## Supplementary information

**Sequencing abasic sites in DNA at single-nucleotide resolution**
Zheng J. Liu, Sergio Martínez Cuesta, Pieter van Delft and Shankar Balasubramanian

### Table of contents

| Section                                                                 | Page |
|-------------------------------------------------------------------------|------|
| General remarks                                                         | 2    |
| Experimental methods                                                    | 2    |
| Data access and analysis                                                | 5    |
| Supplementary Scheme 1: Synthesis of HIPS probe 1                        | 7    |
| Synthesis and characterisation of HIPS probe 1                          | 8    |
| Supplementary Table 1: Sequence of ODNs                                | 11   |
| Supplementary Table 2: Calculated and recorded masses of ODNs by LC-MS | 14   |
| Supplementary Table 3: Read counts of synthetic DNA                     | 15   |
| Supplementary Table 4: Sequencing library information                   | 15   |
| Supplementary Table 5: Encode and GEO datasets used                     | 16   |
| Supplementary Figure 1-7: LC-MS spectra of ODNs and reactions           | 17   |
| Supplementary Figure 8: qPCR standard curves                           | 22   |
| Supplementary Figure 9: Enrichment of AP DNA relative to GCAT DNA       | 23   |
| Supplementary Figure 10: Read alignments of control input libraries    | 23   |
| Supplementary Figure 11: Effects of DNA sonication and extraction       | 24   |
| Supplementary Figure 12: Limit of detection of snAP-seq                 | 25   |
| Supplementary Figure 13: Example of Tapestation analysis                | 26   |
| Supplementary Figure 14: Assessment of SMUG1 excision bias             | 27   |
| Supplementary Figure 15: Detecting individual sites in snAP-seq        | 28   |
| Supplementary Figure 16: Peak calling in snAP-seq data                 | 29   |
| Supplementary Figure 17: Distribution of snAP-seq HeLa peaks           | 29   |
| Supplementary Figure 18-29: $^1$H and $^{13}$C NMR spectra of compounds | 30   |
| Supplementary Figure 30-31: $^1$H and $^{13}$C NMR spectra of HIPS probe | 36   |
| References                                                              | 37   |
General remarks

All solvents and reagents were used as supplied from commercial sources (Sigma Aldrich unless stated otherwise). LC-MS was performed on an Amazon ESI-MS (Bruker) connected to a Dionex UltiMate 3000 UHPLC system (Thermo Fisher Scientific). Data was analysed using Bruker Compass DataAnalysis 4.2. Flash chromatography was carried out using CombiFlash Rf (Teledyne Isco) with puriFlash columns (Interchim). NMR spectra were recorded in CDCl\textsubscript{3} on a Bruker 400 MHz Avance III HD Spectrometer or a 500 MHz DCH Cryoprobe Spectrometer. Collected NMR spectra were processed using MestReNova software. Accurate mass spectra were recorded on a Waters LCT Premier (ESI) spectrometer. DNA concentration was measured using a Qubit 2.0 fluorometer. ODNs were purchased from either Invitrogen or Sigma with HPLC (up to 30 nucleotides) or PAGE (30-105 nucleotides) purification unless stated otherwise.

Experimental methods

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

Generation of double-stranded AP DNA, 5-fU DNA, 5-fC DNA, 5-hmU DNA and GCAT DNA: Reaction volumes (20 μL) contained template ODN (2.5 μg), primer (10 μM), either dGTP, dCTP, dATP, dfUTP (200 μM) for 5-fU DNA, dGTP, dfCTP, dATP, dTTP (200 μM) for 5-fC DNA, dGTP, dCTP, dATP, dhmUTP (200 μM) for 5-hmU DNA, or dNTPs (200 μM) for AP DNA and GCAT DNA, DreamTaq Buffer (2 μL) and DreamTaq polymerase (0.4 μL, 2 U). Samples were heated to 95°C for 30 s, then annealed at the stated temperatures (Table S1) for 60 s then held at 72°C for 10 min, before purification with a GeneJET PCR purification kit according to the manufacturer’s protocol.

LC-MS analysis of short ODNs: LC-MS was performed on a Bruker amaZon system using an XTerra MS C18 column (2.5 μM, 2.1 x 50 mm). Separation was carried out using 5-30 % solvent A (100 mM 1,1,1,3,3,3-hexafluoro-2-propanol, 10 mM triethylamine) in solvent B (methanol) over 25 mins, at a flow rate of 0.2 mL/min. % conversions were calculated by integrating UV signals of starting material and products at 260 nm. The identity of each species was confirmed by ESI-MS.

Sodium hydroxide cleavage: Sodium hydroxide (100 mM) was added to purified ODNs in 50 μL reactions and incubated at 70°C for 15 min. Reactions were immediately quenched with either Tris-HCl (pH 7.0, 5 μL, 1 M) or hydrochloric acid (5 μL, 1 M) and purified with mini quick-spin columns (Roche).
**DNA enrichment for qPCR studies:** Enrichments were based on a reported protocol with modifications. MagneSphere streptavidin magnetic beads (50 μg, Promega), were washed with 1 × binding buffer (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20) (3 × 500 μL) and resuspended in 50 μL 2 × binding buffer (10 mM Tris pH 7.5, 1 mM EDTA, 2 M NaCl, 0.1% Tween 20). Input DNA (1 ng/ODN) and poly dI:dC (2 μg, Thermo Scientific) were mixed and made up to a final volume of 50 μL, and then added to the magnetic beads, before incubation for 15 minutes at room temperature with gentle rotation. Beads were washed with 1X binding buffer (6 × 500 μL), then incubated with NaOH (100 μL, 100 mM) at room temperature for 10 mins. The beads were washed again with NaOH (100 μL, 100 mM) followed by 1X binding buffer (3 × 500 μL) then eluted in NaOH (50 μL, 100 mM) at 70 °C for 15 mins and quenched immediately with Tris-HCl (25 μL, 500 mM, pH 7.0). A fresh sample of pre-washed streptavidin beads (75 μg) was incubated with poly dI:dC (2 μg) and resuspended in 2X binding buffer (75 μL), to which the neutralized DNA eluent was added. The sample was incubated at room temperature for a further 15 mins, before separating from the beads. The recovered DNA was purified using a ssDNA/RNA clean and concentrator (Zymo Research) according to the manufacturer’s instructions and eluted in water (25 μL).

**qPCR quantification of enrichment:** qPCRs were performed using a CFX96 Real-Time System (BioRad), and data was processed using CFX software manager 3.1 (BioRad). Reaction volumes (10 μL) contained enriched DNA (1 μL), Brilliant III ultra-Fast SYBR green qPCR master mix (5 μL, Agilent Technologies), and the corresponding forward and reverse primers (1 μM each). The mixture was subject to qPCR according to the protocol outlined by the manufacturer. The extent of DNA amplification was compared to that of input samples. Primers were designed 3’- to modifications so that any possible strand cleavage would not affect amplification.

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

**SMUG1-snAP-seq of Leishmania major DNA:** Leishmania major DNA (product 30012D, lot 62762024) was purchased from ATCC. DNA was purified before use with a DNeasy blood and tissue kit (Qiagen) according to the manufacturer’s instructions, with the exception that the protein digestion step was omitted. The purified DNA was eluted in Tris-HCl (50 μL, 10 mM, pH 7.4). For SMUG1-AP-seq, DNA (400 ng) was sonicated to an average of 450 bp using a Covaris M220 system, and treated with hSMUG1 (25 U, NEB) in NEBuffer 1 supplemented with BSA (100 μg/mL) at 37°C for 18 h before purification using AMPure XP beads (2.0 X), or a DNA clean & concentrator kit (Zymo Research). Samples were then subjected to snAP-seq, see methods section for detailed protocol.

**UNG-snAP-seq of Leishmania major DNA:** All steps were carried out as described above for SMUG1-snAP-seq, with the exception that SMUG1 enzyme was replaced by UNG (10 U, NEB) and BSA was omitted.
Cell culture: HeLa cells were cultured in DMEM (Gibco, 41965039) supplemented with 10% heat-inactivated fetal bovine serum (Thermo Fisher). Cells were short tandem repeat (STR) genotyped and mycoplasma tested.

siRNA knockdown of APE1: siRNAs were purchased from Dharmaco (ON-TARGETplus human APE1 siRNA, product code J-010237-08-0002 and ON-TARGETplus non-targeting pool, product code D-001810-10-05) and resuspended in water to give 20 μM stocks. For transfections, HeLa cells were seeded in 6-well plates at 100,000-200,000 cells/well and grown overnight. On the second day, cells were transfected with 10 nM of siRNA or control siRNA using Lipofectamine RNAiMAX according to the manufacturer’s instructions. After 96 hr cells were washed once in cold PBS and harvested by scraping into cold PBS (500 μL) followed by centrifugation at 300 × g.

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

DNA isolation for snAP-seq: Cells were harvested as described above and collected as a pellet. DNA was extracted using a Quick-DNA kit (Zymo Research) according to the manufacturer’s guidelines, and genomic lysis buffer was supplemented with TEMPO (20 mM) immediately before each extraction. Purified DNA was eluted in Tris-HCl (10 mM, pH 7.4) and quantified using a Qubit 2.0 fluorometer. DNA (up to 5 μg) was sonicated to an average length of 450 bp using a Covaris M220 system, and processed for snAP-seq. For detailed protocol, see methods section.
Data access and analysis:
Data was deposited in the ArrayExpress database under accession number E-MTAB-7152. The computational code is available in the manuscript's GitHub page (https://github.com/sblab-bioinformatics/snAP-seq).

Sequencing data processing: The quality of raw sequencing reads was evaluated using FastQC v0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality bases were filtered and Illumina TruSeq adapters were trimmed from the read's 3' end using cutadapt v1.12. No reads smaller than 15bp were kept (after adapters and base quality trimming). In the genomic libraries (Leishmania major and HeLa), trimmed reads containing the P7 adapter sequence (GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T) at the 5' end were discarded.

Alignment: bwa v0.7.15-r11404 was used to prepare reference sequences (spike-ins and genomes), which were then used to align trimmed sequencing reads using bwa mem. References from the Sanger Institute and UCSC hg38 were used for Leishmania major and HeLa, respectively. Resulting alignments were cleaned, merged, sorted and indexed using samtools v1.3.1. Duplicate reads in genomic libraries were marked using sambamba v0.6.5. The filtering of alignments involved the removal of unaligned reads, secondary/alternative alignments, PCR duplicates, alignments with a quality score of less than 10, reads overlapping blacklisted regions (HeLa, http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38-human/) and spike-ins. Clean alignments were further split into alignments of R1 reads mapping to the forward and to the reverse strands respectively. These were then converted to tdf format using igvtools v2.3.917. The bamCoverage function available in deeptools v2.4.2-5-f439d22 was used to perform sequencing coverage calculation normalised by RPKM at single-nucleotide resolution (after merging biological replicates).

ODN and spike-in analysis: distribution of reads aligning to reference oligonucleotide sequences and patterns of alignment start sites were obtained using samtools view and standard Unix tools. For each library, the number of reads aligned to each ODN was first normalized to the corresponding input library. Only the reverse strand was used here for normalization due to a large underrepresentation of AP site-containing strands in input libraries (see Supplementary Fig. 9).

Synthetic 5-hmU N-oligo analysis: trimmed reads were deduplicated based on the two randomized N10 stretches flanking the 5-hmU site. Fasta files and tables of nucleotide counts were generated with in-house Python v2.7.12 scripts. Sequence logo representations were generated with the ggseqlogo v0.1 R library (https://cran.rstudio.com/web/packages/ggseqlogo/index.html). Fisher's Exact and Pearson's Chi-squared tests were used to compare nucleotide counts as implemented in the R v3.3.2 programming language.

Calling single-nucleotide AP sites: bedtools v2.27.0 genomecov9 was used to obtain the sequencing coverage of 5' positions for R1 reads aligning to the forward and reverse strands
respectively. Genome-wide modelling and comparative assessment of counts from SMUG1-snAP-seq, UNG-snAP-seq and input libraries (Leishmania major) were performed using negative binomial generalized linear models as implemented in edgeR v3.16.5. High-confidence SMUG1-snAP sites were obtained when selecting single nucleotides with positive log₂ fold-change (SMUG1-snAP-seq vs. input) at an FDR threshold smaller than 10⁻¹⁰. Volcano plots were generated using ggplot2 v2.2.1 (https://cran.r-project.org/web/packages/ggplot2/citation.html). No significant sites were found when comparing UNG-snAP-seq against input libraries in Leishmania major or snAP-seq against input libraries in HeLa cells.

Coverage profiles, base composition, sequence logos, motif analyses and TG enrichment: intersections with 5-hmU and base J peaks obtained in Kawasaki et al., 2017 were performed with bedtools intersect. Fasta files containing sequences flanking the SMUG1-snAP-seq sites by 2, 5 and 10bp were generated using bedtools getfasta. As a control, similar fasta files were obtained for T bases genome-wide and located within the 5-hmU and base J peaks obtained previously. Base composition plots were produced with ggplot2 and coverage profile plots were obtained using the computeMatrix and plotProfile functions available in deeptools. The ggseqlogo library in R was used to generate sequence logo representations and counting of dinucleotides genome-wide and within 5-hmU and base J peaks as well as frequency comparisons were obtained with the compseq tool in EMBOSS v6.6.0.0. The dreme tool as available in meme v4.11.2 was used to extract sequence motifs. Tests and simulations for TG enrichment over TX (X being A, C, G) in the context of SMUG1-snAP-seq sites, 5-hmU and base J peaks, genome-wide and synthetic 5-hmU N-oligo libraries were scripted in Python and visualized with R.

Peak calling: macs2 v2.1.1.20160309 callpeak was used to obtain regions of enriched signal of (SMUG1-)snAP-seq using input libraries as control, with options \( p < 0.00001 \), and --nomodel. To further correct for regions with naturally high coverage of reads, macs2 callpeak with option --nomodel was also used in the input libraries only. Overlaps between the enriched and naturally high regions were obtained using bedtools intersect and subtracted from the regions of enriched signal. Consensus peak regions were obtained between two intersecting (Leishmania major) and three out of four (HeLa) replicates. For visualization of enrichment within a given genomic region, raw read counts within the selected region were normalized by dividing by the total number of reads in the region \( \times 1,000 \).

Testing genomic associations: hg38 gene feature annotations (promoters, 5'UTR, 3'UTR, exons, introns and intergenic regions were extracted from the UCSC’s genes.gtf file using the library GenomicFeatures in R and searching for regular expressions using Python (https://github.com/dariober/bioinformatics-cafe/tree/master/fastaRegexFinder). Computing the significance of overlap between the genomic annotations and AP peaks was performed with the Genomic Association Tester (GAT).
Synthesis and characterisation of HIPS probe 1

Supplementary Scheme 1: Synthesis of HIPS probe 1.
**Ethyl 1-propargylindole-2-carboxylate (3)**

Sodium hydride (60 w% dispersion in oil, 0.315 g, 7.95 mmol) was dissolved in dry dimethylformamide (38 mL). A solution of ethyl indole-2-carboxylate 2 (1.50 g, 7.95 mmol) in dimethylformamide (4.5 mL) was then added dropwise at 0 °C. After stirring for 30 minutes at 0 °C, propargyl bromide (80 w% in toluene, 1.35 mL, 11.9 mmol, 1.5 equiv) was added dropwise and the resulting brown solution stirred for a further 4 hours at 0 °C. Ammonium chloride (sat. solution, 35 mL) was added to quench the reaction, and the mixture poured into brine (35 mL). The mixture was extracted with ethyl acetate (3 × 75 mL), and the combined organic fractions was washed with brine (25 mL), dried over sodium sulfate and concentrated in vacuo. The crude product was purified by column chromatography (0-40 % ethyl acetate in hexane) to give a white solid (1.39 g, 77 %). 1H NMR (400 MHz, CDCl₃) δ 7.72 (d, J = 7.9 Hz, 1H), 7.53 (d, J = 8.4 Hz, 1H), 7.42 (dd, J = 8.4, 7.0 Hz, 1H), 7.38 (s, 1H), 7.21 (dd, J = 7.9, 7.0 Hz, 1H), 4.58 (d, J = 2.5 Hz, 2H), 4.43 (q, J = 7.1 Hz, 2H), 2.28 (t, J = 2.5 Hz, 1H), 1.45 (t, J = 7.1 Hz, 3H) ppm. 13C NMR (101 MHz, CDCl₃) δ 162.0, 138.9, 126.9, 126.2, 125.4, 122.8, 121.1, 111.4, 110.5, 78.7, 72.0, 60.8, 33.9, 14.3 ppm. HRMS (ESI-TOF) calcd for C₁₄H₁₄NO₂ [M+H]+: 228.1025; found: 228.1022.

**[1-(2-Propynyl)-1H-indol-2-yl]methanol (4)**

Lithium aluminium hydride (1.0 M solution, 6.6 mL, 6.6 mmol, 1.2 equiv) was added dropwise to a solution of ethyl 1-propargylindole-2-carboxylate 3 (1.25 g, 5.5 mmol) in diethyl ether (20 mL) at 0 °C. After stirring at room temperature for 2 hours, the mixture was added to a solution of ethyl acetate (70 mL) and water (35 mL), and the layers separated. The organic layer was washed with water (2 × 30 mL), aqueous sodium hydroxide (1M, 20 mL), and brine (20 mL) and dried over sodium sulfate and then concentrated in vacuo. The crude product was purified by column chromatography (0-40 % ethyl acetate in hexane) to give a white solid (0.79 g, 78 %). 1H NMR (400 MHz, CDCl₃) δ 7.62 (d, J = 7.9, Hz, 1H), 7.45 (d, J = 8.2 Hz, 1H), 7.30 (t, J = 8.2 Hz, 1H), 7.16 (t, J = 7.9 Hz, 1H), 6.51 (s, 1H), 5.04 (d, J = 2.5 Hz, 2H), 4.90 (d, J = 6.0 Hz, 2H), 2.31 (t, J = 2.5 Hz, 1H), 1.72 (t, J = 6.0 Hz, 1H), 1.59 (s, 1H) ppm. 13C NMR (101 MHz, CDCl₃) δ 137.7, 137.3, 127.5, 122.5, 121.1, 120.2, 109.4, 102.7, 78.5, 72.4, 57.5, 32.9 ppm. HRMS (ESI-TOF) calcd for C₁₂H₁₂NO [M+H]+: 186.0919; found: 186.0921.

**1-(Prop-2-yn-1-yl)-1H-indole-2-carbaldehyde (5)**

Dess-Martin periodinane (1.87 g, 4.4 mmol) was dissolved in a mixture of pyridine (1 mL) and dichloromethane (8 mL). After stirring for 5 minutes at room temperature, the solution was transferred to a solution of [1-(2-propynyl)-1H-indol-2-yl]methanol 4 (0.75 g, 4.0 mmol) in dichloromethane (4 mL) and the solution stirred for a further 3 hours. The reaction was then quenched by the addition of sodium thiosulphate (10 % aqueous solution, 4 mL) and sodium bicarbonate (sat. solution, 4 mL). The aqueous layer was extracted with
dichloromethane (3 × 30 mL) and the combined organic phases dried over sodium sulfate and concentrated in vacuo. The crude product was purified by column chromatography (0-25 % ethyl acetate in hexane) to give a white solid (0.63 g, 85 %). $^1$H NMR (400 MHz, CDCl$_3$) δ 9.92 (s, 1H), 7.79 (d, $J = 8.1$ Hz, 1H), 7.57 (d, $J = 8.5$ Hz, 1H), 7.50 (dd, $J = 8.5, 6.9$ Hz, 1H), 7.32 (s, 1H), 7.25 (dd, $J = 8.1, 6.9$ Hz, 1H), 5.49 (d, $J = 2.5$ Hz, 2H), 2.29 (t, $J = 2.5$ Hz, 1H) ppm. $^{13}$C NMR (101 MHz, CDCl$_3$) δ 182.7, 140.1, 134.5, 127.4, 126.7, 123.6, 121.5, 118.7, 110.8, 78.2, 72.5, 33.9 ppm. HRMS (ESI-TOF) calcd for C$_{12}$H$_{10}$NO [M+H]+: 184.0762; found: 184.0758.

1-((9H-fluoren-9-yl)methyl) 2-(tert-butyl) hydrazine-1,2-dicarboxylate (6)

To a solution of 1-Boc-1-methylhydrazine (1 mL, 0.985 g, 6.74 mmol) in THF (5 mL) and water (5 mL) was added sodium bicarbonate (1.13 g, 13.5 mmol, 2 equiv) with rapid stirring. A solution of Fmoc-chloride (1.74 g, 6.74 mmol) in THF (5 mL) was then added dropwise, and the reaction mixture stirred at room temperature for a further 1 hr. Ether (10 mL) was added, the organic layer was washed with brine (15 mL), dried over sodium sulfate and concentrated in vacuo to give a yellow oil (1.64 g, 4.46 mmol, 66 %). $^1$H NMR (400 MHz, CDCl$_3$) δ 7.78 (d, $J = 7.5$ Hz, 2H), 7.62 (d, $J = 7.4$ Hz, 2H), 7.43 (t, $J = 7.4$ Hz, 2H), 7.33 (t, $J = 7.5$, 2H), 4.49 (d, $J = 7.0$ Hz, 2H), 4.27 (t, $J = 7.0$ Hz, 1H), 3.16 (br s, 3H), 1.49 (s, 9H) ppm. $^{13}$C NMR (101 MHz, CDCl$_3$) δ 143.6, 141.3, 127.8, 127.1, 125.1, 120.0, 68.0, 65.9, 47.1, 28.2, 25.6 ppm. HRMS (ESI-TOF) calcd for C$_{21}$H$_{25}$N$_2$O$_4$ [M+H]+: 369.1809; found: 369.1819.

(9H-fluoren-9-yl)methyl 2-methylhydrazine-1-carboxylate (7)

1-((9H-fluoren-9-yl)methyl) 2-(tert-butyl) hydrazine-1,2-dicarboxylate 6 (1.2 g, 3.26 mmol) was dissolved in DCM (6 mL) and TFA (2 mL) and stirred at room temperature for 2 h. The solvent was then removed in vacuo, and the resulting oil was dissolved in ethyl acetate (15 mL). Saturated sodium bicarbonate (10 mL) was added, and the precipitate formed was collected, redissolved in DCM (30 mL), washed with saturated sodium bicarbonate (15 mL), dried over sodium sulfate and concentrated in vacuo to give a white solid (0.651 g, 2.43 mmol, 75 %). $^1$H NMR (400 MHz, CDCl$_3$) δ 7.86 – 7.72 (m, 2H), 7.61 (d, $J = 7.4$ Hz, 2H), 7.51 – 7.39 (m, 2H), 7.34 (td, $J = 7.4, 1.2$ Hz, 2H), 4.48 (d, $J = 6.7$ Hz, 2H), 4.26 (t, $J = 6.7$ Hz, 1H), 2.67 (s, 3H) ppm. $^{13}$C NMR (101 MHz, CDCl$_3$) δ 157.1, 143.7, 141.3, 127.8, 127.1, 125.0, 120.0, 67.0, 47.2, 39.3 ppm. HRMS (ESI-TOF) calcd for C$_{16}$H$_{17}$N$_2$O$_2$ [M+H]+: 269.1285; found: 269.1380.
(9H-fluoren-9-yl)methyl 2-methyl-2-(((1-prop-2-yln-1-yl)-1H-indol-2-yl)methyl)hydrazine-1-carboxylate (8)

1-(Prop-2-yln-1-yl)-1H-indole-2-carbaldehyde 5 (200 mg, 1.1 mmol) and (9H-fluoren-9-yl)methyl 2-methylhydrazine-1-carboxylate 7 (351 mg, 1.68 mmol, 1.5 equiv) was dissolved in dichloromethane (15 mL). Sodium triacetoxyborohydride (462 mg, 2.2 mmol, 2 equiv) was then added, and the resulting suspension stirred for 16 h at room temperature under argon before quenching with sodium bicarbonate (sat. aqueous solution, 10 mL). The organic layer was separated and the aqueous layer extracted with dichloromethane (5 x 10 mL). The pooled extracts were dried over sodium sulfate and concentrated in vacuo. The crude product was purified by column chromatography (0-40 % ethyl acetate in hexane) to give a colourless oil (359 mg, 76 %). \(^1\)H NMR (500 MHz, CDCl\(_3\)) \(\delta\) 7.77 (t, \(J = 7.0\) Hz, 2H), 7.64 – 7.46 (m, 3H), 7.46 – 7.34 (m, 3H), 7.34 – 7.26 (m, 2H), 7.26 – 7.21 (m, 1H), 7.11 (dd, \(J = 7.9, 7.0, 1\)H), 6.39 (s, 1H), 5.88 (s, 1H), 5.20 (br s, 2H), 4.43 (br s, 2H), 4.15 (br s, 2H), 2.67 (s, 3H), 2.22 (br s, 1H) ppm. \(^{13}\)C NMR (126 MHz, CDCl\(_3\)) \(\delta\) 155.1, 143.7, 141.3, 137.3, 133.9, 127.7, 127.5, 127.1, 125.1, 125.0, 122.1, 120.7, 120.0, 109.4, 104.2, 79.0, 72.0, 66.6, 47.2, 44.3, 32.9, 28.2 ppm. HRMS (ESI-TOF) calcd for C\(_{26}\)H\(_{27}\)N\(_3\)O\(_2\) [M+H]^+: 436.2020; found: 436.2031.

2-((1-methylhydrazinyl)methyl)-1-(prop-2-yln-1-yl)-1H-indole (1)

(9H-Fluoren-9-yl)methyl 2-methyl-2-(((1-prop-2-yln-1-yl)-1H-indol-2-yl)methyl)hydrazine-1-carboxylate 8 (180 mg, 0.41 mmol) was dissolved in a solution of piperidine (0.65 mL, 6.7 mmol, 16 equiv) in \(N,N\)-dimethylformamide (2.5 mL) and stirred at room temperature for 30 min. The reaction mixture was diluted with ethyl acetate (20 mL), washed with brine (4 x 8 mL), dried over sodium sulphate and concentrated in vacuo. The crude product was purified by column chromatography (0-20 % methanol in dichloromethane) to give a pale yellow oil (47 mg, 54 %). Probe 1 was stored neat at -80 °C until further use. \(^1\)H NMR (400 MHz, CDCl\(_3\)) \(\delta\) 7.58 (d, \(J = 7.9\) Hz, 1H), 7.42 (d, \(J = 8.2\) Hz, 1H), 7.24 (m, 1H), 7.12 (m, 1H), 6.42 (s, 1H), 5.12 (d, \(J = 2.5\) Hz, 2H), 3.84 (s, 2H), 2.53 (s, 3H), 2.25 (t, \(J = 2.5\) Hz, 1H) ppm. \(^{13}\)C NMR (126 MHz, CDCl\(_3\)) \(\delta\) 137.3, 135.0, 127.7, 122.0, 120.6, 120.0, 109.4, 104.0, 77.5, 72.1, 60.1, 48.6, 32.9 ppm. HRMS (ESI-TOF) calcd for C\(_{13}\)H\(_{16}\)N\(_3\) [M+H]^+: 214.1339; found: 214.1342.
| ODN          | Sequence 5′-3′                                          |
|--------------|--------------------------------------------------------|
| U-ODN        | AGC GAC AUA TCT TGT                                     |
| AP-ODN       | AGC GAC A-AP-A TCT TGT                                  |
| 5-fU-ODN     | ATC GCA 5-fUGT A                                       |
| 5-fC-ODN     | ATC G5-fCG CGT A                                       |
| 5-hmU-ODN    | p-ATC GCA 5-hmUGT A                                    |
| Base J-ODN   | GAA C_jG JCG GAG                                       |
| AP DNA template (forward) | CAC ACC GCC AGC CAC AGC AAC GAA CGAP GCA GCG CCC CTC ACG CCA CAG AAC ATC GCA TTT ACG ACG ATT GAT GTA CTA AAT AGT GGG TGG TCG GTT CGC G |
| AP DNA primer | CGC GAA CCG ACC ACC ACC CAC TA (T<sub>anneal</sub> = 60 °C) |
| AP DNA (reverse) | CGC GAA CCG ACC ACC CAC TAT TTA GTA CAT CAA TCG TCG TAA ATG CGA TGT TCT GTG GCG TGA GGG GCG CTG CAC GTT CGT TGC TGT GGC TGG CGG TGT G |
| 5-fU DNA template (reverse) | CTG TTC GCT GGT CTG CTT GCC TGG TTC TCT TGT CTT GCT TGG TTC TCT TGT GTG CTT GGC TTT CCT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A |
| 5-fU DNA primer | TAG CCA TAC TGC CTC GTC CG (T<sub>anneal</sub> = 68 °C) |
| 5-fU DNA (forward) | TAG CCA TAC TGC CTC GTC CGG ACA CG5-fU CAG GAG GAA AGC CAA GAC ACA CAA AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G |
| 5-fC DNA template (reverse) | CCT CAC TCA CCT CCA CCC TCT CAC TAC CTC ACT CTT CCT CCT CTT AAT CAA CTC CCT CAAA CCT CTC CAC CCT CTC ACA AGA TCT GAC TCA GCG TCG GTT CGC A |
| 5-fC DNA primer | TCG CAC ACG CTC AGT CAG GT (T<sub>anneal</sub> = 68 °C) |
| 5-fC DNA (forward) | TCG CAC ACG CTC AGT CAG GTA GAG AT5-fC TAG GAG GGT GGA GAG GTG GTT GGA GAG GGT TAG GAG GAA GAG TGA GTG AGT GAG GAG GAG G |

**ODN**

**Sequence 5′-3′**

- **U-ODN**: AGC GAC AUA TCT TGT
- **AP-ODN**: AGC GAC A-AP-A TCT TGT
- **5-fU-ODN**: ATC GCA 5-fUGT A
- **5-fC-ODN**: ATC G5-fCG CGT A
- **5-hmU-ODN**: p-ATC GCA 5-hmUGT A
- **Base J-ODN**: GAA C_jG JCG GAG

**AP DNA template (forward)**

CAC ACC GCC AGC CAC AGC AAC GAA CGAP GCA GCG CCC CTC ACG CCA CAG AAC ATC GCA TTT ACG ACG ATT GAT GTA CTA AAT AGT GGG TGG TCG GTT CGC G

**AP DNA primer**

CGC GAA CCG ACC ACC ACC CAC TA (T<sub>anneal</sub> = 60 °C)

**AP DNA (reverse)**

CGC GAA CCG ACC ACC CAC TAT TTA GTA CAT CAA TCG TCG TAA ATG CGA TGT TCT GTG GCG TGA GGG GCG CTG CAC GTT CGT TGC TGT GGC TGG CGG TGT G

**5-fU DNA template (reverse)**

CTG TTC GCT GGT CTG CTT GCC TGG TTC TCT TGT CTT GCT TGG TTC TCT TGT GTG CTT GGC TTT CCT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A

**5-fU DNA primer**

TAG CCA TAC TGC CTC GTC CG (T<sub>anneal</sub> = 68 °C)

**5-fU DNA (forward)**

TAG CCA TAC TGC CTC GTC CGG ACA CG5-fU CAG GAG GAA AGC CAA GAC ACA CAA AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G

**5-fC DNA template (reverse)**

CCT CAC TCA CCT CCA CCC TCT CAC TAC CTC ACT CTT CCT CCT CTT AAT CAA CTC CCT CAAA CCT CTC CAC CCT CTC ACA AGA TCT GAC TCA GCG TCG GTT CGC A

**5-fC DNA primer**

TCG CAC ACG CTC AGT CAG GT (T<sub>anneal</sub> = 68 °C)

**5-fC DNA (forward)**

TCG CAC ACG CTC AGT CAG GTA GAG AT5-fC TAG GAG GGT GGA GAG GTG GTT GGA GAG GGT TAG GAG GAA GAG TGA GTG AGT GAG GAG GAG GAG G

**ODN**

**Sequence 5′-3′**
| GCAT DNA template (forward) | GGC CAC CAC CCG CAC ATA CTC TGG TAC GAT TAC GAA CAC AGC CCG ACA CCA CCT CTA ATG AAC GTC GCT TAT AGT GAT TAA CGC CCC GTA GAC ACC ATG G |
|----------------------------|------------------------------------------------------------------------------|
| GCAT DNA primer            | CCA TGG TGT CTA CGG GGC GT \( T_{\text{anneal}} = 60 \, ^\circ \text{C} \)       |
| GCAT DNA (reverse)         | CCA TGG TGT CTA CGG GGC GTT AAT CAC TAT AAG CGA CGT TCA TTA GAG GTG GTG TCG GGC TGT GTT CGT AAT CGT ACC AGA GTA TGT GCG GGT GGT GGC C |
| 5-hmU DNA template (reverse) | CTG TTC GCT GGT CTG CTT GTG CTC TTC TGG CTG TGG TCG TCT TGG TTC TGT CTT GCC TTT CTT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A |
| 5-hmU DNA primer           | TAG CCA TAC TGC CTC GTC CG \( T_{\text{anneal}} = 65 \, ^\circ \text{C} \)       |
| 5-hmU DNA (forward)        | TAG CCA TAC TGC CTC GTC CGG ACA CGA CGA5-hmU CAG GAG GAA AGC CAA GAC ACA CGA ACC AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G       |
| AP DNA2 template (reverse) | CTG TTC GCT GGT CTG CTT GTG CTC TTC TGG CTG TGG TCG TCT TGG TTC TGT CTT GCC TTT CTT CCT GAC GTG TCC GGA CGA GGC AGT ATG GCT A |
| AP DNA2 primer             | TAG CCA TAC TGC CTC GTC CG \( T_{\text{anneal}} = 65 \, ^\circ \text{C} \)       |
| AP DNA2 (forward)          | TAG CCA TAC TGC CTC GTC CGG ACA CGAAP CAG GAG GAA AGC CAA GAC ACA CGA ACC AAG AGA ACC AAG CAA GAC AGA AGA GCA CAA GCA GAC CAG CGA ACA G       |
| P7 adapter (top) (also P7 primer) | Me-GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T |
| P7 adapter (bottom)        | p-GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC-NNNNNN-ATC TCG TAT GCC GTC TTC TGG TTG-spacerC3 |
| P5 adapter (top)           | GAA TGA TAC GGC GAC CCA GAT CTA CAC TCT TTC CCT ACA CGA CGC TCT TCC GAT CT |
### Supplementary Table 1: Sequence of ODNs.

| **P5 adapter (bottom)** | p-GAT CGG AAG AGC G |
|-------------------------|---------------------|
| **Library amplification primers** | Forward: AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GA  |
|  | Reverse: CAA GCA GAA GAC GGC ATA CGA GAT |
| **AP DNA qPCR primers** | Forward: GCC CCT CAC GCC ACA GAA CA  |
|  | Reverse: CGC GAA CCG ACC ACC CAC TA |
| **GCAT DNA qPCR primers** | Forward: GGC CAC CAC CCG CAC ATA CT  |
|  | Reverse: CCA TGG TGT CTA CGG GGC GT |
| **5-fU DNA qPCR primers** | Forward: AAG CAA GGA AGC AAG GCA GA  |
|  | Reverse: TGC TTG GCT GCG TGG TCT CG |
| **5-fC DNA qPCR primers** | Forward: GTA TGG AGG AAT GAG TGT GG  |
|  | Reverse: TAA CTA CCT ATC TAC CAT TC |
| **5-hmU N-oligo template** | GTC TAC CTG AAC GCC GCT GTN NNN NNN NNN 5-hmU NNN NNN NNN NNG TAG TAG TCG ACT AGA CG TCC AAC CAA CGG AAG GGT ATT CGG ACG AGG CAG TAT GGC TA  |
|  | where N indicates randomized bases |
| **5-hmU N-oligo primer** | TAG CCA TAC TGC CTC GTC CG (T<sub>anneal</sub> = 65 °C) |

AP DNA and GCAT DNA templates were purchased from IDT with PAGE purification. 5-hmU DNA template was purchased from Biomers with PAGE purification. 5-hmU N-oligo was purchased from ATDBio with HPLC purification. 5-fU-ODN was synthesized using a protected 5-formyldeoxyuridine phosphoramidite. 5-fC-ODN was purchased from Eurogentec with HPLC purification. All adapters with purchased with double HPLC purification from ATDBio.
| ODN                        | Calculated MW | Calculated ion | ESI-MS found |
|----------------------------|---------------|----------------|--------------|
| U-ODN                      | 4554          | M<sup>3</sup> = 1517 | M<sup>3</sup> = 1516.34 |
| AP-ODN                     | 4459          | M<sup>3</sup> = 1485 | M<sup>3</sup> = 1485.07 |
| AP-ODN + HIPS 1            | 4654          | M<sup>3</sup> = 1550 | M<sup>3</sup> = 1550.04 |
| AP-ODN + HIPS 1 + Biotin PEG3 azide | 5099          | M<sup>3</sup> = 1699 | M<sup>3</sup> = 1697 | M<sup>3</sup> = 1696.99 |
|                           | 5095 (oxidized)| M<sup>3</sup> = 1697 |             |
| AP-ODN β-δ-elimination products | 2165          | M<sup>2</sup> = 1082 | M<sup>2</sup> = 1082.09 |
|                           | 2194          | M<sup>2</sup> = 1096 | M<sup>2</sup> = 1096.07 |
| 5-fU-ODN                   | 3041          | M<sup>2</sup> = 1520 | M<sup>2</sup> = 1519.08 |
| 5-fU-ODN + HIPS 1          | 3236          | M<sup>2</sup> = 1617 | M<sup>2</sup> = 1616.74 |
| 5-fU-ODN + HIPS 1 + Biotin PEG3 azide | 3681          | M<sup>2</sup> = 1840 | M<sup>2</sup> = 1836.92 |
|                           | 3677 (oxidized)| M<sup>2</sup> = 1837 |     |
| 5-fC-ODN                   | 3056          | M<sup>2</sup> = 1527 | M<sup>2</sup> = 1526.69 |
| 5-fC-ODN + HIPS 1          | 3251          | M<sup>2</sup> = 1625 | M<sup>2</sup> = 1624.23 |
| 5-hmU-ODN                  | 3122          | M<sup>2</sup> = 1560 | M<sup>2</sup> = 1560.27 |
| 5-hmU-ODN AP               | 2998          | M<sup>2</sup> = 1498 | M<sup>2</sup> = 1498.33 |
| 5-hmU-ODN + HIPS 1         | 3193          | M<sup>2</sup> = 1596 | M<sup>2</sup> = 1595.89 |
| 5-hmU-ODN + HIPS 1 + Biotin PEG3 azide | 3638          | M<sup>2</sup> = 1819 | M<sup>2</sup> = 1816.02 |
|                           | 3634 (oxidized)| M<sup>2</sup> = 1816 |     |
| 5-hmU-ODN β-δ-elimination product | 1935          | M<sup>1</sup> = 1934 | M<sup>1</sup> = 1934.17 |
| Base J-ODN                 | 4220          | M<sup>3</sup> = 1406 | M<sup>2</sup> = 1405.44 (1) 1405.45 (1-CuAAC) 1405.63 (SMUG1) |

**Supplementary Table 2:** Calculated and recorded masses of short ODNs and reaction products. All ESI-MS were recorded by LC-MS as described in the methods. Biotin-PEG3-azide was introduced by copper catalysis, and both the expected click product, along with an oxidized derivative in which the equivalent of 4 hydrogen atoms are lost are listed. For all three ODNs, the oxidized adduct is found which supports the proposed aromatized structure.
| Library name | DNA source | Reads raw (R1+R2) | Reads aligned (R1+R2) | Enrichment over GCAT DNA |
|--------------|------------|-------------------|----------------------|--------------------------|
| Lmaj_SMUG1_snAP1 | L. major | 45586764 | 18577161 | 122.23 (5-hmU) |
| Lmaj_snAP1 | L. major | 38802616 | 26893969 | 176.61 (AP) |
| Lmaj_input1 | L. major | 37599314 | 20222512 | - |
| Lmaj_SNUG1_snAP2 | L. major | 42681890 | 18523918 | 73.18 (5-hmU) |
| Lmaj_snAP2 | L. major | 56334634 | 25532776 | 101.64 (AP) |
| Lmaj_input2 | L. major | 47286768 | 24184938 | - |
| Lmaj_UNG_snAP1 | L. major | 56617526 | 31364208 | 94.61 (U) |
| Lmaj_UNG_input1 | L. major | 60013790 | 34250759 | - |
| Lmaj_UNG_snAP2 | L. major | 60705944 | 36435348 | 71.77 (U) |
| Lmaj_UNG_input2 | L. major | 47980640 | 24705855 | - |
| APE1-siRNA_snAP1 | HeLa APE1 KD | 356138764 | 209655010 | 260.28 (AP) |
| APE1-siRNA_input1 | HeLa APE1 KD | 366216496 | 271034459 | - |
| APE1-siRNA_snAP2 | HeLa APE1 KD | 440297144 | 213587504 | 257.42 (AP) |
| APE1-siRNA_input2 | HeLa APE1 KD | 283444538 | 221451304 | - |
| APE1-siRNA_snAP3 | HeLa APE1 KD | 488731524 | 320012574 | 332.31 (AP) |
| APE1-siRNA_input3 | HeLa APE1 KD | 43713922 | 329110219 | - |
| APE1-siRNA_snAP4 | HeLa APE1 KD | 440323246 | 273585368 | 283.07 (AP) |
| APE1-siRNA_input4 | HeLa APE1 KD | 456173156 | 341493560 | - |
| Cont-siRNA_snAP1 | HeLa control | 467195776 | 255519837 | 256.02 (AP) |
| Cont-siRNA_input1 | HeLa control | 451662634 | 342185708 | - |
| Cont-siRNA_snAP2 | HeLa control | 487547412 | 325686623 | 226.61 (AP) |
| Cont-siRNA_input2 | HeLa control | 386161754 | 286991318 | - |
| Cont-siRNA_snAP3 | HeLa control | 375936312 | 175522647 | 314.61 (AP) |
| Cont-siRNA_input3 | HeLa control | 376448196 | 273634277 | - |
| Cont-siRNA_snAP4 | HeLa control | 501690906 | 232636858 | 259.13 (AP) |
| Cont-siRNA_input4 | HeLa control | 368563444 | 273634277 | - |

**Supplementary Table 3:** Read counts of synthetic DNA. Raw number of reads aligning to each synthetic DNA sequence after NGS are shown. For further analysis, the raw number of reads were normalized to ensure equal amounts of input DNA, by dividing each value by the corresponding number of reads in the ‘synthetic_input_untreated, reverse’ library. Percentages after normalization are shown in brackets, and were used to generate all pie charts shown.

**Supplementary Table 4:** Sequencing library information for all genomic libraries created in this study.
| Genomic assay  | Accession code |
|---------------|----------------|
| DNase-seq     | ENCF950NDW     |
| FAIRE-seq     | ENCF001UYM     |
| ATAC-seq      | GSM2830381     |
| H3K27ac       | ENCF392EDT     |
| H3K4me3       | ENCF862LUQ     |
| H3K27me3      | ENCF512TQI     |
| H3K9me3       | ENCF712ATO     |

Supplementary Table 5: Encode and GEO datasets used to analyze HeLa snAP-seq peaks.
Supplementary Figure 1: LC-MS UV trace of ODNs before and after reaction with 1 (10 mM) at pH 7.4 for 2 hr at RT.
**Supplementary Figure 2:** LC-MS UV trace of U-ODN. From top to bottom: untreated, after UNG treatment, after UNG treatment followed by reaction with 1, after UNG treatment followed by reaction with 1 and biotinylation.

**Supplementary Figure 3:** LC-MS UV trace of AP-ODN. From top to bottom: alkaline cleavage of unfunctionalized AP-ODN at 70 °C, alkaline cleavage of 1-biotinylated AP-ODN at room temperature, alkaline cleavage of 1-biotinylated AP-ODN at 70 °C.
Supplementary Figure 4: LC-MS UV trace of 5-fU-ODN. From top to bottom: untreated, reaction with 1 after 2 hr, reaction with 1 for 2 hr at RT followed by biotinylation, alkaline cleavage of 1-biotinylated 5-fU-ODN at 70 °C. No elimination products are observed after alkaline cleavage.

Supplementary Figure 5: LC-MS UV trace of 5-fC-ODN. From top to bottom: untreated, reaction with 1 after 2 hr, reaction with 1 for 24 hr at 37 °C followed by biotinylation, alkaline cleavage of 1-biotinylated 5-fC-ODN at 70 °C. No elimination products are observed after alkaline cleavage.
Supplementary Figure 6: LC-MS UV trace of 5-hmU-ODN. From top to bottom: untreated, after treatment with SMUG1 (25 U, NEB) in NEBuffer 1 supplemented with BSA (100 μg/mL) at 37°C for 18 h, after SMUG1 treatment followed by reaction with 1, after SMUG1 treatment followed by reaction with 1 and biotinylation, alkaline cleavage of 1-biotinylated 5-hmU-ODN at 70°C, alkaline cleavage of SMUG1 treated 5-hmU-ODN at 70°C.
Supplementary Figure 7: LC-MS UV trace of Base J-ODN. Top: reaction with 1 after 2 hr, reaction with 1 for 2 hr at RT followed by biotinylation, treatment with SMUG1 (25 U, NEB) in NEBuffer 1 supplemented with BSA (100 μg/mL) at 37 °C for 18 hr. M^3 values correspond to that calculated for the starting ODN in all three cases (see supplementary table 2), small shifts in retention time are due to technical differences between runs.
Supplementary Figure 8: qPCR standard curves. For each primer pair (see Supplementary Table 1 for DNA and primer sequences), the efficiency of amplification is within ±10% of 100%.
Supplementary Figure 9: Enrichment of AP DNA relative to GCAT DNA. All samples were incubated with pre-washed streptavidin beads and then washed with 1X binding buffer (6 × 500 μL). For direct elution of duplex DNA, beads were heated in elution buffer (95 % formamide, 10 mM EDTA) for 5 min. For NaOH wash, complementary DNA strands were first denatured in NaOH (100 mM, 10 min at RT), then the remaining bound DNA was eluted in elution buffer as above. For alkaline-cleavage elution, complementary strands were denatured in NaOH (100 mM, 10 min at RT), and the remaining DNA was eluted in fresh NaOH (100 mM, 15 min at 70°C). Fold enrichment was calculated by comparison of qPCR C_{t} values of each sample relative to input. Mean and S.E.M of triplicate experiments are plotted.

Supplementary Figure 10: Read alignments of control input libraries. Number of sequencing reads after Illumina sequencing aligned to the forward strand of ds-ODNs using TruSeq adapters (Illumina) without enrichment (see 'input library' protocol) are shown. Forward is defined here as the strand bearing the modification (see Supplementary Table 1). a) without HIPS treatment, b) with HIPS treatment and biotinylation. In both cases AP DNA is underrepresented. Reads were normalized to the unmodified reverse strand, to ensure equal input of each ds-ODN duplex.
**Supplementary Figure 11:** Effects of DNA sonication and extraction on snAP-seq enrichment.  
**a)** % recovery of AP and GCAT DNA after streptavidin enrichment. DNA was quantified by qPCR and represented as a percentage of the input DNA. No statistical difference was found in recovery of either AP or GCAT DNA before and after additional sonication and extraction (Kruskal-Wallis test, $p = 0.2964$ (AP) and $p = 0.3607$ (GCAT)). DNA was recovered by alkaline-cleavage (100 mM NaOH, 70 °C) after streptavidin enrichment, and qPCR primers were designed 3’- to the AP site (see Supplementary Table 1 for details).  
**b)** SMUG1-snAP-seq sites detected in *L. major* DNA. Control refers to the 3,200 high-confidence sites detected, where purified *L. major* DNA was sonicated, then SMUG1 treatment is carried out. Here, AP sites were tagged immediately after SMUG1 generation and do not undergo further DNA manipulation. To assess the effects of DNA sonication, high-molecular-weight *L. major* DNA was first treated with SMUG1, then sonicated after the generation of AP sites. 82 % of the 3,200 sites remain detectable after sonication of generated AP sites. To assess the combined effects of DNA extraction and sonication, high-molecular-weight *L. major* DNA was treated with SMUG1, then the resultant AP sites in the high-molecular-weight DNA was subjected to a mock re-extraction and DNA sonication. 67% of the 3,200 sites remain detectable after this treatment. Whilst a proportion of sites are lost during sonication and extraction, the majority remain intact.
Supplementary Figure 12: Limit of detection of snAP-seq. Equal amounts of AP DNA1 (AP DNA in Supplementary Table 1), GCAT DNA, and AP DNA2 with varying % modification, were subjected to snAP-seq. To obtain AP DNA2 with varying modification percentage, AP DNA2 was diluted with DNA of the same sequence where the AP site is replaced by T, where the % modification refers to the AP/T ratio. The raw read counts aligning to the forward strand of each model DNA was normalized to the reverse (unmodified) strand in the corresponding input library, and the enrichment of AP DNA1 and AP DNA2 relative to GCAT DNA is shown. The modification % of AP DNA1 is constant at 100 %, and variations in enrichment of this sequence reflect technical variation in enrichment efficiency. Enrichment of AP DNA2 is observed down to 1 % modification, where an 8-fold enrichment relative to GCAT DNA is found. Sequence information is shown in Supplementary Table 1.
Supplementary Figure 13: Example of Tapestation analysis of DNA samples during library preparation in snAP-seq. Left panel: genomic DNA after sonication (450 bp setting using a Covaris M220 machine). Middle panel: during snAP-seq protocol, after P7 adapter ligation (lanes A1, B1) and after further phosphatase treatment (lanes C1, D1). Right panel: Libraries after PCR-amplification for snAP-seq (lanes A1, B1) and input libraries (C1, D1). The bands between 100 and 200 bp in the final snAP-seq libraries are possibly due to partial adapter dimers, but do not appear to interfere with further sequencing. D1000 screentape and corresponding sample buffer (Agilent) were used for all analyses, except for the visualization of final libraries where high-sensitivity D1000 screentape and buffer were used.
Supplementary Figure 14: Assessment of SMUG1 excision bias. a) Experimental design using synthetic 5-hmU N-oligo. A 100mer ODN containing two randomized N10 stretches flanking a 5-hmU site was used. N-oligo was treated with SMUG1 and labelled with 1 and biotinylated. For ‘input’ library, Truseq adapters were introduced. For ‘snAP-seq’, the standard protocol is followed. As snAP-seq only captures fragments 3’- to the 5-hmU site and therefore any information in the N10 5’- to the 5-hmU site is lost, an additional ‘enriched’ library was prepared. Here, labelled DNA underwent Truseq adapter ligation and the DNA was captured on streptavidin beads followed by washing with 1X binding buffer (6 x 500 μL). The enriched complement was recovered by denaturation (100 mM NaOH, 15 min, RT) and the supernatant was quenched (500 mM Tris-HCl pH 7.0). Libraries from all three conditions were then sequenced. b) Sequence logo of randomized region after Input, Enriched and snAP-seq treatment. To rule out effects of PCR amplification bias, only sequencing reads with non-redundant N10 sequences were used for analysis. A small enrichment for Gp5-hmUpG is observed. c) Comparison of 5-hmUpG enrichment in SMUG1-snAP-seq sites identified in L. major, with technical bias detected from N-oligo experiment. As 5-hmUpG was the strongest enriched motif found after SMUG1-snAP-seq of L. major DNA, this was compared to possible experimental bias caused by SMUG1-snAP-seq. Enrichment is still found to be significant (*p < 0.05) when compared to the extent of 5-hmUpG enrichment of the N-oligo.
Supplementary Figure 15: Detecting individual sites in snAP-seq. Volcano plots of sites detected by calling individual single-nucleotide AP sites on a) two replicates of SMUG1-snAP-seq of *L. major* DNA, b) two replicates of UNG-snAP-seq of *L. major* DNA. Any sites detected with negative log2FC are considered to be false positives detected in the input sample. Above an FDR threshold of $10^{-10}$ in a), no sites are observed for the *L. major* samples and therefore only sites called below this threshold were considered high-confidence for further analysis.

Unlike SMUG1, UNG displays very high specificity for uracil and is not able to excise 5-substituted uracil derivatives such as 5-hmU\textsuperscript{18,19}. No sites were detected after UNG-snAP-seq (FDR < $10^{-10}$), suggesting that the sites identified by SMUG1-snAP-seq are specific to oxidized thymine derivatives. Volcano plot of sites detected in human DNA on c) four replicates of snAP-seq on APE1 siRNA treated HeLa DNA, d) four replicates of snAP-seq on control siRNA treated HeLa DNA and e) four replicates of snAP-seq on APE1 siRNA treated HeLa DNA, analyzed for mitochondrial DNA only and f) four replicates of snAP-seq on control siRNA treated HeLa DNA, analyzed for mitochondrial DNA only. For the HeLa samples in b) and c), no FDR threshold can be set which favours positive over negative log-fold enrichment, indicating that no confident sites can be called. Only reads aligning to the forward reference strand are shown.
Supplementary Figure 16: Peak calling in snAP-seq data. **a, b)** Correlation of normalized read counts in two replicates of *L. major* SMUG1-snAP-seq libraries at the 3,200 detected SMUG1-snAP-seq sites, in the forward and reverse strands of the reference genome respectively. **c)** Overlap of 5-hmU pulldown peaks\(^\text{11}\) and consensus SMUG1-snAP-seq peaks in *L. major* called by MACS2 software\(^\text{14}\) (\(p < 10^{-5}\)). **d)** Overlap of the union of 5-hmU and Base J peaks\(^\text{11}\), and consensus SMUG1-snAP-seq peaks in *L. major* called by MACS2 software (\(p < 10^{-5}\)). The strong overlap between SMUG1-snAP peaks called by MACS2 here with the reported 5-hmU and Base J peaks supports that peak-calling based analysis is also applicable for snAP-seq data, in addition to calling single-nucleotide AP sites. **e)** Overlap of SMUG1-snAP peaks called by MACS2 (\(p < 10^{-5}\)) in two replicates of *L. major*. No consensus peaks were detected in libraries in which SMUG1 treatment was omitted. **f)** Overlap of snAP-seq peaks in four control siRNA-treated HeLa libraries. **g)** Overlap of snAP-seq peaks in four APE1 siRNA-treated HeLa libraries.

Supplementary Figure 17: Distribution of snAP-seq HeLa peaks relative to **a)** chromatin accessibility and **b)** histone modifications expressed as \(\log_2\) (fold change) when compared to randomized sets of peaks obtained through simulation (\(N = 10,000\)). Error bars represent 95% confidence intervals. *\(q < 0.05\)*. See Supplementary Table 5 for details of datasets used.
Supplementary Figure 18: $^1$H NMR spectrum of 3.

Supplementary Figure 19: $^{13}$C NMR spectrum of 3.
Supplementary Figure 20: $^1$H NMR spectrum of 4.

Supplementary Figure 21: $^{13}$C NMR spectrum of 4.
Supplementary Figure 22: $^1$H NMR spectrum of 5.

Supplementary Figure 23: $^{13}$C NMR spectrum of 5.
Supplementary Figure 24: $^1$H NMR spectrum of 6.

Supplementary Figure 25: $^{13}$C NMR spectrum of 6.
Supplementary Figure 26: $^1$H NMR spectrum of 7.

Supplementary Figure 27: $^{13}$C NMR spectrum of 7.
Supplementary Figure 28: $^1$H NMR spectrum of 8.

Supplementary Figure 29: $^{13}$C NMR spectrum of 8.
Supplementary Figure 30: $^1$H NMR spectrum of HIPS probe 1.

Supplementary Figure 31: $^{13}$C NMR spectrum of HIPS probe 1.
References

1. Raiber, E.-A. et al. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* 13, R69 (2012).

2. Nakamura, J., La, D. K. & Swenberg, J. A. 5′-Nicked Apurinic/Apyrimidinic Sites Are Resistant to β-Elimination by β-Polymerase and Are Persistent in Human Cultured Cells after Oxidative Stress. *J. Biol. Chem.* 275, 5323–5328 (2000).

3. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10–12 (2011).

4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

5. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

6. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034 (2015).

7. Thorvalsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192 (2013).

8. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44, W160–W165 (2016).

9. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

10. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140 (2010).

11. Kawasaki, F. et al. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biol.* 18, 23 (2017).
12. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

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

14. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

15. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).

16. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046–2048 (2013).

17. Hardisty, R. E., Kawasaki, F., Sahakyan, A. B. & Balasubramanian, S. Selective Chemical Labeling of Natural T Modifications in DNA. *J. Am. Chem. Soc.* **137**, 9270–9272 (2015).

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

19. Moréra, S. *et al.* Biochemical and structural characterization of the glycosylase domain of MBD4 bound to thymine and 5-hydroxymethyluracil-containing DNA. *Nucleic Acids Res.* **40**, 9917–9926 (2012).