Mapping of DNA damage genome-wide at nucleotide resolution by circle-damage-sequencing

Seung-Gi Jin, Dean Pettinga, Jennifer Johnson and Gerd P. Pfeifer*

Center for Epigenetics, Van Andel Institute, Grand Rapids, Michigan 49525, USA

*Corresponding author:

Gerd P. Pfeifer
Center for Epigenetics
Van Andel Institute
Grand Rapids, MI 49525
USA
616-234-5398
gerd.pfeifer@vai.org

ABSTRACT

To establish relationships between mutations, for example in cancer genomes, and possible mechanisms linked to DNA damage, it is necessary to know at what sequence positions of the genome the damage occurs. However, it has been challenging to specifically map DNA damage at the nucleotide level of resolution and genome-wide with high sensitivity. Here, we describe a new method, which we named circle damage sequencing (circle-damage-seq), to accomplish this goal. The method is based on circularization of DNA molecules and DNA damage-selective cleavage of the circularized DNA followed by adapter ligation and sequencing. Based on the design of this approach, only DNA damage-containing molecules are sequenced. We conducted proof-of-principle studies to show that mapping of ultraviolet B-induced cyclobutane pyrimidine dimers (CPDs) can easily be achieved and show a specific tetranucleotide sequence context for CPDs (5'PyPy<>PyT/A) with no further sequence enrichment outside of this context. Our approach shows strongly reduced levels of CPDs near transcription start sites and a spike of this damage near the transcription end sites of genes. We then show that 1,N^6-etheno-deoxyadenosine DNA adducts formed after treatment of cells with the lipid peroxidation product 4-hydroxynonenal can be mapped genome-wide at adenine positions within a preferred sequence context of 5'TAC/G3'. The circle-damage-seq method can be adapted for a variety of DNA lesions for which specific excision enzymes are available.

INTRODUCTION

Cancer genome sequencing has identified thousands of somatic mutations in different types of human cancer [1-3]. These mutations provide a valuable source of information as to the potential mutagenic events that have occurred and have created the mutations, sometimes decades before the tumor develops. In some cases, convincing links between an exposure and cancer mutations have been established, best exemplified by typical sunlight-induced mutations (C to T mutations at dipyrimidine sequences) that are prevalent in melanoma and in non-melanoma skin cancers [4], by smoking-related G to T transversion mutations that are common in lung cancers from smokers but not nonsmokers [5-7], and by aristolochic acid-associated mutations in urothelial carcinomas [8,9]. Several human cancers carry unusual and characteristic mutational signatures of completely unknown origin [5,10,11]. Some of these signatures are thought to arise from specific DNA damage or from defective DNA repair processes, either following exposure of tissues to carcinogens or derived from endogenous chemical reactions [9,12].

Many DNA-reactive agents are characterized by a unique spectrum of DNA damage they produce in exposed cells or tissues that is often directly linked to the mutational spectrum. DNA damage mapping is possible at the nucleotide level within specific genes by PCR-based techniques [13-16]. Today, it is of greater interest to develop methodology that can be used to map DNA damage at single base resolution and genome-wide. One key challenge with this type of analysis is the development of DNA damage-specific and sensitive methods.

Several such methods exist but all have certain limitations. At a lower level of resolution, antibodies against DNA lesions can be used to immunoprecipitate the damaged DNA followed by high-throughput sequencing of the collected DNA fragments [17,18]. However, these methods do not achieve single-base resolution, as it would be required for interpreting mutational spectra linked to DNA damage.

One method is based on the release of small DNA fragments, about 20-30 nucleotides in length, by the nucleotide excision repair (NER) process within cells. These oligonucleotides can be purified from exposed cells and are then sequenced using high-throughput sequencing (excision repair sequencing, or XR-seq) [19,20]. This method is very

specific and sensitive but is limited to DNA lesions that are processed by NER and hence cannot easily detect smaller base lesions. Other methods rely on the ligation of oligonucleotides into a DNA strand break, either a single-strand break or a double-strand break [21-26]. These methods often use secondary DNA breakage, such as by sonication, to create a random DNA break in addition to the specific DNA break produced from the base damage, which may increase background and sequencing load. Because of the limitations often associated with currently used methods, we have been developing a new approach, which we named circle-damage-sequencing (or circle-damage-seq). This method is used in combination with DNA damage-specific repair enzymes and allows the detection and mapping of DNA base damage at the single base level of resolution and genome-wide.

RESULTS

Overview of circle-damage-sequencing (circle-damage-seq)

DNA lesions may form as a result of radiation exposure or after exposure of cells to exogenous or endogenous reactive chemicals. In cells, the lesions can be recognized and repaired by lesion-specific DNA repair enzymes with high fidelity and efficiency. We employed these DNA repair systems to perform high-throughput mapping of various DNA lesions in damaged cells at single-nucleotide resolution on a next-generation sequencing (NGS) platform. In this approach, we adapted the circle-seq method originally based on circularization reactions to map rare genomic CRISPR/Cas9 off-target cleavage sites in vitro [27]. We refer to this novel DNA damage-selective mapping method as “circle-damage sequencing” (circle-damage-seq) (Fig. 1).

We isolated genomic DNA from cells that either have an endogenous base modification or have been treated with a DNA damaging agent. In order to obtain a DNA damage spectrum after an exposure, DNA needs to be harvested before adducts can be repaired by intracellular repair activity. We treated cells with DNA damaging agents and

immediately harvested the cells. Using sonication or restriction enzyme digestion, we converted the purified DNA to an average of 300 to 400 base pair size, a suitable range for DNA circularization and downstream NGS analysis. Circularization of DNA fragments smaller than 100 bp is limited. After data analysis, we obtained more uniform genomic mapping results from random shearing by sonication compared to restriction enzyme cutting, due to sequence bias of the restriction enzymes.

After the fragmentation, we performed the DNA circularization step in circle-damage-seq using 1 µg of DNA per 200 µl of reaction volume with T4 DNA ligase to minimize ligation between linear DNA fragments. As a key step in circle-damage-seq, the circularization step efficiently converts genomic DNA to a suitable size of circular DNA molecules containing modified DNA bases in random positions for next-generation sequencing. In addition to this feature, during circularization, T4 DNA ligase repairs DNA single strand breaks that are present endogenously or are caused by mechanical stress or other factors during sample preparation. Thus, the method avoids mapping most DNA single-strand breaks that are often present as background and false-positive reads in DNA adduct mapping procedures. In the next step, the remaining non-circular DNA fragments with a free end in the DNA pool are removed by treatment of the circularized DNA with an exonuclease mixture, plasmid-safe ATP-dependent DNase (Fig. 1A). Then, in order to produce double-strand breaks at DNA adduct bases within the circularized DNA, we initially convert the DNA with a lesion-specific DNA repair enzyme such as a DNA glycosylase to generate abasic (AP) sites at the positions of the damaged base. Following excision of the damaged base by the DNA glycosylase, we use AP endonuclease 1 (APE1) together with the AP lyase activity of a DNA glycosylase to cleave the AP sites, creating a one nucleotide gap in the DNA at the position of the damaged base. Next, we cleave and remove the single nucleotide on the opposite strand of the gapped DNA molecules with single-strand-specific S1 nuclease thus creating a double strand break at the DNA damage position (Fig. 1B).

As another feature of circularization, these double-strand breaks at DNA adduct positions within circularized DNAs only have one sequencing adaptor ligatable break position that can be converted into libraries suitable for next-generation sequencing. Circle-damage-seq uniquely enables sequencing of both sides of the single cleavage site

in one DNA molecule using paired-end sequencing. Closed circular DNA without any DNA adduct bases is excluded from the breakage and ligation steps resulting in a lesion-selective DNA library enrichment by PCR (Fig. 1C). Notably, sequencing reads obtained from circle-damage-seq show a unique aligned read pattern, divergent reads with a single base gap in the center representing the damaged DNA base, due to the fact that each paired-end read is derived from a single opened circular DNA molecule (Fig. 1C). Thus, circle-damage-seq provides high selectivity for DNA damage mapping at single nucleotide resolution with low background and dramatically reduces the required sequencing depth and cost of this approach.

Mapping of DpnI cleavage sites in the *E.coli* genome

We initially tested the feasibility of circle-damage-seq by analyzing the endogenous DNA modification *N6*-methyladenine (6mA). Recent work has suggested that 6mA may play important roles in gene regulation in prokaryotic and even in mammalian genomes, but 6mA is at or below the limit of detection in mammals when using currently known techniques [28-30]. In the *E.coli* genome, 6mA is formed at 5'GATC sequences and is distributed genome-wide in dam(+) but not in dam(-) strains due to the activity of the Dam DNA methyltransferase [31]. The DpnI restriction enzyme specifically recognizes 5'G(6mA)TC sequences with methylated adenine and cuts between G(6mA) and TC but does not cleave unmethylated 5'GATC sites (Fig. S1A). To test the applicability and sensitivity of circle-damage-seq to detect 6mA, we mixed *E.coli* DNA from two different strains, dam(+) and dam(-), to levels of 0%, 2% and 10% of dam(+) DNA in a background of dam(-) DNA representing 0%, 2% and 10% of GATC sequences being methylated at adenine in input DNA. We subjected the mixture to the circle-damage-seq procedure. DNA double-strand breakage at 6mA bases within circularized DNA was achieved through DpnI treatment which generates a blunt end by splitting the target DNA between G(6mA) and TC sequences. The converted DNA was ligated and PCR-amplified to produce sequencing libraries (Fig. S1B).

We aligned the sequencing reads to the *E.coli* K-12/MG1655 reference genome. The aligned reads were initially filtered to retain the paired reads that are divergent but have no gaps. DpnI should only cleave G(6mA)TC site and leave the beginning of the

reads with 5'TC at both paired divergent reads in the sequencing output. Thus, we focused on the 5'GATC tetranucleotides context positioned at the center of divergent reads and examined the frequency of 5'GATC compared to 5'NNNN background sequences. Each library was also normalized by library size. As expected, the frequency of divergent reads with 5'GATC derived from DpnI cuts was proportional to the amount of dam(+) input DNA in the sequenced libraries (Fig. S1C). Also, the ratio of reads with 5'GATC in 2% and 10% of dam(+) libraries were significantly higher than in the 0% dam(+) control library, suggesting that circle-damage-seq is a quantitative and sensitive method to map DNA modifications in the genome. We mapped all forward reads (75 bp length) derived by circle-damage-seq to the *E.coli* reference genome and identified the enrichment and distribution of peaks along the genomic sequence. As shown at Fig. S1D, the visualized peak patterns in the IGV viewer showed the same proportional enrichment to the amount of dam(+) on the reference genome and the beginning of reads with the TC dinucleotide context, defining 6mA sites at GATC sequences.

Circle-damage-seq for genome-wide mapping of UVB-induced cyclobutane pyrimidine dimers (CPDs) at single-base resolution

UVB-induced cyclobutane pyrimidine dimers (CPDs) at dipyrimidine sequences can be specifically recognized and cleaved by T4 endonuclease V (also known as T4-PDG) [32]. Encouraged by the DpnI mapping results, we extended circle-damage-seq to map UVB induced CPD distribution genome-wide in mammalian cells. Mouse embryo fibroblast cells (MEFs) were irradiated with different doses of UVB. We confirmed the formation of CPDs in the genomic DNA and their effective cleavage with combined treatment of the UVB-irradiated DNA with T4-PDG and the single-strand specific S1 nuclease. This combined treatment leads to the formation of double-stranded DNA breakage at CPD sites which is easily apparent by dose-dependent DNA cleavage on non-denaturing agarose gels (Fig. S2).

To perform circle-damage-seq, we fragmented the isolated genomic DNAs from UVB treated cells to an average of 300 bp size of DNA by using sonication and then performed the circularization step. As shown in Fig. 2A, to create double-strand breaks at the DNA lesions, we initially used the CPD-specific DNA glycosylase, T4-PDG, which

has AP lyase activity, and the AP endonuclease APE1, to cleave the 5' side of the sugar-phosphate backbone at pyrimidine dimers. We followed this treatment by incubation with *E.coli* photolyase under long-wave UVA light to remove dimerized pyrimidine bases that remain after the DNA glycosylase incision. Next, S1 nuclease treatment generated ligatable double-strand breaks within the circularized DNA. Note that the combined treatment creates a single nucleotide gap representing the 5' pyrimidine of the dimer.

We aligned the sequencing reads from circle-damage-seq to the mm10 mouse reference genome. To evaluate the divergent read patterns, we converted the mapped reads obtained from 1,000 J/m^2 of UVB-irradiated MEF cells to the bam file format and visualized the reads at single-base resolution using the IGV viewer. Indeed, we observed the divergent paired reads with a single base gap in the center and confirmed the matching bases in the gaps as pyrimidines on the reference sequence that have another pyrimidine base right next to it in the 3' position. We observed this pattern with a high frequency at 5'TT and 5'TC sites and at lower frequency at 5'CT and 5'CC dinucleotide sequences. This result is consistent with the known specificity of CPD formation at dipyrimidine sequences (Fig. 2B and Fig. S3A) [4]. We also observed a dipyrimidine sequence context on the opposite DNA strands of the reference genome, confirming that circle-damage-seq can detect DNA damage on either DNA strand (Fig. 2C). These divergent reads were distributed relatively uniform across the genome and can be displayed at the gene resolution level as shown, for example, in Figure 2 and in Figure S3B.

Previous work has focused mainly on the dinucleotide specificity of genomic UV damage [4,26,33-35]. First, we confirmed that CPDs were most frequent at 5'TT and 5'TC dipyrimidines followed by 5'CC and 5'CT. We next analyzed trinucleotide sequences that reflect the sequence-specificity of CPD formation using approximately 200 million aligned, divergent reads with single nucleotide gaps. The most frequent sequence contexts considering the base flanking the dimerized pyrimidines at the 3' side were 5'TTT and 5'TTA and these were followed by 5'TCT and 5'TCA (Fig. 3A). Thus, T or A are the preferred bases 3' to the most abundant CPDs (5'TT and 5'TC). The most highly enriched trinucleotide sequences considering the 5' neighboring base of the CPDs were 5'TTT and 5'CTT, followed by 5'CTC and 5'TTC (Fig. 3B) showing a preference for having another

pyrimidine 5' to a pyrimidine dimer. Divergent reads with single base gaps were >300 times less frequent in non-treated (NT) cells than in UVB-irradiated cells (Fig. 3). There was no particular enrichment of sequencing reads associated with such trinucleotide sequences in the non-UVB-treated control samples (Fig. 3A, 3B, bottom panels).

We then analyzed a potential tetranucleotide sequence context for CPD formation (Fig. S4). For the two bases 5' of the CPDs (5'NNPy<>Py), we observed the most frequent contexts as 5'TTTT, 5'ATTT and 5'CTTT (Fig. S4A). For the two bases 3' of the CPDs (5'Py<>PyNN), we observed the most frequent contexts as 5'TTTA, 5'TTAA and 5'TTTT (Fig. S4B), again reflecting a preference of T or A as the base neighboring a CPD at the 3' side. It should be noted that 5'CG sequences are 5-6-fold underrepresented in mammalian genomes and this underrepresentation has not been corrected in the current analysis.

To identify possible broader sequence motifs for CPD formation, we obtained a base enrichment on twenty nucleotide regions around the gapped base-pair (Fig. 3C). Position 11 represents the nucleotide in the gap between divergent reads (the 5' pyrimidine of the CPD). We observed a dominant enrichment of 5'TT or 5'TC contexts at positions 11 and 12 representing the pyrimidine dimers. In this motif, T or C were enriched in position 10 and T and A bases were enriched in position 13. Surprisingly, however, no further sequence enrichment was observed outside of this tetranucleotide context (positions 10 to 13). Our findings further define the enrichment of specific trinucleotide and tetranucleotide sequences as most susceptible to CPD formation in UVB-irradiated cells and define a new tetranucleotide sequence preference for this type of UV damage. This data confirms the high sensitivity and specificity of circle-damage-seq.

We performed sequencing of up to 0.8 billion paired-end reads from a single circle-damage-seq library from cells irradiated with 1000 J/m^2 of UVB. Figure 4 shows increasing read density as a function of the number of reads obtained. Next, we examined the global distribution of CPDs detected by circle-damage-seq in the mouse genome. We observed that CPD formation in cells is chiefly uniform along the genome and is largely dependent on DNA sequence where 5'TY sequences show the strongest accumulation of CPDs.

Interestingly, we observed low levels of CPD formation at many gene promoter regions (Fig. 5 and Fig. S5). Promoter regions are often G+C-rich and provide less opportunity for T-containing CPDs to form. Indeed, plotting of the G+C content of genes shows a pronounced increase near transcription start sites, which is reverse to the CPD plot (Fig. 5B, 5C). Examples of specific genes are shown in Figure 5D and in Figure S5. In addition to sequence-based reduction of CPDs at promoters, this finding is also consistent with previous observations reported a number of years ago for single gene promoters for which we showed that transcription factors most often repress CPD formation at many sites within promoter regions [14,36], and as also seen in more recent reports [25,26,35,37,38].

The meta-gene profile of CPD formation along all genes in the mouse genome not only shows an overall dip of CPDs near the transcription start sites (TSS) but reveals a striking spike of CPD formation near the transcription end sites (TES) of genes (Fig. 5; Fig. S6). It is likely that the enhanced susceptibility of these regions to CPD formation is related to DNA sequence features near the TES. A preferred TT-containing CPD sequence context is inherent to human and mouse polyadenylation signals. The two preferred sequences are 5'AAUAAA, which is present in almost 60% of mouse genes and the second most preferred one is 5'AUUAAA, which is found in ~16% of mouse genes [39] (Fig. 5E). Sequences surrounding these motifs are additionally T/U-rich and include U-rich or U/G-rich elements that are located 20–40 nt downstream of the cleavage site. There is also widespread alternative polyadenylation in human and mouse genes [39].

Genome-wide mapping of etheno-dA adducts using the circle-damage-seq method

To demonstrate that circle-damage-seq can be applied to a variety of DNA adducts, we mapped the distribution of 1,N^6-etheno-deoxyadenosine (etheno-dA) adducts in 4-hydroxynonenal (4-HNE) treated human liver cells. As one of the major lipid peroxidation products, the 4-HNE aldehyde forms exocyclic etheno-dA DNA adducts, which are known to have mutagenic and carcinogenic properties [40-42]. These DNA adducts are detectable in human liver using sensitive mass spectrometry-based methods and have therefore been linked to liver carcinogenesis [43]. However, no methods are currently available to map etheno-dA adduct positions in the genome at single-nucleotide resolution.

We treated an immortalized human hepatocyte cell line, THLE-2, with 40 μM 4-HNE for 10 minutes to induce etheno-dA DNA adducts. We isolated the genomic DNA and fragmented it using NlaIII restriction enzyme, which cleaves at 5'CATG sequences and leaves a sticky DNA end to enable effective circularization of DNA fragments. As shown in Figure 6A, to create double-strand breaks at these DNA lesions, we used the monofunctional DNA glycosylase, hAAG that cleaves the *N*-glycosidic bond of etheno-dA [44]. After hAAG incision, we used APE1 plus an AP lyase to create the adduct-specific nucleotide gaps and followed this treatment with S1 nuclease digestion to generate a ligatable double-strand break at the etheno-dA sites within the circularized DNA. The break points were labeled with an Illumina sequencing adaptor, and the library was sequenced with paired-end reads to obtain about 15 million reads.

The paired sequencing reads were aligned to the hg19 human reference genome. We observed evenly distributed divergent paired reads genome-wide from the 4-HNE treated liver cell samples, but we hardly found any divergent reads from control cell samples not treated with 4-HNE (Fig. 6B). In addition to obtaining the expected divergent read pattern, our bioinformatics analysis showed that the vast majority (96.5%) of the reference DNA bases matching to the single base gaps between divergent reads were adenines or thymines (depending on the targeted DNA strand) (Fig. 6C and 6D), as expected from etheno-dA DNA adducts. This data further confirms that circle-damage-seq is a highly sensitive technique applicable for genome-wide mapping of various DNA lesions at single-nucleotide resolution whenever a suitable DNA repair enzyme is available to specifically excise the lesion.

We then analyzed the sequence context in which the etheno-dA adducts are formed. The logo plots confirm the prevalence of A or T as the gapped nucleotide (Fig. 6). This base pair is also enriched moderately in the non-treated (NT) sample, although there were very few such gapped, divergent reads. Potentially, this low background may arise from rare endogenous 3-methyladenine bases, which are also excised by AAG DNA glycosylase [45]. Examining the 5' and 3' nearest neighbor bases in the single-strand sequence context of adenines, we noticed an enrichment of T as the 5' base and of C or G as the 3' base, which suggests that 5'T<u>A</u>C/G is a preferred sequence context for etheno-dA DNA adduct formation in human cells.

DISCUSSION

We describe a new method that can be used to analyze modified DNA bases in mammalian cells at the single nucleotide level of resolution and genome-wide. The method depends on the ability to convert the modified base into a DNA single-strand break. Cleavage with S1 nuclease on the opposite DNA strand creates an adduct-specific ligatable double strand break. DNA glycosylase enzymes which operate at the initial step of the base excision repair pathway [46] can easily be adapted for this technique. Monofunctional DNA glycosylase cleave the N-glycosidic bond of the modified base and leave behind an AP site, which is further cleaved by AP endonuclease in cells. Bifunctional DNA glycosylases additionally cleave the AP site through their AP lyase activity after removal of the modified base [47]. The combined cleavage of DNA lesions with DNA glycosylases, APE1, AP lyase and S1 nuclease creates ligatable ends and a single nucleotide gap, which is then assigned to the modified base after genome sequencing. A variety of DNA glycosylases may be employed for circle-damage-seq. Pyrimidine dimer specific DNA glycosylase and AAG protein were employed here to map CPDs and etheno-dA adducts, respectively. The oxidized guanine-specific enzyme OGG1 could be used for mapping of 8-oxo-deoxyguanosine in DNA (8-oxo-dG). DNA glycosylases of the NEIL family, e.g. NEIL1, may also be used for mapping of 8-oxo-dG. These enzymes additionally recognize other oxidized guanine derivatives including spiroiminodihydantoin and guanidinohydantoin [48,49]. Oxidized pyrimidine bases may be excised by NTH family proteins. Uracil arising in DNA from cytosine deamination could be excised and mapped using uracil DNA glycosylases.

It should also be possible to adopt the nucleotide excision repair (NER) complex for mapping of DNA adducts, in particular bulky adducts, which are excellent substrates for this repair pathway. When the NER system of *E. coli* (UvrABC complex) is used, the incisions on the damaged DNA occur, dependent on the exact lesion to be processed, approximately seven bases 5’ to the lesion and four bases 3’ to the lesion [50,51]. In some

cases, for example for excision of benzo[*a*]pyrene-derived N2-guanine adducts, the positions of the excision are well defined [52], which should permit base-resolution mapping of these DNA adducts. The longer single-stranded gap remaining after excision with NER enzymes and removal of the damaged fragment should be a good substrate for S1 nuclease leading to an 11-nucleotide gap of each divergent read after circle-damage-seq.

Modified DNA bases not only may arise from DNA damage but they can represent endogenous modifications with an epigenetic function. Examples are the *N6*-methyladenine bases, which we have mapped here at DpnI sites (Fig. S1). Other important modified bases are derived from cytosine or from 5-methylcytosine (5mC) and include a series of oxidized 5mC bases in the form of 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [53,54]. There are restriction enzymes that can recognize and distinguish these modified cytosine bases and could be used for their mapping. One example is the enzyme AbaSI, which recognizes glucosylated 5hmC with high specificity when compared to 5mC or C, and cleaves the target sequences at a narrow range of distances away from the recognized 5hmC [55]. The rare DNA bases 5fC and 5caC, which are thought to be intermediate bases in DNA demethylation reactions initiated by the TET family of 5mC oxidase proteins, can be removed from DNA by thymine DNA glycosylase (TDG) [53,56,57]. Therefore, TDG could be used to map 5fC and 5caC at base resolution using circle-damage-seq.

The advantages of circle-damage-seq over other DNA damage mapping techniques are several-fold. (i) The single incision point of the modified base and the divergent reads emanating from this break position give a clear indication where the damage is located. (ii) Background signals are minimized by removal of DNA molecules that did not undergo circularization. (iii) Single strand breaks that also contribute to background are removed during circular ligation. (iv) The method represents adduct-specific sequencing because DNA circles or linear molecules without adducts are not sequenced, thus contributing to lower background and lower sequencing costs.

We have used circle-damage-seq for deriving information about the precise DNA sequence contexts of CPDs and etheno-dA adducts in exposed cells. This goal can easily be achieved with relatively low to moderate sequencing depth. However, for most DNA adducts it is still challenging to achieve a saturating sequencing density, meaning that all

base positions of the genome are covered by enough reads to quantitate the base damage at each nucleotide position. Even when we used 0.8 billion reads for mapping of CPDs, we have not achieved this required depth. The reason is that dipyrimidines, the targets of CPD formation, occur on average at every 4th base position on each DNA strand giving rise to approximately 800 million theoretical CPD positions per strand and haploid mammalian genome. It is likely that DNA sequencing costs will further decline in the future which will make the circle-damage-seq technology more generally affordable. When such high sequencing depth is achieved, it should also become feasible to follow precise DNA repair kinetics at every damaged base position of the genome over time and create high resolution DNA repair maps for different types of damaged bases.

Even with the current sequencing depth, we were able to clearly define the nearest neighbor sequence specificity of CPDs and observed that it is limited to a tetranucleotide context, which had not been appreciated previously. The formation of excited states in DNA by UV radiation is a key mechanism leading to photochemical base dimerization [58]. These excited states can be delocalized over two or more stacked DNA bases and can be of either Frenkel exciton ($\pi\pi^*$) type, of a charge transfer type or represent a mix of both [59-61]. The degree of sequence-specific base delocalization of excited states as well as the exact mechanisms how excited states undergo relaxation are still unclear, but our data suggest that delocalization does not extend beyond a tetranucleotide range.

Furthermore, we observe a pronounced reduction of CPD formation near the TSS and promoter regions of many genes (Fig. 5). The lower UV damage susceptibility of these critical gene control regions may be related to at least three phenomena: (i) DNA sequences at the TSS and within promoters are often G+C-rich, which provides a lower frequency of 5'TT or 5'TC sequences compared to the rest of the genome. These sequences are also generally unmethylated and less prone to CPD formation at 5-methylcytosine bases [34,62] (ii) Although a few transcription factors can enhance CPD formation when bound to DNA [14,25,35,38], most of these factors actually protect the DNA from forming pyrimidine dimers, presumably because the bound proteins provide more rigidity to their target DNA thus preventing the needed flexibility to form pyrimidine dimers [14,25,26,36,37]. Perhaps multi-protein complexes consisting of general transcription factors involved in transcription initiation will also contribute to protection of DNA from UV

damage. (iii) Sequences near the TSS are some of the most rapidly repaired regions of the genome [63]. Although we isolated DNA immediately after irradiation, some repair may have already taken place. The exclusion of CPDs from these critical gene control regions should permit the maintenance of functional gene expression programs in UV-exposed cells. We also observed a distinct enhancement of CPD formation at the transcription end sites (Fig. 5). This enhanced damage formation, most likely due to an A/T-rich sequence context near polyadenylation sites, might represent a barrier to transcript processing and could impact cell survival if the damage is not effectively repaired. This phenomenon may contribute to UVB-induced cytotoxicity.

The circle-damage-seq method will allow the definition of precise sequence contexts for DNA adduct formation, even at moderate sequencing depth, as we have shown here for CPDs and etheno-dA adducts. The comparison of these sequence contexts, once defined for a variety of mutagenic DNA adducts, with cancer genome sequencing data from hundreds or thousands of patients should be useful to identify candidate agents that may have an etiological link to specific human malignancies.

REFERENCES

1. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* **15**, 585-98 (2014).
2. Martincorena, I. & Campbell, P.J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-9 (2015).
3. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
4. Pfeifer, G.P. Mechanisms of UV-induced mutations and skin cancer. *Genome Instability & Disease* **1**, 99-113 (2020).
5. Pfeifer, G.P. How the environment shapes cancer genomes. *Curr Opin Oncol* **27**, 71-7 (2015).
6. Alexandrov, L.B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622 (2016).
7. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
8. Rosenquist, T.A. & Grollman, A.P. Mutational signature of aristolochic acid: Clue to the recognition of a global disease. *DNA Repair (Amst)* **44**, 205-11 (2016).
9. Kucab, J.E. *et al.* A Compendium of mutational signatures of environmental agents. *Cell* **177**, 821-836 e16 (2019).
10. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246-59 (2013).
11. Hollstein, M., Alexandrov, L.B., Wild, C.P., Ardin, M. & Zavadil, J. Base changes in tumour DNA have the power to reveal the causes and evolution of cancer. *Oncogene* **36**, 158-167 (2017).
12. Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis* **30**, 763-70 (2015).
13. Denissenko, M.F., Pao, A., Tang, M. & Pfeifer, G.P. Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* **274**, 430-2 (1996).
14. Pfeifer, G.P., Drouin, R., Riggs, A.D. & Holmquist, G.P. Binding of transcription factors creates hot spots for UV photoproducts in vivo. *Mol Cell Biol* **12**, 1798-804 (1992).
15. Pfeifer, G.P., Drouin, R., Riggs, A.D. & Holmquist, G.P. In vivo mapping of a DNA adduct at nucleotide resolution: detection of pyrimidine (6-4) pyrimidone photoproducts by ligation-mediated polymerase chain reaction. *Proc Natl Acad Sci U S A* **88**, 1374-8 (1991).
16. Li, M., Ko, T. & Li, S. High-resolution Digital Mapping of N-Methylpurines in Human Cells Reveals Modulation of Their Induction and Repair by Nearest-neighbor Nucleotides. *J Biol Chem* **290**, 23148-61 (2015).
17. Garcia-Nieto, P.E. *et al.* Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J* **36**, 2829-2843 (2017).
18. Zavala, A.G., Morris, R.T., Wyrick, J.J. & Smerdon, M.J. High-resolution characterization of CPD hotspot formation in human fibroblasts. *Nucleic Acids Res* **42**, 893-905 (2014).

19. Hu, J., Adar, S., Selby, C.P., Lieb, J.D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev* **29**, 948-60 (2015).
20. Hu, J. *et al.* Genome-wide mapping of nucleotide excision repair with XR-seq. *Nat Protoc* **14**, 248-282 (2019).
21. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10**, 361-5 (2013).
22. Canela, A. *et al.* DNA breaks and end resection measured genome-wide by end sequencing. *Mol Cell* **63**, 898-911 (2016).
23. Li, W. & Sancar, A. Methodologies for detecting environmentally induced DNA damage and repair. *Environ Mol Mutagen* **61**, 664-679 (2020).
24. Mingard, C., Wu, J., McKeague, M. & Sturla, S.J. Next-generation DNA damage sequencing. *Chem Soc Rev* (2020).
25. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat Commun* **9**, 2626 (2018).
26. Mao, P., Smerdon, M.J., Roberts, S.A. & Wyrick, J.J. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **113**, 9057-62 (2016).
27. Tsai, S.Q. *et al.* CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat Methods* **14**, 607-614 (2017).
28. Wu, T.P. *et al.* DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329-33 (2016).
29. O'Brown, Z.K. *et al.* Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**, 445 (2019).
30. Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B. & Nestor, C.E. No evidence for DNA N (6)-methyladenine in mammals. *Sci Adv* **6**, eaay3335 (2020).
31. Marinus, M.G. & Morris, N.R. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J Bacteriol* **114**, 1143-50 (1973).
32. Radany, E.H. & Friedberg, E.C. A pyrimidine dimer-DNA glycosylase activity associated with the v gene product of bacterophage T4. *Nature* **286**, 182-5 (1980).
33. Cadet, J., Grand, A. & Douki, T. Solar UV radiation-induced DNA Bipyrimidine photoproducts: formation and mechanistic insights. *Top Curr Chem* **356**, 249-75 (2015).
34. Lindberg, M., Bostrom, M., Elliott, K. & Larsson, E. Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light. *Proc Natl Acad Sci U S A* **116**, 20411-20417 (2019).
35. Premi, S. *et al.* Genomic sites hypersensitive to ultraviolet radiation. *Proc Natl Acad Sci U S A* **116**, 24196-24205 (2019).
36. Tornaletti, S. & Pfeifer, G.P. UV light as a footprinting agent: modulation of UV-induced DNA damage by transcription factors bound at the promoters of three human genes. *J Mol Biol* **249**, 714-28 (1995).
37. Hu, J., Adebali, O., Adar, S. & Sancar, A. Dynamic maps of UV damage formation and repair for the human genome. *Proc Natl Acad Sci U S A* **114**, 6758-6763 (2017).

38. Elliott, K. *et al.* Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLoS Genet* **14**, e1007849 (2018).
39. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**, 201-12 (2005).
40. Basu, A.K., McNulty, J.M. & McGregor, W.G. Solution conformation and mutagenic specificity of 1,N6-ethenoadenine. *IARC Sci Publ*, 325-33 (1999).
41. Lee, S.H., Oe, T. & Blair, I.A. 4,5-Epoxy-2(E)-decenal-induced formation of 1,N(6)-etheno-2'-deoxyadenosine and 1,N(2)-etheno-2'-deoxyguanosine adducts. *Chem Res Toxicol* **15**, 300-4 (2002).
42. Sodum, R.S. & Chung, F.L. Stereoselective formation of in vitro nucleic acid adducts by 2,3-epoxy-4-hydroxynonanal. *Cancer Res* **51**, 137-43 (1991).
43. Linhart, K.B. *et al.* The generation of carcinogenic etheno-DNA adducts in the liver of patients with nonalcoholic fatty liver disease. *Hepatobiliary Surg Nutr* **4**, 117-23 (2015).
44. Saparbaev, M., Kleibl, K. & Laval, J. Escherichia coli, Saccharomyces cerevisiae, rat and human 3-methyladenine DNA glycosylases repair 1,N6-ethenoadenine when present in DNA. *Nucleic Acids Res* **23**, 3750-5 (1995).
45. Engelward, B.P. *et al.* Base excision repair deficient mice lacking the Aag alkyladenine DNA glycosylase. *Proc Natl Acad Sci U S A* **94**, 13087-92 (1997).
46. Beard, W.A., Horton, J.K., Prasad, R. & Wilson, S.H. Eukaryotic Base Excision Repair: New Approaches Shine Light on Mechanism. *Annu Rev Biochem* **88**, 137-162 (2019).
47. Krokan, H.E., Standal, R. & Slupphaug, G. DNA glycosylases in the base excision repair of DNA. *Biochem J* **325 (Pt 1)**, 1-16 (1997).
48. Hailer, M.K., Slade, P.G., Martin, B.D., Rosenquist, T.A. & Sugden, K.D. Recognition of the oxidized lesions spiroiminodihydantoin and guanidinohydantoin in DNA by the mammalian base excision repair glycosylases NEIL1 and NEIL2. *DNA Repair (Amst)* **4**, 41-50 (2005).
49. Hazra, T.K. *et al.* Oxidative DNA damage repair in mammalian cells: a new perspective. *DNA Repair (Amst)* **6**, 470-80 (2007).
50. Kisker, C., Kuper, J. & Van Houten, B. Prokaryotic nucleotide excision repair. *Cold Spring Harb Perspect Biol* **5**, a012591 (2013).
51. Sancar, A. Mechanisms of DNA excision repair. *Science* **266**, 1954-6 (1994).
52. Tang, M.S., Zheng, J.B., Denissenko, M.F., Pfeifer, G.P. & Zheng, Y. Use of UvrABC nuclease to quantify benzo[a]pyrene diol epoxide-DNA adduct formation at methylated versus unmethylated CpG sites in the p53 gene. *Carcinogenesis* **20**, 1085-9 (1999).
53. He, Y.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-7 (2011).
54. Pfeifer, G.P., Szabo, P.E. & Song, J. Protein Interactions at Oxidized 5-Methylcytosine Bases. *J Mol Biol* (2019).
55. Sun, Z. *et al.* High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep* **3**, 567-76 (2013).

56. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334-8 (2011).
57. Hashimoto, H., Hong, S., Bhagwat, A.S., Zhang, X. & Cheng, X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res* **40**, 10203-14 (2012).
58. Banyasz, A. *et al.* Electronic excited states responsible for dimer formation upon UV absorption directly by thymine strands: joint experimental and theoretical study. *J Am Chem Soc* **134**, 14834-45 (2012).
59. Schreier, W.J., Gilch, P. & Zinth, W. Early events of DNA photodamage. *Annu Rev Phys Chem* **66**, 497-519 (2015).
60. Takaya, T., Su, C., de La Harpe, K., Crespo-Hernandez, C.E. & Kohler, B. UV excitation of single DNA and RNA strands produces high yields of exciplex states between two stacked bases. *Proc Natl Acad Sci U S A* **105**, 10285-90 (2008).
61. Markovitsi, D. UV-induced DNA Damage: The Role of Electronic Excited States. *Photochem Photobiol* **92**, 45-51 (2016).
62. You, Y.H. & Pfeifer, G.P. Similarities in sunlight-induced mutational spectra of CpG-methylated transgenes and the p53 gene in skin cancer point to an important role of 5-methylcytosine residues in solar UV mutagenesis. *J. Mol. Biol.* **305**, 389-399 (2001).
63. Tu, Y., Tornaletti, S. & Pfeifer, G.P. DNA repair domains within a human gene: selective repair of sequences near the transcription initiation site. *EMBO J* **15**, 675-83 (1996).
64. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
65. Faust, G.G. & Hall, I.M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503-5 (2014).
66. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
67. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 1-34 (2014).
68. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718-9 (2011).
69. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645-3647 (2017).

METHODS

Cell culture

Mouse embryonic fibroblast (MEF) cells from C57BL/6 mouse embryos were obtained from ATCC (Cat. SCRC-1008), grown in DMEM high glucose (Invitrogen; Carlsbad, CA) containing 15% fetal bovine serum and used at passage number <5. THLE-2 cells derived from primary human liver epithelial cells were obtained from ATCC (Cat. CRL-2706) and grown in BEGM medium (Lonza; Walkersville, MD) supplemented with 10% fetal bovine serum, 5 ng/ml EGF and 70 ng/ml phosphoethanolamine on the culture plates precoated with a mixture of 0.01 mg/ml fibronectin, 0.03 mg/ml collagen type I and 0.01 mg/ml BSA.

UVB irradiation, 4-HNE treatment and genomic DNA isolation

To irradiate cells with ultraviolet B (UVB) light, MEF cells at 80-90% confluence on 10 cm culture plates were washed with 1x PBS and irradiated in PBS with various doses (600-1500 J/m^2) of UVB. The UVB lamp (ThermoFisher Scientific, UVP 3UV lamp) has a peak spectral emission at 302 nm. The UVB dose was determined using a UVX radiometer with a UVB probe (Ultraviolet Products; Upland, CA). To induce 1,N^6-etheno-deoxyadenosine (etheno-dA) adducts in human cells, the THLE-2 liver cells at 80-90% confluence on 10 cm culture plates were washed with 1x PBS and treated with 40 μM of 4-HNE (Cayman; Ann Arbor, MI) in PBS for 10 min at 37°C.

After the irradiation or chemical treatment, the cells were immediately trypsinized and collected, and then genomic DNA was isolated using Quick-DNA Miniprep Plus Kit (Zymo Research; Irvine, CA) according to the manufacturer's instruction manual.

Circle-damage-seq mapping of CPDs

DNA preparation and circularization

To prepare fragmented genomic DNA, the DNA from UVB-irradiated cells was sheared to an average length of 300 bp by sonication with a Covaris E220 sonicator (Covaris; Woburn, MA) under the following conditions: peak incident power (W), 140; duty factor, 10%; cycles per burst, 200 times, 80 seconds, and then the fragmented DNAs were purified with DNA Clean & Concentrator-5 Kit (Zymo Research, DCC-5) according to the

manufacturer's protocol. DNA was eluted in 42 µl of 10 mM Tris-HCl pH 7.5. Next, the sonicated DNA was end-repaired to prepare blunt-ended DNA. One microgram of the sonicated DNA was incubated in 1x T4 DNA ligase buffer (NEB; Ipswich, MA) with components; 50 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, 1 mM ATP and 10 mM DTT and with 1 µl (2 U/µl) of RNase H (NEB), 1 µl (3 U/µl) of T4 DNA polymerase (NEB), 1 µl (10 U/µl) of T4 polynucleotide kinase (NEB) and 1 µl of 10 mM dNTPs in a final volume of 50 µl at 24°C for 30 min, and then the reaction was heat-inactivated by further incubation at 75°C for 20 min. The DNA was purified using DCC-5 Kit (Zymo Research) and eluted in 100 µl of 10 mM Tris-HCl, pH 7.5. The purified DNA was quantitated using Nanodrop (Thermo Fisher Scientific).

To circularize the DNA, 1 µg of the blunt-ended DNA was further incubated in 1x T4 DNA ligase buffer (NEB) in 1x T4 DNA ligase buffer (NEB) with 4 µl (400 U/µl) of T4 DNA ligase (NEB) in a final volume of 200 µl at 16°C overnight. Then, the reaction mixture was cleaned up with DCC-5 kit and eluted in 42 µl of 10 mM Tris-HCl, pH 7.5. To remove non-circularized linear DNA from the circularized DNA pool, the ligase-treated DNA was incubated with 2 µl (10 U/µl) of Plasmid-Safe ATP-Dependent DNase (Lucigen; Middleton, WI) in 1x reaction buffer (33 mM Tris-acetate, pH 7.5, 66 mM potassium acetate, 10 mM magnesium acetate, 0.5 mM DTT and 1 mM ATP) in a final volume of 50 µl at 37°C for 30 min and then at 70°C for 20 min to stop the reaction. Then, the DNA was cleaned up with 90 µl (1.8x) of AMPure XP beads (Beckman Coulter; Indianapolis, IN) and eluted in 35 µl of 10 mM Tris-HCl, pH 8.0.

Cleavage of DNA at CPD sites

To generate double-strand breaks at CPD sites, the circularized DNA was treated with the following DNA repair enzymes. First, to specifically incise DNA at CPD positions, the circularized DNA was incubated in 1x NEBuffer 4 reaction buffer (NEB) (50 mM potassium acetate, 20 mM Tris-acetate, pH 7.9, 10 mM magnesium acetate, 1 mM DTT) and 0.4 µl of 20 mg/ml BSA with 1 µl (10 U/µl) of T4 endonuclease V (NEB, T4-PDG) and 1 µl (10 U/µl) of APE1 (NEB) in a final volume of 40 µl at 37°C for 20 min. Then, the DNA was cleaned up with 72 µl (1.8x) of AMPure XP beads and eluted in 25 µl of 10 mM Tris-HCl, pH 7.5. To revert the dimerized pyrimidines remaining after T4-PDG incision, the DNA

from above was treated with 3 µl (0.25 µg/µl) of *E. coli* phrB photolyase (Novus Biologicals; Centennial, CO) in 1x reaction buffer (50 mM Tris-HCl, pH 7.0, 50 mM NaCl and 10 mM DTT) in a final volume of 50 µl. The reaction tube was placed at a distance of 15 cm under a UVA lamp (ThermoFisher Scientific) that has a peak spectral emission at 365 nm and incubated under UVA illumination for 90 min at room temperature. Then, the DNA was cleaned up with 90 µl (1.8x) of AMPure XP beads and eluted in 32 µl of 10 mM Tris-HCl, pH 8.0. To cleave the opposite strand of the nicked DNA, the DNA from above was incubated with 1 µl (5 U/µl) of single-strand specific S1 nuclease (ThermoFisher Scientific) in 1x reaction buffer (40 mM sodium acetate, pH 4.5, 300 mM NaCl and 2 mM $ZnSO_4$) in a final volume of 40 µl for 4 min at room temperature. We stopped the reaction by adding 2 µl of 0.5 M EDTA and 1 µl of 1M Tris-HCl, pH 8.0 to the reaction mixture and further incubated the mixture for 10 min at 70°C. The DNA sample was cleaned up with 72 µl (1.8x) of AMPure XP beads and eluted in 48 µl of 10 mM Tris-HCl, pH 8.0.

Adapter ligation

The double-strand cleaved DNA was subsequently A-tailed in 1x NEBNext dA-tailing reaction buffer (NEB) (10 mM Tris-HCl, pH 7.9, 10 mM $MgCl_2$, 50 mM NaCl, 1 mM DTT) and 0.2 mM dATP, with 2 µl (5 U/µl) of Klenow fragment exo- (NEB) in a final volume of 55 µl by incubation at 37°C for 30 min, and then the reaction was heat-inactivated by further incubation at 70°C for 30 min. To perform adaptor ligation, 30 µl of NEBNext Ultra II Ligation Master Mix, 1 µl of NEBNext Ligation Enhancer (NEB), and 2 µl of 1.5 µM NEBNext adaptors for Illumina sequencing (NEB): 5'-phos-GATCGGAAGAGCAC-ACGTCTGAACTCCAGTC/ideoxyU/ACACTCTTTCCTACACGACGCTCTTCCGATC*T-3' and 5'-phos-GATCGGAAGAGCACACGTCTGAACTCCAGTC/ideoxyU/AC-ACTCTTTCCTACACGACGCTCTTCCGATC*C-3' (*; phosphorothioate linkage) were added to the dA-tailing reaction mixture in a total volume of 93 µl and incubated at 20°C for 15 min, and followed by incubation with 3 µl (1 U/µl) of USER enzyme (NEB) at 37°C for 20 min. Then the adaptor ligated circle-damage-seq library was cleaned up with 96 µl (1x) of AMPure XP beads and size-selected (300 to 1,000 bp) with 0.5 volumes of AMPure XP beads and eluted in 25 µl of 10 mM Tris-HCl, pH 8.0.

Library amplification and sequencing

To amplify the circle-damage-seq library by PCR, the eluted library was mixed with 25 µl of NEBNext Ultra II Q5 Master mix (2x), 2.5 µl of 10 µM NEBNext i5 and 2.5 µl of 10 µM NEBNext i7 primers (NEB) in a final volume of 50 µl. The amplification reaction mixture was incubated at 98°C for 30 sec, and then 12 cycles of PCR at 98°C for 10 sec and 65°C for 75 sec were performed, followed by a final extension step at 65°C for 5 min. The PCR product was purified with 50 µl (1x) of Ampure XP beads and eluted in 20 µl of 10 mM Tris-HCl, pH 8.0. The purified library was quantitated using a Qubit 3.0 fluorometer and ds DNA HS Assay kit (Thermo Fisher Scientific). The size distribution of the circle-damage-seq library was determined using Bioanalyzer (Agilent; Santa Clara, CA), and the circle-damage-seq library was quantified using KAPA Library Quantification kit (Kapa Biosystems) according to the manufacturer's instruction. The circle-damage-seq library was sequenced as 150 bp paired-end sequencing runs on an Illumina Hiseq2500 platform to obtain between 10 million and 800 million reads. From the 800 million read sample, approximately 200 million paired aligned reads were retained specifically as divergent read pairs with single base gaps after removal of duplicates.

Circle-damage-seq mapping of etheno-dA adducts

To prepare fragmented DNAs, NlaIII restriction enzyme (NEB) that leaves 5'CATG overhangs was used for digestion of the 4-HNE treated genomic DNA. We initially optimized the enzyme digestion to produce an average length of 300 bp, and 2 µg of genomic DNA was treated with 1 µl (5 U/µl) of NlaIII in a total volume of 50 µl at 37°C for 2 min. The sheared DNA was purified using DCC-5 Kit and eluted in 100 µl of 10 mM Tris-HCl, pH 7.5. The purified DNA was quantitated using Nanodrop.

To circularize the sheared DNA, 1 µg of NlaIII-digested, sticky-ended DNA was incubated in 1x T4 DNA ligase buffer with 4 µl (400 U/µl) of T4 DNA Ligase (NEB) in a final volume of 200 µl at 16°C overnight. Then, the reaction mixture was cleaned up with DCC-5 kit and eluted in 42 µl of 10 mM Tris-HCl, pH 7.5. To remove non-circularized linear DNA from the circularized DNA pool, the ligase treated DNA was incubated with 3 µl (10 U/µl) of Plasmid-Safe ATP-Dependent DNase in a final volume of 50 µl at 37°C for

30 min and then at 70°C for 20 min to stop the reaction. Then, the DNA was cleaned up with 90 µl (1.8x) of AMPure XP beads and eluted in 27 µl of 10 mM Tris-HCl, pH 8.0.

To generate double-strand breaks at etheno-dA adduct sites, the circularized DNA was initially treated with human alkyl adenine DNA glycosylase (hAAG, NEB) to cleave the *N*-glycosidic bond of etheno-dA bases. The circularized DNA was incubated in 1x ThermoPol Reaction buffer (provided by NEB for the hAAG reaction) with components; 20 mM Tris-HCl, pH 8.8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$ and 0.1% Triton X-100, with 1 µl (10 U/µl) of hAAG in a final volume of 30 µl at 37°C for 30 min. Then, the DNA was cleaned up with 54 µl (1.8x) of AMPure XP beads and eluted in 35 µl of 10 mM Tris-HCl, pH 7.5. Subsequently, the hAAG treated DNA was incubated in 1x NEBuffer 4 reaction buffer with 1 µl (10 U/µl) of T4 PDG (AP lyase) and 1 µl (10 U/µl) of APE1 (to double-incise the AP sites) in a final volume of 40 µl at 37°C for 20 min, and the nicked DNA was cleaned up with 72 µl (1.8x) of AMPure XP beads and eluted in 32 µl of 10 mM Tris-HCl, pH 7.5. To cleave the opposite strand of the nicked DNA, the DNA was incubated with 1 µl (5 U/µl) of single-strand specific S1 nuclease in 1x reaction buffer (see above) in a final volume of 40 µl for 4 min at room temperature. Then, the reaction was stopped by adding 2 µl of 0.5 M EDTA and 1 µl of 1M Tris-HCl, pH 8.0 to the reaction mixture. We further incubated the mixture for 10 min at 70°C, followed by purification with 72 µl (1.8x) of AMPure XP beads. We eluted the DNA in 48 µl of 10 mM Tris-HCl, pH 8.0. circle-damage-seq library preparation and amplification by PCR were performed as described above using dA-tailing and ligation of the (5'-phos-CGGTGGACCGATGATC/ideoxyU/ATCGGTCCACCG*T-3') hairpin linker. The circle-damage-seq library for etheno-dA mapping was sequenced using 75 bp paired-end reads on an Illumina NextSeq500 platform to obtain 15 million read.

Circle-damage-seq mapping of DpnI cuts in the *E. coli* genome

As a pilot experiment, we initially performed the mapping of DpnI cut sites to optimize the circle-damage-seq method. To test the sensitivity of circle-damage-seq towards different percentages of DpnI-cleaved sites, we mixed *E. coli* DNAs from two different strains, dam(+) and dam(-) to prepare 0%, 2% and 10% of dam(+) DNA in a background of dam(-) input DNA. We subjected this mixture to the circle-damage-seq procedure as described

above. Briefly, to prepare fragmented DNAs, RsaI and AluI restriction enzymes (NEB) that leave blunt ends were used to digest 1 µg of *E. coli* genomic DNA. The cleaved DNAs were subjected to A-tailing using NEBNext Ultra II End Prep kit (NEB) and followed by ligation of an annealed hairpin adaptor (5'-phos-CGGTGGACCGATGATC/ideoxyU/ATCGGTCCACCG*T-3') using NEBNext Ultra II Ligation Master mix. The adaptor-ligated DNAs were treated with USER enzyme and circularized with T4 DNA ligase followed by treatment with Plasmid-Safe ATP-Dependent DNase. To generate double-strand breaks at DpnI cleavage sites, the circularized *E. coli* DNAs were incubated in 1x CutSmart buffer (NEB) with 1 µl (20 U/µl) of DpnI restriction enzyme (NEB) in a final volume of 20 µl at 37°C for 1 hour. Then, circle-damage-seq library preparation and amplification by PCR were performed as described above. The circle-damage-seq library was sequenced as 75 bp paired-end reads on an Illumina NextSeq500 platform to obtain ~15 million reads.

Bioinformatics analysis

All libraries were hard-clipped from the 3' end to 2x50 nucleotide length using *Trim_Galore* (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Reads were then adapter-trimmed and quality-trimmed (phred <20) in a single pass using *Trim_Galore* default parameters. Trimmed reads were aligned using *BWA-MEM* [64] with default parameters. Duplicate alignments were marked and removed with the 'removeDups' *SAMBLASTER* [65] option. *SAMtools* [66] was then used to remove reads with low mapping quality (MAPQ <20), and to sort reads by genomic coordinates.

Divergently aligned read pairs (pairs facing away from each other) with a single nucleotide gap between reads were selected by selecting SAM records with TLEN of 3. TLEN equals the distance between the mapped end of the template and the mapped start of the template [66]. Records with TLEN=3 represent divergently aligned reads with 1 nucleotide between mates. Sequences for the 3 nucleotides representing the last nucleotide of each read and the gapped base were identified using *BEDtools* [67].

When treating cells with UVB to produce CPDs, if the gapped nucleotide was T or C in the (+) strand, the damage would have occurred on the (+) strand, whereas a gapped A or G in the (+) strand would indicate that the damage must have occurred on the (-)

strand. For 4-HNE treatment producing etheno-dA adducts, if the gapped nucleotide was A in the (+) strand, the damage would have occurred on the (+) strand, whereas a gapped T would indicate that damage must have occurred on the paired A in the (-) strand. Gapped G or C found in libraries produced following 4-HNE treatments provide no inference about the strand in which damage occurred. As seen in Fig. S1, the DpnI-treatment protocol probes 6mA formed at 5'GATC sequences from dam(+) *E.coli* producing sequencing libraries with no gap between divergently aligned read pairs. As such, alignment records with TLEN=2 were selected.

Chromosome name, start position, stop position, and inferred strand (+/-) for all selected read pair loci were written to a BED file, sorted, and indexed with *tabix* [68] for downstream analyses. Logo plots were drawn using ggseqlogo [69].

Data availability

All data are available from the GEO database (submitted on October 20, 2020 with immediate public access).

Code availability

All code for this project is available at https://github.com/deanpettinga/CDseq.

Acknowledgements

This work was supported by NIH grant CA228089 to GPP.

Author contributions

S.G.J. and G.P.P. conceptualized and designed the project. S.G.J. and J.J. performed experiments. D.P., S.G.J. and G.P.P. performed data analysis. G.P. and S.G.J. wrote the manuscript. All authors commented on the manuscript.

Competing interests

The authors declare no competing interests.

FIGURE LEGENDS:

Figure 1:

Outline of circle-damage-sequencing (circle-damage-seq).

A. Genomic DNA is isolated from DNA damaging agent-treated cells and is sheared to obtain average size fragments of 300 base pairs. DNA is then circularized using T4 DNA ligase. Remaining non-circularized DNA fragments are removed using an exonuclease cocktail. The circular DNAs are collected after column-based purification.

B. DNA lesions (indicated by arrows) within a DNA circle are removed by specific DNA base excision repair enzymes producing AP sites, which are then processed by the AP endonuclease, APE1, and an AP lyase. Single-strand- and nick-specific S1 nuclease cleaves the opposite strand of the nicked DNAs and leaves ligatable double-strand breaks with single base gaps. Then the sequencing adaptors are ligated to the DNA ends. Of note, in the circle-damage-seq process, non-damaged circular DNAs are not cleaved and therefore will not be sequenced.

C. During DNA library preparation, DNA molecules opened at the circles after adduct-specific cleavage are ligated to sequencing adaptors and can be selectively enriched by PCR, representing DNA damage-specific sequence enrichment. After paired-end sequencing and alignment to the reference genome, there will be a unique aligned read pattern, divergent paired-end reads with single base gaps (corresponding to the modified base) in the center between forward (blue arrow) and reverse (red arrow) reads. Each divergent read is derived from a single opened circular DNA molecule.

Figure 2:

Genome-wide mapping of CPD lesions and visualization at single nucleotide resolution.

A. CPD lesions at dipyrimidine sequences were induced into the MEF genome by UVB irradiation. To create double-strand breaks at CPDs (e.g., T=T; thymine dimer, red), CPD

specific-DNA glycosylase (T4-PDG) and APE1 coupled enzyme reactions were initially used and followed by *E.coli* photolyase treatment to create ligatable, nicked DNA. To create double strand breaks, S1 nuclease was used to generate a free 5' phosphate and 3' hydroxyl group at cleaved DNA ends, which are ligated and sequenced.

B. Snapshot of the IGV viewer displays the characteristic divergent paired reads of CPD mapping by circle-damage-seq aligned to the mm10 mouse reference genome at single-base resolution (chr11:69,588,587-69,588,647). Red segments represent reads mapped to the plus strand (forward) and blue segments represent reads mapped to the plus strand (reverse). Each paired-end read is derived from a single DNA circle and is displayed as a single divergent pair. Green arrows indicate single-base gaps matching pyrimidine bases on the reference sequence corresponding to the 5' base of a pyrimidine dimer.

C. Snapshot of the IGV viewer represents CPD sites mapped by circle-damage-seq at an aligned region (chr6:114,103,889-114,103,928) at single-base resolution.

Figure 3:

CPD mapping by circle-damage-seq shows CPD sequence specificity.

A. Distribution plot of trinucleotide sequences (5'YYN3') undergoing CPD formation in UVB-irradiated cells (blue). NT, non-treated control (orange).

B. Distribution plot of trinucleotide sequences (5'NYY3') undergoing CPD formation in UVB-irradiated cells (blue). NT, non-treated control (orange).

C. Logo plot shows the accumulated nucleotide composition relative to mapped CPD positions along the reference genome. Position 11 represents the nucleotide in the gap between divergent reads, and position 11 and 12 represent UVB-damaged dipyrimidine sequences undergoing CPD formation. The total height of the stack represents the conservation of that position (measured in bits) using Shannon entropy. Within each stack, the height of each letter represents the relative frequency of that nucleic acid base.

Figure 4:

Coverage comparison of CPD lesions relative to sequencing depth.

The snapshot of the IGV viewer shows the read density maps for 9 kb windows (chr19:56,427,800-56,436,800) of the mm10 mouse reference genome for the *Casp7* gene. Each read depth was obtained from circle-damage-seq libraries prepared with 1000 J/m^2 of UVB dose-irradiated cells. The divergent, paired reads of CPD mapping by circle-damage-seq were displayed as paired blue and red segments. 10 M; 10 million reads, 86 M; 86 million reads, and 0.8 B; 0.8 billion reads.

Figure 5:

Gene-level distribution of UVB-induced CPDs shows reduced levels near the transcription start sites and a spike of CPDs near the transcription end sites.

A. Heat map of CPD distribution along all genes of the mm10 mouse genome. CPD coverage is sorted from high (top) to low (bottom). The CPD signals were mapped and binned in 50 bp windows from 3 kb upstream of the TSS, then normalized relative to gene length over the gene bodes to the TES, and 3 kb downstream of the TES.

B. Profile of CPD distribution along all genes.

C. Profile of the G+C content (%) along all genes (50 bp windows).

D. Example of a typical CPD sequence read distribution along the *Htatip2* gene. This gene shows a reduced frequency of CPDs near the TSS (blue arrow) and a spike near the TES (red arrow).

E. Logo plot of sequences near the transcription end sites of all mouse genes (+/- 50 nucleotides surrounding the transcription end sites (TES). Note the general A/T-richness and the pronounced enhancement of A/T near the polyadenylation signals 10 to 30 nucleotides upstream of the actual transcript end, which is at position 50.

Figure 6:

Genome-wide mapping of etheno-dA adducts using circle-damage-seq.

A. Etheno-dA adducts were induced in the THLE-2 genome by treatment with 40 μM of 4-HNE. To create double-strand breaks at etheno-dA (A with red symbol), the etheno-dA specific-DNA glycosylase hAAG and APE1/AP lyase coupled enzyme reactions were used to create nicked DNA. Then, S1 nuclease was used to generate a free 5' phosphate and 3' hydroxyl group at cleaved DNA ends, which are ligated and sequenced.

B. The snapshot of the IGV viewer displays the divergent, paired reads of etheno-dA mapping by circle-damage-seq. The distribution map shows 6.6 kb windows (chr10:93,940,000-93,946,591) of the hg19 human reference genome. Each paired-end read derived from a single DNA molecule was displayed as a single divergent pair (paired blue and red segments). 4-HNE; 4-HNE treated THLE-2 sample. NT; non-treated control THLE-2 sample showing almost no paired, divergent reads with single base gaps.

C. The snapshot of the IGV viewer represents etheno-dA sites mapped by circle-damage-seq at single-base resolution (chr17:8,302,652-8,302,754). Red segments represent reads mapped to the plus strand (forward) and blue segments represent reads mapped to the plus strand (reverse). Green arrows indicate single-base gaps matching etheno-dA bases in the reference sequence.

D. Percentages of nucleotides positioned in the gap between divergent reads. 'G + C' and 'A + T' indicate the combined percentages of G and C or A and T in the gaps, respectively.

E. Logo plot bit scores of DNA bases modified in 4-HNE-treated cells and in non-treated (NT) controls. This analysis considers both DNA strands.

F. Logo plot bit scores of modified adenine bases in 4-HNE-treated cells and in non-treated (NT) controls. This analysis is specific for adenine bases in the gapped positions and reflects the single strand-specific DNA sequence context.

Supplementary Figures:

Figure S1:

Mapping of N6-methyladenine at GATC sites.

A. The agarose gel shows DpnI cleavage of DNA from dam(+) *E. coli* cells but not from dam(-) cells.

B. Outline of the method to map DpnI cleavage sites. After circularization, DNA is cleaved at methylated GATC sites using DpnI. Linkers are ligated and the opened molecules are sequenced. Divergent, paired reads indicate the positions of methylated adenines.

C. Ratio of cleaved 5'GATC sites relative to 5'NNNN sequences displayed as paired, divergent reads.

D. Example of read stacks at methylated GATC sequences.

Figure S2:

Double-strand cleavage at UVB-induced CPDs.

Cells were irradiated with different doses of UV from a UVB light source. DNA was isolated and cleaved with the indicated enzymes. The cleaved DNA molecules were separated on a non-denaturing agarose gel. Double cleavage with T4 endonuclease V and S1 nuclease leads to dose-dependent double-strand break formation.

Figure S3:

Example of CPD mapping by circle-damage-seq.

A. Divergent reads (shown in red and blue colors for the forward and reverse reads) with single nucleotide gaps indicate the positions of UVB-induced CPDs.

B. Lower resolution view of divergent reads from UVB-treated cells.

Figure S4:

UVB-induced CPD tetranucleotide sequence specificity.

A. Analysis of the two bases 5' to a CPD. The sequence context is 5'NNY<>Y.

B. Analysis of the two bases 3' to a CPD. The sequence context is 5'Y<>YNN.

Figure S5:

Reduced CPD levels at transcription start sites.

These genes with G+C-rich promoters show the typical low CPD frequencies near the transcription start sites (TSS) (blue arrows indicate the direction of transcription). **A.** *Dnmt3a*. **B.** *Mpp6*. **C.** *Plxnd1*. **D.** *Mbd2*.

Figure S6:

High CPD levels near transcription end sites.

These examples show an increased level of CPDs near transcription end sites (TES) (red vertical arrows). **A.** *Zfp800*. **B.** *Clec2l*. **C.** *Zfp777*. **D.** *Hoxb9*. The low levels of CPDs near the TSS are also indicated (blue arrows).

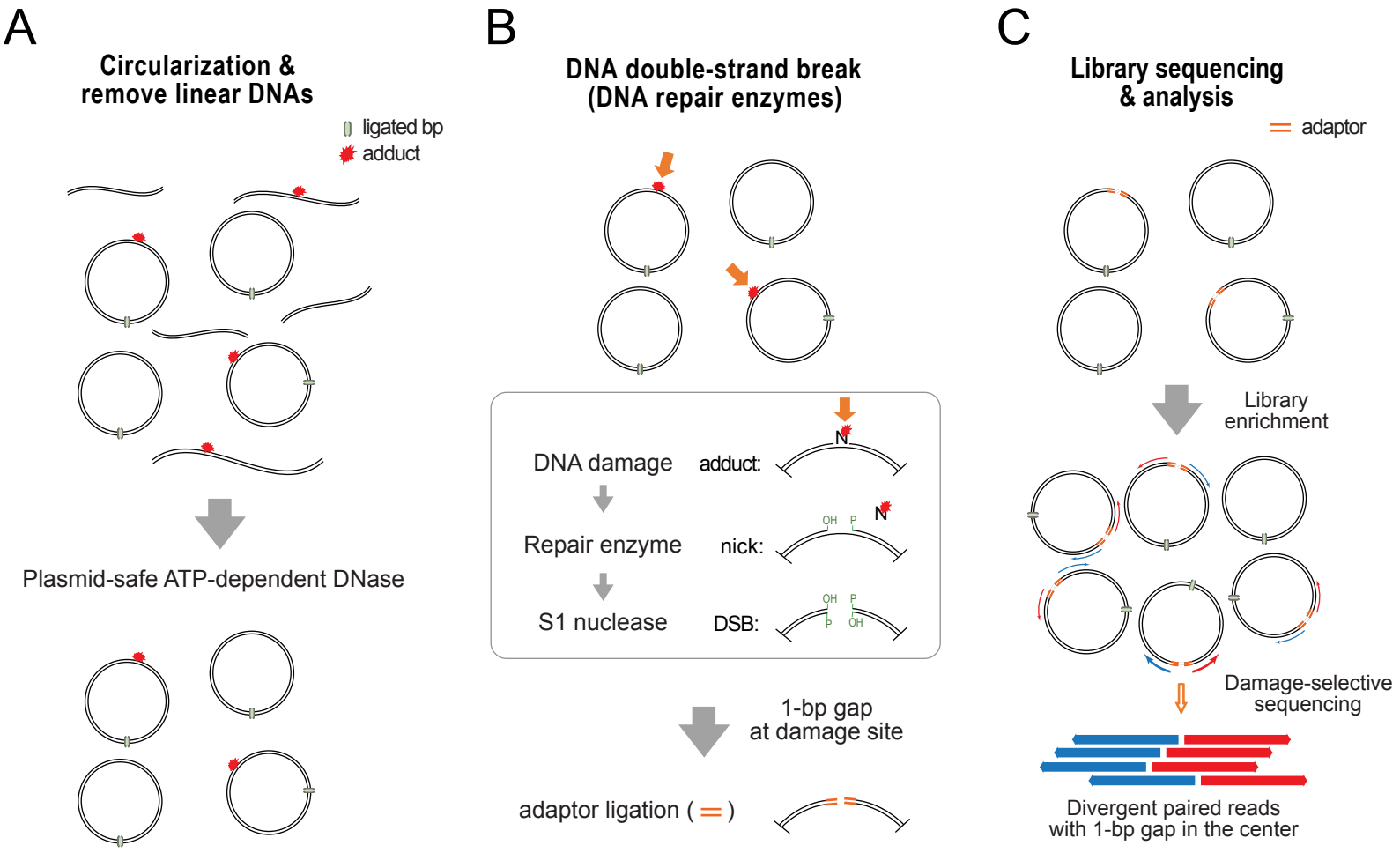
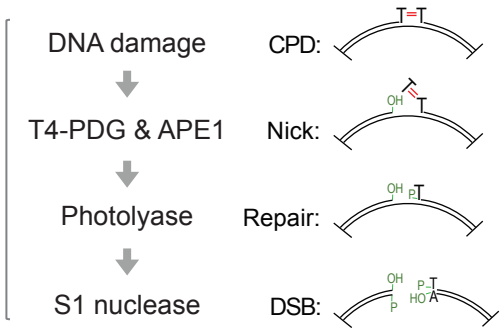
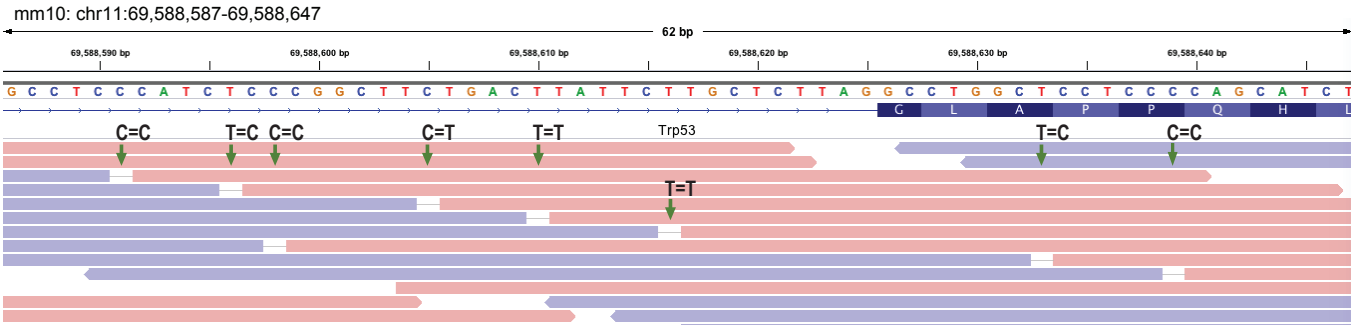

A
Circularization &
remove linear DNAs
ligated bp
adduct
Plasmid-safe ATP-dependent DNase
B
DNA double-strand break
(DNA repair enzymes)
DNA damage
adduct:
N
Repair enzyme
nick:
OH
P
N
S1 nuclease
DSB:
OH
P
P
OH
1-bp gap
at damage site
adaptor ligation (=)
C
Library sequencing
& analysis
adaptor
Library
enrichment
Damage-selective
sequencing
Divergent paired reads
with 1-bp gap in the center

Figure 1

A


DNA damage
T4-PDG & APE1
Photolyase
S1 nuclease
CPD:
Nick:
Repair:
DSB:

B


mm10: chr11:69,588,587-69,588,647
62 bp
C=C
T=C
C=C
C=T
T=T
Trp53
T=C
C=C
T=T

C

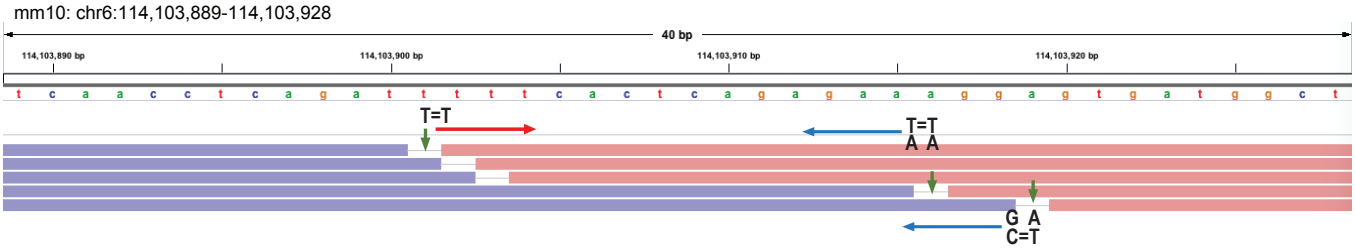

mm10: chr6:114,103,889-114,103,928
40 bp
T=T
T=T
A A
G A
C=T

Figure 2

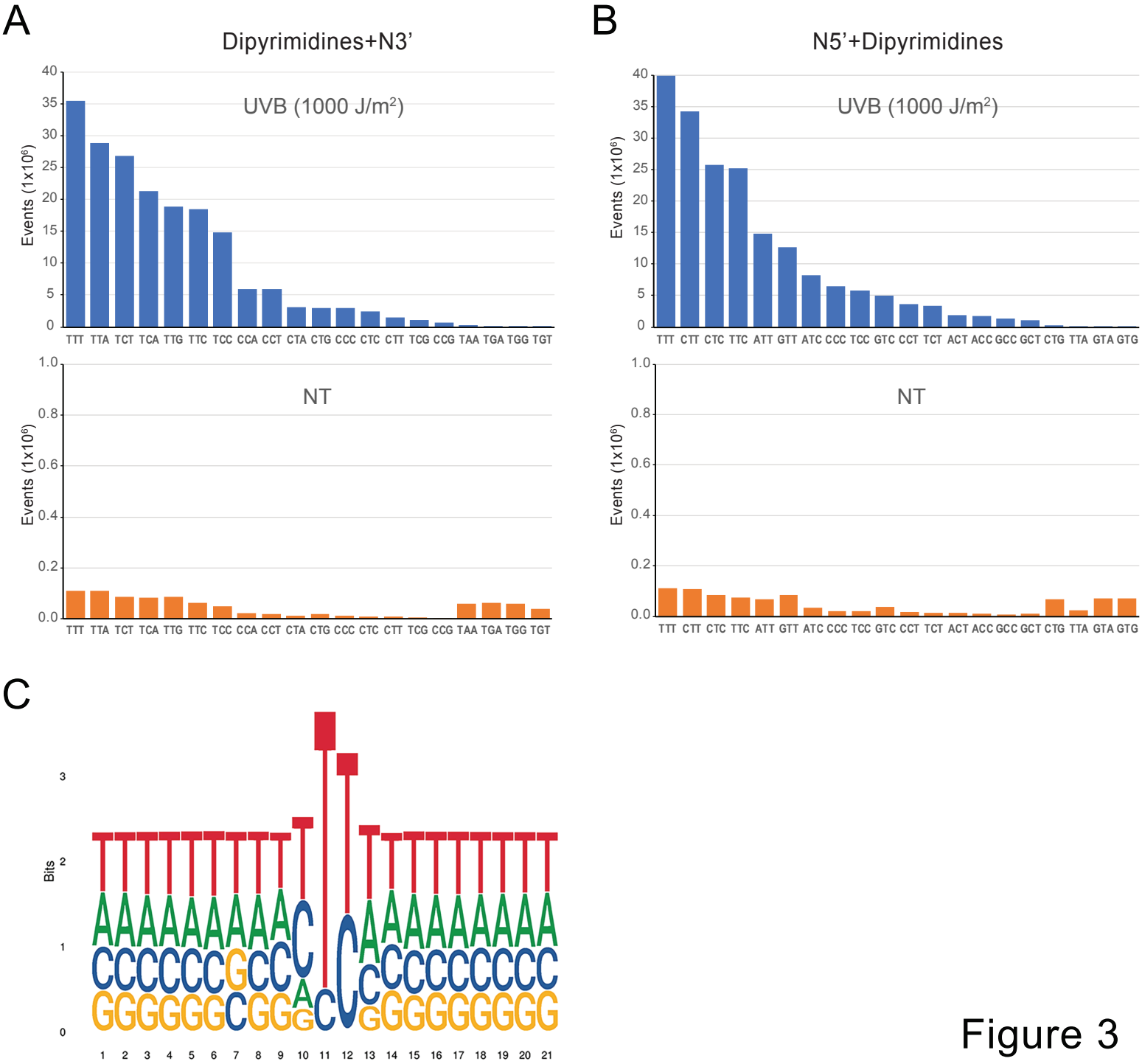

A
Dipyrimidines+N3'
UVB (1000 J/m2)
Events (1x10^6)
TTT TTA TCT TCA TTG TTC TCC CCA CCT CTA CTG CCC CTC CTT TCG CCG TAA TGA TGG TGT
NT
B
N5'+Dipyrimidines
UVB (1000 J/m2)
TTT CTT CTC TTC ATT GTT ATC CCC TCC GTC CCT TCT ACT ACC GCC GCT CTG TTA GTA GTG
C
Bits

Figure 3

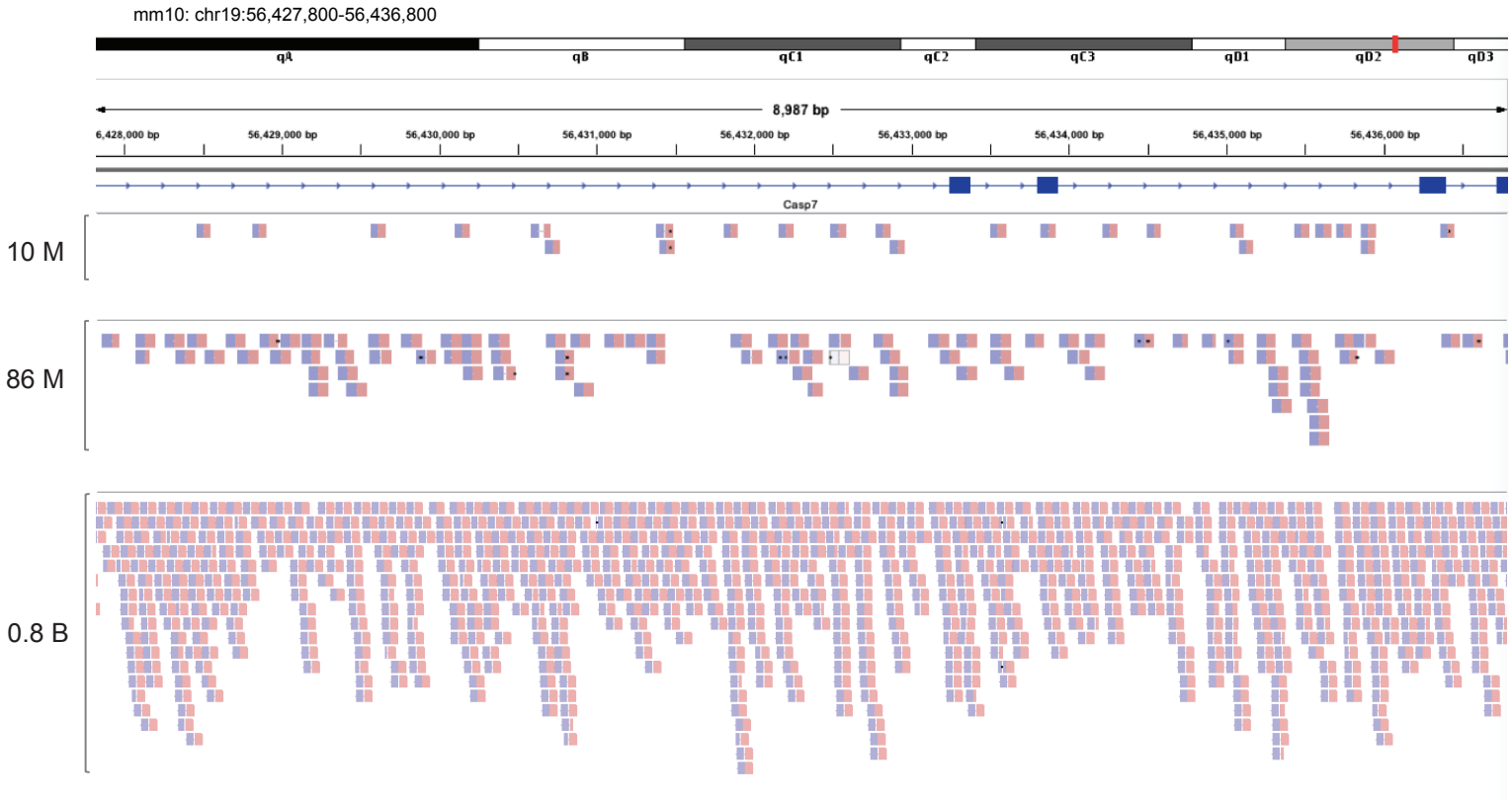

mm10: chr19:56,427,800-56,436,800
qA
qB
qC1
qC2
qC3
qD1
qD2
qD3
8,987 bp
6,428,000 bp
56,429,000 bp
56,430,000 bp
56,431,000 bp
56,432,000 bp
56,433,000 bp
56,434,000 bp
56,435,000 bp
56,436,000 bp
Casp7
10 M
86 M
0.8 B

Figure 4

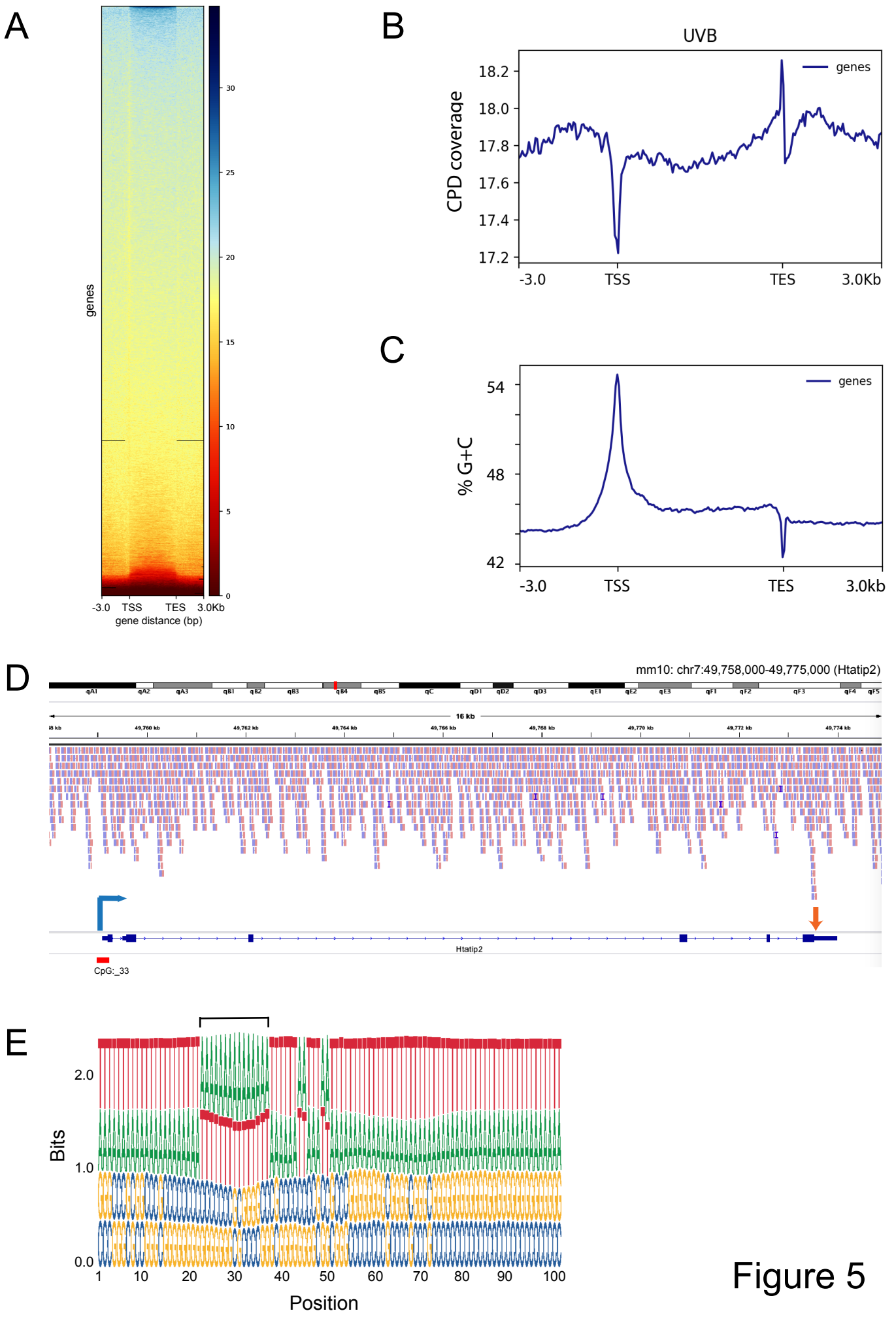

A
genes
-3.0
TSS
TES
3.0Kb
gene distance (bp)
B
UVB
CPD coverage
genes
18.2
18.0
17.8
17.6
17.4
17.2
-3.0
TSS
TES
3.0Kb
C
% G+C
genes
54
48
42
-3.0
TSS
TES
3.0kb
D
mm10: chr7:49,758,000-49,775,000 (Htatip2)
16 kb
Htatip2
CpG:_33
E
Bits
2.0
1.0
0.0
1
10
20
30
40
50
60
70
80
90
100
Position

Figure 5

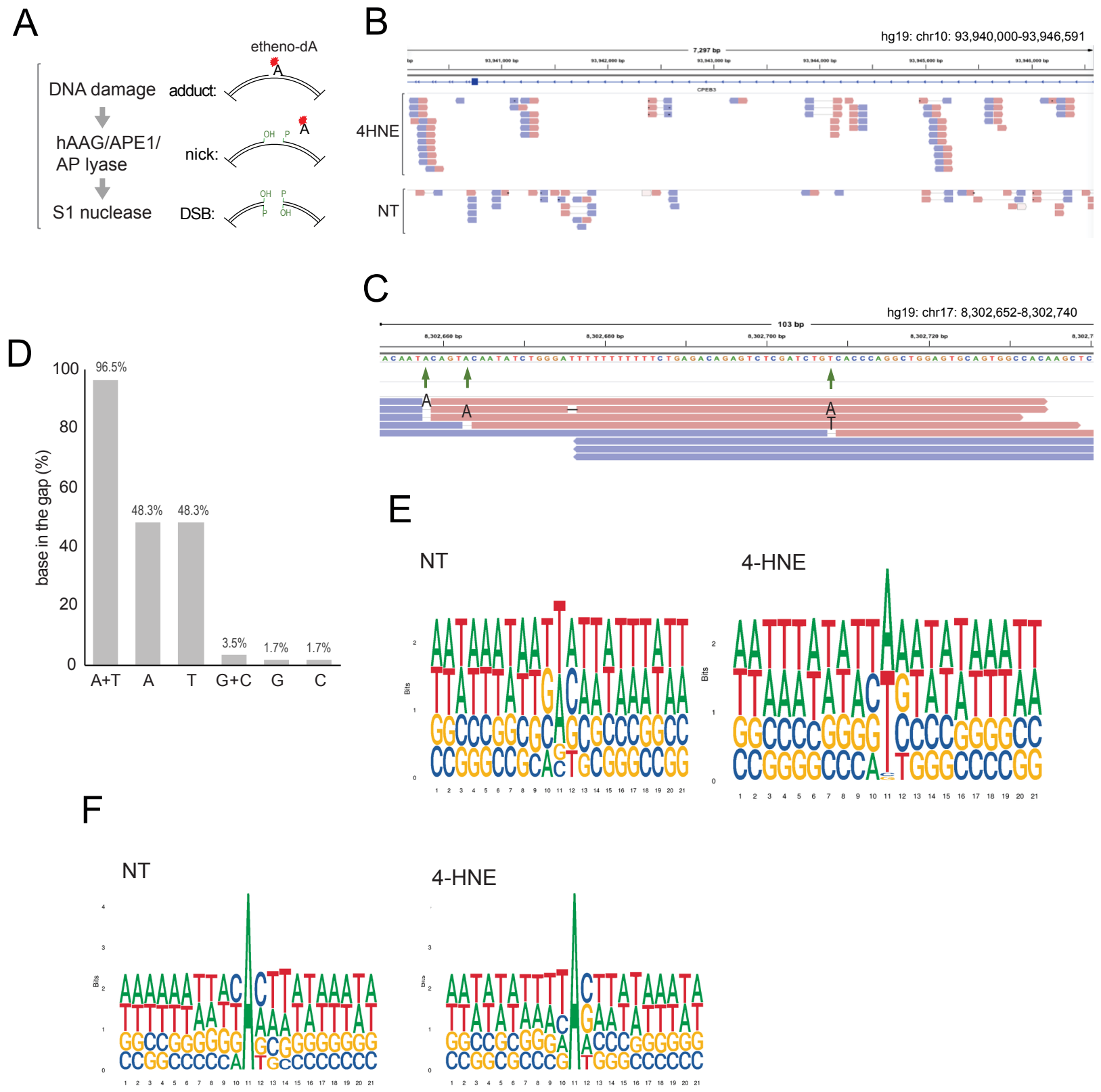

A
etheno-dA
DNA damage
adduct:
hAAG/APE1/
AP lyase
nick:
S1 nuclease
DSB:
B
hg19: chr10: 93,940,000-93,946,591
4HNE
NT
C
hg19: chr17: 8,302,652-8,302,740
D
base in the gap (%)
96.5%
48.3%
48.3%
3.5%
1.7%
1.7%
A+T
A
T
G+C
G
C
E
NT
4-HNE
F
NT
4-HNE

Figure 6

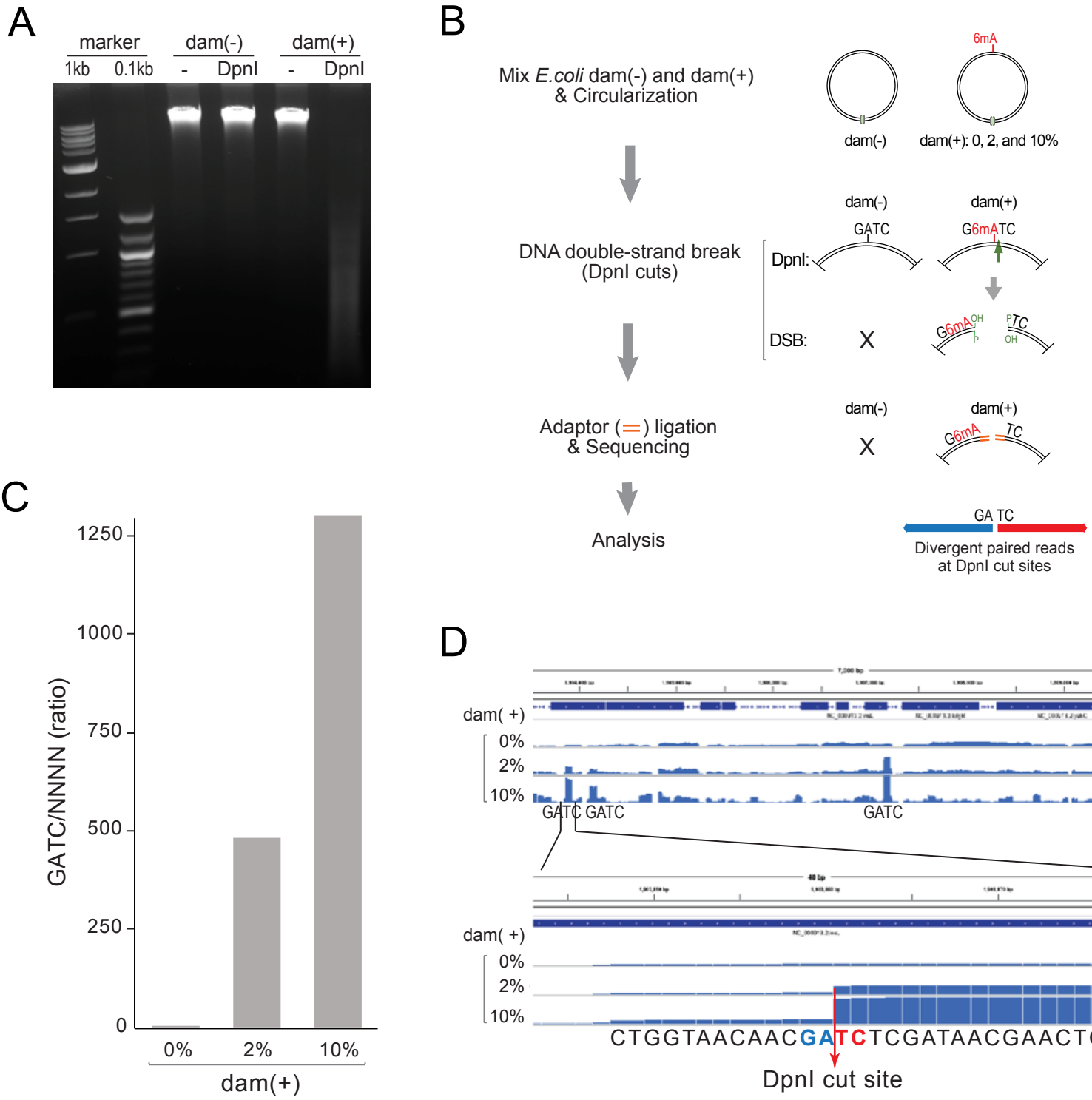

A
marker
dam(-)
dam(+)
1kb
0.1kb
-
DpnI
-
DpnI
B
Mix E.coli dam(-) and dam(+)
& Circularization
6mA
dam(-)
dam(+): 0, 2, and 10%
DNA double-strand break
(DpnI cuts)
dam(-)
dam(+)
GATC
G6mATC
DpnI:
DSB:
X
Adaptor (=) ligation
& Sequencing
dam(-)
dam(+)
X
Analysis
GA TC
Divergent paired reads
at DpnI cut sites
C
GATC/NNNN (ratio)
1250
1000
750
500
250
0
0%
2%
10%
dam(+)
D
dam(+)
0%
2%
10%
GATC GATC
GATC
dam(+)
0%
2%
10%
CTGGTAACAACGATCTCGATAACGAACTGGA
DpnI cut site

Figure S1

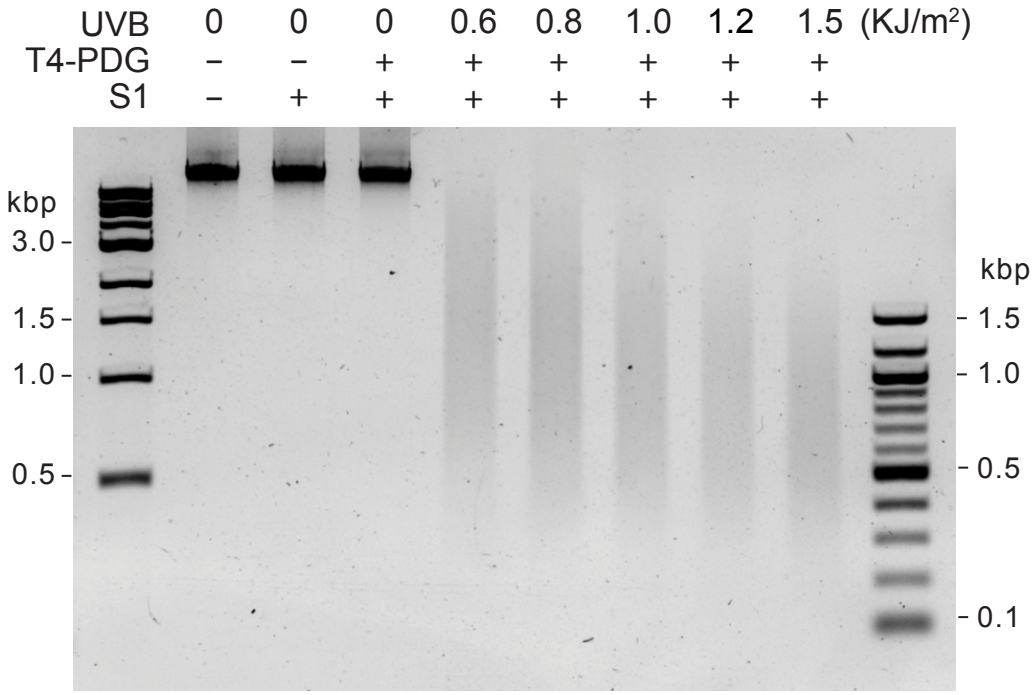
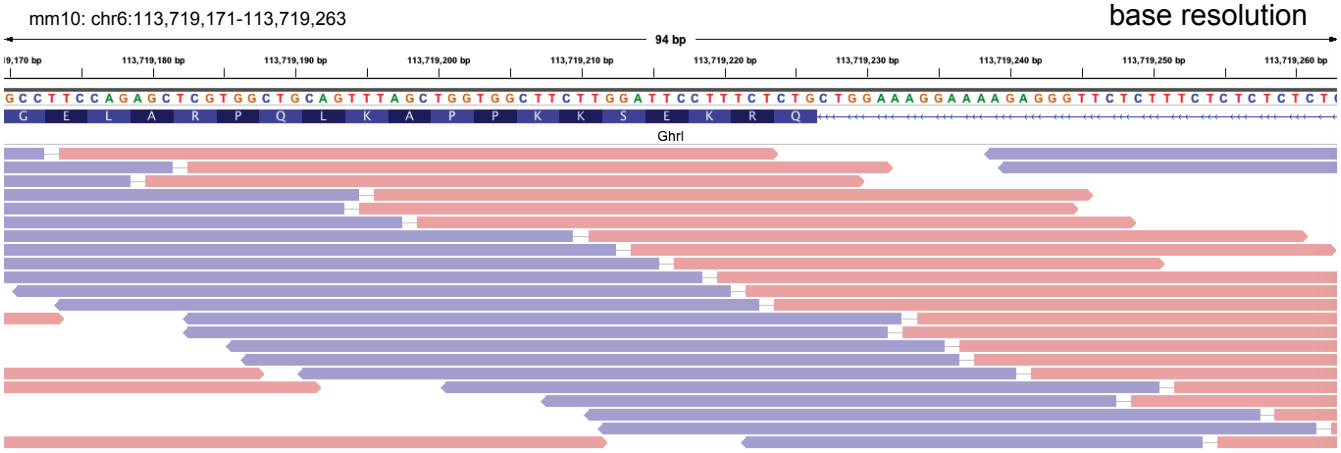


Figure S2

A


mm10: chr6:113,719,171-113,719,263
base resolution
94 bp
113,719,180 bp
113,719,190 bp
113,719,200 bp
113,719,210 bp
113,719,220 bp
113,719,230 bp
113,719,240 bp
113,719,250 bp
113,719,260 bp
G E L A R P Q L K A P P K K S E K R Q
Ghrl

B

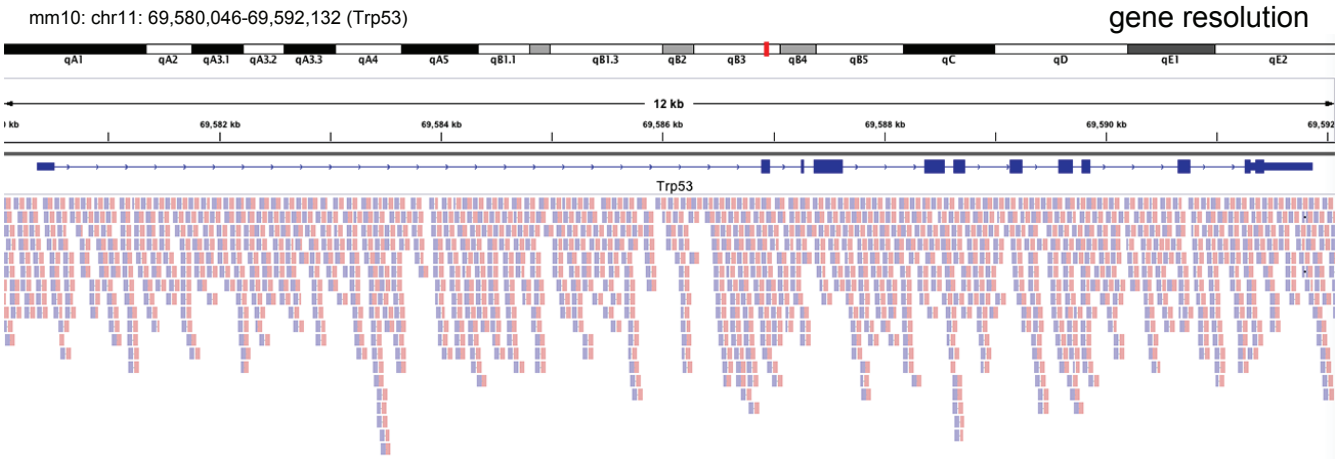
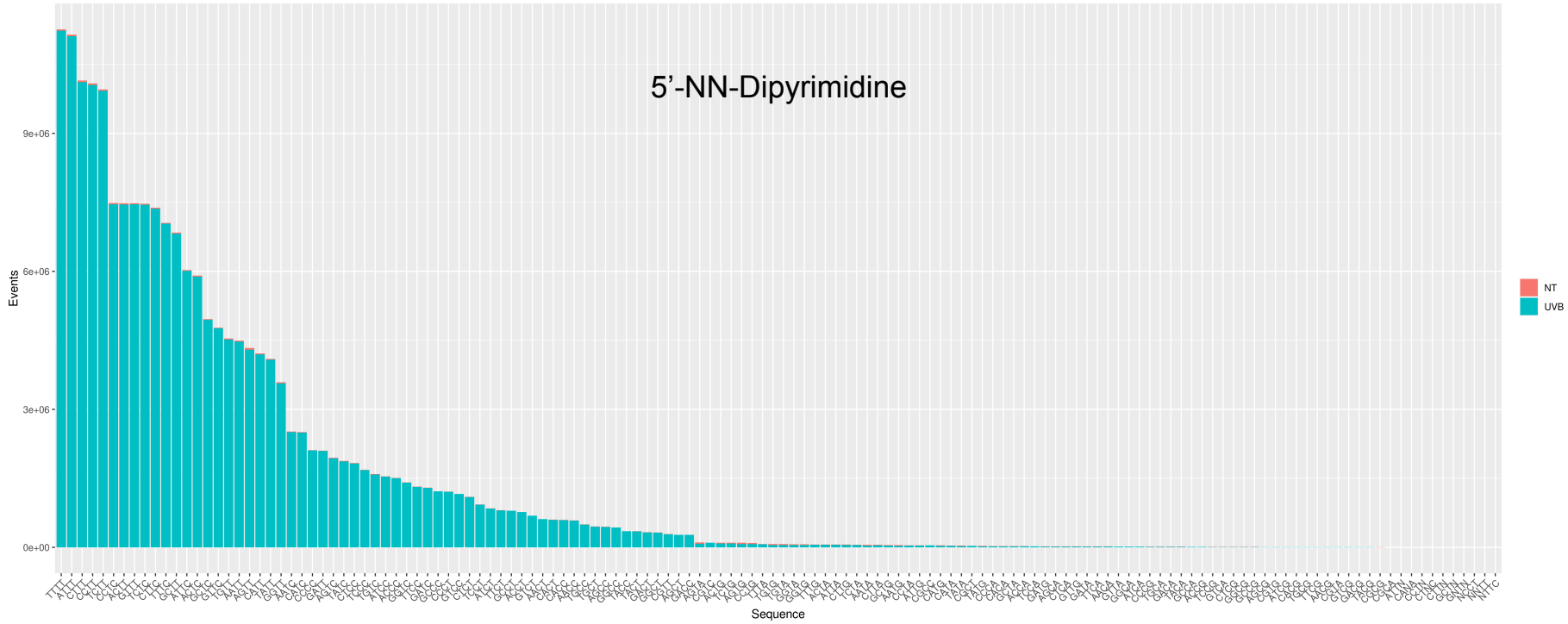

mm10: chr11: 69,580,046-69,592,132 (Trp53)
gene resolution
qA1
qA2
qA3.1
qA3.2
qA3.3
qA4
qA5
qB1.1
qB1.3
qB2
qB3
qB4
qB5
qC
qD
qE1
qE2
12 kb
69,582 kb
69,584 kb
69,586 kb
69,588 kb
69,590 kb
Trp53

Figure S3

A


5’-NN-Dipyrimidine
Events
9e+06
6e+06
3e+06
0e+00
Sequence
NT
UVB

B

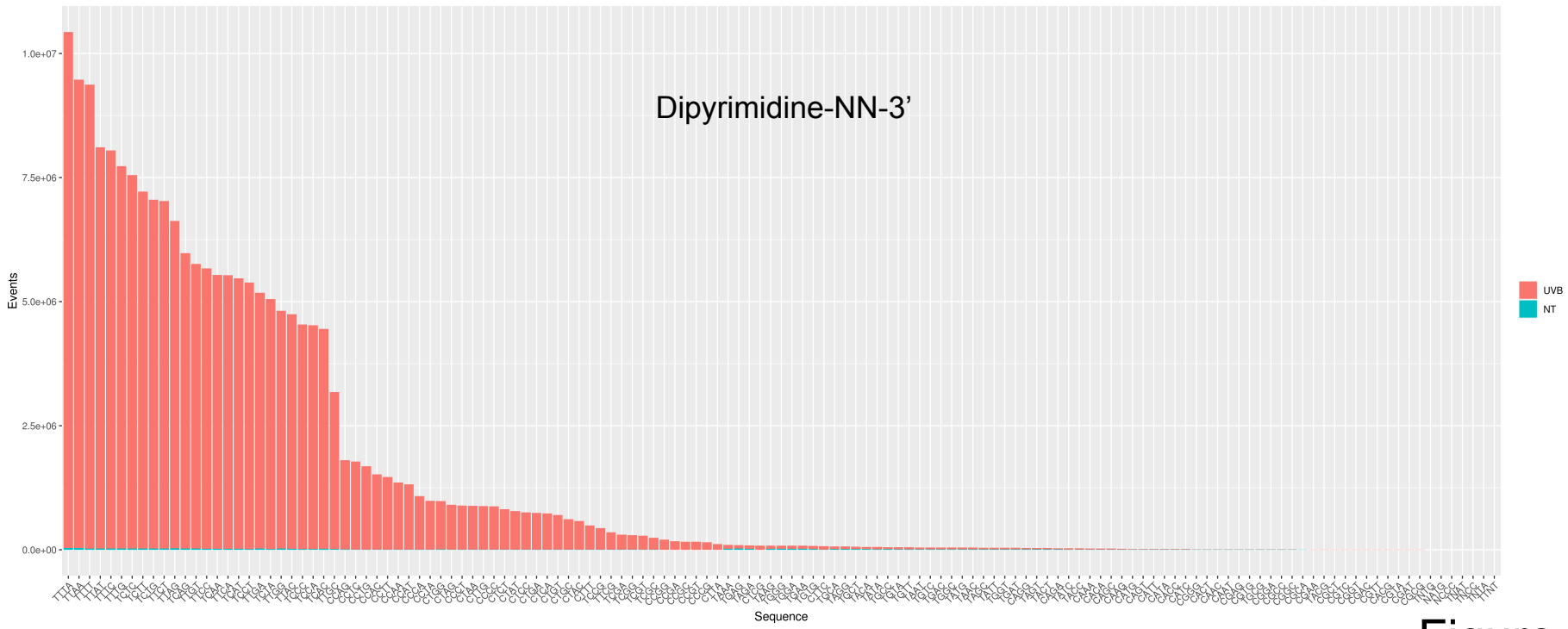
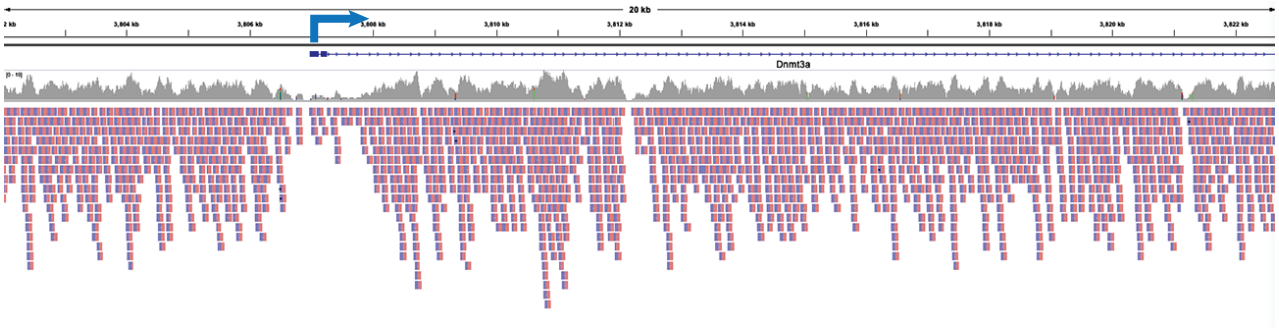

Dipyrimidine-NN-3’
Events
1.0e+07
7.5e+06
5.0e+06
2.5e+06
0.0e+00
Sequence
UVB
NT

Figure S4

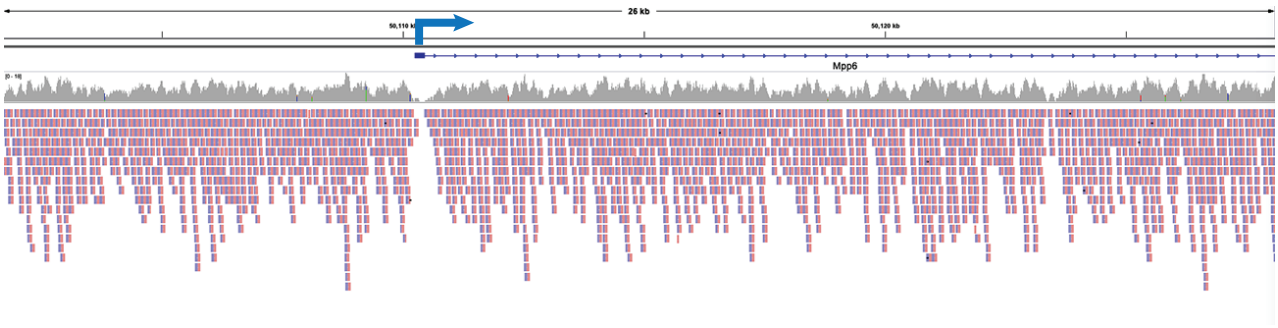
A

mm10: chr12:3,801,994-3,822,676 (Dnmt3a)



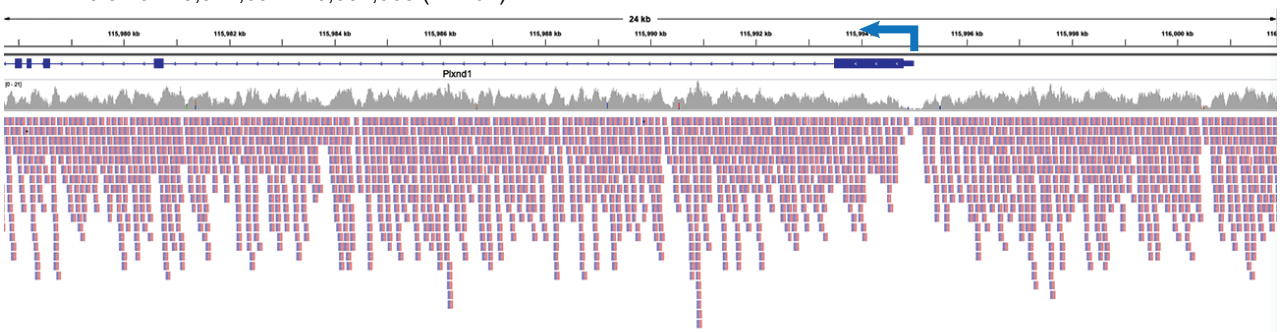
B

mm10:chr6:50,101,684-50,128,108 (Mpp6)



C

mm10:chr6:115,977,692-116,001,909 (Plxnd1)



D

mm10: chr18:70,560,530-70,582,258 (Mbd2)

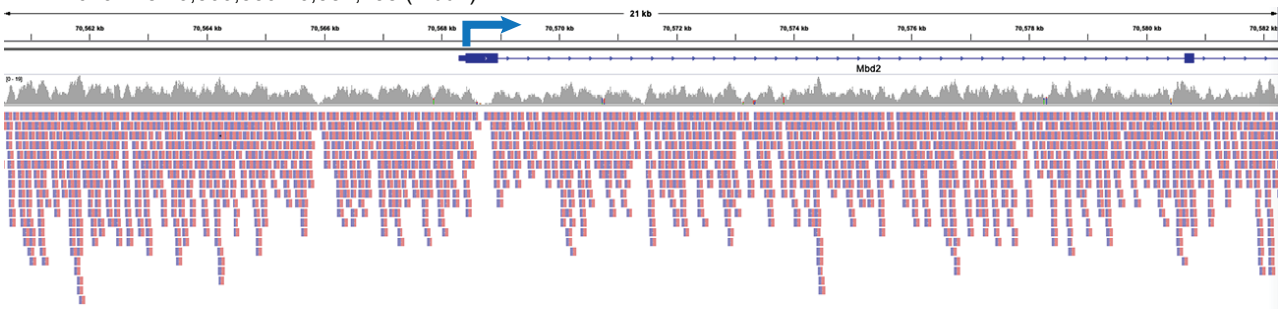
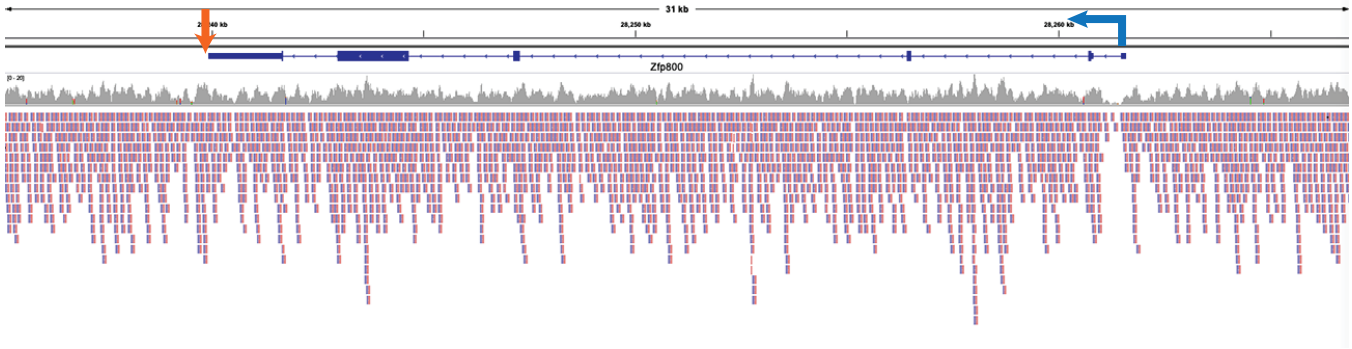


Figure S5

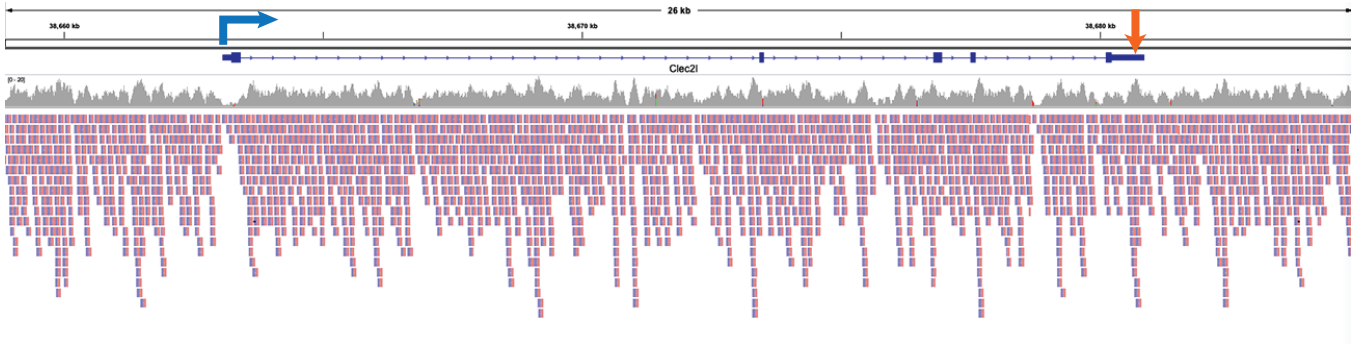
A

mm10: chr6:28,235,097-28,266,900 (Zfp800)



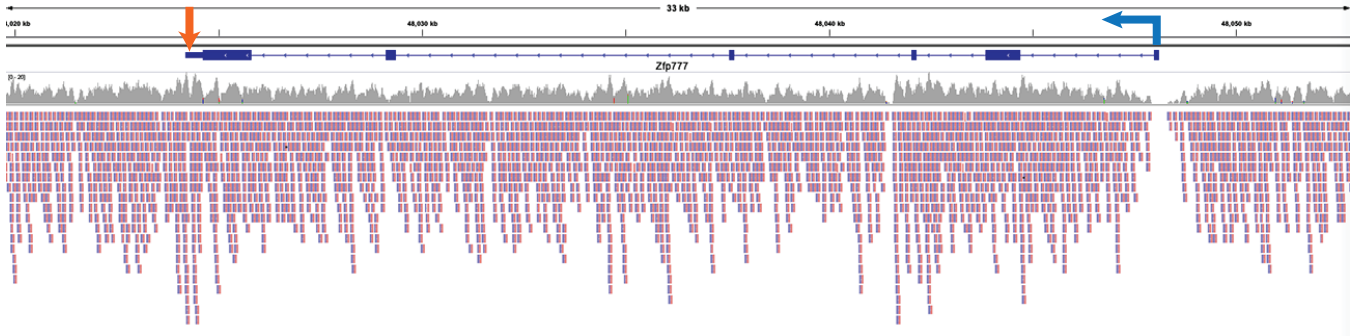
B

mm10: chr6:38,658,850-38,684,916 (Clec2l)



C

mm10: chr6:48,019,753-48,052,836 (Zfp777)



D

mm10: chr11:96,270,204-96,278,458 (Hoxb9)

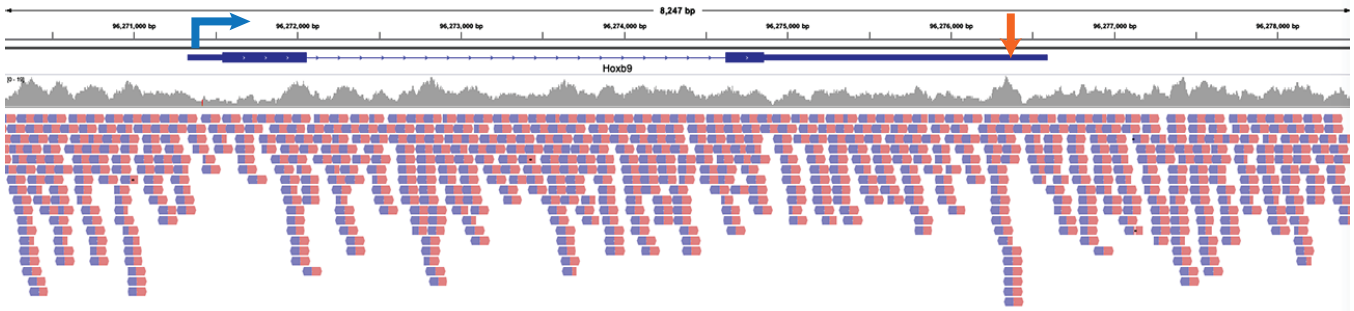


Figure S6