 hr to obtain quantitatively labelled DNA. Alkaline-cleavage, followed by gel electrophoresis again confirmed that fragmentation is only observed for the biotinylated HIPS-AP adduct (Fig. 5). We then subjected the HIPS treated ds-ODNs containing an AP site, 5-fU, 5-fC and an unmodified GCAT sequence to pulldown on magnetic streptavidin beads. Mild sodium hydroxide treatment (100 mM, room temperature) allowed denaturation of the non-modified single strand off the beads. Under these conditions, biotinylated HIPS-AP sites are largely stable (Fig. 4). Elution of the modified strand was then achieved by alkaline-cleavage. Primers were designed 3' to the modified site in each model ODN so that amplification would not be affected by any possible cleavage. Quantification of recovered DNA was carried out by qPCR, confirming that around 100-fold selectivity was obtained for AP sites relative to both 5-fC and GCAT DNA (Fig. 6a). However, this step alone was insufficient for enrichment relative to 5-fU DNA, most likely due to disruption of the biotin-streptavidin interaction under such harsh alkaline conditions. Importantly, whilst the released 5-fU DNA still bears the biotin moiety, the released fragments 3'- to AP sites are no longer biotinylated. Thus, a second round of reverse enrichment was carried out by incubating the neutralized alkaline-cleavage eluent with a fresh sample of streptavidin beads. Almost quantitative amounts of AP DNA remained in the eluent, whilst less than 5 % of 5-fU DNA was recovered (Fig. 6b). Therefore, high selectivity for AP sites is achievable after two rounds of enrichment.

With a method for the selective enrichment of AP sites in hand, we next turned our attention to adapting this into a mapping strategy (Fig. 7). First, a modified P7 adapter sequence is ligated onto both ends of DNA sequences. This P7 adapter contains a 5'-OMe modification on the top strand and a 3'-spacer on the bottom to prevent self-ligation. DNA is then enriched in two rounds on streptavidin beads, then a primer extension is performed on enriched single-stranded fragments. A final ligation using a P5 adapter generates sequenceable fragments. The 5'-OMe modification of the P7 adapter also functions as a protecting group during the second P5 adapter ligation for any non-AP derived sequences that have may have been carried through the workflow to further enhance selectivity.

AP, 5-fU, 5-fC and GCAT ds-ODNs (100-105 bp length, randomly designed sequences) were subjected to our AP-seq protocol and sequenced. Over 95% of total reads obtained by AP-seq aligned to the modified strand of AP ds-ODN (Fig. 8a). In a control experiment, where the same input DNA was subjected to standard Illumina library preparation without enrichment, the AP-strand was heavily underrepresented, accounting for less than 2% of reads (Fig. 8b). In addition, almost all the reads aligned to the AP ds-ODN were truncated directly adjacent to the AP site, whilst truncation at either the 5-fU or 5-fC modifications were minimal. The combination of biotin-streptavidin enrichment with selective DNA fragmentation results in even higher selectivity for AP-DNA and is typically around 500-1000-fold over 5-fU and 5000- to 10,000-fold over 5-fC. With respect to the whole fragment, selectivity is typically 50- to 100-fold over 5-fU and over 200-fold relative to 5-fC and GCAT DNA. Thus, AP-seq can successfully enrich for fragments in which the first base sequenced is directly 3' to an AP site to provide both nucleotide-resolution and strand-specific information.

Many DNA glycosylases have been identified that can often display remarkable levels of selectivity for their given nucleobase substrates. AP-seq can also be applied in combination with these enzymes for the mapping of different base modifications. UDG, for example can be used to selectively generate AP sites representing genomic uracil sites⁴. With the genomic distribution of uracil largely unexplored, AP-seq can also be adapted to generate a map of these sites.

3. *T-modification (hmU) mapping in Leishmania major*

In the *Leishmania major* genome, the DNA base modification 5-hydroxymethyluracil (5-hmU) is known to replace approximately 0.01 % of all thymine residues^{20, 21}. 5-hmU is associated with the hypermodified residue Base J, which plays a key role during transcription in the *L. major* genome. The human glycosylase SMUG1 is able to excise 5-hmU in DNA to generate an abasic site, as well other thymine modifications including 5-formyluracil (5-fU), uracil and

5-hydroxyuracil. We extended AP-seq to study the distribution of T-modifications in *L. major* genomic DNA. Sonicated genomic DNA was treated with SMUG1 and the resultant AP sites were mapped.

5 After alignment of the SMUG1-AP-seq sequencing reads, enriched peaks appear which have a characteristic stacked appearance whereby the first nucleotides are aligned. These sharp increases in coverage can be used to detect individual SMUG1-sensitive, modified thymine sites when compared to input DNA. A total of 3200 high confidence sites were called by SMUG1-AP-seq across two replicates at an FDR threshold of 10^{-10} . Defining the
10 start position of sequencing read 1 as position '1', we analysed the base composition of the position '0', which corresponds to the captured AP site. Over 98 % of called sites correspond to a thymine in the reference genome. In the absence of SMUG1 treatment, no significant sites were called. Together, this provides indication that the signals observed here by SMUG1-AP-seq are highly specific to any AP sites generated by SMUG1.

15

We compared our dataset to a previously reported map of 5-hmU in *L. major*²⁰. Out of the 139 reported 5-hmU enriched regions, 76 % overlap with the high-confidence sites called by SMUG1-AP-seq. Our results show that at the single nucleotide level, SMUG1-sensitive sites are highly clustered within these broad stretches and are also enriched in TpG (hmUpG)
20 motifs (Figure 9). To assess whether any motif enrichment was due to inherent sequence-context bias of the SMUG1 enzyme, synthetic DNA containing a single 5-hmU site flanked by 10 randomized bases at either end was subjected to SMUG1-AP-seq. The hmUpG enrichment was found to still be significant when compared to any enzyme bias.

25 As SMUG1 is able to excise a number of thymine derivatives, as a proof of principle we also showed with synthetic ODNs that 5-fU can be selectively excluded from this list by chemical treatment with methoxyamine. Nucleophilic probes such as methoxyamine were reacted with the aldehyde group in 5-fU, as well as any pre-existing AP sites which are not generated by SMUG1 treatment so that neither of these sites are enriched after AP-seq. After standard
30 AP-seq, only pre-existing AP sites are enriched. After SMUG1-AP-seq, both pre-existing AP sites and SMUG1-generated AP sites are enriched (5-fU and 5-hmU out of the chosen modifications) (Figure 10). Here, the SMUG1-sensitive sites can be determined as any sites that appear in SMUG1-AP-seq but not AP-seq. Alternatively, both 5-fU and AP enrichment is lost after methoxyamine treatment, which increases the specificity for 5-hmU. Although not
35 studied here, the enrichment for other SMUG1 substrates such as uracil and 5-hydroxyuracil should not be affected by this approach.

4. *Uracil Mapping*

The deamination of cytosine gives rise to uracil in DNA. This C to U transition is mutagenic and can occur spontaneously or enzymatically by AID and APOBEC proteins. For example, C to U conversions have been suggested to be important in a number of pathways, such as immunoglobulin class switch recombination²², processive demethylation of epigenetic markers²³ and hypermutation signatures associated with cancer^{24, 25}. Whilst SMUG1 displays activity for a range of thymine modifications, UNG (e.g. from *E. coli*) is known to show much higher specificity for uracil but not further C5-substituted thymine derivatives²⁶. As proof of principle for selective uracil mapping, a synthetic dsODN containing a single uracil site was pooled together with 5-fC containing DNA along with unmodified DNA, and the pool was treated with UNG and purified. The resultant DNA was then treated with HIPS probe and mapped using AP-seq. After sequencing, analysis of the relative read counts of each ODN shows good enrichment for uracil (Figure 11), demonstrating the utility of this method for mapping this modification.

5. *8-oxoG mapping*

8-oxoguanine (8-oxoG) is a common lesion that can form from guanine under oxidative stress. Genomic levels are reported to be elevated in cancer and other diseases^{27, 28}. The human glycosylase hOGG1 is able to excise 8-oxoG, as well as further oxidized derivatives e.g. 2,6-diamino-4-hydroxy-5-formamidopyrimidine (FapyG) and also 8-oxoadenine to generate an abasic site²⁹. hOGG1 is a bifunctional glycosylase, which in addition to glycosylase activity, is also capable to a lesser extent of AP-lyase activity. The product of the latter step is a beta-eliminated AP site, which is incompatible with AP-seq. It is reported that the lyase activity of hOGG1 can be reduced at high magnesium concentration³⁰. Using synthetic ODNs, we have shown successful enrichment for 8-oxoG containing DNA when hOGG1 incubation is carried out in high magnesium concentration (20 mM) (Figure 12). In addition, HIPS probe 1 was directly supplemented into the enzymatic reaction buffer in this case, in an attempt to directly trap any AP intermediates formed before lyase activity can occur. With the implications of 8-oxoG in different disease states, as well as the more recent suggestions that 8-oxoG can act as an epigenetic signalling modification in mammalian cells³¹, this is a further extension of AP-seq.

6. *Mapping of endogenous abasic sites in the human genome*

We used AP-seq to directly investigate the distribution of endogenous AP sites in HeLa DNA. As AP sites are chemically labile at high pH^{33,19}, whilst at low pH the generation of additional AP site artefacts *via* depurination is accelerated^{1,34}, to maintain accurate and sensitive detection, it is important to ensure that none of the DNA processing steps

preceding chemical labelling alter the AP site landscape. The two key steps between cell harvesting and DNA labelling used here are DNA extraction and sonication, and the effect of both of these on the pulldown efficiency was assessed. Synthetic AP and GCAT DNA (synthetic oligos) were subjected to each of these steps and the extent of enrichment was followed by qPCR. Consistent enrichment levels were observed in both conditions when compared to untreated DNA (Fig. 13a), suggesting that these treatments are sufficiently mild and compatible with the method without significant introduction of artefacts.

In mammalian cells, APE1 is the main endonuclease to initiate BER at AP sites, accounting for over 95 % of AP endonuclease activity³⁵. To further understand the distribution of AP damage in human DNA, we used siRNA mediated knockdown of APE1 in HeLa cells to study the AP landscape in genomic DNA before repair by the BER pathway. Western blot analysis confirmed that around 90% knockdown of APE1 was achieved after a 96-hour transfection period when compared to cells treated with control siRNA (Fig. 13b).

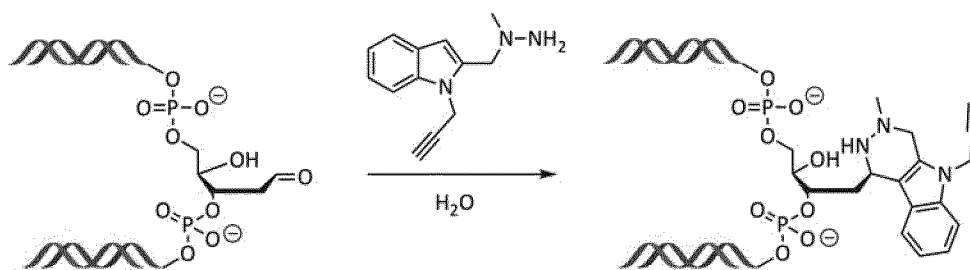
After AP-seq, we did not observe a significant build-up of sequencing reads at the single-nucleotide level. In contrast to 5-hmU sites in *L. major*, which are likely to be installed enzymatically at thymine sites with sequence specificity, we do not find such a distribution for AP damage. Instead, peak-calling identified 16,835 enriched peaks in the cells treated with control siRNA when compared to input DNA. Only high-confidence peaks that appear in at least three out of four replicates were considered here. Upon depletion of APE1, the number of peaks increased to 27,516. In addition, 75 % of control siRNA peaks also appeared in the APE1 depleted samples (Fig. 13d).

Next, the high-confidence AP peaks were analysed in the context of gene structures (Fig. 13e). For the siRNA control, a weak AP enrichment is observed in intergenic regions. This pattern changes for the APE1 deficient cells, where a modest enrichment in regulatory and transcribed regions including promoters, 5' and 3' UTRs and exons is observed instead ($q < 0.05$).

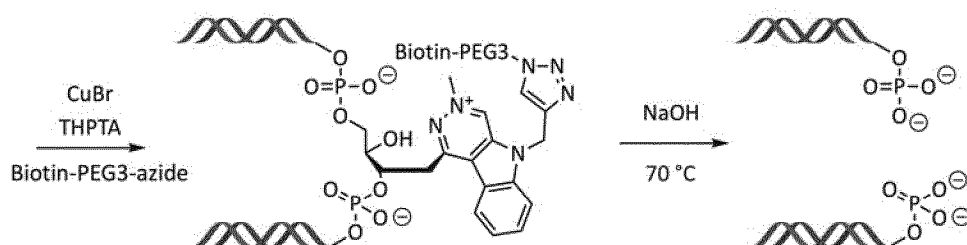
In conclusion, we have demonstrated that the Hydrazino-*iso*-Pictet-Spengler reaction can be used to selectively enrich for AP sites. After selective cleavage of the biotin handle followed by library preparation, AP-seq can be used to reveal the location of AP sites at high resolution. This method is useful in exploring the significance of abasic sites in DNA damage and repair, and can also be applied more widely to the study of a variety of DNA modifications through the use of glycosylases.

Reactions

Reaction 1



Reaction 2



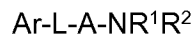
References

- (1) Lindahl, T.; Nyberg, B. *Biochemistry (Mosc.)* 1972, 11 (19), 3610.
- (2) Nakamura, J.; La, D. K.; Swenberg, J. A. *J. Biol. Chem.* 2000, 275 (8), 5323.
- 5 (3) Boiteux, S.; Radicella, J. P. *Arch. Biochem. Biophys.* 2000, 377 (1), 1.
- (4) Lindahl, T.; et al *J. Biol. Chem.* 1977, 252 (10), 3286.
- (5) Maiti, A.; Drohat, A. C. *J. Biol. Chem.* 2011, 286 (41), 35334.
- (6) He, Y.-F.; et al. *Science* 2011, 333 (6047), 1303.
- (7) Chastain, P. D. et al. *FASEB J.* 2006, 20 (14), 2612.
- 10 (8) Chastain, P. D.; Nakamura, J.; Rao, S.; Chu, H.; Ibrahim, J. G.; Swenberg, J. A.; Kaufman, D. G. *FASEB J.* 2010, 24 (10), 3674.
- (9) Nakamura, J.; Walker, V. E.; Upton, P. B.; Chiang, S.-Y.; Kow, Y. W.; Swenberg, J. A. *Cancer Res.* 1998, 58 (2), 222.
- (10) Kurisu, S. et al. *Nucleic Acids Res. Suppl.* 2001 2001, No. 1, 45.
- 15 (11) Raiber, E.-A.; et al *Genome Biol.* 2012, 13 (8), R69.
- (12) Hardisty, R. E. et al *J. Am. Chem. Soc.* 2015, 137 (29), 9270.
- (13) Ide, H.; Akamatsu, K.; Kimura, Y.; Michiue, K.; Makino, K.; Asaeda, A.; Takamori, Y.; Kubo, K. *Biochemistry (Mosc.)* 1993, 32 (32), 8276.
- (14) Pfaffeneder, T.; Spada, F.; Wagner, M.; Brandmayr, C.; Laube, S. K.; Eisen, D.;
- 20 Truss, M.; Steinbacher, J.; Hackner, B.; Kotljarova, O.; Schuermann, D.; Michalakakis, 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. *Nat. Chem. Biol.* 2014, 10 (7), 574.
- (15) Iurlaro, M.; McInroy, G. R.; Burgess, H. E.; Dean, W.; Raiber, E.-A.; Bachman, M.;
- 25 Beraldi, D.; Balasubramanian, S.; Reik, W. *Genome Biol.* 2016, 17, 141.
- (16) Zhang, H.-Y.; Xiong, J.; Qi, B.-L.; Feng, Y.-Q.; Yuan, B.-F. *Chem. Commun.* 2015, 52 (4), 737.
- (17) Agarwal, P.; Kudirka, R.; Albers, A. E.; Barfield, R. M.; de Hart, G. W.; Drake, P. M.; Jones, L. C.; Rabuka, D. *Bioconjug. Chem.* 2013, 24 (6), 846.
- 30 (18) Liu, C.; Wang, Y.; Zhang, X.; Wu, F.; Yang, W.; Zou, G.; Yao, Q.; Wang, J.; Chen, Y.; Wang, S.; Zhou, X. *Chem. Sci.* 2017.
- (19) Lhomme, J.; Constant, J. F.; Demeunynck, M. *Biopolymers* 1999, 52 (2), 65.
- (20) Kawasaki, F. et al. *Genome Biol.* 18, 23 (2017).
- (21) Bullard, W., Lopes da Rosa-Spiegler, J., Liu, S., Wang, Y. & Sabatini, R. *J. Biol. Chem.* 289, 20273–20282 (2014).
- 35 (22) Park, S.-R. *Immune Netw.* 12, 230–239 (2012).
- (23) Franchini, D.-M. et al. *PLOS ONE* 9, e97754 (2014).

- (24). Seplyarskiy, V. B. *et al.*. *Genome Res.* 26, 174–182 (2016).
- (25). Lada, A. G. *et al.* *Biol. Direct* 7, 47 (2012).
- (26). Schormann, N., Ricciardi, R. & Chattopadhyay, D. *Protein Sci. Publ. Protein Soc.* 23, 1667–1685 (2014).
- 5 (27). Nakabeppu, Y. *Int. J. Mol. Sci.* 15, 12543–12557 (2014).
- (28). Markesbery, W. R. & Lovell, M. A. *Antioxid. Redox Signal.* 8, 2039–2045 (2006).
- (29). Boiteux, S. & Radicella, J. P.. *Arch. Biochem. Biophys.* 377, 1–8 (2000).
- (31). Morland, I., Luna, L., Gustad, E., Seeberg, E. & Bjørås, M. *DNA Repair* 4, 381–387 (2005).
- 10 (32). Fleming, A. M., Ding, Y. & Burrows, C. J. *Proc. Natl. Acad. Sci.* 114, 2604–2609 (2017).
- (33). Sugiyama, H. *et al.* *Chem. Res. Toxicol.* 7, 673–683 (1994).
- (34). An, R. *et al.* *PLOS ONE* 9, e115950 (2014).
- (35). Masani, S., Han, L. & Yu, K.. *Mol. Cell. Biol.* 33, 1468–1473 (2013).

Claims:

1. A method of labelling a nucleic acid containing an abasic (AP) site, the method comprising;
- 5 providing a nucleic acid comprising an AP site,
reacting the nucleic acid with an aldehyde probe of formula A, such that the probe covalently binds to the AP site to produce a nucleic acid labelled with the probe,
wherein the probe of formula A is:



where -Ar is optionally substituted aryl, such as heteroaryl, such as indolyl,
-L- is alkylene, such as methylene,
-A- is $-\text{CR}^3\text{R}^4-$, $-\text{N}(\text{R}^5)-$ or $-\text{O}-$, such as $-\text{N}(\text{R}^5)-$, where each of $-\text{R}^3$ and $-\text{R}^4$ is
15 independently hydrogen or alkyl, and $-\text{R}^5$ is hydrogen or alkyl, such as alkyl,
 $-\text{R}^1$ is hydrogen,
 $-\text{R}^2$ is hydrogen or alkyl, such as hydrogen,
and salts, solvates and protected forms thereof.

2. A method according to claim 1 comprising isolating the labelled nucleic acid.

3. A method of labelling a nucleic acid containing an abasic (AP) site, the method comprising;
- 25 providing a nucleic acid comprising an AP site,
reacting the nucleic acid with an aldehyde probe, such that the probe covalently binds to the AP site to produce a nucleic acid labelled with the probe,
optionally oxidising the nucleic acid labelled with the probe,
optionally isolating the labelled nucleic acid, and
selectively cleaving the nucleic acid at the AP site to produce nucleic acid fragments.

4. A method of isolating nucleic acids containing AP sites, the method comprising;
- 30 providing a population of double-stranded nucleic acids,
contacting the population with an aldehyde probe, such that the probe covalently binds to AP sites to produce nucleic acid strands labelled with the probe, and optionally
35 oxidising the nucleic acid strands labelled with the probe,
isolating labelled nucleic acid strands from the population of double-stranded genomic nucleic acids,

selectively cleaving the isolated nucleic acid strands at the APs sites that are covalently bound to the aldehyde probe to produce a population of nucleic acid fragments, and

isolating the nucleic acid fragments.

5

5. A method according to any one of the preceding claims wherein the nucleic acids are DNA molecules.

10

6. A method according to claim 5 wherein the DNA molecules are genomic DNA molecules.

7. A method according to claim 6 wherein the genomic DNA molecules are obtained from a sample of cells from an individual.

15

8. A method according to any one of claims 3-7 wherein the probe is a hydrazine probe, including a hydrazide probe, or a hydroxylamine probe.

20

9. A method according to claim 8, wherein the aldehyde probe is a probe of formula A, as defined in claim 1.

10. The method according to claim 9, wherein -Ar is optionally substituted heteroaryl.

11. The method according to claim 10, wherein the heteroaryl is a nitrogen-containing heteroaryl.

25

12. The method according to claim 10, wherein the heteroaryl is optionally substituted indolyl.

30

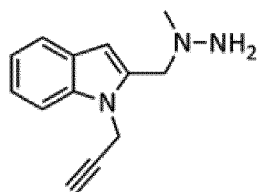
13. The method according to any one of claims 11 or 12, wherein -L- is connected to a nitrogen-containing aromatic ring.

14. The method according to claim 13, wherein the heteroaryl group bears the group -L-A-NR¹R² on a carbon aromatic ring atom.

35

15. The method according to claim 14, wherein the aryl group bears the group -L-A-NR¹R² on a carbon aromatic ring atom that is α to the nitrogen aromatic ring atom.

16. The method according to any one of claims 9-15, wherein a functional group is connected to -Ar either directly or *via* a linker, where the functional group is selected from amine, halo, hydroxyl, thiol, carboxyl and activated carboxyl, alkenyl, alkynyl, nitro, azide and maleimide, and the linker group is selected from alkylene, heteroalkylene, cycloalkylene, heterocycloalkylene, arylene, alkylene-arylene (aralkylene), and heteroalkylene-arylene, and additionally or alternatively, -Ar is substituted with one or more groups selected from alkyl, heteroalkyl, cycloalkyl, heterocycloalkyl, aryl, alkyl-aryl (aralkylene), and heteroalkyl-aryl.
17. The method according claim 16, wherein the aryl group is optionally substituted with alkynyl or azide, either directly or *via* a linker.
18. The method according to claim 17, wherein the aryl group is substituted with alkynyl, such as propargyl, and is optionally further substituted.
19. The method according to any one of claims 9 to 18, wherein -L- is methylene.
20. The method according to any one of claims 9 to 19, wherein -A- is -N(R⁵)- or -O-.
21. The method according to claim 20, wherein -A- is -N(R⁵)-.
22. The method according to claim 21, wherein -R⁵ is hydrogen or methyl.
23. The method according to any one of claims 8 to 22, wherein -R⁵ is methyl.
24. The method according to any one of claims 8 to 22, wherein -R⁵ is hydrogen.
25. The method according to any one of claims 8 to 24, wherein -R² is hydrogen or methyl.
26. The method according to claim 25, wherein -R² is hydrogen.
27. A method according to claim 9 wherein the probe has the formula B:



28. A method according to claim 8, wherein the probe comprises a hydrazine group, including a hydrazide.
- 5 29. A method according to claim 28, where the probe is selected from N-biotinyl-hydrazine, biotinamidocaproyl hydrazide, biotin-dPEG₃-hydrazide and alkyne hydrazide.
30. A method according to claim 8, wherein the probe comprises a hydroxylamine group.
- 10 31. A method according to claim 30, where the probe is selected from the Aldehyde Reactive Probe (ARP), biotin-dPEG₃-oxyamine and biotin-dPEG₁₁-oxyamine.
32. A method according to any one of the preceding claims wherein the probe is reacted with the AP sites at pH 6-7.4.
- 15 33. A method according to any one of the preceding claims wherein the probe forms an adduct at the AP site.
34. A method according to any one of the preceding claims wherein the nucleic acid fragments correspond to nucleotide sequences 5' and 3' of an AP site.
- 20 35. A method according to any one of the preceding claims wherein the method comprises coupling a capture tag to the probe.
- 25 36. A method according to any one of the preceding claims wherein a capture tag is coupled to the probe before reaction with the nucleic acid or nucleic acids.
37. A method according to any one of claims 1 to 36 wherein a capture tag is coupled to the probe after reaction with the nucleic acid or nucleic acids, and the coupling reaction optionally oxidises the nucleic acid labelled with the probe.
- 30 38. A method according to any one of claims 1 to 37 wherein the nucleic acid labelled with the probe is oxidised.
- 35 39. A method according to any one of the preceding claims wherein the method comprises ligating a first sequencing adapter to the ends of the nucleic acid or nucleic acids, wherein the first sequencing adapter comprises a non-ligatable end.

40. A method according to claim 39 wherein the non-ligatable end of the first sequencing adapter comprises a blocked 5' terminus and a 3' overhang

5 41. A method according to any one of claims 2 to 40 wherein the labelled nucleic acid strands are isolated by contacting the population with an immobilised binding member that specifically binds to the labelled nucleic acids and isolating the labelled nucleic acid strands that are bound to the immobilised binding member.

10 42. A method according to claim 41 wherein the immobilised binding member specifically binds to a capture tag coupled to the labelled nucleic acid.

43. A method according to claim 42 wherein the capture tag is biotin.

15 44. A method according to claim 43 wherein the binding member is streptavidin.

45. A method according to any one of claims 41 to 44 wherein the binding member is immobilised on a magnetic bead.

20 46. A method according to any one of claims 41 to 45 wherein the method comprises eluting the labelled nucleic acid strands from the immobilised binding member.

47. A method according to claim 45 wherein the isolated nucleic acid strands are eluted and selectively cleaved simultaneously.

25

48. A method according to any one of claims 3 to 47 wherein isolated nucleic acid strands are selectively cleaved by subjecting the strands to basic conditions.

30 49. A method according to claim 47 wherein the isolated nucleic acid strands are selectively cleaved by subjecting the strands to 0.01 M to 1 M NaOH at 50-90°C.

50. A method according to any one of claims 3 to 49 wherein the nucleic acid fragments are isolated by separating the fragments from labelled nucleic acids.

35 51. A method according to claim 46 wherein the nucleic acid fragments are contacted with an immobilised binding member that binds to labelled nucleic acid and strands and nucleic acid fragments that do not bind to the immobilised binding member are recovered.

52. A method according to any one of claims 3 to 51 comprising annealing an extension primer to the nucleic acid fragments, and extending the extension primer along the unlabelled nucleic acid fragments to produce double stranded fragments.

5

53. A method according to any one of claims 3 to 51 comprising annealing an adapter oligonucleotide to a target genomic sequence in the unlabelled nucleic acid fragments, wherein the adapter oligonucleotide comprises a first sequencing adapter and a targeting sequence that hybridises to the target genomic sequence, and

10

extending the adapter oligonucleotide along the unlabelled nucleic acid fragments comprising the target genomic sequence to produce double stranded nucleic acid fragments comprising the first sequencing adapter.

15

54. A method according to claim 52 or 53 comprising ligating one or both of a first and a second sequencing adapter to the double stranded fragments to produce a population of adapted nucleic acid fragments.

20

55. A method according to any one of claims 3 to 54 comprising amplifying the nucleic acid fragments.

56. A method according to any one of claims 3 to 55 comprising sequencing the nucleic acid fragments.

25

57. A method of mapping abasic (AP) sites in genomic nucleic acids comprising; providing a population of double-stranded genomic nucleic acids, contacting the population with an aldehyde probe, such as a probe of formula A as defined in claim 1, such that the probe covalently binds to AP sites in nucleic acid strands within the population and labels the nucleic acid strands with the probe,

30

ligating a first sequencing adapter to both ends of the double-stranded genomic nucleic acids in the population, wherein first sequencing adapter comprises a non-ligatable end,

isolating nucleic acid strands labelled with the probe from the population of double-stranded genomic nucleic acids,

35

cleaving the isolated nucleic acid strands at probe bound AP sites to produce a population of unlabelled nucleic acid fragments,

isolating the unlabelled nucleic acid fragments,

annealing an extension primer to the first sequencing adapter sequence of the unlabelled nucleic acid fragments,

extending the extension primer along the unlabelled nucleic acid fragments to produce double stranded nucleic acid fragments with a non-adapted end,

5 ligating a second sequencing adapter to the non-adapted end of the double stranded nucleic acid fragments to produce a population of adapted double stranded nucleic acid fragments,

amplifying the adapted double stranded nucleic acid fragments to produce a population of amplified nucleic acid fragments, and,

10 sequencing the population of amplified nucleic acid fragments, wherein the sequences of the amplified nucleic acid fragments correspond to the sequences 3' of AP sites in the population of double-stranded genomic nucleic acids.

58. A method of mapping abasic (AP) sites in genomic nucleic acids comprising a target
15 sequence, the method comprising;

providing a population of double-stranded genomic nucleic acids,

contacting the population with an aldehyde probe, such as a probe of formula A as defined in claim 1, such that the probe covalently binds to AP sites in nucleic acid strands within the population and labels the nucleic acid strands with the probe,

20 isolating nucleic acid strands labelled with the probe from the population of double-stranded genomic nucleic acids,

cleaving the isolated nucleic acid strands at probe bound AP sites to produce a population of unlabelled nucleic acid fragments,

isolating the unlabelled nucleic acid fragments,

25 annealing an adapter oligonucleotide to a target genomic sequence in the unlabelled nucleic acid fragments, wherein the adapter oligonucleotide comprises a first sequencing adapter and a targeting sequence that hybridises to the target genomic sequence,

extending the adapter oligonucleotide along the unlabelled nucleic acid fragments comprising the target genomic sequence to produce double stranded nucleic acid fragments with a non-adapted end, said fragments comprising the first sequencing adapter,

30 ligating a second sequencing adapter to the non-adapted end of the double stranded nucleic acid fragments to produce a population of adapted double stranded nucleic acid fragments,

amplifying the adapted double stranded nucleic acid fragments to produce a
35 population of amplified nucleic acid fragments, and,

sequencing the population of amplified nucleic acid fragments,

wherein the sequences of the amplified nucleic acid fragments correspond to the

sequences 3' of AP sites in double-stranded genomic nucleic acids in the population that comprise the target genomic sequence.

59. The method according to claim 57 or claim 58 comprising oxidising the nucleic acid strands labelled with the probe.
60. The method according to any one of claims 57 to 59 comprising coupling a capture tag to the probe.
61. A method according to any one of claims 57 to 60 wherein one of the first and second sequencing adapters is a P7 sequencing adapter and the other of the first and second sequencing adapters is a P5 sequencing adapter.
62. A method according to any one of claims 3 to 61 comprising determining the location of AP sites in the population of nucleic acids from the sequences of the nucleic acid fragments.
63. A method according to any one of the preceding claims wherein the AP sites are endogenous AP sites.
64. A method according to any one of claims 3 to 62 wherein the AP sites are introduced into the population of nucleic acids by a method comprising;
treating the population of nucleic acids with a glycosylase specific for a target modified base or base mismatch to replace the target modified base or base mismatch in the nucleic acids with an AP site.
65. A method according to claim 64 comprising silencing endogenous AP sites in the population of nucleic acids before treatment with the glycosylase.
66. A method according to claim 64 comprising identifying endogenous AP sites in a control population of nucleic acids and subtracting the identified endogenous AP sites from the total AP sites identified in the glycosylase-treated population to identify the AP sites introduced by the glycosylase
67. A method according to any one of claims 64 to 66 wherein the modified base is uracil and the glycosylase is Uracil-DNA-glycosylase (UNG/UDG).

68. A method according to any one of claims 64 to 66 wherein the modified base is alkylpurine and the glycosylase is AlkC or AlkD.

69. A method according to any one of claims 64 to 66 wherein the modified base is 5-hydroxymethyluracil or 5-formyluracil and the glycosylase is single-strand selective monofunctional uracil DNA glycosylase (SMUG1).

70. A method according to any one of claims 64 to 66 wherein the modified base is 5-formylcytosine or 5-carboxycytosine or the base mismatch is a G/T mismatch, and the glycosylase is Thymine DNA glycosylase (TDG).

71. A method according to any one of claims 64 to 66 wherein the modified base is 8-oxoguanine, 2,6-diamino-4-hydroxy-5-formamidopyrimidine, or 8-oxoadenine, and the glycosylase is 8-oxoguanine glycosylase (OGG1).

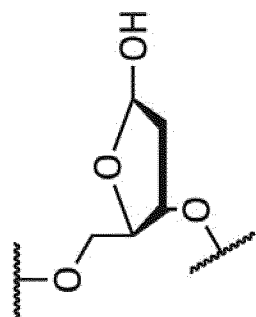
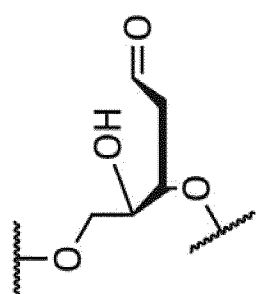
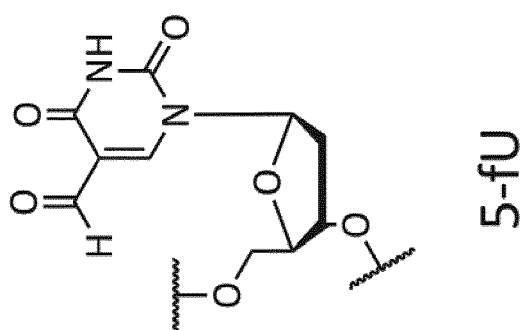
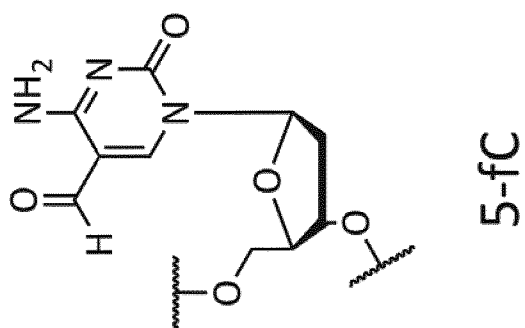
72. A method according to any one of claims 64 to 71 wherein the target modified base or base mismatch is located at positions in the nucleic acids or nucleic acid identified as AP sites.

73. A probe of formula A, as defined in claim 1.

74. Use of a probe according to claim 73 for labelling an abasic (AP) site in a nucleic acid.

75. A kit for use in labelling AP sites; isolating nucleic acids containing AP sites; or mapping AP sites, wherein the kit comprises an aldehyde probe, such as a probe of formula A according to claim 73.

Figure 1



AP

Figure 2

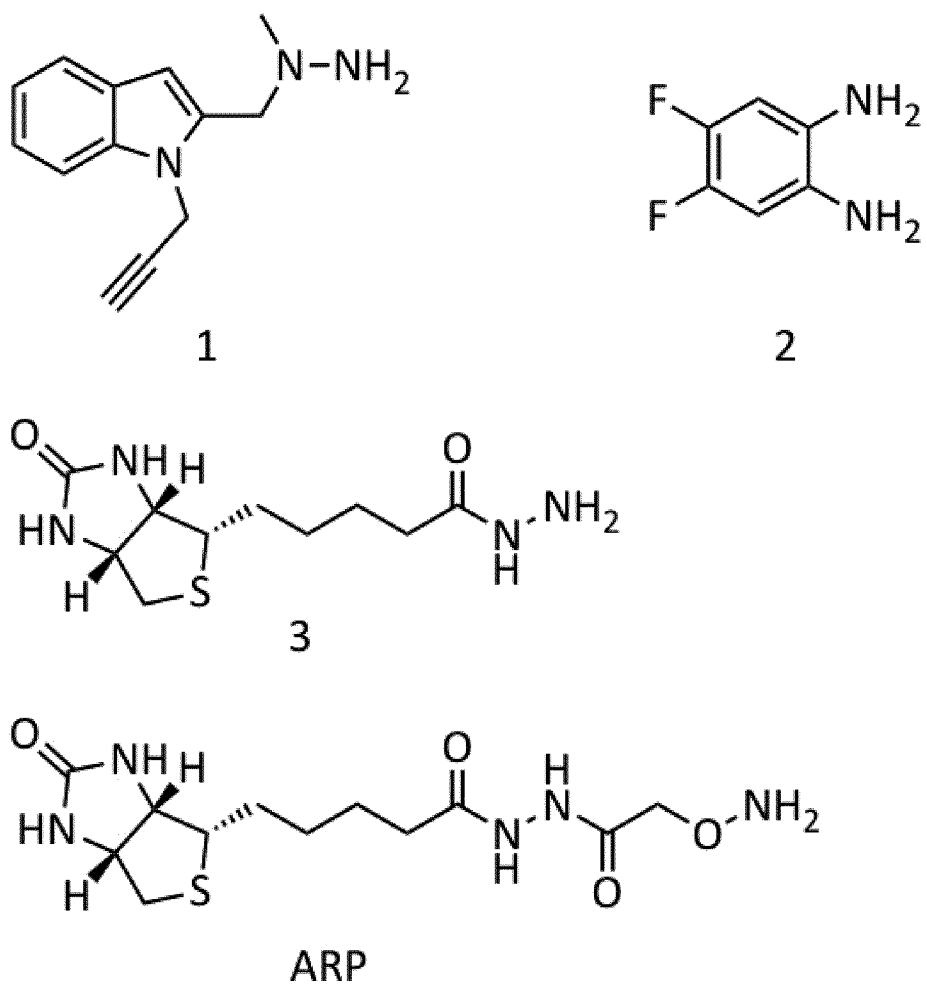
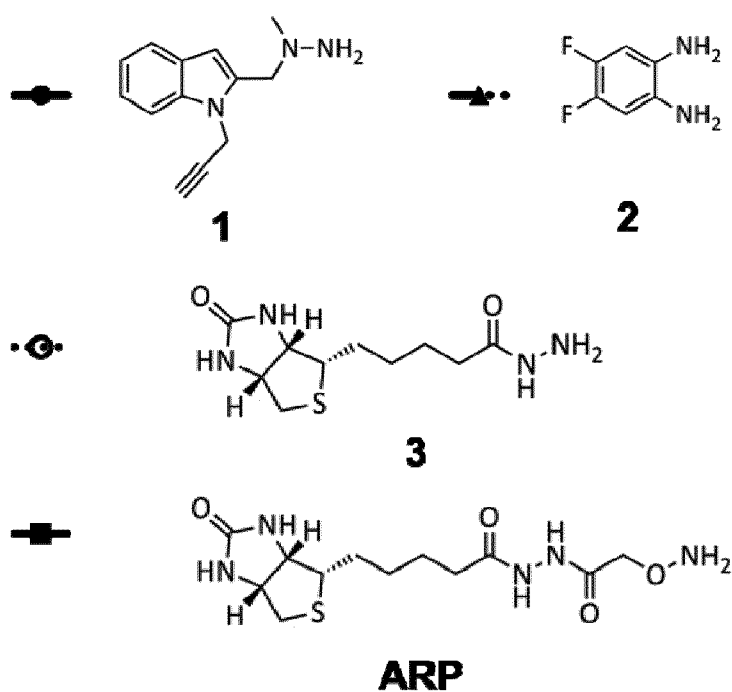
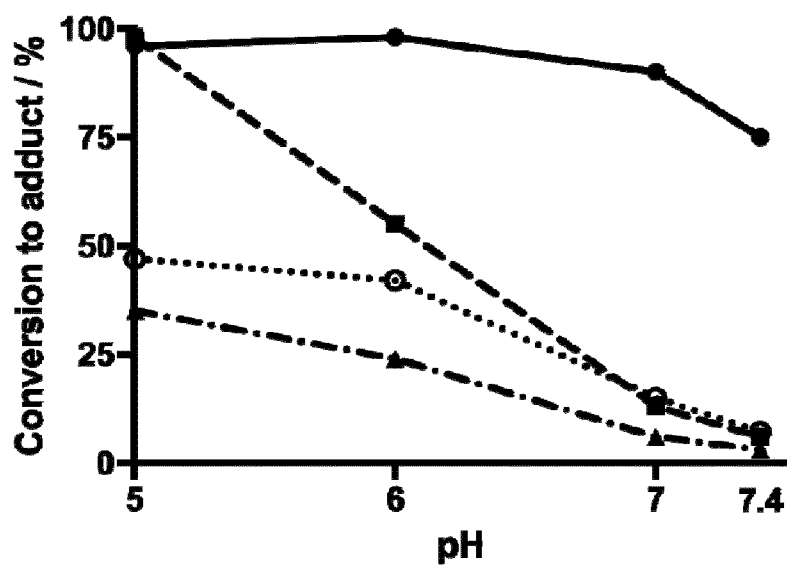


Figure 3

(A)



(B)

ODN1:	5'-AGC GAC A- <u>U</u> -A TCT TGT-3'
AP-ODN1:	5'-AGC GAC A- <u>AP</u> -A TCT TGT-3'
fU-ODN2:	5'-ATCGCA- <u>fU</u> -GTA-3'
fC-ODN3:	5'-ATCG- <u>fC</u> -GCGTA-3'

Figure 4

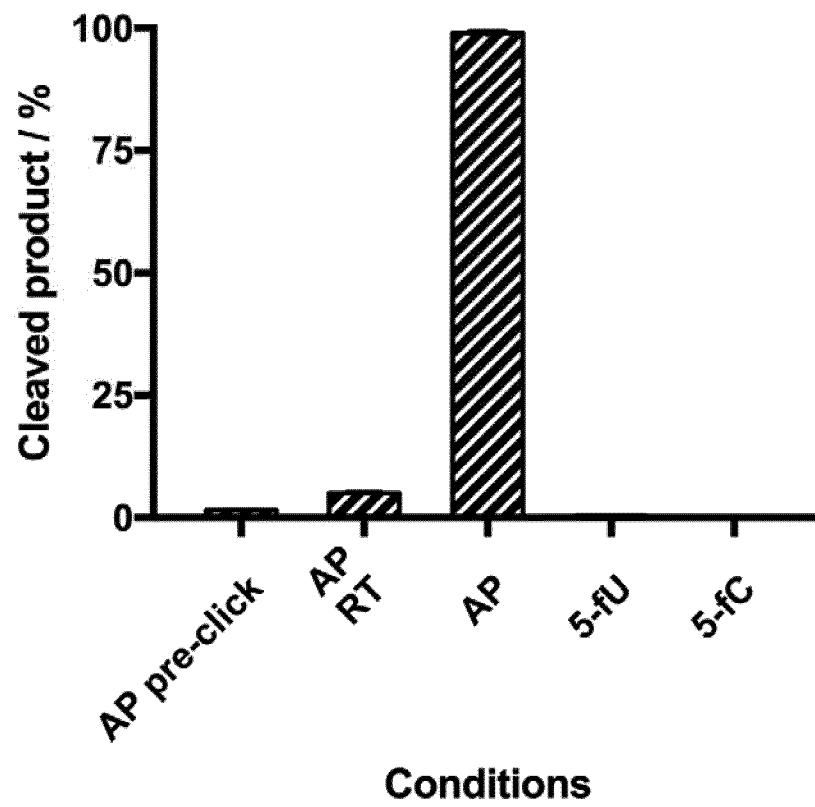


Figure 5

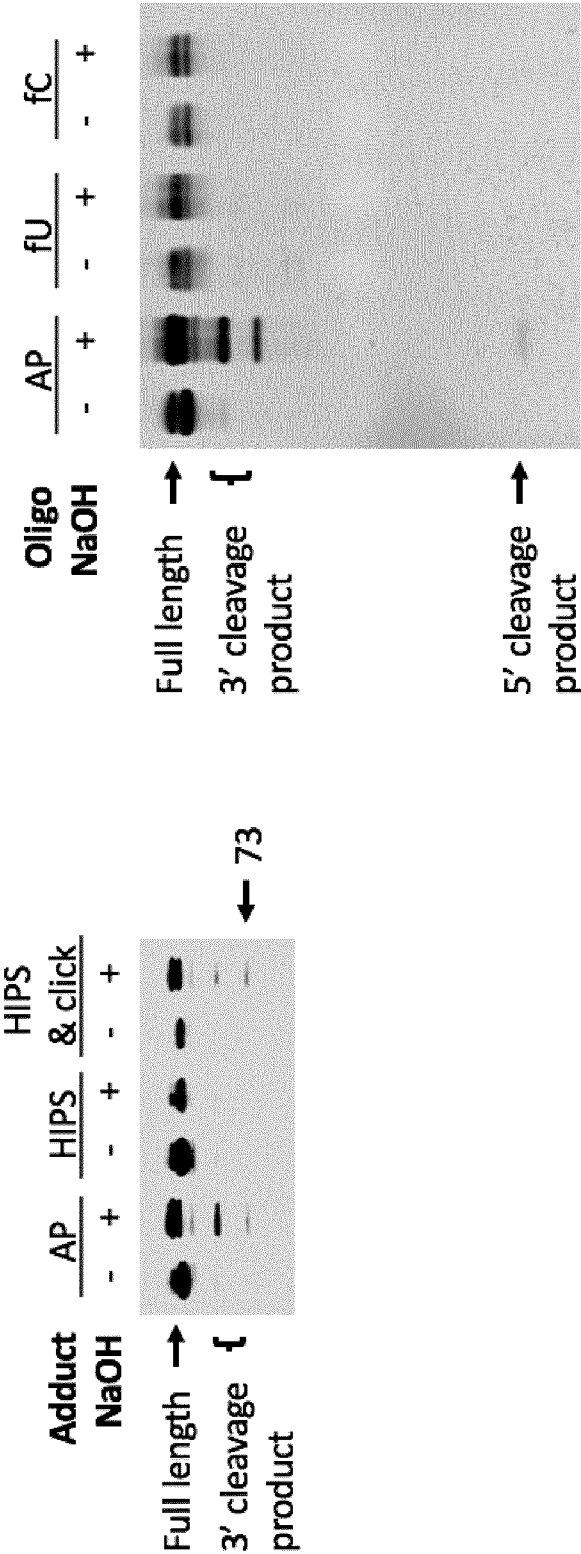
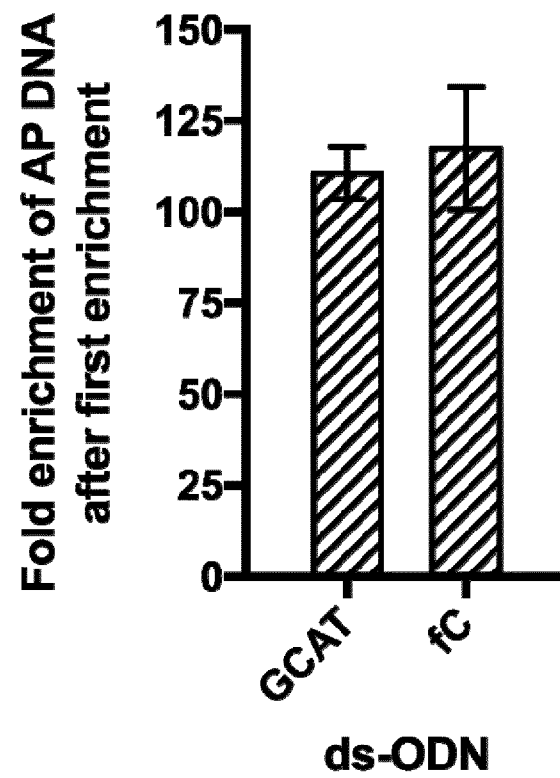


Figure 6

(A)



(B)

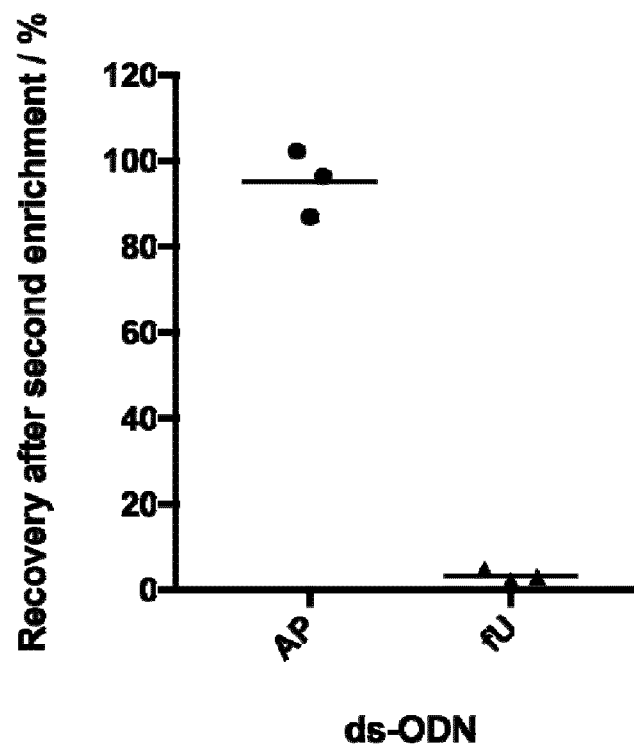


Figure 7

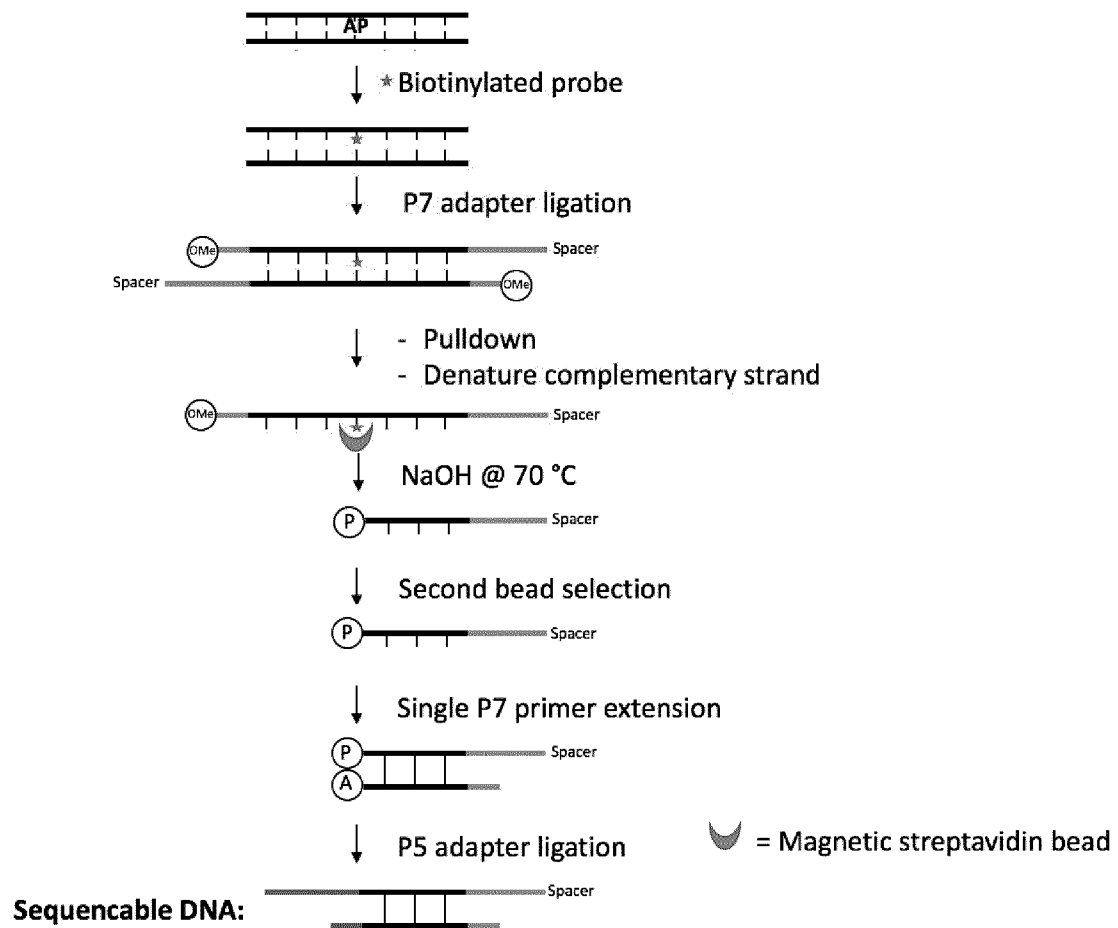
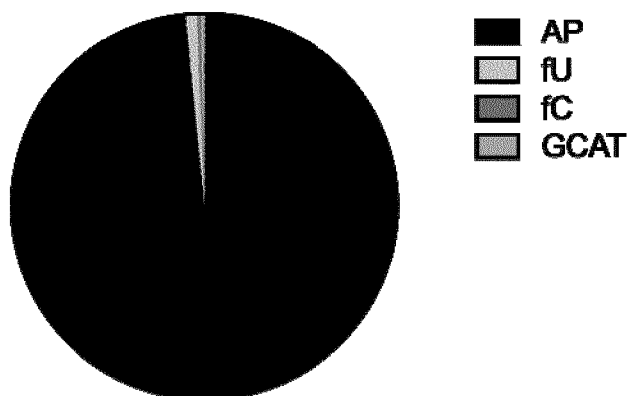


Figure 8

(A)



(B)

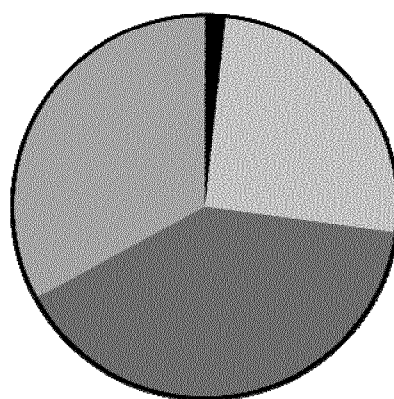
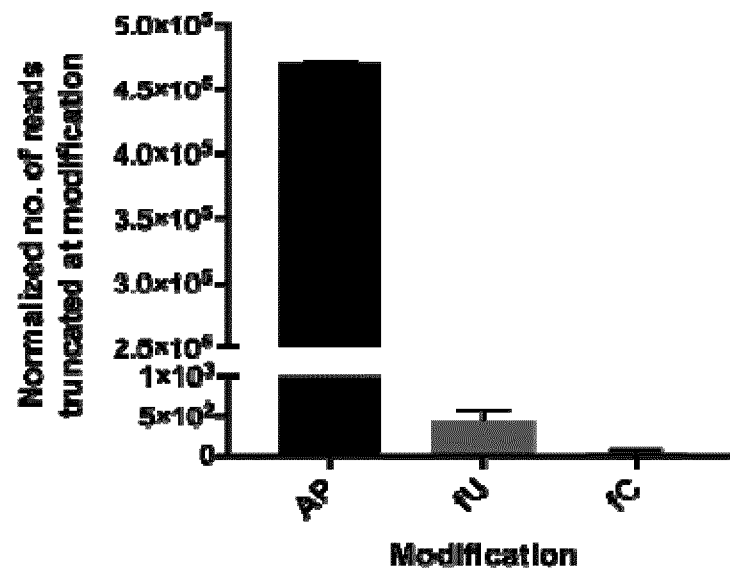


Figure 8 cont.

(C)



(D)

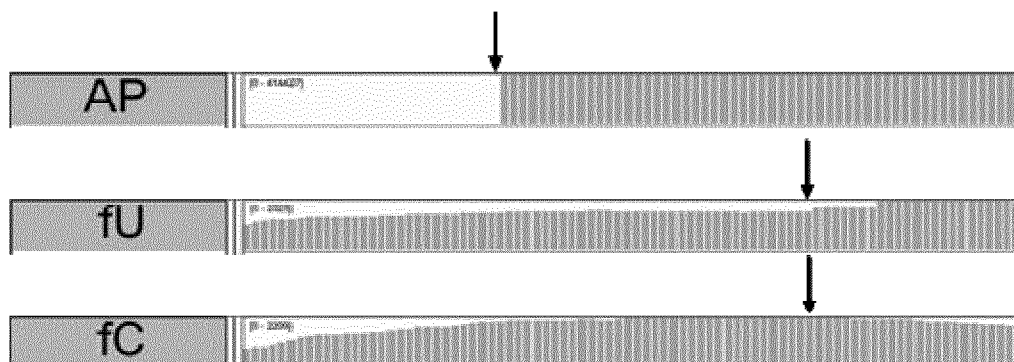


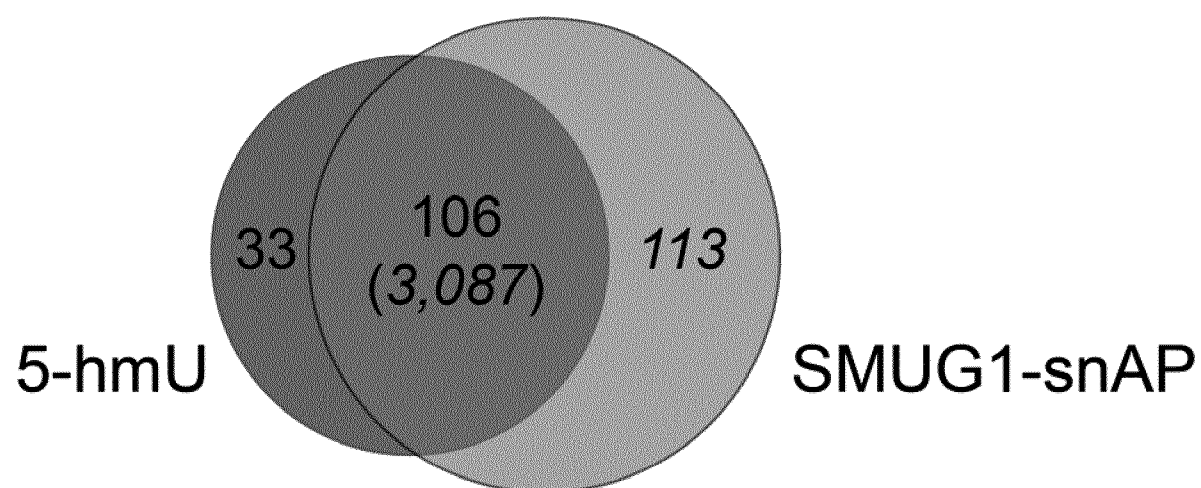
Figure 9

Figure 10

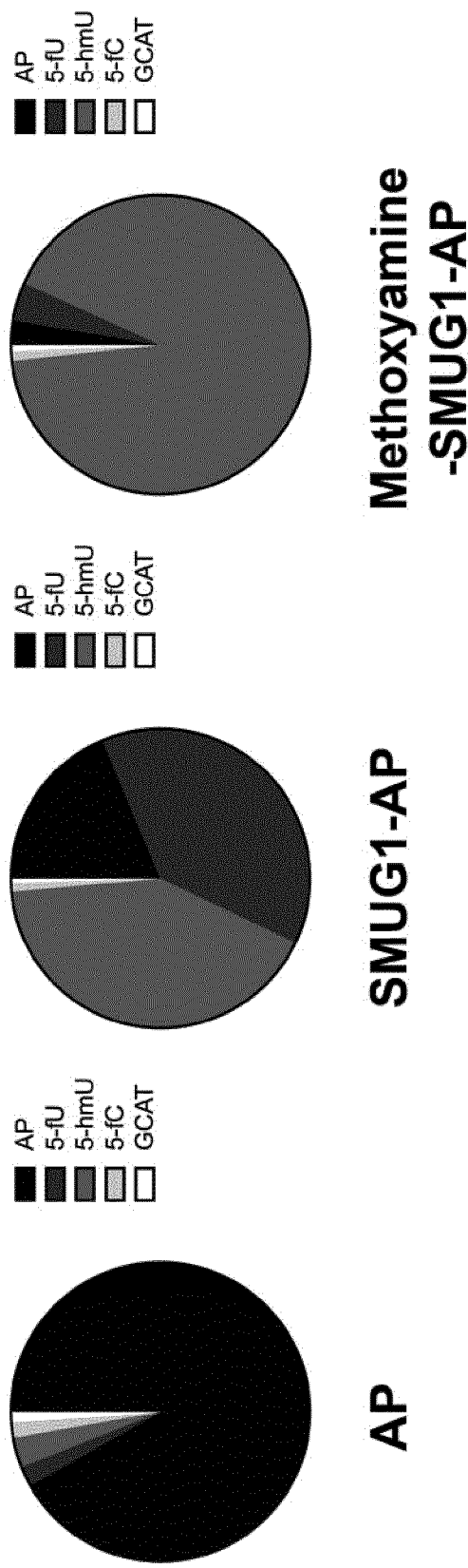


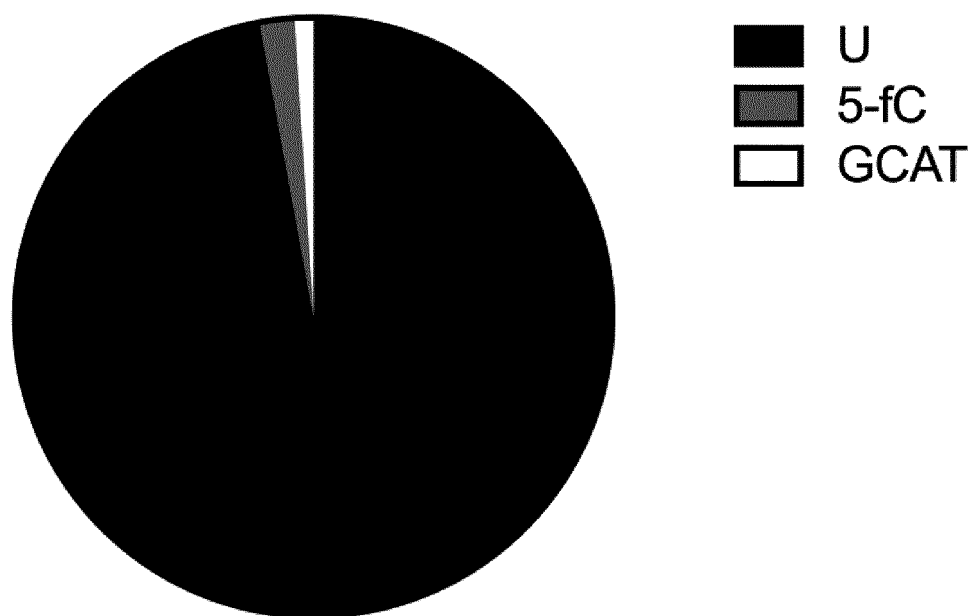
Figure 11

Figure 12

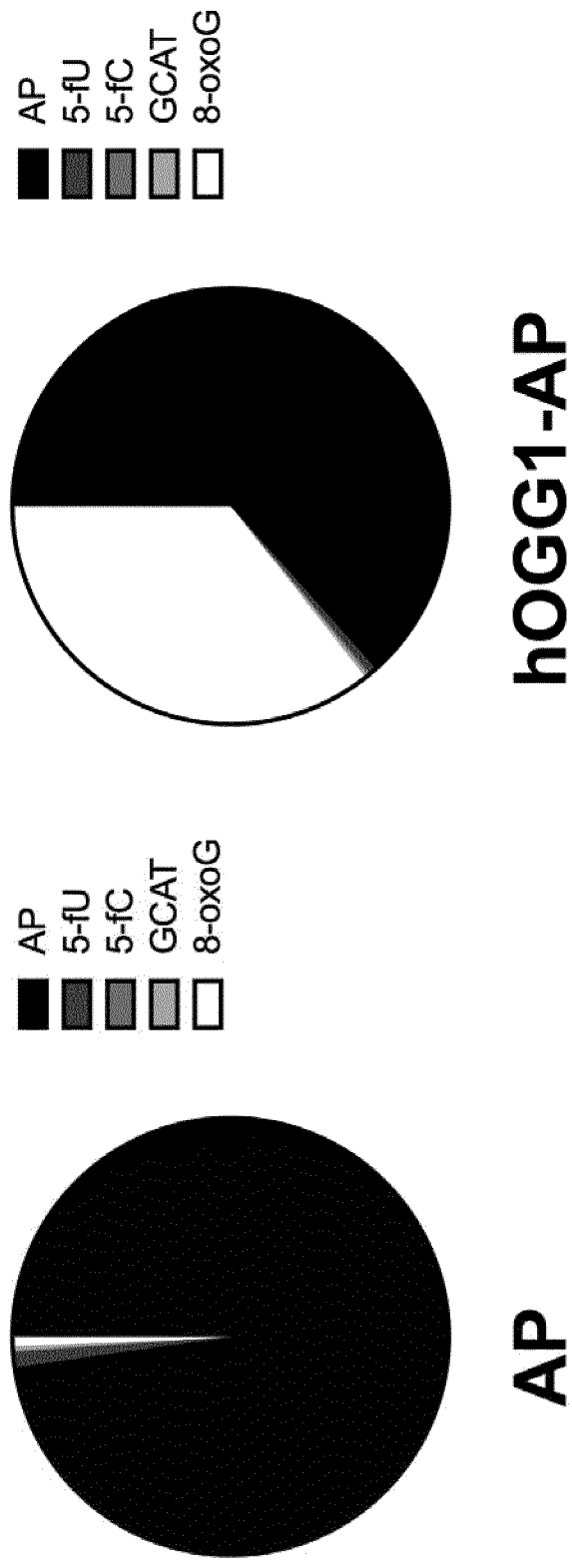
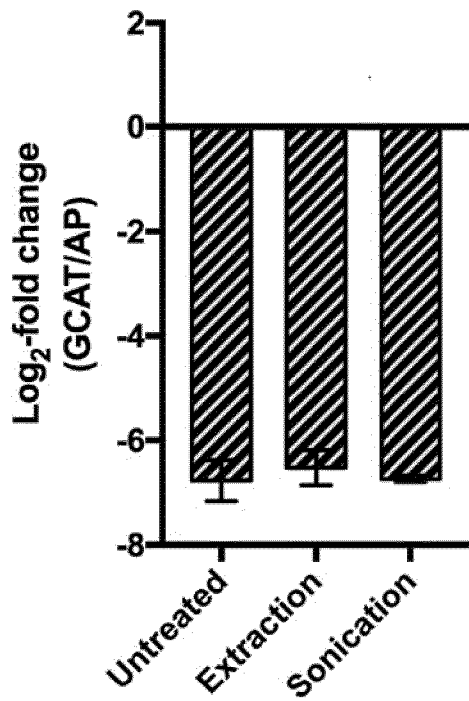


Figure 13

(A)



(B)

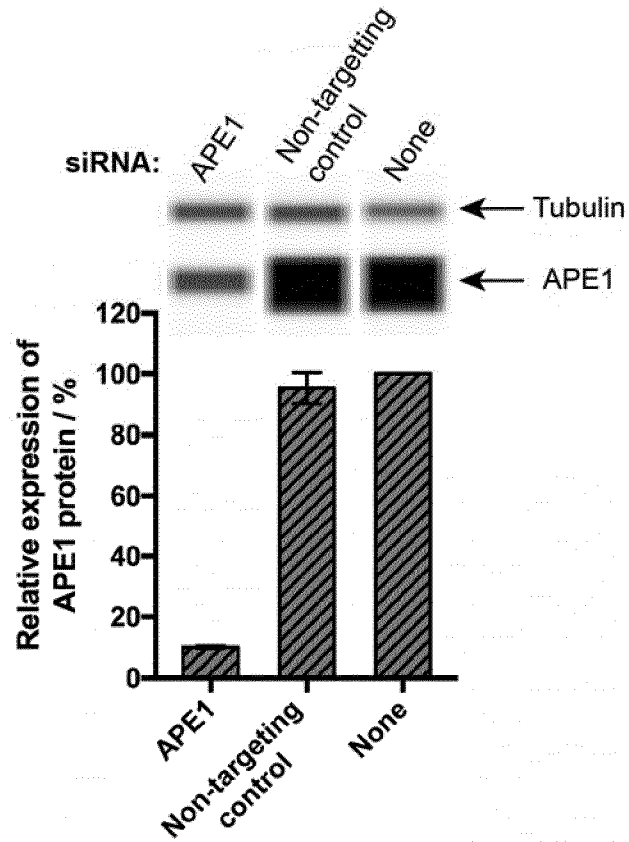
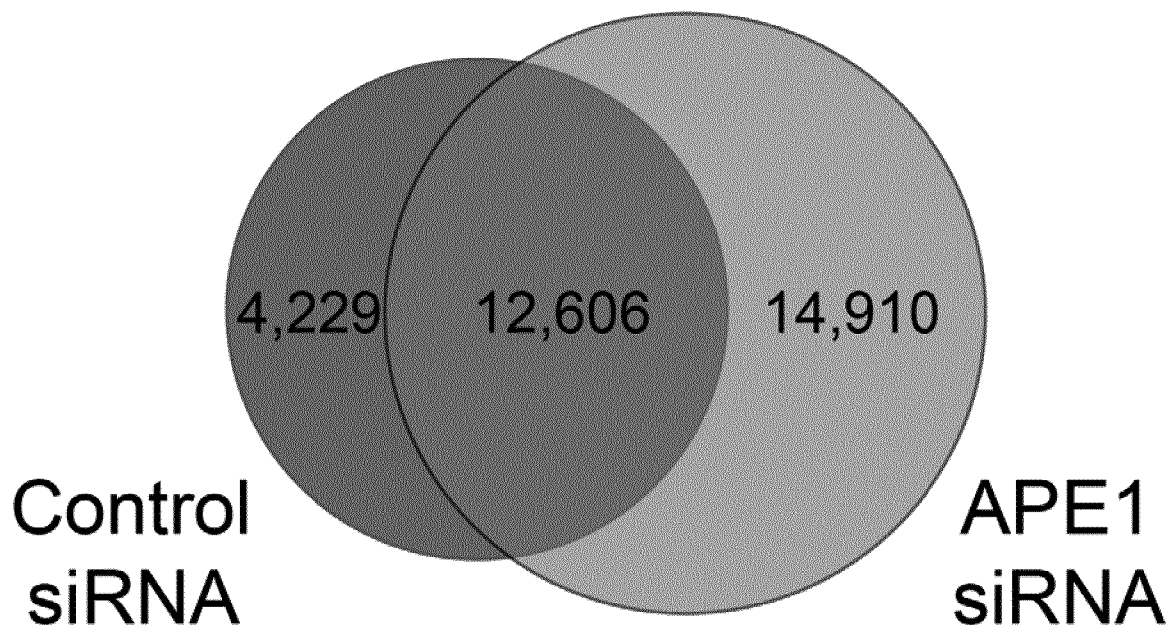


Figure 13 cont.

(C)



(D)

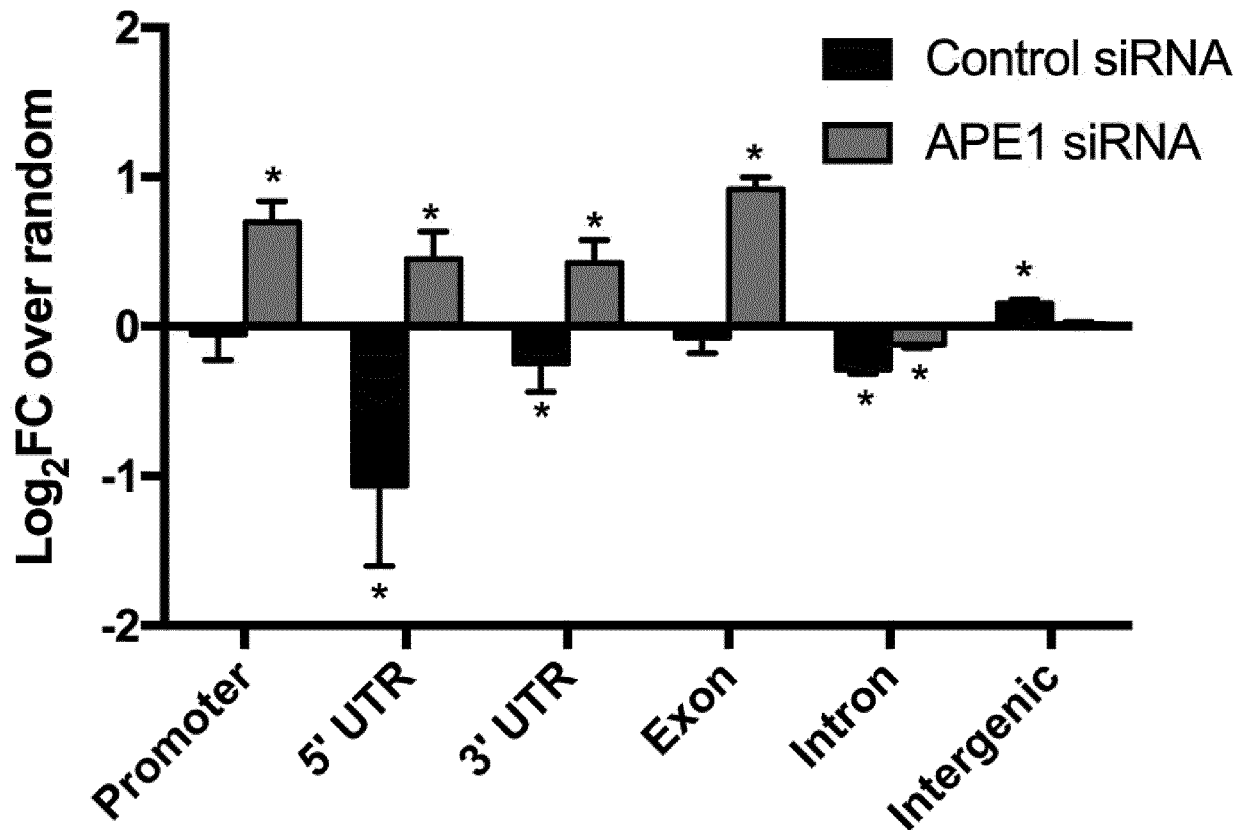
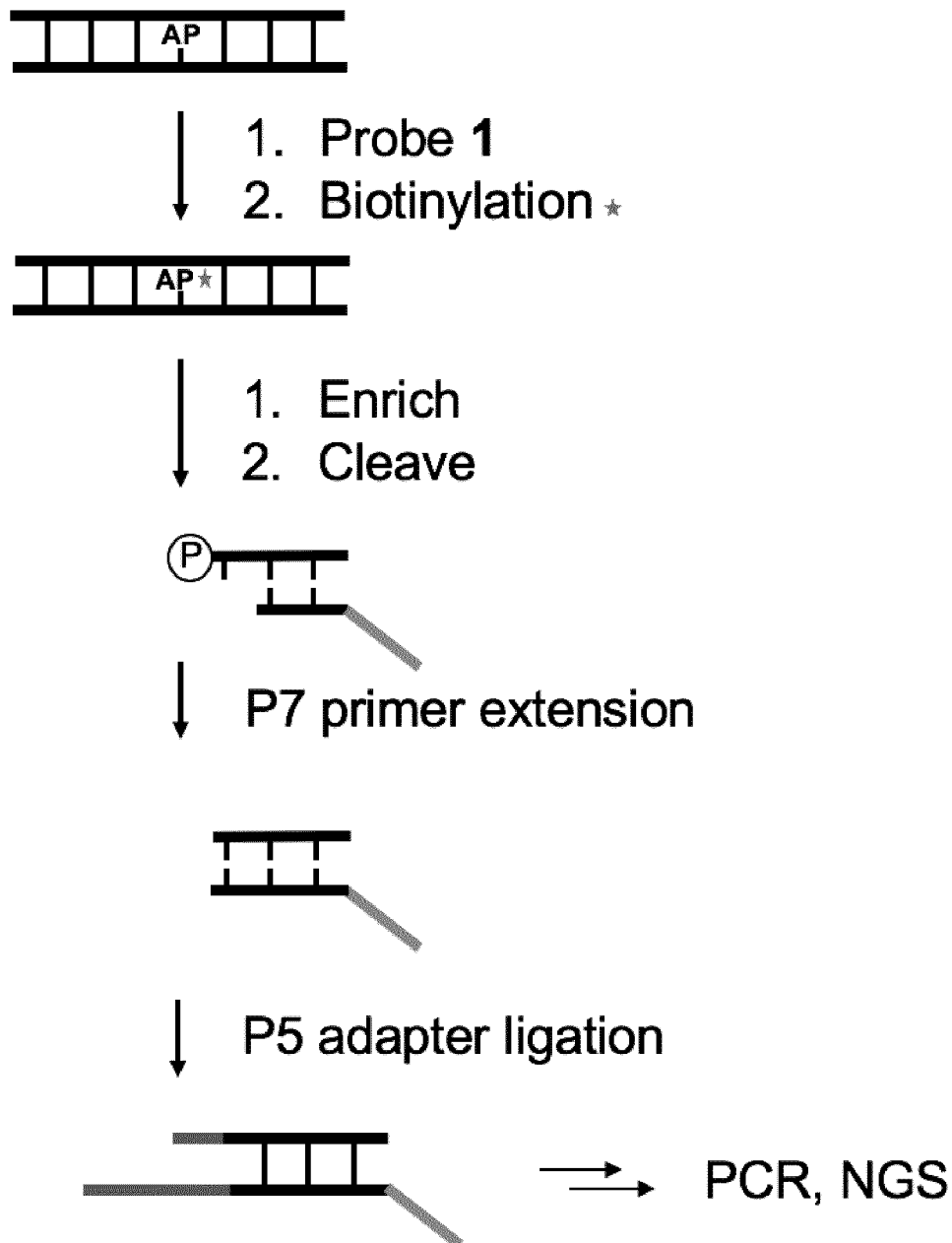


Figure 14



INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2019/070255

A. CLASSIFICATION OF SUBJECT MATTER
INV. C07D401/00 C12Q1/6806
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C07D C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data, BIOSIS, CHEM ABS Data, Sequence Search, EMBASE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MAKRIGIORGOS G M ET AL: "FLUORESCENT LABELLING OF ABASIC SITES: A NOVEL METHODOLOGY TO DETECT CLOSELY-SPACED DAMAGE SITES IN DNA", INTERNATIONAL JOURNAL OF RADIATION BIOLOGY, TAYLOR & FRANCIS, GB, vol. 74, no. 1, 1 January 1998 (1998-01-01), pages 99-109, XP002925462, ISSN: 0955-3002, DOI: 10.1080/095530098141762 page 101, figure 1A; page 101 item 2.5; page 103, item 3.4 and figure 5 ----- -/--	1-3,5, 8-10,16, 19,20, 25,26, 30,32, 33,63, 73-75



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

25 October 2019

Date of mailing of the international search report

06/11/2019

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Hennard, Christophe

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2019/070255

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	NAOSHI KOJIMA ET AL: "Construction of Highly Reactive Probes for Abasic Site Detection by Introduction of an Aromatic and a Guanidine Residue into an Aminooxy Group", JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, vol. 131, no. 37, 23 September 2009 (2009-09-23), pages 13208-13209, XP055373076, US ISSN: 0002-7863, DOI: 10.1021/ja904767k	1,2, 5-10,12, 15-19, 21-26, 30,31, 35-38, 73-75
Y	figure 1; page 13209 -----	39-56
X	WO 2016/168467 A2 (UNIV CASE WESTERN RESERVE [US]) 20 October 2016 (2016-10-20) claims; [0064]; [0085] -----	1,2, 5-11,13, 14,16, 19,20, 25,26, 30,38
X	US 2004/005614 A1 (KURN NURITH [US] ET AL) 8 January 2004 (2004-01-08) claims 1, 9-13, 31; examples; figure 3 -----	3-8, 28-34, 73-75
X	US 6 174 680 B1 (MAKRIGIORGOS GERASSIMOS M [US]) 16 January 2001 (2001-01-16)	73-75
Y	claims; examples 3 and 4; column 20, lines 16-34 -----	39-56
A	BRETT GYARFAS ET AL: "Mapping the Position of DNA Polymerase-Bound DNA Templates in a Nanopore at 5 ? Resolution", ACS NANO, vol. 3, no. 6, 23 June 2009 (2009-06-23), pages 1457-1466, XP055148440, ISSN: 1936-0851, DOI: 10.1021/nn900303g abstract -----	1-75

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2019/070255

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2016168467 A2	20-10-2016	US 2018127830 A1 WO 2016168467 A2	10-05-2018 20-10-2016
US 2004005614 A1	08-01-2004	AU 2003279697 A1 CA 2486283 A1 EP 1573056 A2 JP 4551216 B2 JP 2005534304 A JP 2010011872 A US 2004005614 A1 WO 2004011665 A2 ZA 200409152 B	16-02-2004 05-02-2004 14-09-2005 22-09-2010 17-11-2005 21-01-2010 08-01-2004 05-02-2004 28-11-2007
US 6174680 B1	16-01-2001	NONE	