Published in final edited form as:
Nat Chem. 2019 June 06; 11(7): 629–637. doi:10.1038/s41557-019-0279-9.

Sequencing abasic sites in DNA at single-nucleotide resolution

Zheng J. Liu[1], **Sergio Martínez Cuesta**[1,2], **Pieter van Delft**[1], and **Shankar Balasubramanian**[1,2,3,*]

[1]Department of Chemistry, University of Cambridge, Cambridge, UK

[2]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

[3]School of Clinical Medicine, University of Cambridge, UK

Abstract

In DNA, the loss of a nucleobase by hydrolysis generates an abasic site. Formed as a result of DNA damage, as well as a key intermediate during the base excision repair pathway, abasic sites are frequent DNA lesions which can lead to mutations and strand breaks. Here we present snAP-seq, a chemical approach that selectively exploits the reactive aldehyde moiety at abasic sites to reveal their location within DNA at single-nucleotide resolution. Importantly, the approach resolves abasic sites from other aldehyde functionalities known to exist in genomic DNA. snAP-seq was validated on synthetic DNA, then applied to two separate genomes. We studied the distribution of thymine modifications in the *Leishmania major* genome by enzymatically converting these modifications into abasic sites followed by abasic site mapping. We also apply snAP-seq directly to HeLa DNA to provide a map of endogenous abasic sites in human genomic DNA.

Cellular DNA is constantly subject to both endogenous and exogenous sources of damage. Abasic sites, also known as AP sites, are one of the most common forms of damage products. The spontaneous hydrolysis of nucleobases generates an estimated 10,000 AP sites per cell, per day1. AP site levels can be further elevated in the presence of exogenous damage agents, such as reactive oxygen species (ROS) or alkylating agents2–4.

AP sites are also generated enzymatically *in vivo* as an essential intermediate within the base excision repair (BER) pathway5. DNA base lesions such as 8-oxoguanine and uracil are

[*]Correspondence should be addressed to S.B. (sb10031@cam.ac.uk).

Data availability
Sequencing data are available in the ArrayExpress database under accession number E-MTAB-7152.

Code availability
We have released all the computational code in the accompanying manuscript's GitHub page (https://github.com/sblab-bioinformatics/snAP-seq).

Author contribution
Z. J. L., P. V. D. and S. B. designed the study. Z. J. L. performed the experiments. S. M. C. performed the computational analysis. All authors analysed and interpreted the data. Z. J. L. and S. B. wrote the manuscript, with contribution from all authors.

Competing interests statement
S. B. is a founder, adviser and shareholder of Cambridge Epigenetix Ltd.

normally efficiently excised by glycosylase enzymes to generate an AP site, which is further processed by the BER pathway[6,7]. More recently, BER and abasic sites have also been suggested to play a role in the maintenance of epigenetic DNA base modifications. Two pathways proposed to remove the DNA epigenetic marker 5-methylcytosine (5-mC), active demethylation[8,9] and processive demethylation[10], rely on separate glycosylase enzymes to convert key DNA bases at or around 5-mC to abasic sites. Subsequent repair of the AP site by the BER pathway is suggested to restore unmodified cytosine into the genome. Together, these pathways highlight the further importance of BER beyond the repair of DNA damage.

Abasic sites can be generated from DNA base modifications *in vitro* using enzyme chemistry. A variety of glycosylase enzymes have been identified, which excise a single or set of structurally related modified nucleobases to form an AP site[11]. These glycosylases can therefore be used *in vitro* to convert DNA modifications into highly reactive AP sites that can potentially be subsequently detected. The exceptional specificity offered by glycosylases, even when a substrate is greatly outnumbered by the canonical DNA bases, is particularly useful when studying rare modifications which may otherwise be difficult to target selectively. This approach has been used to quantify uracil levels in genomic DNA[12], and to enrich for DNA fragments containing uracil and 8-oxoguanine to facilitate the mapping of these modifications by sequencing[13,14].

When AP sites are not successfully repaired and instead persist in the genome, these lesions can lead to the stalling of polymerases and DNA strand breaks, as well as mutations due to the misincorporation of bases opposite the non-coding AP site[15,16]. Mutations in APE1, the primary endonuclease responsible for initiating AP site repair in mammals, have been associated with an increased risk of cancer and altered prognosis[17,18]. Despite the severe consequences of persistent AP sites, there is still a lack of understanding regarding their formation, persistence and location within genomic DNA. Fluorescent immunostaining of abasic sites in DNA fibers has suggested that AP sites formed in the presence of exogenous damage can be clustered and non-random[19,20], however little is known about where abasic sites form in the genome and the sequence contexts in which they occur.

Methods to detect global levels of abasic sites have been described which utilize *O*-(biotinylcarbazoylmethyl) hydroxylamine, also known as ‘aldehyde-reactive probe’ (ARP). In the ring-open form, the aldehyde group of AP sites can react with hydroxylamines such as ARP. The amount of ARP bound to DNA, and therefore number of reactive sites, can then be quantified by detection of the biotin moiety within ARP using dot-blot or ELISA assays[21,22]. A major drawback to using hydroxylamine condensation for the labelling of abasic sites is that other reactive aldehydes which have been identified in DNA can result in cross-reactivity[23]. In particular, formylated-bases, 5-formylcytosine (5-fC) and 5-formyluracil (5-fU) have been shown to react with ARP[24,25]. Quantitative mass spectrometry measurements have revealed that the natural abundance of these bases can actually be higher than that of AP sites, with around 10 and 2.5 per 10^6 bases for 5-fC and 5-fU respectively measured in mouse embryonic stem cells, compared to 0.9 AP sites[26–28]. Thus, there remains a chemoselectivity challenge to discern abasic sites in the presence of these naturally occurring bases, in order to correctly analyze them in genomic DNA.

Herein, we describe a method for the selective detection of DNA abasic sites in the presence of formylated-bases 5-fC and 5-fU, as well as canonical G, C, A and T bases. We demonstrate that the approach can provide a high-resolution map of AP sites within DNA that comprises abasic sites or formylated-bases with high chemical selectivity. We extend this method to investigate the distribution of thymine modifications in *Leishmania major* genomic DNA at single-nucleotide resolution, by mapping abasic sites generated enzymatically by the glycosylase SMUG1. We also apply snAP-seq directly to DNA from HeLa cells and reveal a relationship between AP sites and regions containing genes when APE1 enzyme is depleted.

Results

Chemical tagging of aldehyde residues in DNA

The 2’-deoxyribose ring at DNA abasic sites is in equilibrium between the cyclic hemiacetal and acyclic form, where in the latter a reactive aldehyde group is exposed (Fig. 1a). In addition to AP sites, the naturally occurring base modifications 5-fC and 5-fU also provide sources of reactive aldehydes in DNA and thus the selectivity of aldehyde tagging is crucial when studying any of these sites. Hydrazine- and hydroxylamine-based probes have been used for the tagging of aldehyde residues in proteins and nucleic acids25,29–31. However, the resulting hydrazones can be hydrolytically unstable, whilst oxime formation can require acidic conditions or nucleophilic catalysis which promote additional AP site formation, or DNA cleavage1,32. More recently, the Hydrazino-*iso*-Pictet-Spengler (HIPS) reaction has been shown to form stable adducts at formylglycine residues within proteins at near-neutral pH33. We envisaged that the HIPS reaction would be compatible with AP sites under mild conditions. We designed HIPS probe **1** which bears an alkyne handle for ease of subsequent functionalization *via* a copper(I)-catalyzed azide-alkyne cycloaddition (CuAAC) (Fig. 1b). AP-ODN, a synthetic DNA oligomer containing a single abasic site, reacts readily with probe **1** in 2 h without side products (Supplementary Fig. 1). The reactivity of 5-fC has been shown to be reduced at near-neutral pH, and it is possible to achieve the chemoselective reaction of substituted hydrazines with 5-fU over 5-fC above pH 724. Under our HIPS reaction conditions (pH 7.4), reactivity with 5-fC-ODN was low (1%) whilst the reactivity of 5-fU-ODN was comparable to that of AP-ODN (Supplementary Fig. 1). Therefore, a way to further discriminate between these two aldehydes is necessary.

Selective isolation of DNA abasic sites

During sample preparation for next-generation sequencing (NGS), genomic DNA is broken into smaller fragments suitable for sequencing. To enrich for fragments containing a DNA feature of interest relative to bulk genomic DNA, a tag such as biotin can be selectively introduced. Streptavidin beads are then used to capture tagged fragments. After NGS, a pile-up of sequencing reads is expected at loci where the target accumulates. This enrichment-based sequencing has been demonstrated for other DNA modifications24,25,34. We adopted this approach for snAP-seq, by incorporating a biotin moiety onto **1**-labelled DNA *via* a CuAAC reaction. The resulting adducts with an AP site, 5-fU or 5-fC were found to differ in their stabilities under alkaline conditions. Unfunctionalized AP sites undergo DNA strand-scission *via* β or β-δ elimination at high pH35,36. The biotinylated **1**-AP adduct is quantitatively fragmented *via* the same chemistry upon heating in basic conditions. In contrast, the analogous adducts on 5-fU or 5-fC did not show signs of degradation (Fig. 1c). Most importantly, elimination of the DNA backbone at the AP adduct generates two truncated DNA fragments. The fragment 3'- to the adduct no longer contains biotin and instead has a 5'-phosphorylated end. Elimination allows the biotin linker to be selectively cleaved off from DNA strands harboring AP sites, thus separating these strands from streptavidin beads (Fig. 1d), whereas DNA fragments comprising 5-fU remains biotinylated. Therefore, a degree of cross-reactivity of HIPS probe **1** with 5-fU or 5-fC DNA modifications can be tolerated and filtered out by the methodology, to isolate DNA fragments derived exclusively from AP sites.

To evaluate the selectivity of our approach in double-stranded DNA (dsDNA), synthetic dsDNA oligomers (100-105 bp) containing a single AP site, 5-fU or 5-fC and an oligomer containing only canonical bases to represent bulk genomic DNA (GCAT DNA), were treated with **1** and then biotinylated. DNA was captured on streptavidin beads, followed by mild alkaline treatment (100 mM NaOH, room temperature) to denature and release unmodified complementary DNA strands. Under these conditions, biotinylated **1**-AP adducts were confirmed to be stable towards elimination (Fig. 1c). Elution of bound DNA was then achieved under alkaline-cleavage conditions (100 mM NaOH, 70°C). qPCR quantification of recovered DNA confirmed that around 100-fold selectivity was obtained for AP DNA relative to both 5-fC and unmodified GCAT DNA (Fig. 2a). Although the 5-fU-**1** adduct does not undergo DNA fragmentation under alkaline-cleavage conditions (Fig. 1c), near-quantitative recovery of 5-fU DNA occurred after this step, likely due to disruption of the biotin-streptavidin interaction under harsh conditions. As released 5-fU DNA remains biotinylated, these strands were removed by further incubation with streptavidin beads. Whilst AP DNA remains in the supernatant during this step, 5-fU DNA is successfully recaptured and less than 5% is recovered (Fig. 2b). Therefore, high selectivity for AP-derived DNA is achievable after two successive rounds of enrichment.

Mapping abasic sites in double-stranded DNA

With a chemical method for the selective enrichment of AP sites in hand, we turned our attention to combining this with NGS to create a strategy for mapping AP sites in genomic DNA (Fig. 2c). First, a custom P7 adapter is introduced to DNA sequences by ligation. DNA is enriched on streptavidin beads, then released by alkaline-cleavage. Undesired recovery of 5-fU DNA is removed using streptavidin, followed by primer extension on recovered fragments. Ligation of the P5 adapter and PCR amplification then generates the sequencing library. The stepwise introduction of sequencing adapters ensures that only the 5'-phosphates selectively exposed at AP sites by elimination are available for ligation. Sequencing libraries are therefore selective for DNA derived from captured AP sites, and non-specific recovery of untagged DNA is reduced.

To validate that this enrichment strategy is compatible with NGS, sequencing libraries were prepared using equal input amounts of AP, 5-fU, 5-fC and GCAT DNA. 98% of total reads obtained after snAP-seq aligned to AP DNA (Fig. 2d). Furthermore, 95% of reads begin exactly 1 nucleotide after the AP position, resulting in a stacking of reads immediately after the AP site. Thus, snAP-seq allows single-base resolution detection of AP sites with chemical selectivity. In a control experiment, where the same input DNA was sequenced without streptavidin enrichment, the AP-strand was heavily underrepresented (1.6% of total reads) (Fig. 2d, Supplementary Fig. 9). Here, AP DNA is not efficiently amplified by PCR due to the stalling of high-fidelity polymerases at AP sites16. In contrast, the snAP-seq protocol (Fig. 2c) cleaves the DNA at the abasic site and sequences the bases immediately adjacent, preventing polymerase stalling during PCR amplification.

Mapping thymine-modifications in the *Leishmania major* genome

In the *Leishmania major* genome, the DNA base modification 5-hydroxymethyluracil (5-hmU) replaces approximately 0.01% of all thymine (T) residues37,38. 5-hmU is associated with the hypermodified base 5-(β-glucopyranosyl) hydroxymethyluracil (base J)37,39,40, which also occurs in this genome. Studies on the quantification and mapping of both of these modifications have revealed potentially interesting roles in *L. major* and other trypanosomatid genomes38,39. The mapping methods used in these studies allows the identification of broad regions in which modifications are enriched, however the precise loci and sequence context of modifications remains unclear. The glycosylase SMUG1 can be used biochemically to excise 5-hmU from genomic DNA to generate an abasic site, as well as other thymine modifications including 5-fU, uracil and 5-hydroxyuracil41. In contrast, base J is not a substrate (Supplementary Fig. 7). Therefore, we envisaged exploiting snAP-seq in combination with SMUG1 to map 5-hmU at single base resolution in the *L. major* genome. Genomic *L. major* DNA was sonicated, treated with SMUG1 to replace T-modifications with AP sites, then snAP-seq was applied.

After alignment of the sequencing reads obtained by SMUG1-snAP-seq, pile-ups of sequenced reads appear. Reads within these enriched regions are aligned with a common sequencing start-site, corresponding to the base immediately 3’- to SMUG1-generated AP sites (Fig. 3). These sharp increases in coverage were used to detect individual SMUG1-sensitive, modified thymine sites when compared to input DNA. A total of 3,200 high confidence sites were called by SMUG1-snAP-seq across two technical replicates (FDR < 10^{-10}).

Defining the start position of sequencing read 1 as position ‘1’ in the forward strand, and ‘-1’ in the reverse strand, we analyzed the base composition of the position ‘0’, which represents the captured AP site. Over 98% of called sites correspond to a thymine in the reference genome which is the canonical precursor for 5-hmU (Fig. 4a), suggesting that this method is highly accurate and artefacts arising from the non-specific pile-up of reads is low. In the absence of SMUG1 treatment, no significant sites were called, suggesting any endogenous AP sites in these samples do not accumulate at specific loci. Comparison of the relative read counts around the 3,200 SMUG1-senstive sites shows that the SMUG1 untreated libraries closely resemble input DNA (Fig. 4b), further confirming that the signal observed by SMUG1-snAP-seq is specific to enzymatic AP site generation.

We compared our dataset to a previously reported, low-resolution map of 5-hmU in *L. major*38. Out of the 139 reported 5-hmU enriched regions, 76% overlap with high-confidence SMUG1-snAP-seq sites (Fig. 5a). Our results also show that at the single nucleotide level, SMUG1-sensitive sites are highly clustered within these broad stretches (Fig. 4c). Low-resolution mapping38 previously identified an enrichment of G-rich motifs, TpT dinucleotides and longer T-stretches within the broad 5-hmU peaks, however the exact sequence context and strand specificity is lost. Motif analysis using the sites identified here using SMUG1-snAP-seq shows a strong enrichment within TpG dinucleotides (Fig. 5b), as well as a preference for G and T bases on the 5'- side of 5-hmU (GpT and TpT) (Fig. 5c).

As SMUG1-snAP-seq fragments DNA at captured AP sites, multiple SMUG1-generated AP sites which occur in the same individual DNA strand are likely to be represented as only one site, corresponding to the one closest to the 3'- end. The identification of closely clustered SMUG1-snAP sites (Fig. 3) suggests that partially incomplete reactivity may be aiding the identification of dense SMUG1-snAP sites.

As SMUG1 recognizes a number of substrates, we also carried out UNG-snAP-seq by treating *L. major* DNA with UNG then applying snAP-seq. Unlike SMUG1, UNG displays very high specificity for uracil and is not able to excise 5-substitued uracil derivatives such as 5-hmU42. 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 (Supplementary Fig. 14).

Mapping endogenous abasic sites in the human genome

Next, we used snAP-seq to directly investigate the distribution of endogenous AP sites in the human genome. As HeLa cells are well-characterized and easy to manipulate, we chose this cell line for our study. AP sites are chemically labile at high pH35,36, whilst at low pH the generation of additional AP site as artefacts *via* depurination is accelerated1,43. Therefore, to maintain accurate and sensitive detection it is important to ensure that all DNA processing steps preceding chemical tagging do not alter the AP site landscape. DNA extraction and sonication are two key steps used here between cell harvesting and the HIPS reaction, and the effect of both steps on the pulldown efficiency was assessed. Synthetic AP and GCAT DNA were subjected to each of these steps and the extent of enrichment was measured by qPCR. No significant change in enrichment was observed compared to untreated DNA ($p >$ 0.9999), suggesting that these steps do not introduce detectable artefacts (Fig. 6a).

In mammalian cells, APE1 is the main endonuclease to initiate repair at AP sites, accounting for over 95% of AP endonuclease activity44. To further understand the distribution of AP damage, we used siRNA-mediated knockdown of APE1 to study the AP landscape in genomic DNA before completion of repair by BER. Western blot analysis confirmed that around 90% knockdown of APE1 protein was achieved compared to cells treated with non-targeting control siRNA (Fig. 6b).

After snAP-seq, we were not able to detect nucleotides with AP accumulation (Supplementary Fig. 14). The results immediately tell us that AP sites do not accumulate at single-nucleotide sites in the genome across the population of HeLa cells, and contrasts with 5-hmU sites in *L. major* which are likely to be installed enzymatically at specific thymine sites. Given the additional potential for oxidative damage in mitochondrial DNA, we looked to see whether damage accumulates preferentially here. No individual sites were detected, despite the greater sequencing depth at 200X for mitochondria, compared to 10X for the nuclear genome (Supplementary Fig. 14).

We therefore assessed whether AP sites accumulate within stretches of the genome by identifying peaks, rather than single-sites, that are enriched compared to input DNA by integrating signals in both libraries in windows45. For cells treated with control-siRNA, a total of 14,110 peaks were detected, where only high-confidence peaks that appear in at least three out of four replicates were considered. This increases to 25,080 peaks with APE1 knockdown, 10,387 (74% of siRNA-control peaks) of which are in common (Fig. 6c). This suggests that even when BER is active, some AP sites persist long enough in the genome to be captured chemically, whilst the lifetime of others is affected by the presence of APE1.

Next, high-confidence AP peaks were analyzed in the context of gene features (Fig. 6f). For the siRNA control, AP peaks are weakly enriched only within intergenic regions of the genome ($q < 0.05$). The pattern changes with APE1 knockdown, where an enrichment of AP peaks is observed at regulatory and transcribed regions including promoters, 5’ and 3’ UTRs and exons ($q < 0.05$). Peaks in common between the two siRNA treatments are weakly enriched in promoters, exons and intergenic regions, however the majority of peaks (62%) are within intergenic regions.

The distribution of AP peaks was also studied in relation to both chromatin accessibility and histone modifications. An overall enrichment of peaks, after both control and APE1 siRNA treatment, is observed for genomic locations associated with open chromatin, suggesting that these regions are more prone to damage (Supplementary Fig. 16, $q < 0.05$). An enrichment of peaks is observed at areas of the genome associated with repressive histone marks (H3K27me3 and H3K9me3). A weaker enrichment is also seen at activating histone marks (H3K27ac and H3K4me3) for APE1 deficient cells, which is not observed for cells treated with control siRNA (Supplementary Fig. 16).

We analyzed the average base composition of the nucleotide immediately 5’- to the start of all sequencing reads (position ‘0’) globally across all reads within sequencing libraries in an unbiased manner (Fig. 6g). For both control- and APE1 siRNA-treated cells, an enrichment in the purines, guanine and adenine, is seen for the base at position 0 relative to input libraries ($p < 0.05$).

Discussion

Using synthetic DNA, we demonstrate that the chemistry of snAP-seq can detect the position of AP sites with high specificity. Enrichment is observed for AP sites relative to unmodified DNA, as well as the naturally occurring aldehyde-containing bases, 5-fC and 5-fU. Whilst some methods of detecting AP sites may be confounded by the presence of these bases2,20,22,46, snAP-seq targets AP sites with high selectivity. In addition to enrichment over input DNA, a site-specific DNA cleavage step generates an increase in sequencing coverage exactly 1 nucleotide after captured AP sites, and thus the approach can reveal AP sites at single-nucleotide resolution.

We demonstrate that snAP-seq is easily extendable to map base-modifications for which a DNA glycosylase is available. This is exemplified in the SMUG1-snAP-seq analysis of *Leishmania major* DNA, which reveals that regions previously shown to be enriched in T-modifications such as 5-hmU are densely modified. The high-resolution data generated here also shows that within these broad regions the distribution of T-modifications is not random, and 5-hmU occurs preferentially in TpG dinucleotides. Despite the role of 5-hmU as an intermediate during base J biosynthesis, our data show that at steady-state detectable levels of 5-hmU enrichment persist within the *L. major* genome. The development of high-resolution mapping techniques has potential in further elucidating the role and function of such base modifications.

In addition to T-modification mapping *via* AP site generation by SMUG1, snAP-seq can be coupled with other glycosylases. UNG6, hOGG17 or TDG9 may also find interesting applications in different biological contexts when combined with this method.

We applied snAP-seq to elucidate endogenous abasic sites in the human genome, using DNA from HeLa cells. In addition to studying BER competent cells, we depleted cellular levels of APE1 to investigate the preferences of AP site formation in genomic regions before repair. We find that AP sites do not accumulate site-specifically at the single-nucleotide level, even when APE1 is depleted. This suggests a level of stochasticity for the exact site of damage across a population of cells, such that stretches of the genome more prone to damage can be identified. Alternatively, knockdown of APE1 alone may not be sufficient to observe a substantial accumulation of AP sites at a given location. APE2 is also present in mammals which may become more active in the absence of APE1 activity. Similar to our findings, methods capable of detecting DNA double strand breaks at nucleotide-resolution have found that this type of damage also accumulates within broadened peaks47.

Using peak-based analysis, a high degree of overlap is seen between peaks in control- and APE1 siRNA-treated cells. AP peaks detected in control cells accumulate even in the presence of APE1, suggestive of a basal level of damage which exists largely in intergenic regions of the genome. Upon knockdown of APE1, an additional 14,693 peaks are detected. AP damage here is found to be enriched within genic and regulatory regions, suggesting that APE1 may be responsible for repairing damage in these regions. This general trend is in line with that observed for the oxidative DNA damage marker, 8-oxoguanine, where levels increase in promoters and 5’- and 3’-UTRs upon knockout of the repair enzyme OGG1 in mouse embryonic fibroblasts48. Our results also suggest that in both APE1 depleted and control cells, there is a global preference for AP sites derived from depurination compared to depyrimidination (Fig. 6g). This is consistent with *in vitro* studies which show that purines are less stable than pyrimidines towards hydrolysis49.

In summary, we present a versatile method in which both endogenous abasic sites, as well as those selectively generated *in vitro* can be mapped at high resolution. We investigated both of these classes of abasic sites in two different genomes. Our method was combined with SMUG1 glycosylase treatment to provide the first single-nucleotide resolution map of 5-hmU in the *L. major* genome. Future applications of snAP-seq and glycosylase-coupled variations of this method will be valuable for studies on both DNA damage and base modifications.

Methods

Reaction with **1** and biotinylation: Purified genomic DNA was treated with **1** (10 mM) in sodium phosphate buffer (40 mM, pH 7.4) at room temperature for 2 hr, then purified using a mini-quick spin oligo column (Roche) according to the manufacturer's instructions. The eluted DNA was incubated with CuBr (250 μM), THPTA (1.25 mM) and biotin-PEG3-azide (500 μM) at 37°C for 2 h. Samples were purified using a pre-washed Amicon Ultra-0.5 mL 10K centrifugal filter (500 μL water) and washed on the filter with water (450 μL) and Tris-HCl buffer (450 μL, 10 mM, pH 7.4) and eluted in Tris-HCl buffer (50 μL).

P7 adapter ligation: Sequencing adapters were ligated onto labelled DNA using a NEBNext Ultra II DNA library preparation kit according to the manufacturer's instructions, with the exception that the adapter was replaced with custom P7 adapter (2.5 μL). Ligated DNA was then treated with shrimp alkaline phosphatase (NEB, 3 U) in CutSmart Buffer (NEB) for 30 min at 37°C, before purification with AMPure XP beads (1.4 X volume) and eluted in Tris-HCl (10 mM pH 7.4, 48 μL).

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

Primer extension: Reaction volumes (30 μL) contained purified ssDNA, dNTPs (200 μM), P7 primer (1 μM), NEBuffer 2 and was heated to 95 °C for 1 min, annealed at 65 °C for 30 s and held at 37 °C for 30 min, at which point Klenow fragment (3'->5'- exo-, NEB, 2 U) was added. The synthesized dsDNA was purified using a DNA clean and concentrator-5 kit (Zymo Research) according to the manufacturer's instructions and eluted in Tris-HCl (10 mM, pH 7.4).

P5 adapter ligation: DNA (22.5 μL), Blunt/TA ligase master mix (25 μL, NEB) and custom P5 adapter (2.5 μL) were incubated at 20 °C for 30 min. Libraries were purified with AMPure XP beads (1.5 X volume) and eluted in Tris-HCl (10 mM, pH 7.4) before amplification using a Q5 hot start high-fidelity master mix (NEB) with library amplification primers (10 μM each). Libraries were quantified using a KAPA library quantification kit and sequenced on either an Illumina MiSeq, or NextSeq machine.

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

Data access and analysis: raw sequencing reads were processed and alignments generated and analyzed using Unix tools, R and Python scripts, see details available in the SI. We have released all the computational code in the accompanying manuscript's GitHub page (https://github.com/sblab-bioinformatics/snAP-seq). The raw sequencing data and all relevant snAP-seq sites and peaks have been deposited in the ArrayExpress database at EMBL-EBI under accession number E-MTAB-7152 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7152).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

S.B. is a senior investigator of the Wellcome Trust (grant no. 209441/Z/17/Z). The Balasubramanian laboratory is also supported by core funding from Cancer Research UK (C14303/A17197). Z. J. L. is supported by Pembroke College, Cambridge and the Herchel Smith Fund. P. V. D. was funded by a Marie Curie Fellow of the European Union (grant no. FP7-PEOPLE-2013-IEF/624885).

References

1. Lindahl T, Nyberg B. Rate of depurination of native deoxyribonucleic acid. Biochemistry. 1972; 11:3610–3618. [PubMed: 4626532]
2. Wang Y, et al. Direct detection and quantification of abasic sites for in vivo studies of DNA damage and repair. Nucl Med Biol. 2009; 36:975–983. [PubMed: 19875055]
3. Kidane D, Murphy DL, Sweasy JB. Accumulation of abasic sites induces genomic instability in normal human gastric epithelial cells during Helicobacter pylori infection. Oncogenesis. 2014; 3:e128. [PubMed: 25417725]
4. Nakamura J, La DK, Swenberg JA. 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. 2000; 275:5323–5328. [PubMed: 10681505]
5. Krokan HE, Bjørås M. Base Excision Repair. Cold Spring Harb Perspect Biol. 2013; 5
6. Lindahl T, Ljungquist S, Siegert W, Nyberg B, Sperens B. DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. J Biol Chem. 1977; 252:3286–3294. [PubMed: 324994]
7. Boiteux S, Radicella JP. The Human OGG1 Gene: Structure, Functions, and Its Implication in the Process of Carcinogenesis. Arch Biochem Biophys. 2000; 377:1–8. [PubMed: 10775435]
8. He Y-F, et al. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science. 2011; 333:1303–1307. [PubMed: 21817016]
9. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013; 502:472–479. [PubMed: 24153300]

10. Franchini D-M, et al. Processive DNA Demethylation via DNA Deaminase-Induced Lesion Resolution. PLOS ONE. 2014; 9:e97754. [PubMed: 25025377]
11. Jacobs AL, Schär P. DNA glycosylases: in DNA repair and beyond. Chromosoma. 2012; 121:1–20. [PubMed: 22048164]
12. Lari S-U, Chen C-Y, Vertéssy BG, Morré J, Bennett SE. Quantitative determination of uracil residues in Escherichia coli DNA: Contribution of ung, dug, and dut genes to uracil avoidance. DNA Repair. 2006; 5:1407–1420. [PubMed: 16908222]
13. Riedl J, Fleming AM, Burrows CJ. Sequencing of DNA Lesions Facilitated by Site-Specific Excision via Base Excision Repair DNA Glycosylases Yielding Ligatable Gaps. J Am Chem Soc. 2016; 138:491–494. [PubMed: 26741640]
14. Shu X, et al. Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA. Nat Chem Biol. 2018; 14:680. [PubMed: 29785056]
15. Loeb LA, Preston BD. Mutagenesis by Apurinic/Apyrimidinic Sites. Annu Rev Genet. 1986; 20:201–230. [PubMed: 3545059]
16. Boiteux S, Guillet M. Abasic sites in DNA: repair and biological consequences in Saccharomyces cerevisiae. DNA Repair. 2004; 3:1–12. [PubMed: 14697754]
17. AlMutairi F, et al. Association of DNA Repair Gene APE1 Asp148Glu Polymorphism with Breast Cancer Risk. Dis Markers. 2015; 2015
18. Lirussi L, et al. APE1 polymorphic variants cause persistent genomic stress and affect cancer cell proliferation. Oncotarget. 2016; 7:26293–26306. [PubMed: 27050370]
19. Chastain PD, Nakamura J, Swenberg J, Kaufman D. Nonrandom AP site distribution in highly proliferative cells. FASEB J. 2006; 20:2612–2614. [PubMed: 17068113]
20. Chastain PD, et al. Abasic sites preferentially form at regions undergoing DNA replication. FASEB J. 2010; 24:3674–3680. [PubMed: 20511393]
21. Kubo K, Ide H, Wallace SS, Kow YW. A novel sensitive and specific assay for abasic sites, the most commonly produced DNA lesion. Biochemistry. 1992; 31:3703–3708. [PubMed: 1567824]
22. Kurisu S, et al. Quantitation of DNA damage by an aldehyde reactive probe (ARP). Nucleic Acids Res Suppl. 2001; 2001:45–46.
23. Ide H, et al. Synthesis and damage specificity of a novel probe for the detection of abasic sites in DNA. Biochemistry. 1993; 32:8276–8283. [PubMed: 8347625]
24. Hardisty RE, Kawasaki F, Sahakyan AB, Balasubramanian S. Selective Chemical Labeling of Natural T Modifications in DNA. J Am Chem Soc. 2015; 137:9270–9272. [PubMed: 25946119]
25. 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. 2012; 13:R69. [PubMed: 22902005]
26. Pfaffeneder T, et al. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. Nat Chem Biol. 2014; 10:574–581. [PubMed: 24838012]
27. Pfaffeneder T, et al. The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. Angew Chem. 2011; 123:7146–7150.
28. Rahimoff R, et al. 5-Formyl- and 5-Carboxydeoxycytidines Do Not Cause Accumulation of Harmful Repair Intermediates in Stem Cells. J Am Chem Soc. 2017; 139:10359–10364. [PubMed: 28715893]
29. Zhang H, Li X, Martin DB, Aebersold R. Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat Biotechnol. 2003; 21:660–666. [PubMed: 12754519]
30. Wu P, et al. Site-specific chemical modification of recombinant proteins produced in mammalian cells by using the genetically encoded aldehyde tag. Proc Natl Acad Sci. 2009; 106:3000–3005. [PubMed: 19202059]
31. Rashidian M, Dozier JK, Lenevich S, Distefano MD. Selective labeling of polypeptides using protein farnesyltransferase via rapid oxime ligation. Chem Commun. 2010; 46:8998–9000.
32. Lindahl T, Andersson A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. Biochemistry. 1972; 11:3618–3623. [PubMed: 4559796]

33. Agarwal P, et al. Hydrazino-Pictet-Spengler Ligation as a Biocompatible Method for the Generation of Stable Protein Conjugates. Bioconjug Chem. 2013; 24:846–851. [PubMed: 23731037]

34. Xia B, et al. Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. Nat Methods. 2015; 12:1047–1050. [PubMed: 26344045]

35. Sugiyama H, et al. Chemistry of thermal degradation of abasic sites in DNA. Mechanistic investigation on thermal DNA strand cleavage of alkylated DNA. Chem Res Toxicol. 1994; 7:673–683. [PubMed: 7841347]

36. Lhomme J, Constant JF, Demeunynck M. Abasic DNA structure, reactivity, and recognition. Biopolymers. 1999; 52:65–83. [PubMed: 10898853]

37. Bullard W, Lopes da Rosa-Spiegler J, Liu S, Wang Y, Sabatini R. Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. J Biol Chem. 2014; 289:20273–20282. [PubMed: 24891501]

38. Kawasaki F, et al. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. Genome Biol. 2017; 18:23. [PubMed: 28137275]

39. Reynolds D, et al. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. Nucleic Acids Res. 2014; 42:9717–9729. [PubMed: 25104019]

40. van Luenen HGAM, et al. Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania. Cell. 2012; 150:909–921. [PubMed: 22939620]

41. Masaoka A, et al. Mammalian 5-Formyluracil–DNA Glycosylase. 2. Role of SMUG1 Uracil–DNA Glycosylase in Repair of 5-Formyluracil and Other Oxidized and Deaminated Base Lesions. Biochemistry. 2003; 42:5003–5012. [PubMed: 12718543]

42. 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. 2014; 23:1667–1685.

43. An R, et al. Non-Enzymatic Depurination of Nucleic Acids: Factors and Mechanisms. PLOS ONE. 2014; 9:e115950. [PubMed: 25546310]

44. Masani S, Han L, Yu K. Apurinic/Apyrimidinic Endonuclease 1 Is the Essential Nuclease during Immunoglobulin Class Switch Recombination. Mol Cell Biol. 2013; 33:1468–1473. [PubMed: 23382073]

45. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

46. Kow YW, Dare A. Detection of Abasic Sites and Oxidative DNA Base Damage Using an ELISA-like Assay. Methods. 2000; 22:164–169. [PubMed: 11020331]

47. Lensing SV, et al. DSBCapture: *in situ* capture and sequencing of DNA breaks. Nat Methods. 2016; 13:855–857. [PubMed: 27525976]

48. Ding Y, Fleming AM, Burrows CJ. Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. J Am Chem Soc. 2017; 139:2569–2572. [PubMed: 28150947]

49. Lindahl T, Karlstrom O. Heat-induced depyrimidination of deoxyribonucleic acid in neutral solution. Biochemistry. 1973; 12:5151–5154. [PubMed: 4600811]

50. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinforma Oxf Engl. 2011; 27:1653–1659.

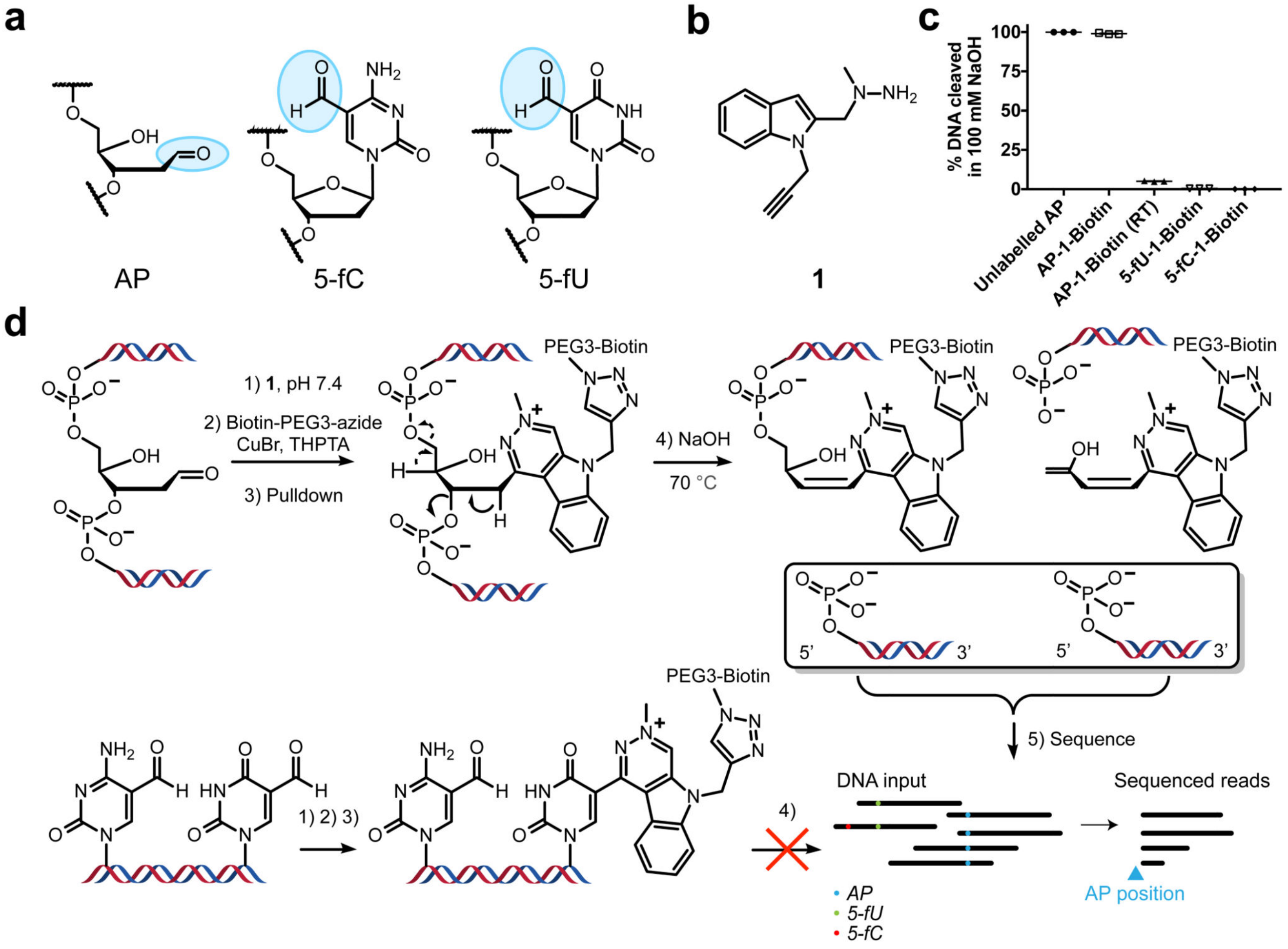

Figure 1.

Chemical tagging of DNA abasic sites. **a**) Structures of naturally occurring aldehyde residues found in DNA. **b**) Structure of HIPS probe **1**. **c**) Stability of AP-ODN and probe **1**-labelled ODNs after biotinylation, to alkaline-cleavage (100 mM NaOH, 15 min). Reactions were carried out at 70°C except where labelled RT, where room temperature was used. 5-fC-1-biotin was obtained by extending the reaction time with **1** to 24 h. Reactions were followed by LC-MS. Results from three independent replicates are shown. **d**) Workflow of chemical enrichment of abasic sites using probe **1**. Tagging of aldehyde residues followed by biotinylation through CuAAC and streptavidin pulldown enriches for all reacted aldehydes, whilst the removal of the biotin tag by alkaline-cleavage occurs only at AP sites. The truncated DNA fragments with 5'-phosphate termini are then recovered and prepared for next-generation sequencing. Sequencing reads stack up immediately after captured AP sites due to the site-specific DNA cleavage step *via* base-catalyzed elimination.

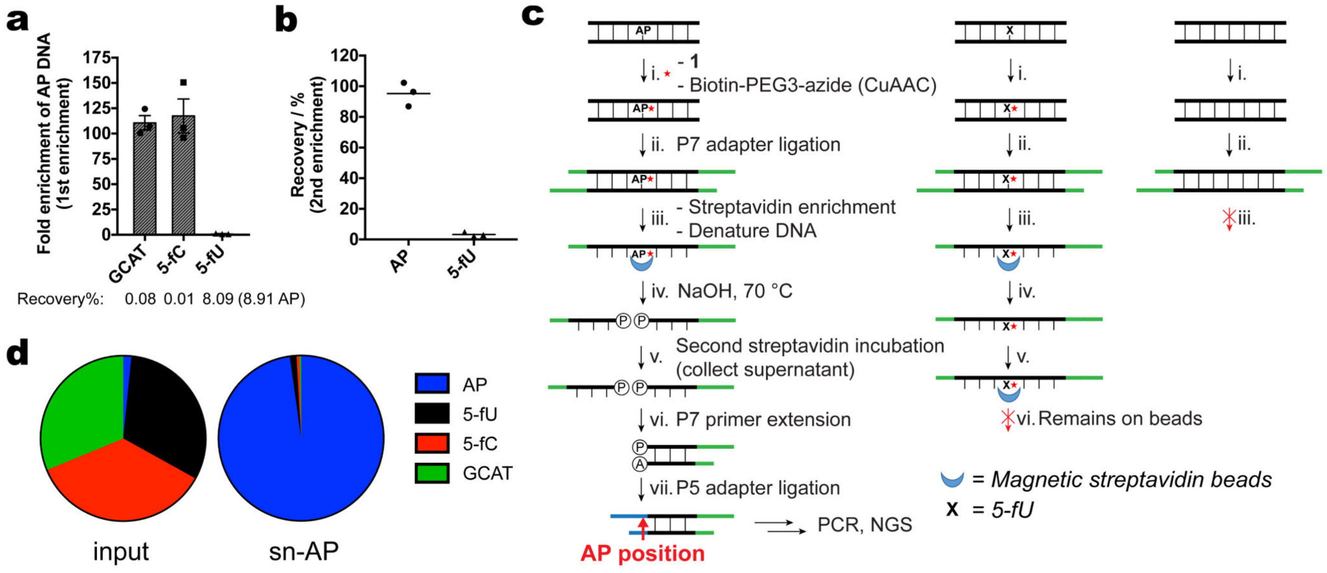

Figure 2.

Outline of snAP-seq and analysis of synthetic DNA spike-ins. **a**) Enrichment of AP DNA relative to GCAT, 5-fC and 5-fU DNA, after first streptavidin enrichment. DNA was recovered under alkaline-cleavage conditions (100 mM NaOH, 70°C), before neutralization and purification. DNA was quantified by qPCR, and normalized to input DNA. The recovery of each DNA sequence was compared to AP DNA, and represented as a fold-enrichment of AP DNA. Apart from GCAT DNA, all primers were designed 3’- to each modification and therefore expected cleavage site to allow DNA amplification. Mean ± S.E.M. of three replicates are shown. **b**) Recovery of DNA sequences after second streptavidin enrichment. AP DNA remains unbound and is recovered from the supernatant, whilst 5-fU DNA is captured on the beads. DNA was quantified by qPCR and normalized to the amount of DNA recovered after first streptavidin enrichment. Results from three replicates are plotted. **c**) Library-preparation workflow of snAP-seq. **d**) Distribution of reads aligning to each model DNA sequence after next-generation sequencing. The number of reads of each ODN was normalized to the corresponding reverse strand in the input sample (see Supplementary Table 3). Only the modification-containing, forward strands of each ODN are analyzed.

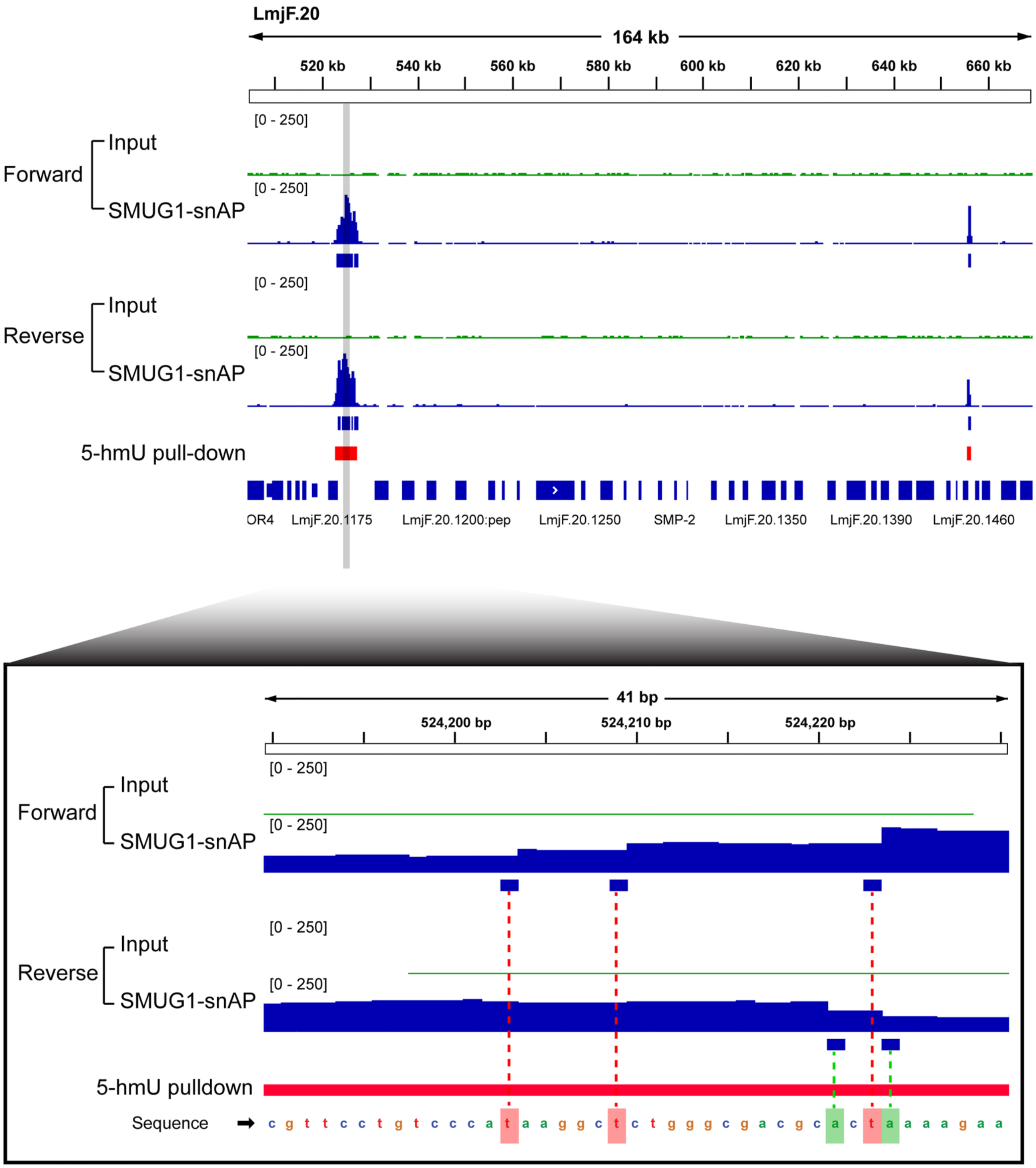

Figure 3.

Representative genome browser view of sites called by SMUG1-snAP-seq in the *Leishmania major* genome. Sharp increases in coverage are observed in snAP-seq data (blue) when compared to input samples (green), which begin one nucleotide after captured abasic sites. These locations are therefore identifiable at single-nucleotide resolution upon alignment of reads to the reference genome.

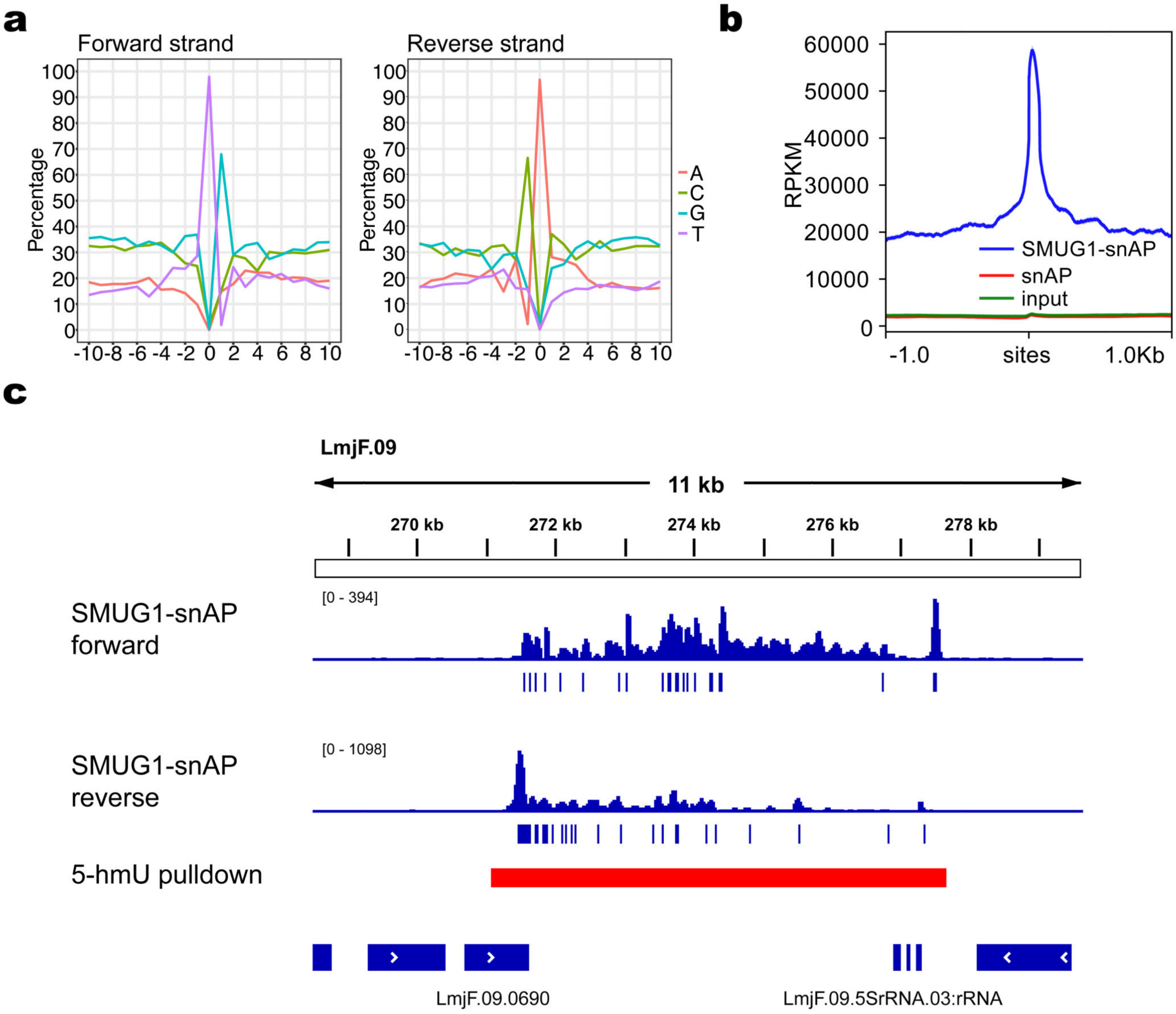

Figure 4.

SMUG1-snAP-seq sites in the *L. major* genome. **a**) Plot of base composition around identified AP sites in the forward and reverse strands of the reference genome. Position 0 corresponds to the called SMUG1-AP site. **b**) Comparison of sequencing coverage around the 3,200 high-confidence SMUG1-AP sites in SMUG1 treated, SMUG1 untreated and input libraries. **c**) Representative genome browser view of SMUG1-snAP-seq sites and 5-hmU pulldown peaks. SMUG1-sensitive sites are clustered densely within 5-hmU enriched peaks.

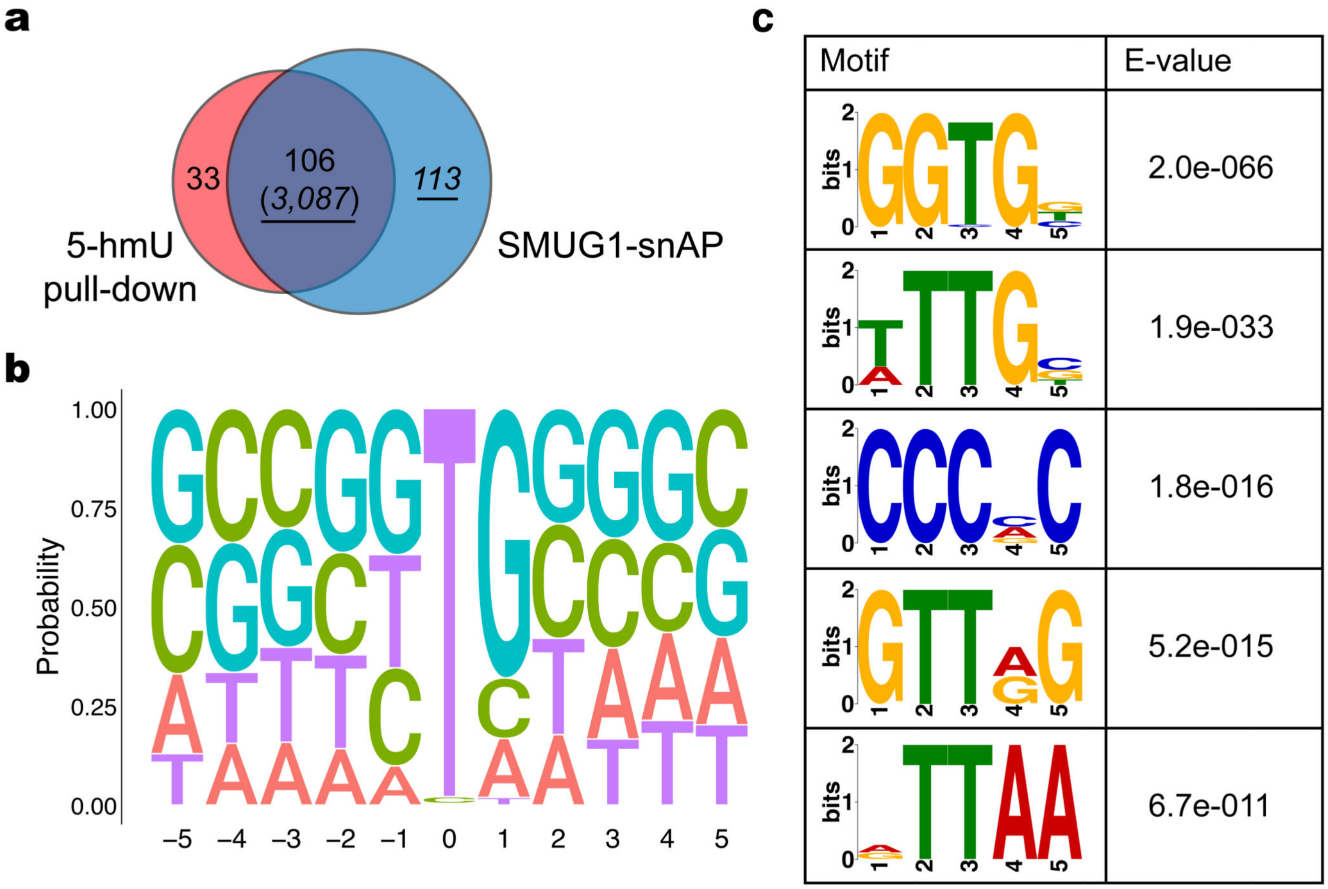

Figure 5.

Analysis of SMUG1-snAP-seq sites in the *Leishmania major* genome. **a**) Overlap of high-confidence SMUG1-snAP-seq sites (underlined) with 5-hmU enriched regions. **b**) Sequence logo plot of base composition 5 bases upstream and downstream of high-confidence SMUG1-snAP sites (base ‘0’). **c**) Enriched motif sequences obtained using DREME50.

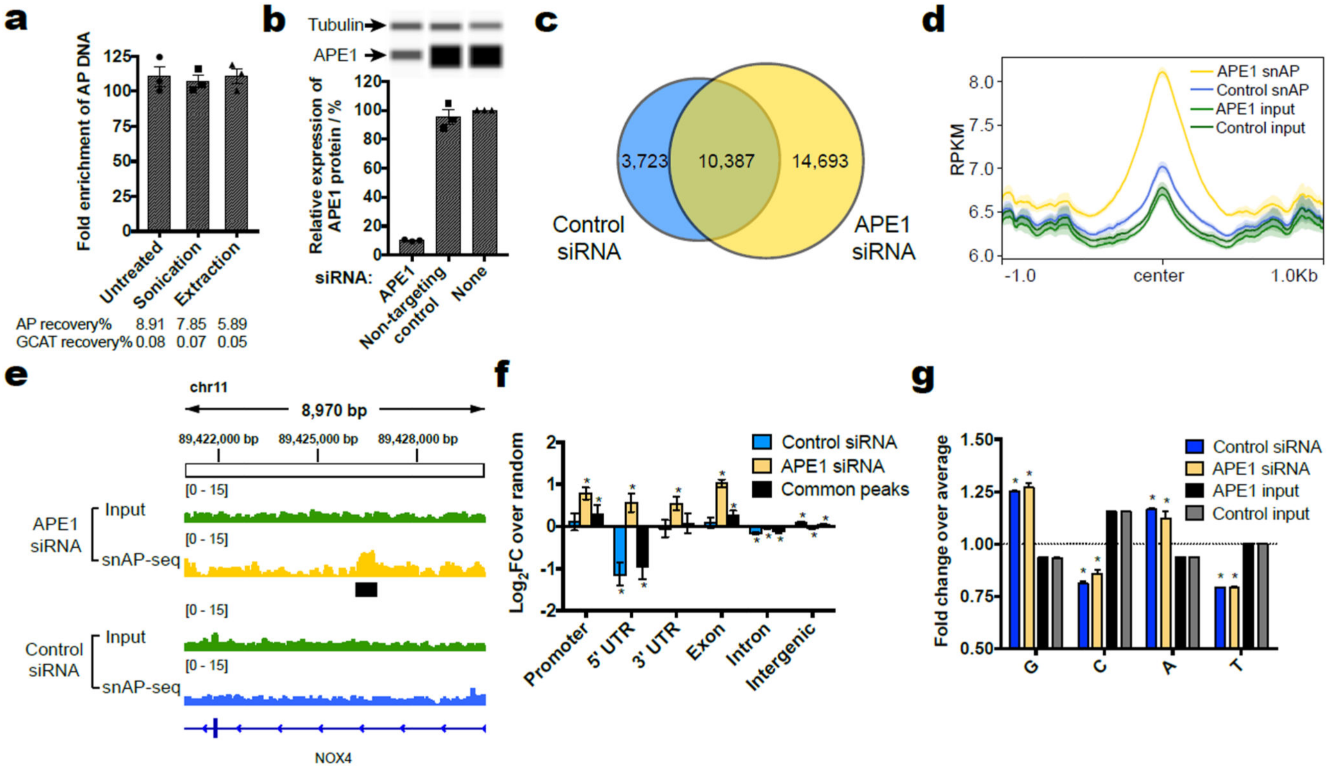

Figure 6.

Mapping of AP sites in human (HeLa) DNA. **a**) Enrichment of AP DNA relative to GCAT DNA determined by qPCR quantification, with and without DNA sonication and mock re-extraction. Mean ± S.E.M. of three replicates are shown. **b**) Western blot of APE1 protein with siRNA knockdown. Mean ± S.E.M. of three replicates are shown. **c**) Overlap of snAP-seq peaks called in HeLa cells treated with control or APE1 siRNA. Four replicates across two independent biological samples were used for both conditions. Only high-confidence peaks that appear in at least three out of four replicates were used45. **d**) Example genome browser view of peaks called after snAP-seq. Normalized read counts are shown. Black bars represent called peaks. **e**) Comparison of sequencing coverage around high-confidence snAP-seq peaks detected for APE1 siRNA-treated cells. **f**) Relative enrichment of snAP-seq peaks in different genomic regions, 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, $*q < 0.05$. **g**) Relative enrichment of bases at the ‘0’ position, directly 5’- to read start sites. Mean and S.E.M of four replicate libraries are shown, $*p < 0.05$ (two-way ANOVA, Sidak’s multiple comparisons test).