Nucleic Acids Research, 2025, **53**, gkaf048
https://doi.org/10.1093/nar/gkaf048
Advance access publication date: 11 February 2025
Genome integrity, repair and replication

The epigenetic landscape shapes smoking-induced mutagenesis by modulating DNA damage susceptibility and repair efficiency

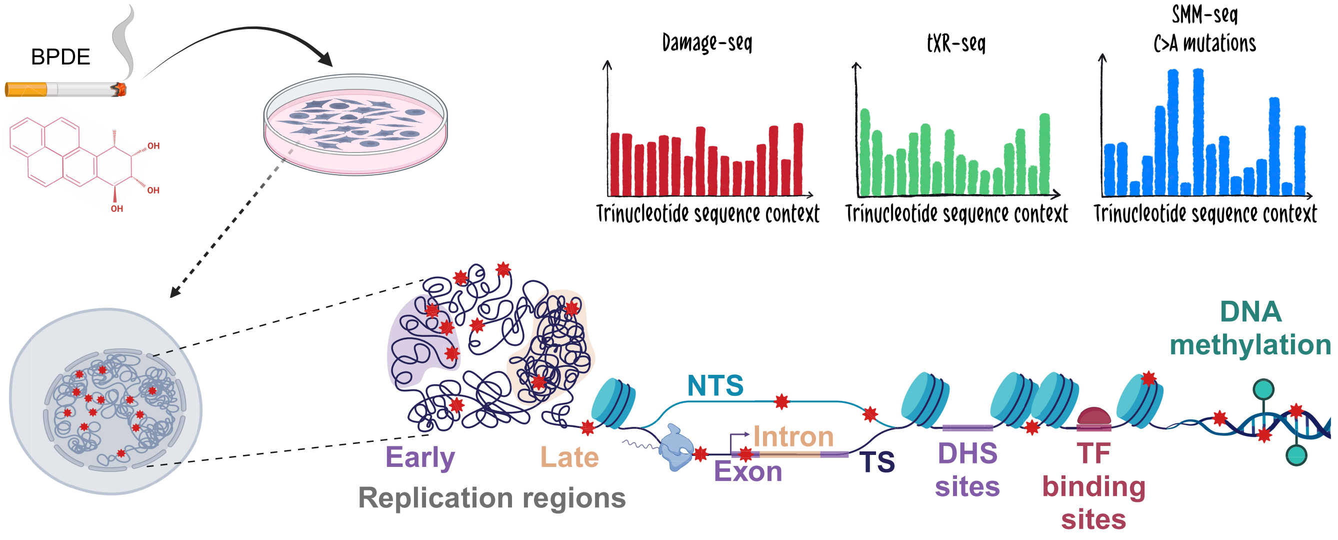
Elisheva E. Heilbrun[1], Dana Tseitline[1], Hana Wasserman[2], Ayala Kirshenbaum[1], Yuval Cohen[1], Raluca Gordan[3,4,5], Sheera Adar[1,*]

[1]Department of Microbiology and Molecular Genetics, The Institute for Medical Research Israel-Canada, The Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 9112102, Israel
[2]Program in Computational Biology and Bioinformatics, Duke University School of Medicine, Durham, NC 27708, United States
[3]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27708, United States
[4]Department of Computer Science, Duke University, Durham, NC 27708, United States
[5]Department of Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA
[*]To whom correspondence should be addressed. Email: sheera.adar@mail.huji.ac.il

Abstract

Lung cancer sequencing efforts have uncovered mutational signatures that are attributed to exposure to the cigarette smoke carcinogen benzo[a]pyrene. Benzo[a]pyrene metabolizes in cells to benzo[a]pyrene diol epoxide (BPDE) and reacts with guanine nucleotides to form bulky BPDE adducts. These DNA adducts block transcription and replication, compromising cell function and survival, and are repaired in human cells by the nucleotide excision repair pathway. Here, we applied high-resolution genomic assays to measure BPDE-induced damage formation and mutagenesis in human cells. We integrated the new damage and mutagenesis data with previous repair, DNA methylation, RNA expression, DNA replication, and chromatin component measurements in the same cell lines, along with lung cancer mutagenesis data. BPDE damage formation is significantly enhanced by DNA methylation and in accessible chromatin regions, including transcribed and early-replicating regions. Binding of transcription factors is associated primarily with reduced, but also enhanced damage formation, depending on the factor. While DNA methylation does not appear to influence repair efficiency, this repair was significantly elevated in accessible chromatin regions, which accumulated fewer mutations. Thus, when damage and repair drive mutagenesis in opposing directions, the final mutational patterns appear to be dictated by the efficiency of repair rather than the frequency of underlying damages.

Graphical abstract



Received: October 30, 2024. **Revised:** January 10, 2025. **Editorial Decision:** January 13, 2025. **Accepted:** January 20, 2025

Introduction

Smoking is a well-established driver of cancer mutagenesis. Tobacco smoke is a complex mixture of thousands of chemicals, at least 60 of which are carcinogenic [1–3]. These include the highly carcinogenic benzo[a]pyrene, a polyaromatic hydrocarbon that is metabolized in cells to the reactive form benzo[a]pyrene diol epoxide (BPDE). BPDE exerts its carcinogenic potential by reacting primarily with the N^2 position of guanines to form a bulky DNA adduct (BPDE-dG) [1–3]. BPDE adducts alter the helical structure of the DNA resulting in major consequences to the cells: First, they block transcription, resulting in a transcriptional shutdown that compromises cellular function and survival [4–6]. Second, they block the replicative DNA polymerases and lead to elevated mutagenesis [7–9]. Most affected are tissues directly exposed to the smoke such as those of the lung, respiratory system, head, and neck [3]. Seminal studies on the mutations of the p53 tumor suppressor genes in the 1990's linked BPDE exposure to specific cancer-driving mutations [10]. Recent cancer genome sequencing efforts discovered specific mutational signatures that are linked to smoking [11]. Treating cell lines with BPDE re-created these mutational signatures [12, 13] indicating that BPDE is indeed a major driver of smoke-related mutagenesis.

In human cells, the major pathway for BPDE-dG adduct repair is nucleotide excision repair (NER) [14, 15]. The process of repair is divided into three major steps: (i) Recognition of the damage, which can occur either directly (in general repair), or by a stalled RNA polymerase in a transcription-coupled manner (in transcription-coupled repair); (ii) Incision 3′ and 5′ of the damage, removing a nucleotide stretch of 24–32 nt and leaving a single stranded gap; (iii) Gap-filling DNA synthesis and ligation to restore intact double-stranded DNA. Inactivating mutations in the key NER pathway genes result in Xeroderma Pigmentosum, a severe genetic syndrome of high cancer susceptibility [16]. In addition, genome-wide association studies suggest attenuated repair resulting from polymorphism (and not inactivation) of excision repair genes can lead to enhanced lung cancer risk [17–20].

Epigenetic factors can influence damage formation and NER efficiency – and by that the degree of mutational burden. Advances in the genome-wide mapping of DNA damage and repair have boosted our understanding of the determinants of damage formation and NER [21]. The majority of these studies focused on ultraviolet (UV)-induced damages, primarily the cyclobutane pyrimidine dimers (CPDs [22–26]). The major determinant of CPD damage formation is the frequency of the target pyrimidine dimers within a sequence [26, 27]. While overall chromatin accessibility does not strongly influence damage formation [23], the rotational setting of the nucleosomes or binding of specific transcription factors (TFs) does modulate damage formation [25,28–33]. Outward-facing rotational positions in nucleosomes and ETS binding sites exhibit higher damage levels, which is thought to be due to the bending of the DNA into favorable angles for dimer formation [24, 25, 28, 29, 32, 34, 35].

NER efficiency is highly heterogenic and is strongly influenced by the chromatin state. Due to transcription-coupled repair (TC-NER), actively transcribed genes are preferentially repaired. This preferential repair is exclusive to the transcribed strand, on which a stalled RNA polymerase recruits the repair machinery. High-resolution mapping of the excised oligos released during excision by excision-repair sequencing (XR-seq [22]) revealed this preferential repair also occurs at sites of bi-directional transcription in promoters and enhancers. Nucleosome binding, on the other hand, prevents the access of repair factors to the DNA and inhibits global genome NER (GG-NER). As a result, the active and accessible regions of the genome in cells are preferentially repaired, essentially prioritizing regions necessary for cell function [36].

While mutational hotspots were long considered to be the product of phenotypic selection, the seminal study of Gerd Pfeifer and colleagues showed that targeted BPDE adduct sensitivity at specific p53 sites could also be an important driving force [37]. Still, BPDE adduct formation is considerably less characterized compared to other NER-substrates, such as damages induced by UV radiation and adducts induced by the chemotherapy drug cisplatin [21]. Interestingly, an *in vitro* study reported that nucleosome binding decreases BPDE adduct formation, specifically near the dyad [38].

DNA methylation, which in humans occurs on the 5-methyl position of cytosines within CpG pairs, can enhance BPDE-adduct formation on the adjacent G. This was reported in studies of specific sequence contexts, primarily the p53 gene, both *in vitro* in purified DNA [39–44] and in experiments in cells [39, 45]. However, the effect of DNA methylation on damage formation depended on the sequence context [39, 44], and not all sites of elevated BPDE damage also exhibited elevated mutagenesis [42, 46, 47]. Thus, the extent to which DNA methylation affects BPDE mutagenesis is unclear.

BPDE-dG repair, measured genome-wide in human cells by tXR-seq [48], is higher on the transcribed strand and in accessible chromatin regions, similar to the other NER-substrate damages [22, 36, 49]. This study also revealed an enrichment of CpG dinucleotides within the excised reads; however, it could not determine whether this was due to higher damage or preferential repair of the CpG sites, and if this was due to DNA methylation [48]. A recent study mapped BPDE-dG adducts in cells exposed to low doses of BPDE over a 24-h period [50]. Under this long exposure, damage formation and repair co-occur. Thus, their individual contribution to the damage profile cannot be isolated, complicating the interpretation of the results.

Here we applied the single-nucleotide resolution Damage-seq method [23, 49] to map the initial BPDE-dG adduct formation and to identify sites of elevated damage sensitivity. We then applied single-molecule mutation sequencing (SMM-seq [51]) to identify the BPDE-dG-induced mutations in the same cells. Integrating the damage and mutagenesis data we generated with previous repair, methylation, expression, replication timing, and chromatin component measurements in the same experimental system, along with lung cancer mutagenesis data, we delineate the determinants of damage and repair and their relative contribution to the final mutagenic outcome of BPDE exposure.

Materials and methods

Reagents

Cell culture reagents, including Dulbecco's modified Eagle's medium (DMEM) (01–055-1A), RPMI 1640 medium, and all media supplements were from Biological Industries, Beit Haemek, Israel. BPDE (#477) was purchased from MRIGlobal, Kansas City, MI, USA. Reagents for DNA analysis, included DNeasy® Blood Tissue kit (69 504, Qiagen, Hilden, Germany), the QuantiFluor® dsDNA System (E2670, Promega Corporation, Madison, WI, USA), and the G50 spin columns (GE healthcare). For immunodetection and enrichment of BPDE-dG damages, nitrocellulose membrane (Cytiva, 10 600 003, Marlborough, MA, USA), the anti-BPDE antibody (Santa Cruz Biotechnology, Inc., sc-52624, Dallas, TX, USA), Dynabeads Protein G beads (#10004D), and M280 Sheep Anti-rabbit IgG (#11203D) from Invitrogen Waltham, Massachusetts, USA, horseradish peroxidase-conjugated anti-mouse secondary antibody (NA931, Cytiva, Marlborough, MA, USA), Enhanced Chemiluminescence (ECL™ Prime Western Blotting System, Cytiva, RPN2236, Marlborough, MA, USA) and SYBR™ Gold Nucleic Acid Gel Stain (Invitrogen, S11494, Carlsbad, CA, USA)

Biological resources

The study was performed in the GM12878 lymphoblast cell line (Coriell Repository, Camden, NJ, USA) and A549 lung adenocarcinoma CCL-185™ cell line (ATCC, Manassas, VA, USA).

Cell culture and treatment

GM12878 cells were grown in RPMI medium supplemented with 15% Fetal Bovine Serum (FBS), 4 mM glutamine, 100 units/ml penicillin, and 0.25 mg/ml streptomycin. A549 cells were grown in DMEM supplemented with 10% FBS, 2 mM glutamine, 1 mM sodium pyruvate, 100 units/ml penicillin, and 0.25 mg/ml streptomycin. Mycoplasma was monitored every 3–4 months.

BPDE treatment

GM12878 cells were grown to 700 000 cells/ml in a T75 flask, and A549 cells were grown to ~80% confluence in a 150 mm dish. For damage treatment, the cells were incubated with media containing 25 μM of BPDE for 2 h. Cells were collected immediately, followed by genomic DNA extraction by DNeasy® Blood & Tissue kit, and quantified by the QuantiFluor® dsDNA System following manufacturers' protocols.

For *in vitro* treatment of genomic DNA, 3 μg of genomic DNA was incubated with 3 μM BPDE in a final volume of 30 μl at 37 degrees for 2 h. Treated DNA was purified through a G50 spin column (GE) and subjected to Damage-seq.

Damage-seq

Damage-seq was performed as previously described without biotin purification after the primer extension step [49,52]. Briefly, genomic DNA was sheared by sonication with Bioruptor Pico sonicator to generate fragments averaging 300 bp in length. For polymerase chain reaction (PCR)-amplified DNA, after sonication and ligation of the Ad1 adapter, 10 ng of the ligation product were PCR-amplified with primers Pu/Pi for 10 cycles. The PCR products were purified and subject to BPDE treatment as described above. Damaged DNA immunoprecipitation was performed with Anti-BPDE Antibody (8E11, Santa-Cruz Biotechnology, sc-52624), using 2 μl per μg DNA using 10 μl each of protein G and anti-rabbit dynabeads. Library quality was assessed using the Agilent 4200 TapeStation. Qualified libraries were pooled and sequenced on Illumina NovaSeq 6000 or NextSeq 550 sequencers. Reads were processed following the steps mentioned in Hu *et al.* Reads containing the Ad1 adapters were discarded by cutadapt (version 3.5) and were aligned to hg38 genome using bowtie1 (version 1.3.1). Then, Picard MarkDuplicate (version 2.26.10; http://broadinstitute.github.io/picard) was used to remove read duplicates. Next, unique reads in BED format were further filtered with Bedtools (version v2.27.1) and custom BASH scripts. Reads from replicates of the same condition were merged for further analyses. Read counts obtained after each step of the analysis for each experimental replicate are detailed in Supplemental Table S1. For replicate correlation plots, a 10 kb windows bed file was created using bedtools makewindows (version 2.31.0). Read counts over these windows was calculated for each replicate using bedtools coverage and Spearman correlation coefficients were calculated using the corrplot R package (version 0.95).

Immuno-dot blot assay

DNA extraction was performed using the DNeasy Blood & Tissue Kit. For each sample, 500 ng of DNA was applied per well in duplicate technical replicates and transferred to a nitrocellulose membrane via vacuum using the Bio-Dot apparatus (Bio-Rad, 1706 545, Rishon LeZion, Israel). The membrane was subsequently baked at 75°C for 60 min in a Bio-Rad Gel Dryer model 583. After blocking with 5% milk, the membrane was incubated with the primary anti-BPDE antibody, diluted 1:500. Following incubation with an HRP-conjugated secondary antibody, the damage signal was visualized using the Enhanced Chemiluminescence. The amount of genomic DNA loaded on the membrane was quantified using SYBR™ Gold Nucleic Acid Gel Stain, and the damage signal was normalized relative to the SYBR-Gold signal using Bio-Rad's Image Lab version 6.1 software.

In vitro mutagenesis assay

GM12878 cells were cultured in T25 flasks with 5 ml of media containing either 0.125 μM BPDE, or Dimethyl sulfoxide (DMSO), over a total of 12 passages. Cells were counted every 2 days, and were split into new flasks at a concentration of 300 000 cells/ml under the condition that they completed at least 1.5 replication cycles. BPDE was freshly dissolved in media to the desired concentration from a 1 mM BPDE stock solution in DMSO for each passage. Cells were harvested approximately every 5–7 replication cycles, and DNA was extracted as previously described. DNA from GM12878 cells prior to treatment, DMSO-treated cells, and 0.125 μM BPDE-treated cells at the 2-week timepoint (~11 replication cycles) was sequenced using SMM-seq [51] performed by Mutagentech. In short, library preparation included fragmentation of DNA using restriction enzymes, size selection for reduced representation, and rolling circle-based linear amplification, to create multiple copies of a single original DNA molecule (for both strands), then conventional sequencing library was prepared and DNA was sequenced using llumina NovaSeq instrument using 150 paired-end mode.

VCF files were obtained from Mutagentech following sequencing, alignment to the hg38 genome and variant calling using GATK [53]. Using the untreated sample sent for sequencing, background mutations were filtered using the bcftools isec command [54]. Filtered VCF were further analyzed using the MutationalPatterns package in R [55] to create a count table of six single base substitution (SBS) types.

P-value between DMSO-treated and BPDE-treated replicates was calculated using the chi-square statistical analysis.

Web sites/data base referencing

For comparative analyses, genomic data was obtained from the TCGA (https://www.cancer.gov/ccg/research/genome-sequencing/tcga), ENCODE (https://www.encodeproject.org/), cBioPortal (https://www.cbioportal.org/), COSMIC (https://cancer.sanger.ac.uk/signatures/), Zenodo (https://zenodo.org/records/556775#.XrfJJagzaUl) and GEO (https://www.ncbi.nlm.nih.gov/geo/) databases.

XR-seq data analysis

Genome-wide maps of NER for BPDE in the GM12878 cell line were obtained from GEO (accession number GSE97675). The sequencing reads were extracted, processed, and mapped to the human genome following the steps outlined in [49]. To avoid biases in the normalization of repair to damage, the read length in XR-seq was reduced to 3 nt, based on the identified guanine-enriched sites, by taking the −1 and +1 flanking nucleotides.

Comparative data analysis

Read counts for each genomic feature were obtained using bedtools coverage. Average profiles for each element were generated using the bedtools intersect command and the Bioconductor package genomation (version 1.36.0). The curation of the different genomic features is detailed below.

Curation of active and accessible chromatin regions: Coordinates of DNase I hypersensitivity sites (Narrow peak format) for GM12878 (ENCSR000EMT) and A549 (ENCSR000ELW) cell lines were downloaded from ENCODE. Overlapping sites were merged, retaining only the longer regions using the bedtools (version 2.31.0) cluster command. Additionally, DNase hypersensitivity sites (DHS) overlapping with genes were removed using the bedtools intersect command, resulting in a total of 14 841 sites for GM12878 and 34 995 sites for A549.

DHS in normal lung tissue were identified using data from 17 lung tissue samples obtained from ENCODE (Supplementary Table S2). The DHS summits were recalculated by merging the peaks across all samples and then determining the summit of each peak as the point of maximum signal coverage.

Chromatin state annotations from ChromHMM for GM12878, A549 and lung tissue were retrieved from ENCODE (ENCSR988QYW, ENCSR283FYU and ENCFF361HLB, respectively).

Curation of methylated DNA data: Whole genome bisulfite sequencing (WGBS) data for CpG methylation in GM12878 (ENCSR890UQO) and A549 (ENCSR481JIW) cell lines were retrieved from ENCODE. Only CpG sites with a read depth >5 and with <15% variance in methylation scores between replicates were retained for analysis, resulting in 17 326 394 sites for GM12878 and for 31 610 739 sites for A549.

CpG island coordinates were obtained from the UCSC Table Browser for the hg38 genome assembly, comprising 31 448 sites. For each CpG island, the average methylation score (calculated from all CpGs within the island), the standard deviation (SD), and the coverage (the fraction of CpGs within the island that have methylation information) were determined. To classify CpG islands as methylated or unmethylated, only CpG islands with a methylation score SD ≤10 and a coverage fraction ≥5 were considered. Methylated CpG islands were defined as those with an average methylation score >50, while unmethylated CpG islands had an average score of 50 or less.

WGBS data from lung tissue were obtained from ENCODE (ENCFF992DYS and ENCFF453HAD datasets). Only CpG sites with a read depth >5 in both datasets and with <10% variance in methylation scores between them were retained for analysis, resulting in 30 403 746 CpG sites. To compare mutation rates between methylated and unmethylated CpG sites, methylated sites were defined as those with an average score of ≥70%, while unmethylated sites were defined as those with an average score of ≤30%. For each group, the fraction of overlapping C > A mutations was calculated as the number of mutations within the group divided by the proportion of that group out of all CpG sites.

Curation of active TF-binding sites: Active TF-binding sites for the GM12878 cell line were curated from binding site calls based on 286 non-redundant TF motif clusters previously reported in Vierstra *et al.* [56] using reference genome assembly hg38, encompassing 2179 total motifs for 702 distinct human TF proteins. The called sites were intersected with genome-wide DNA accessibility data (DNase-seq) from GM12878 cells, downloaded from ENCODE [ENCSR000EMT] and processed as described above. For each TF motif cluster, the binding site calls in accessible DNA were ranked by their motif scores using MOtif Occurrence Detection Suite [57], and the top 50% of sites with the highest scores were used for further analysis.

Curation of early and late replicating regions: Constitutive replication origins across multiple cell lines were obtained from Guilbaud *et al.* [58]. Intervals of 10 kb centered on the origin midpoints were created to define the regions of interest. In cases where neighboring origins were <10 kb apart, only the longer origins were kept, leaving 15 637 unique origins. Early and late constitutive replicating regions were obtained from [59].

Gene annotations: The annotation file for 28 712 protein-coding genes was retrieved from the UCSC Table Browser (RefSeq, assembly hg38). In cases of multiple gene variants, the longest transcript was retained. Genes that overlapped or were located within 6 kb upstream of neighboring genes were removed using the bedtools overlap and bedtools closest commands. Exon and intron annotation files for these genes were retrieved by uploading the list of genes to the UCSC Table Browser. To avoid biases from splicing junctions, 100 bases were removed from each end of the introns, and 10 bases were removed from each end of the exons.

Analysis of BPDE-dG signal at TF-binding sites

Active TF-binding sites were curated as described above. TF motif clusters with <5000 binding site calls were filtered out, as the low number of sites, when intersected with the BPDE Damage-seq data, resulted in too few BPDE-dG lesions to identify statistically significant trends. After this filtering step, 181 motif clusters covering 618 human TFs were selected for further analysis. For each TF motif cluster, the binding sites were extended 15 bp downstream and upstream of the motif center. For each position in these binding site regions, we counted the number of BPDE-dG adducts at that position, on each strand of the motif, and compared these counts to the expected number of BPDE-dG adducts according to a background model of BPDE-dG formation in accessible DNA. Briefly, we modeled the formation of BPDE-dG adducts as a stochastic Poisson process that consists of discrete, independent rare events where event frequency is dependent on sequence context around a central guanine.

We compared BPDE-dG formation rates in trimers versus pentamers and selected a pentamer-based Poisson model after observing significant variation in BPDE-dG formation among pentamers sharing the same central trimer (Supplementary Fig. S1). BPDE-dG formation rates for all 256 NNGNN pentamers were calculated by intersecting GM12878 BPDE-dG damage-seq data with GM12878 accessible regions using BEDTools (v2.31.0) and dividing total damages by the number of occurrences for each pentamer. We calculated the expected number of BPDE-dG adducts over n occurrences of given pentamer, p as $r_p n_p = E[X_p] = \lambda_p$, where r_p is the rate of BPDE-dG formation for pentamer p. By considering the BPDE-dG formation rate of each pentamer as its own independent Poisson distribution, we leveraged the property for sums of Poisson-distributed random variables [60] to estimate the cumulative amount of damage separately for both strands at each position in the TF-binding site region. The predicted BPDE-dG signal for each strand was then scaled by multiplying the BPDE-dG estimates by the average ratio of observed to predicted BPDE-dG signal in the immediate flanking regions around the TF motif, computed separately for each side of the motif. After scaling, a P-value for the observed BPDE-dG signal at each position in the TF-binding site region was calculated using the Poisson distribution implementation from the scipy.stats module (v1.14.0) in Python. To control for the batch effects observed in GM12878 Damage-seq data (Supplemental Fig. S2), the BPDE-dG predictions for cellular DNA and naked DNA (nDNA) conditions were modeled separately for combined replicates 1 and 2, and combined replicates 3 and 4 (Supplemental Table S3), and the trends were further analyzed for consistency, as described below.

Generation of the TF-binding site BPDE-dG heat map

P-values for the observed Damage-seq signal in BPDE treated GM12878 cells (replicates 1 and 2) were calculated per position for both motif and motif-complement strands of each TF motif cluster. Correction for multiple hypothesis testing was performed using the Benjamini–Hochberg procedure implemented in the statsmodel.stats.multitest (v0.14.2) Python module, and a false discovery rate (FDR) cutoff of 0.01 was used for significance. The same process was applied for TF motif clusters in the Damage-seq from GM12878 DNA treated *in vitro* (replicates 1 and 2) and P-values were corrected for a final FDR cutoff of 0.1. Positions within the TF motif cluster region that demonstrated either a significant enrichment or reduction of BDPE-dG signal (corrected P-value < .01) that was also observed at the same position in the nDNA condition (corrected P-value < .1) were considered false positives and omitted. For further confidence in our results, we then repeated the above process in parallel for cellular and *in vitro* treated DNA conditions in replicates 3 and 4, and we retained only positions in TF-binding site regions with concordance between the two replicate subgroups.

To best describe the magnitude of the differential BPDE-dG levels in the presence of active TF binding, we then determined for each significant position the BPDE-dG Z-score difference between the cellular DNA and nDNA conditions for both strands of each TF motif cluster. We then generated a heatmap of the ΔZ-scores to summarize the magnitude and directionality of the putative effects of TF binding on BPDE-dG formation (Fig. 3C).

Mutation data analysis

Whole genome sequencing (WGS) data of mutational profiles in lung cancer (TCGA-LUAD and TCGA-LUCS) was downloaded from the TCGA database. Smoking status and clinical information were obtained from cBioPortal [61]. VCF files were filtered to retain only SBSs, and further restricted to C > A transversions, which are strongly associated with BPDE exposure. All filtered mutations from the selected samples were merged and used for downstream analyses. Only data from patients (63 individuals) with a confirmed history of smoking were included (a total of 4015 719 mutations). To eliminate sequence context bias, mutation counts were normalized to the occurrence of their respective target trinucleotides using custom scripts.

Creation of damage, repair, and mutagenesis trinucleotide context profiles

Mutagenesis data was obtained by merging the two BPDE-treated samples and filtering for C > A and G > T mutations. Trinucleotide context was extracted using the MutationalPatterns package. A BED file containing regions sequenced by SMM-seq (according to AluI restriction enzymes) were obtained from Mutagenetech. Sequencing was done in 150 bp paired-end mode, so regions were trimmed to include only 150 nt at the start and end of the region. Coordinates were lifted over from the hg19 genome to the hg38 genome using CrossMap [62], and chrY and chrM reads were filtered out. BED files containing damage and repair data from Damage-seq and XR-seq experiments (previously described) were intersected with the bed file containing regions sequenced by SMM-seq using the bedtools intersect command. Bedtools getfasta command was used to get the sequence of these intervals for further analysis. Data was filtered for 3 nt-long reads containing G > T or C > A in the second position and plotted for relative frequencies of each trinucleotide in the pyrimidine context (C > A only). The SBS4 mutational signature in numerical form was downloaded from the COSMIC website (https://cancer.sanger.ac.uk/signatures/sbs/sbs4/). Entries containing C > A mutations were filtered and the relative frequency of those was calculated and plotted. Cosine similarity was calculated using the lsa package in R and heatmap created using corrplot package.

To assess the reduction of BPDE-induced mutations around DHS, the number of overlapping SMM-seq-detected C > A mutations within a 3 kb interval of the DHS midpoint was calculated using the bedtools intersect command. To evaluate statistical significance, the same analysis was repeated on 1000 iterations of randomly selected cytosines (Cs) from the same regions sequenced by SMM-seq. P-values were calculated based on the number of iterations (out of 1000) where the number of overlapping mutations in the random set was equal to or smaller than the number of overlapping BPDE-induced mutations.

Machine learning models for damage and repair

The genome was divided into 500 bp non-overlapping windows using bedtools makewindows. Features of gene presence, CpG islands, transposons, promoters, and enhancers were converted into binary values (0 or 1), representing their absence or presence within each 500 bp window. For quantitative features, including sequence context, CpG methylation, BPDE damage, DNase I hypersensitivity, and RNA expression, counts over windows were calculated. Nucleotide composition of each genomic window was calculated by bedtools nuc. Gene coordinates were obtained from Ensembl's hg38 genome assembly. Gene features were then mapped to the genomic windows using bedtools coverage -S retaining information on whether a window overlapped with a gene and its location relative to the transcription start sites (TSS). Gene-associated windows were flagged as '1′, while non-gene-associated windows were marked as '0'. Promoter regions were defined as 3 kb upstream of the TSS. These regions were compared to the genomic windows using bedtools coverage, and binary promoter features were created (presence = '1', absence = '0'). Transposon data were downloaded from the UCSC RepeatMasker tracks and merged with the genomic windows using bedtools coverage. Transposon presence was flagged as '1', and absence as '0' for each window. Enhancer coordinates were downloaded from Zenodo and overlap with the genomic windows was established by bedtools intersect. Windows overlapping enhancers were marked as '1' and absence as '0'.

DNase sequencing data from two replicates were merged to create a unified dataset. The coverage values were computed using bedtools coverage, quantifying DNase signal within each window. For constitutive replication timing data from [59], two BED files representing early and late replicating regions were interesected with the genomic windows using 'bedtools intersect'. Genomic windows were categorized as '1' for early replication, '2' for late replication, and '0' for regions without timing information. RNA-seq data was obtained from [6]. Replicate BAM files were merged and counts over windows calculated with bedtools coverage -S, since expression on the coding strand is expected to influence repair on the non-coding/transcribed strand. BPDE Damage-seq (from this study) and XR-seq [48] coverage values were computed using bedtools coverage -S.

Damage and repair data were classified into three categories: Class 0 (No damage/repair), Class 1 (Low, 1–10 count coverage), Class 2 (High damage/repair, >10 count coverage). Balanced sampling was applied to ensure that each category had an equal representation, with 100 000 windows drawn from each class using awk.

For model generation, continuous features, including CpG methylation, DNase hypersensitivity were scaled using StandardScaler from scikit-learn to ensure comparability across different scales. The damage and repair datasets were divided into training and testing sets with an 80–20 split. Stratified sampling was applied to ensure a balanced representation across damage and repair classes. The lazypredict library (https://github.com/shankarpandala/lazypredict) was used to evaluate various machine learning models. GridSearchCV was employed to fine-tune hyperparameters, optimizing model performance based on accuracy, precision, recall, and F1 scores.

The XGBoost classifier [63] was trained on the preprocessed dataset, and its performance was evaluated using accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) area under the curve (AUC). Performance was assessed on both the training and testing sets to ensure generalization.

Statistical analyses

All experiments were performed in at least two biological replicates. Statistical analyses were performed using R. The statistical tests used for each analysis are detailed in the figure legends.

Results

Single nucleotide resolution mapping of BPDE-dG adducts reveals enrichment of damage at sites of CpG methylation

We applied the high-sensitivity Damage-seq method [49,52] on genomic DNA isolated from GM12878 lymphoblast and human A549 lung cancer cell lines treated with 25 μM BPDE for 2 h (Fig. 1A). These cell lines are ENCODE cell lines and were chosen due to the abundance of publicly available genomic data generated from them [64]. For brevity, and since previous XR-seq BPDE repair maps were generated only for GM12878 [48], the main figures present data from this cell line. In Damage-seq, single-stranded fragments of damaged DNA were isolated from cells using an anti-BPDE-dG antibody, and the damage site was identified at single nucleotide resolution as the site where a DNA polymerase was blocked *in vitro*. Thus, in the ensuing sequencing reads, the DNA adduct was expected to be in the −1 position relative to the read start [49]. Indeed, in both GM12878 and A549 cells, and across all experimental replicates, we observe a strong enrichment of Gs at the −1 position (Fig. 1B and Supplementary Fig. S3A–C) relative to an input DNA control. Analysis of the sequence context of these BPDE-dG adduct sites indicates enrichment of C in the position 5′ to the damaged G (Fig. 1C and Supplemental Fig. S3D), regardless of the nucleotide at the 3′ position. This enrichment is consistent with higher damage formation at methylated CpGs. We stratified the genome into quartiles of DNA methylation levels based on bisulfite sequencing data (ENCODE [64]) from the same cell lines. Damage counts correlated with the methylation state of the CpGs in the genome (Fig. 1D and Supplementary Fig. S3E). Since this could be an indirect correlation, driven by different cellular or chromatin states of the methylated DNA in the genome, we repeated our experiments with DNA isolated from cells and treated with 3 μM of BPDE for 2 h *in vitro*. This *in vitro* dose yielded similar damage levels to those observed in cells (Supplementary Fig. S3F). A similar enrichment of BPDE-dG adducts in methylated regions was observed in *in vitro* treated DNA. To test whether this enrichment was directly due to DNA methylation, we amplified sonicated genomic DNA by 10 rounds of PCR to dilute DNA methylation, and then performed the *in vitro* BPDE treatment. After amplification, the enrichment of damage in methylated genomic regions was lost (Fig. 1F and Supplementary Fig. S3E and G).

The previous study of BPDE dG repair reported enrichment of CpG sequences in the XR-seq sequencing reads. In XR-seq, the excised oligos containing the damages are isolated by im-

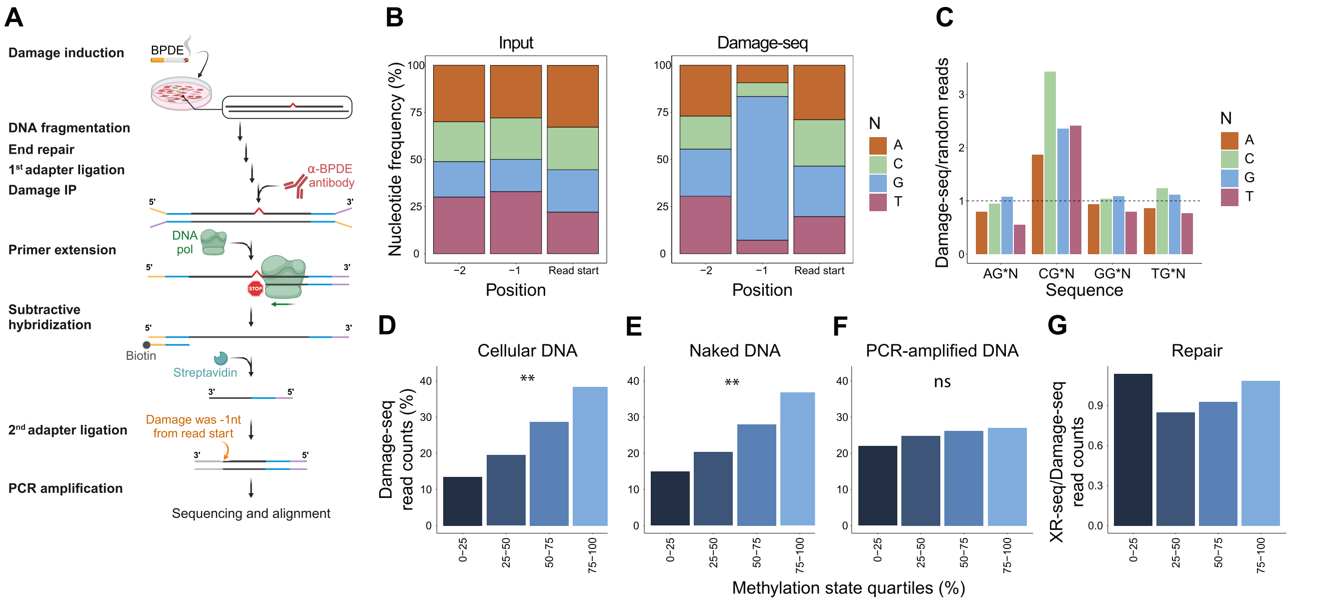


Figure 1. Mapping BPDE-dG adducts at single nucleotide resolution. (**A**) Schematic of the Damage-seq technique performed in GM12878 and A549 cell lines in this study. (**B**) Nucleotide composition at the read start and the two positions immediately upstream of it in input GM12878 DNA (left) or Damage seq (right) from the combined replicates of GM12878 cells treated with 25 μM BPDE. Enrichment of G indicates successful single-nucleotide resolution mapping of the damages. (**C**) Analysis of the sequences flanking the damaged dG (marked by '*') in GM12878 Damage-seq data reveals enrichment of C upstream of the damage position. To control for differences in sequence context in the genome, the Damage-seq frequencies are normalized to the sequence contexts of the same number of randomly selected Gs. (**D**) The percent of BPDE-dG damage read counts falling into each of the methylation state quartiles based on bisulfite sequencing data from BPDE-treated GM12878 cells. (**E**) Same as panel (D), except Damage-seq was performed on naked GM12878 DNA treated with 3 μM BPDE *in vitro*. (**F**) Same as panel (E), except genomic DNA was first sonicated, ligated to adapters, and subjected to 10 cycles of whole genome amplification by PCR prior to *in vitro* damage. (**G**) Analysis of repair of BPDE-dG adducts measured by XR-seq in GM12878 cells, after normalization to the underlying damage frequencies, over the different methylation states. ** $P < .01$, Kruskal Walis test with Benjamini–Hochberg correction.

munoprecipitation and sequenced. Since the half-life of these excised oligos in the cells is relatively short (∼30 min), XR-seq provides a snapshot of repair efficiency. This previous study conducted XR-seq at an early timepoint (1 h) after damage induction; thus, it represents sites of preferential initial repair. However, sites of elevated repair could reflect higher damage levels rather than higher repair efficiency. To test whether the repair was independently affected by the DNA methylation status of the damage, we normalized the XR-seq count data from GM12878 cells by the underlying damage levels. Repair levels did not differ significantly between the different methylation states in the genome, indicating DNA methylation sensitizes the genome to damage but did not significantly affect repair efficiency (Fig. 1G).

Preferential BPDE damage formation and repair in active and accessible chromatin regions

To investigate the effect of different chromatin states on the formation of BPDE-dG adducts, we used chromatin state annotations generated by the chromHMM model based on histone modification data collected in the GM12878 and A549 cells [64]. Damage formation in cells treated with BPDE is higher in transcriptionally active and accessible chromatin states, and lower in repressed and heterochromatic chromatin (Fig. 2A and Supplementary Fig. S4A). This enriched damage formation was not observed in nDNA from cells treated *in vitro*, indicating it is the result of chromatin accessibility in cells rather than features of the DNA itself (Fig. 2B and Supplementary Fig. S4B). BPDE-dG repair was also reported to be higher in active chromatin. To assess whether this is attributed to the higher damage levels, we normalize repair in GM12878 cells to the underlying damage levels. Normalized repair was still significantly enriched in active and accessible chromatin states (Fig. 2C).

To specifically investigate the role of chromatin accessibility, we profiled cellular damage levels at DHS sites in the same cell lines (Fig. 2D and Supplementary Fig. S4C). Damages are highly enriched at DHS peaks, and a periodic profile of damages is observed flanking the peak suggesting effects of adjacent nucleosomes. However, this enrichment is lost in *in vitro* treated DNA (Fig. 2E and Supplementary Fig. S4D). Repair normalized to the underlying damage was still highly enriched at DHS sites (Fig. 2F). Taken together, these results indicate both damage and repair are elevated in active and accessible chromatin regions.

Given the strong effect of both DNA methylation and chromatin status on damage formation, we investigated damage and repair in CpG islands. CpG islands are genomic regions of high CpG density, but the majority of CpG islands are unmethylated and within accessible chromatin [65, 66]. Damage formation in BPDE treated cells was higher in these regions than in *in vitro* treated DNA, and was further reduced if DNA methylation was first diluted by PCR (Fig. 2G and Supplementary Fig. S4E). Thus, both chromatin accessibility and DNA methylation contribute independently to the damage levels in these regions. When separating the CpG islands into methylated and unmethylated based on bisulfite sequencing data in the same cells, damage levels are higher in methylated CpG islands compared to unmethylated islands in

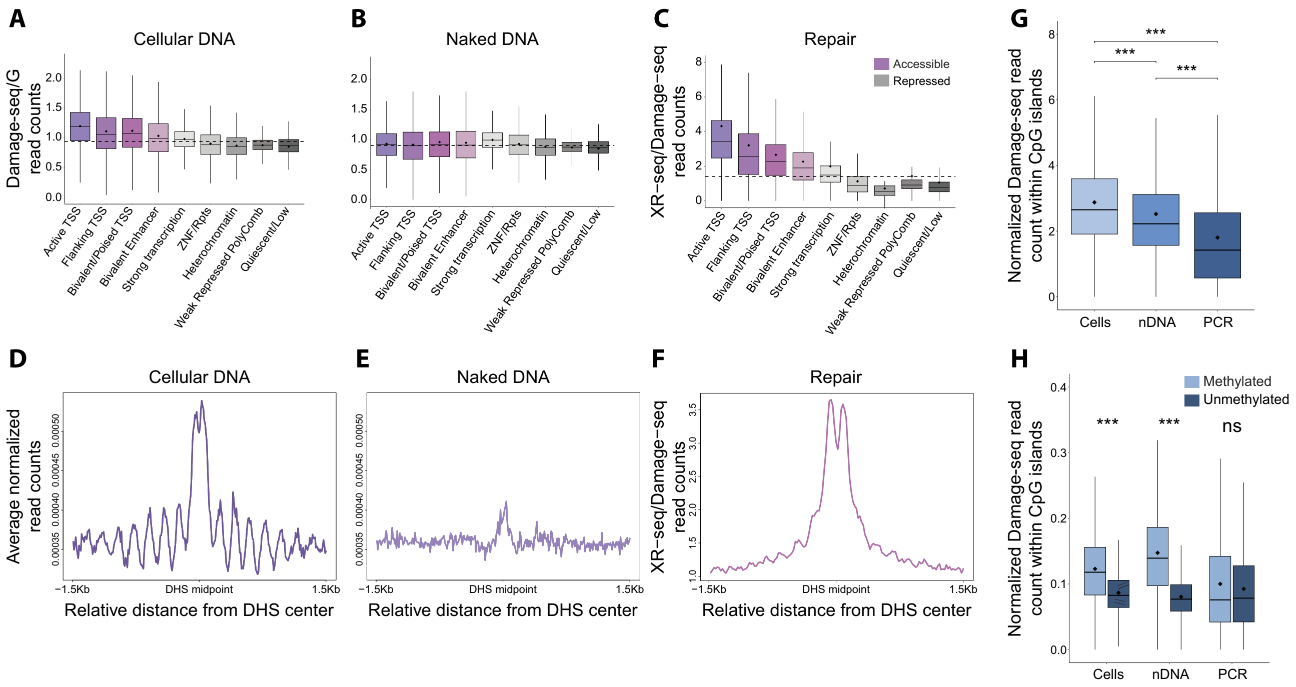


Figure 2. Preferential formation of BPDE-dG adducts in functionally active and accessible chromatin. (**A**) BPDE-dG Damage-seq read counts frequencies from GM12878 cells treated with 25 μM BPDE for 2 h over different chromatin states identified by the chromHMM algorithm. To control for sequence context effects, read counts were normalized to the underlying G nucleotide frequencies and by the total read depth. (**B**) As in panel (A), except GM12878, genomic DNA was treated with 3 μM BPDE for 2 h *in vitro*. (**C**) As in panel (A), except shown are the repair counts measured by XR-seq after normalizing to the underlying damage content. (**D**) Average density profile of BPDE-dG damage counts from GM12878 cells in the 3 kb flanking the midpoint of DHS peaks. Counts were normalized to the total read depth. (**E**) As in panel (D), except plotted is Damage-seq data from *in vitro* treated genomic DNA. (**F**) As in panel (D), except plotted is the repair signal obtained by XR-seq after normalization to the underlying damage. (**G**) BPDE-dG Damage-seq read count frequencies (per kb) normalized to total read depth over CpG islands. Compared are BPDE-dG Damage-seq results from GM12878 treated cell (cells), *in vitro* treated naked DNA (nDNA), and *in vitro* treated PCR-amplified DNA (PCR). (**H**) Same as panel (G), except CpG islands were divided into methylated and unmethylated. Boxes represent the range between 25th and 75th percentile, the line represents the median and the diamond the mean. Outliers were discarded for the presentation. ***$P < .001$, based on Wilcoxon signed-rank test with Bonferroni correction. n.s., not significant.

DNA from cells or DNA treated *in vitro*, but this difference is lost when the methylation is diluted by PCR (Fig. 2H and Supplementary Fig. S4F).

Transcription factor binding modulates damage formation of BPDE-dG adducts

Within the profiles of damage and repair at DNase hypersensitive sites we observed a local dip in signal within the peak midpoint (Fig. 2D and F). We hypothesized this dip could be due to TF binding at these sites. We therefore profiled damage and repair at active TF-binding sites. These sites were based on 286 non-redundant TF motif clusters previously reported in Vierstra *et al.* [56], encompassing 2179 total motifs for 702 distinct human TF proteins. Active TF motif sites were defined based on overlap with a DNase hypersensitive site in the same cell line (see the ‘Materials and methods’ section). To specifically isolate the effect of TF binding, the Damage-seq signal from cells was normalized to damage levels *in vitro*. Aggregating the binding sites of all 702 TFs, both BPDE-dG damage and repair were depleted at the center of TF-binding sites (Fig. 3A and B and Supplementary Fig. S5).

Investigating the effect of binding of specific TFs on damage formation is highly sensitive to the underlying sequence context (Supplementary Fig. S1). We therefore used a pentamer-based Poisson model to calculate multiple-test corrected *P*-values for the damage counts at each position across the motif, for both the forward and reverse strands. Z-score differences (ΔZ) between the damage signals in cells versus *in vitro* were calculated for each position (Fig. 3C and Supplementary Table S3). Depending on the TF and the position within the binding sites, we found both enrichment and depletion of BPDE-dG signals, suggesting that TF binding can both inhibit and stimulate damage formation. Generally, there appears to be more TF inhibition than induction of BDPE-dG adduct formation, with CTCF, NRF1, NFY, and ETS motif clusters exhibiting the widest inhibitory effects. This reduced damage formation is especially pronounced for CTCF, for which the G-rich motif strand shows a large depletion of BPDE-dG signal compared to what is expected based on DNA sequence alone (Fig. 3D).

Effects of transcription and replication timing on BPDE-dG damage and repair

Both active transcription and DNA replication are DNA-templated processes that are inhibited by BPDE damage formation, but could also directly influence genome sensitivity. To isolate the effect of active transcription on BPDE damage formation, we normalized the damage levels from cells by those in *in vitro*-treated DNA. Damages are enriched at the TSS of protein-coding genes (Fig. 4A and Supplementary Fig. S6A), likely due to enhanced chromatin accessibility. There does not appear to be a major difference in damage levels between the

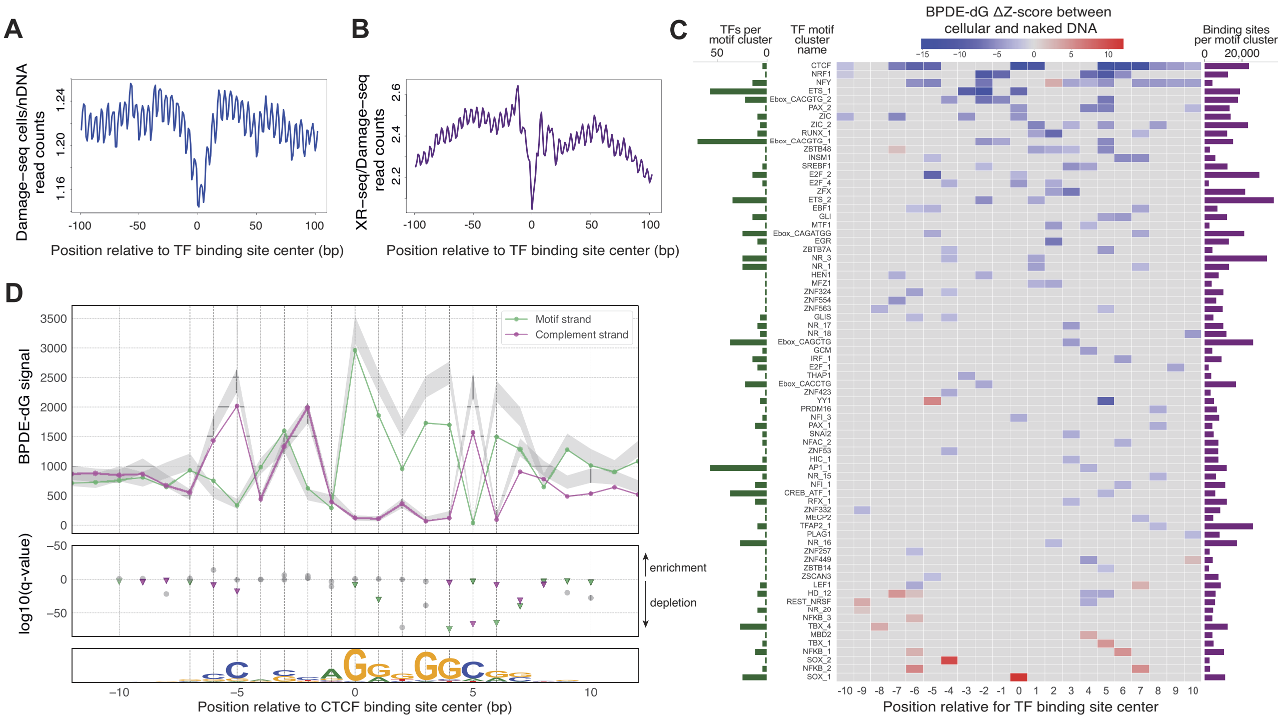


Figure 3. Modulation of BPDE-dG damage formation at TF-binding sites. (**A**) Average read density profiles over active TF-binding sites of BPDE-dG Damage-seq in GM12878 cells normalized by Damage-seq from *in vitro* treated nDNA. TF-binding sites were selected from accessible DNA regions of the human genome, and did not overlap any coding regions. (**B**) Similar to panel (A), but showing the average density of BPDE-dG repair (XR-seq signal) normalized to the underlying damage levels. (**C**) Heat map depicting differential BPDE-dG levels in the presence of active TF binding. The BPDE-dG Z-score difference between Damage-seq signals from cellular versus *in vitro*-treated DNA conditions was determined for each position and for both strands of each TF motif cluster. At each position, the ΔZ-score with the largest magnitude (either the motif or the motif-complement strand) is the cell value, and each row is a specific TF motif cluster. The bar plots on each side of the heatmap illustrate the number of distinct TFs mapped to each motif cluster (left) and the number of binding sites attributed to each motif cluster (right). See the 'Materials and methods' section for details. (**D**) Example of the full BPDE-dG analysis for the CTCF motif cluster, represented in the top row of the heatmap. The top panel shows the observed BPDE-dG signal for the motif and motif-complement strands. The predicted BPDE-dG signal ±4 SDs were calculated with a pentamer Poisson model and are represented by the shaded gray region. The middle panel shows the $\log_{10}$ transformation of the corrected *P*-values (i.e. q-values). Marker colors correspond to strand and arrow directions indicate either enrichment (up) or depletion (down) of the BPDE-dG signal. Gray markers are positions that are insignificant or considered a false-positive after comparison with the nDNA condition. The bottom panel shows a sequence logo of the position weight matrix for the CTCF motif cluster sequences (i.e. the putative CTCF binding sites) used in the analysis.

transcribed and non-transcribes strands. While in GM12878 there is a small preference for the transcribed strand, this is not observed in A549 cells (Supplementary Fig. S6B). As previously reported, repair of BPDE adducts, even after normalization of damage levels, is significantly enriched ($P < .0001$) on the transcribed strand of genes due to transcription-coupled repair (Fig. 4B). Thus, active transcription does not strongly affect damage formation but enhances the removal of damages from the transcribed strands.

We previously reported that NER of UV-induced CPDs is more efficient in gene exons compared to introns [27]. Both the formation and repair of BPDE-dG adducts is enhanced in exons compared to introns (Fig. 4C and D, and Supplementary Fig. S6C). This enhanced damage and repair is consistent with the elevated accessibility observed in gene exons (Fig. 4E).

Smoking-associated mutagenesis, specifically the BPDE-induced signature SBS4, is higher in late-replicating regions [59]. To investigate the effect of replication timing on damage formation and repair, we used constitutive early and late replicating regions identified by Yaakov et. al. in both normal and cancer cell types [59]. Both BPDE-dG damage formation and NER were elevated in early compared to late replicating regions (Fig. 4F and G, and Supplementary Fig. S6D). Furthermore, both damage and repair exhibit a local peak in average density surrounding a set of constitutive early replicating origins identified by Guilbaud *et al.* [58] (Fig. 4H and I, and Supplementary Fig. S6E). Early replicating regions and the early firing constitutive origins of replication are also characterized by more accessible chromatin (Fig. 4J and K), which could contribute to both damage-sensitivity and repair efficiency.

A machine learning model identifies DNA accessibility as the strongest predictor of damage and repair

Our results identify multiple genomic features influencing both BPDE-dG damage formation and repair efficiency. To estimate the relative importance and predictive power of the different features, we divided damage and repair data over genomic windows of 500 nt for each DNA strand into three categories (no damage/repair, medium levels of damage/repair, and high levels of damage/repair) and tested four classification models to compare their predictive abilities on 100 000 windows for each category: Support vector machine [67],

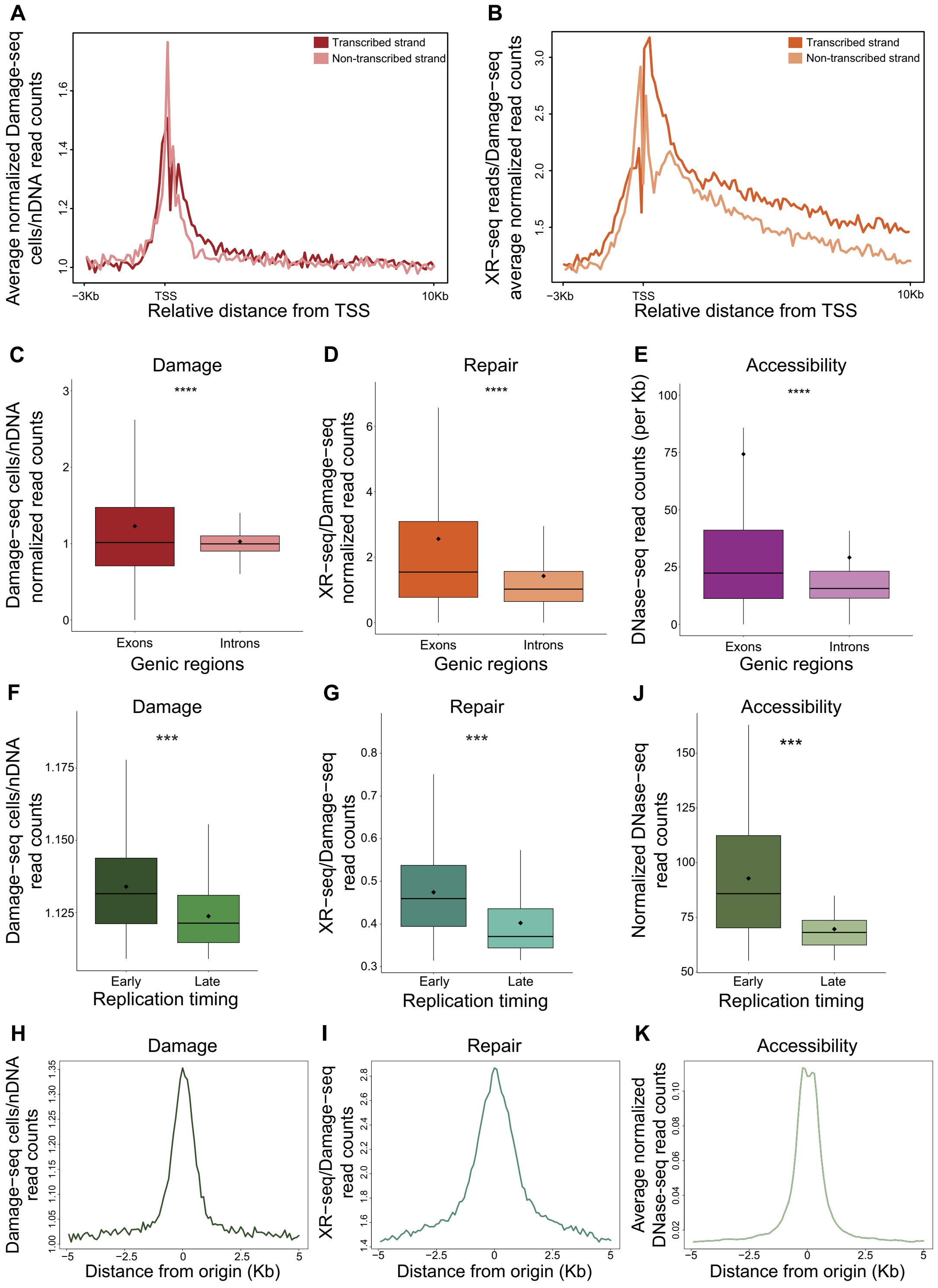


Figure 4. Effects of transcription and replication on BPDE-dG damage and repair. (**A**) Average read density profiles over genes of BPDE-dG Damage-seq in GM12878 cells normalized by Damage-seq from *in vitro* treated naked DNA (nDNA). Data are plotted separately for the transcribed (TS) and non-transcribed (NTS). TSS, transcription start site. (**B**) Similar to panel (A), except plotted is the average density of BPDE-dG repair normalized to the underlying damage levels. (**C**) Damage-seq read frequencies from GM12878 treated cells normalized by Damage-seq from *in vitro* treated DNA (nDNA) calculated over both strands of exons and introns. (**D**) Similar to panel (C), except plotted is repair normalized by the underlying damage levels. (**E**) Similar to panel (D), except plotted is the DNase-hypersensitivity read count reflecting chromatin accessibility. (**F**) BPDE-dG Damage-seq frequencies in GM12878 cells normalized by Damage-seq from *in vitro* treated DNA (nDNA) in early versus late replicating regions. (**G**) Similar to panel (F), except plotted are repair rates normalized by the underlying damage levels. (**H**) Average read density profiles surrounding constitutive early-firing origins of BPDE-dG Damage-seq in GM12878 cells normalized by Damage-seq from *in vitro* treated DNA (nDNA). (**I**) Similar to panel (H), except plotted is repair normalized by the underlying damage levels. (**J**) Similar to panel (F), except plotted is the DNase-hypersensitivity read count reflecting chromatin accessibility. (**K**) Similar to panel (H), except plotted is the average DNase-hypersensitivity read density reflecting chromatin accessibility. Boxes represent range between 25th and 75th percentile, the line represents the median and the diamond the mean. Outliers were discarded for the presentation. ***$P < .0001$, ***$P < .001$ based on Wilcoxon signed-rank test with Bonferroni correction.

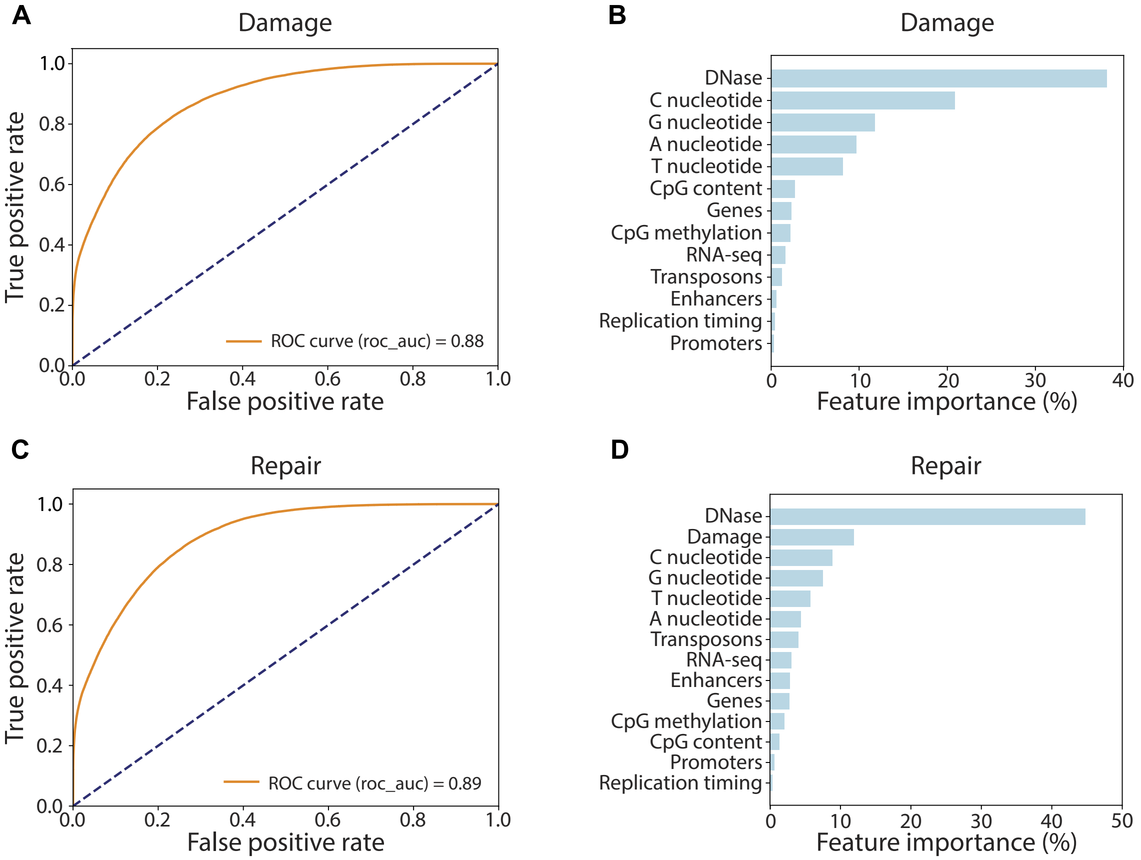


Figure 5. A machine learning model for estimating the relative importance of different genomic features in damage and repair. (**A**) ROC curve and (**B**) feature importance graph determined by the XGboost model for BPDE-dG damages in GM12878 cells. Bar plot provide a ranked list of features based on their importance in the model. Each feature's contribution (the percentage of the overall prediction process) was calculated during the training process. (**C**,**D**) As in panels (A) and (B), except for the XGboost model for repair, in which the BPDE-dG Damage-seq levels measured in GM12878 cells were included as an additional feature.

Logistic regression [68], Random Forest [69], and XGBoost [63] classifier. The features included both the cell-specific features of DNase hypersensitivity, RNA expression, and CpG methylation, as well as shared characteristics such as genomic sequence composition, constitutive replication timing, and gene, promoter, enhancer, and transposon locations. The models were evaluated based on standard performance metrics such as accuracy, precision, and recall (Supplementary Table S4). Of the classification models tested, the XGBoost classifier outperformed the others, achieving the highest predictive accuracy. As a result, XGBoost [63] was selected for further analysis (Fig. 5). Based on the model's feature importance scores, DNase hypersensitivity is the strongest predictor of BPDE-dG damage formation in GM12878 cells, followed by sequence context, CpG methylation and gene expression (Fig. 5A and B). In A549, where there are generally higher levels of DNA methylation, this feature is more prominent than gene expression (Supplementary Fig. S7). In the analysis of repair (Fig. 5C and D), damage levels generated in this study were integrated as a genomic feature. Once more, DNase hypersensitivity had the highest feature importance score, followed by the damage sensitivity of the regions. Thus, damageability strongly influences the repair profiles and must be taken into account in analyzing repair.

BPDE-induced mutagenesis reflects the sequence preferences of damage formation but its rate is determined primarily by repair efficiency

Replication across BPDE-dG adducts results in the misincorporation of A nucleotides and in G > T or C > A transversion mutations. To directly measure mutagenesis under the same experimental system used for damage and repair mapping, we treated GM12878 cells with a low dose of BPDE (0.125 μM) for two weeks (~11 population doublings) and then submitted the genomic DNA from treated and control (DMSO-treated) cells to single-molecule mutation sequencing (SMM-seq [51]; Supplementary Fig. S8A). SMM-seq is an error-corrected sequencing approach that sensitively identifies subclonal mutations within a population of cells, covering ~20% of the human genome. BPDE treatment resulted in over a three-fold increase in mutation accumulation after two weeks of treatment, with ~700 mutations in each of the two experimental replicates. Of these, ~57% were C > A SBSs characteristic of BPDE exposure (Fig. 6A, and Supplementary Fig. S8B and C).

Trinucleotide sequence context analysis of all possible base substitutions in BPDE treated cells found the pattern had the highest cosine similarity to the COSMIC SBS4 mutational signature, which is associated with tobacco smoking (Fig. 6B). C > A mutations were enriched in the CCA, CCC, CCT and

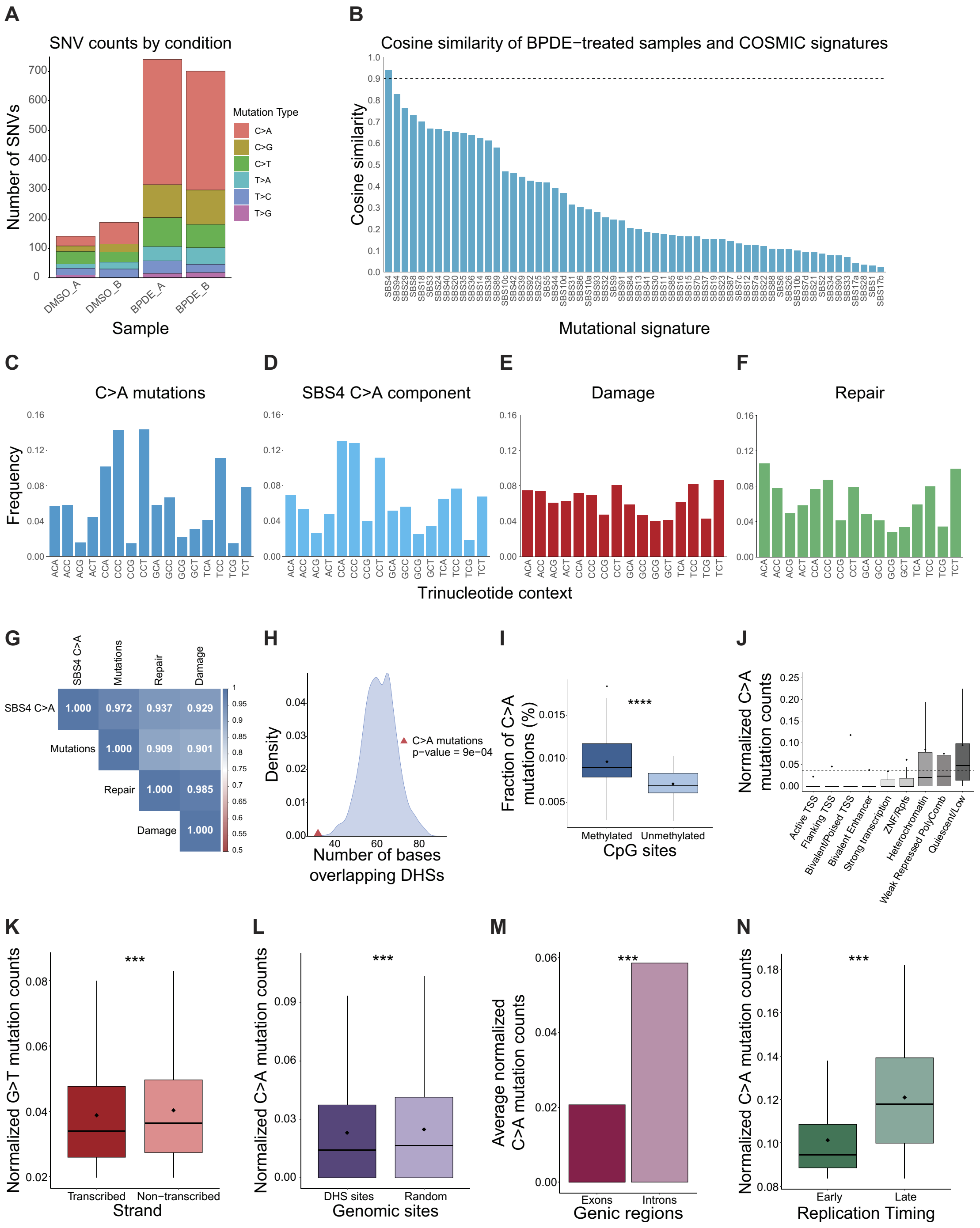


Figure 6. Genomic analysis of BPDE-induced mutagenesis. (**A**) Summary of the SBS mutation counts obtained in SMM-seq of GM12878 cells treated over two weeks with 0.125 μM BPDE or DMSO vehicle control. (**B**) Cosine similarity score between the SBS trinucleotide profile of the SMM-seq data from BPDE-treated cells and the different COSMIC SBS signatures. (**C**) Trinucleotide sequence context frequencies of the C > A mutations identified in SMM-seq of BPDE-treated cells. (**D**) The trinucleotide sequence context frequencies that compose the C > A substitutions in the COSMIC SBS4 signature. (**E**) The reverse complement of the trinucleotide frequencies of BPDE-dG damages in GM12878 treated cells. (**F**) The reverse complement of the trinucleotide frequencies of Gs identified in the XR-seq reads, reflecting sites of BPDE-dG repair in GM12878 cells. (**G**) Pairwise cosine similarities were calculated for the trinucleotide frequencies for C > A mutations or G nucleotide damage and repair. (**H**) Only 32/828 of the SMM-seq C > A mutations fall within the 3 kb regions of DHS peaks (bottom triangle). This is lower than every one of one thousand iterations of randomly selecting 828 G nucleotides within the SMM seq regions. (**I**) Frequency of C > A mutations in lung cancer samples, compared between methylated versus unmethylated CpGs. (**J**) Frequency of C > A mutations in lung cancer samples, normalized to the underlying trinucleotide sequence composition, over the different chromHMM states identified in normal lung samples, shows significant depletion in accessible and active chromatin states. (**K**) Frequency of G > T lung cancer mutations (reflecting the mutated base in the template strand), normalized by the underlying trinucleotide frequency, compared between the transcribed and non-transcribed strands of genes. (**L**) Similar to panel (J), except compared are DHS peak regions and randomly selected non-accessible regions. (**M**) Comparison of the trinucleotide-normalized C > A mutation counts from lung cancers in exons and introns of genes. (**N**) Similar to panel (J), except compared are constitutive early and late replicating regions. Boxes represent range between 25th and 75th percentile, the line represents the median and the diamond the mean. Outliers were discarded for the presentation. ****P < .0001, ***P < .001 based on Wilcoxon signed-rank test with Bonferroni correction.

TCC sequences (Fig. 6C). Very similar trinucleotide mutation patterns are observed in SBS4 signature (Fig. 6D), with a cosine similarity score of 0.972. Analysis of the reverse complement of the sequence context of the damaged dG in the XR-seq and Damage-seq data from GM12878 also gave similar profiles of C > A mutations, with high cosine similarity scores (0.909 and 0.901, respectively, Figs. 6E–G).

Damage and repair have very similar sequence composition. However, our genome-wide analyses indicate they are differently affected by genomic features. We expected that regions in the genome with high damage sensitivity, and/or low repair efficiency would harbor higher rates of mutations. However, in regions where both damage and repair were high, i.e. accessible chromatin, gene exons, and early-replicating regions, it was unclear which would exert the stronger influence on mutagenesis.

SMM-seq only produced a total of 828 C > A mutations, limiting our ability to compare different genomic regions with high confidence. Only 32 mutations occurred within 3 kb of a DNase hypersensitivity midpoint in GM12878. By comparing the data to 1000 iterations of randomly selected C nucleotides from the SMM-seq sequenced regions, we found that these mutations were significantly depleted in accessible regions (Fig. 6H).

We therefore analyzed the distribution of C > A mutations identified in WGS of lung cancer samples from smokers by the Cancer Genome Atlas project (TCGA). Higher C > A mutation rates were observed at methylated CpGs compared to unmethylated CpGs, consistent with higher damage formation but similar repair efficiencies at these sites (Fig. 6I). Comparing mutagenesis across the different chromatin states of lung cancer tissues, active and accessible chromatin regions display significantly lower mutation rates (Fig. 6J). The lower mutation rates on the transcribed strand of genes (Fig. 6K) are attributed to transcription-coupled repair. To study the effect of chromatin accessibility on lung cancer mutagenesis, we used DNase hypersensitivity measurements from normal lung samples. A small but statistically significant difference was observed, with fewer mutations mapped to accessible regions compared to non-accessible regions (Fig. 6L). Similarly, mutagenesis was lower in gene exons compared to introns (Fig. 6M), and early compared to late replicating regions (Fig. 6N). Thus, in the accessible regions, which exhibited higher damage formation but also higher repair efficiency, lower mutagenesis rates are observed.

Discussion

BPDE-dG adducts belong to the category of bulky, helix distorting, DNA damages. This category also include UV-induced CPD and (6-4) pyrimidine-pyrimidone photoproduct [(6-4)PP], and adducts induced by the chemotherapy drug Cis-diamminedichloroplatinum (cisplatin) [70]. Genome-wide mapping of these damages by CPD-seq and Damage-seq indicates that while the rotational setting of the nucleosomes affects damage formation, different chromatin states of chromatin accessibility showed overall similar damage levels [23, 71]. With BPDE damage, however, chromatin accessibility significantly enhanced damage formation. DNA methylation moderately sensitized cytosine-containing dimers (TC or CC) to CPD damage formation after UV-B irradiation (and not after UV-C), and did not appear to strongly influence cisplatin adduct formation. Thus, compared to the previously studied NER substrates, the analysis of the effects of BPDE damages on mutagenesis is significantly more complex.

Using a genomic approach, comparing data sets from multiple sources rather than performing experiments to map chromatin components, damage, and repair simultaneously could theoretically introduce inter-lab variability and miss certain effects. By performing the experiments in the same cell lines and under the same growth conditions as the external data sources, we aimed to minimize such variabilities.

Here, we present the first study where BPDE-induced DNA damages, DNA repair, chromatin and genomic features, and damage-induced mutagenesis were measured and compared in the same experimental system of GM12878 cells (Fig. 7).

BPDE induces conformationally distinct adducts, which may exhibit different damage formation and repair rates. The most common adduct formed by the (±)-anti-BPDE exposure used in this study is the (+)-trans-N^2-BPDE-dG [72, 73]. It is also the preferential adduct recognized by the antibody used in both Damage-seq and XR-seq protocols [74, 75], and therefore our results likely represent primarily the damageability and repair of this conformation.

Methylated CpGs accumulate higher damage levels (Fig. 1). This is likely due to increased intercalative binding of BPDE to sites of methylated CpGs [76] and enhanced reactivity of the guanine due to the base-paired 5meC placing the N^2 position in a favorable orientation for a nucleophilic attack [73]. A previous report using damaged plasmids indicated that DNA methylation could both enhance or repress excision repair efficiency, depending on the sequence context [77]. However, we did not find an effect of the methylation status on excision repair efficiency in XR-seq data from the genome. Our analysis of cancer mutagenesis finds higher C > A mutations in methylated CpGs (Fig. 6). It is important to note that for the cancer mutagenesis analyses, we used DNA methylation data from normal tissues, as DNA methylation patterns could alter during cancer development. In fact, there are reports that BPDE exposure alters DNA methylation patterns in cells [78–83]. Future studies could investigate these complex interactions in experimental models of tumor development.

Accessible chromatin regions are more sensitive to BPDE damage formation, but also more efficiently repaired. These include regulatory regions in the genome, gene exons and early replicating regions. Both our SMM-seq results and cancer mutagenesis data indicate that accessible regions accumulate less mutations. This lower mutation frequency could be directly due to the effects of chromatin accessibility, but could also be due to selective pressure at important functional regions, especially exons. Still, this observation suggests that repair, rather than damageability, could be a stronger determinant of the final mutagenic patterns. While cells have a limited ability to control their exposure to damaging agents, they can activate checkpoint mechanisms to extend the time available for repair in order to restrict their mutagenic outcomes. BPDE damages were reported to stabilize nucleosomes *in vitro* [84]. It will thus be interesting to investigate whether there are accessory mechanisms that specifically facilitate the repair of nucleosomal templates carrying BPDE.

To our knowledge, this is the first study of the effect of TF binding on BPDE damage formation. For the majority of TFs investigated, including CTCF and ETS-family TFs, binding reduced damage formation across multiple positions within their binding sites. This is markedly different from what was previously observed for UV-induced damage, which

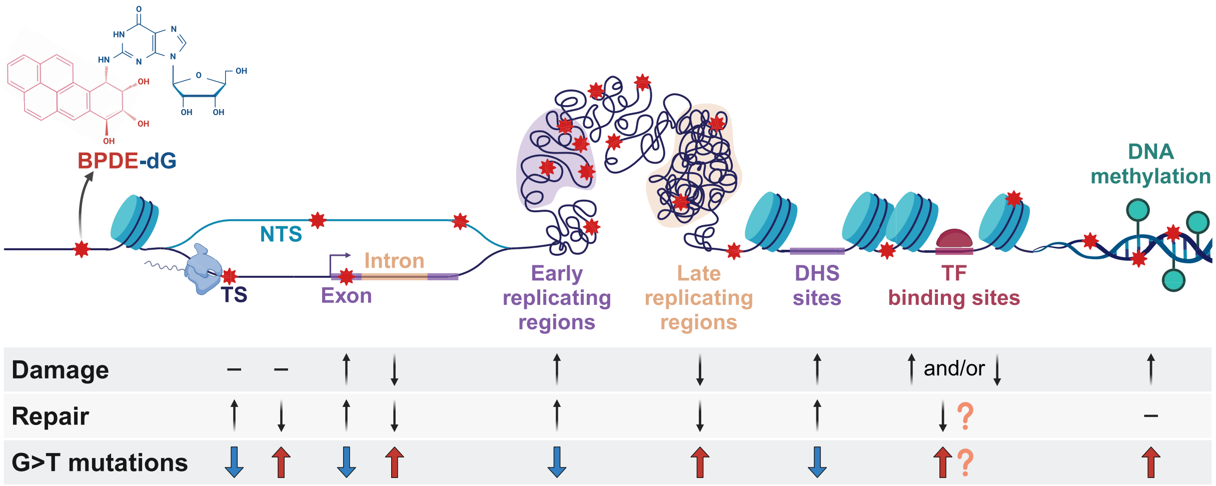


Figure 7. A model of the differential contribution of BPDE-dG damage and repair to mutagenesis across the human genome. Damage sensitivity and repair efficiency both influence the final mutagenic pattern. At methylated CpGs, elevated damage sensitivity drives higher mutagenesis. At accessible genomic regions, elevated repair appears to be a stronger determinant than higher damage sensitivity in the final mutagenic profile. This model applies only to passenger mutations that are not subject to selective pressure.

was highly enriched at specific positions within sites of active TF binding, especially for CTCF and ETS proteins [23, 24, 34, 35, 85]. This trend is consistent with the hypothesis that BPDE has reduced access to DNA that is actively bound by TFs. Interestingly, a minority of TFs exhibited enhanced damage formation at certain positions within the binding sites, highlighting that the effect of TF binding on damage formation is not uniform across TF families. Furthermore, the effect may not even be uniform in different positions within the binding site of a specific TF. This was reported for UV damage formation in the CTCF motif [85], and is also observed in the YY1 and HD 12 clusters (Fig. 3C), where BPDE-dG damage is elevated at one position, but repressed at another. Thus, analysis of the effect of TF binding on damage formation requires careful analysis on an individual TF and position basis to avoid convoluting the inhibition versus the stimulation of damage formation.

Repair profiles at TF-binding sites indicate sites of localized decrease in repair, which is consistent with previous reports for other NER-substrates [86–88]. However, these results should be interpreted with care, as the effect of BPDE-dG damage on TF binding has not yet been determined. While two *in vitro* studies reported that BPDE modification can increase the binding of SP1, E2F1 and E2F4 to DNA [89, 90], it remains to be investigated whether this holds true in cells, whether the binding is strong enough to have an effect on DNA repair, and whether other TFs also interact with their target sites after BPDE adducts have formed. Future studies will need to investigate this question in order to allow for a comprehensive modeling of the fate of BPDE damages at TF-binding sites.

An intriguing question is the effect of the three-dimensional organization of chromatin on damage and repair. Two studies have reported higher UV damage and lower repair rates in the periphery of the nucleus [91, 92]. It will be interesting in future studies to test the effect of nuclear architecture on BPDE damage and repair.

BPDE-dG damage and repair presented a similar trinucleotide distribution to the mutagenic signature produced in cell lines and in cancer samples. This similarity indicates that the sequence context of both repair and mutations is dictated primarily by preferences in BPDE-dG damage formation. However, our results indicate the frequency of these mutations across the genome could be influenced by repair efficiency. While cancer sequencing efforts have been focused on functional, actionable, mutations, this new approach of analysis of the passenger mutations could provide information on the cancer cell state. Analysis of NER capacity in peripheral blood indicates it differs between individuals, and thus could have an additive effect on lung cancer risk in smokers [93]. This insight is important in analyzing the SBS4 mutational signature in cancer samples. While the existence of this signature indicates exposure to cigarette smoke, its prevalence and genomic distribution could reflect repair efficiency and thus be used a therapeutic and prognostic biomarker.

Acknowledgements

The authors would like to thank Dr Avital Parnas, Elnatan Bitensky, Hadar Golan Berman, and Prof. Itamar Simon for fruitful discussion and technical assistance, Dr Abed Nasereddin and Dr Idit Shiff from the Core Research Facilities at the Faculty of Medicine in the Hebrew University for support in sequencing protocols.

Author contributions: E.E.H. and S.A. conceptualized the project, E.E.H., D.T., and Y.C. performed the investigation, E.E.H., D.T., H.W., and A.K. performed formal analysis, E.E.H. and S.A. wrote the original draft and all authors reviewed and edited the manuscript, S.A. and R.G. supervised the work and acquired funding.

Supplementary data

Supplementary data are available at NAR online.

Conflict of interest

None declared.

Funding

Funding to support this work was provided by the Israel Science Foundation grant 482/22, the Israel Cancer Research Fund program grant 24-102-PG (S.A), US–Israel Binational Science Foundation grant BSF-2019272, and National Science Foundation BSF-NSF grant 2 324 614 (R.G. and S.A.). E.E.H. is the recipient of the Israel Council for Higher Education Scholarship.

Data availability

All raw and processed sequencing data generated in this study have been submitted to the Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) under accession number PRJNA1177470 and PRJNA1179438.

References

1. Munnia A, Giese RW, Polvani S *et al.* Bulky DNA adducts, tobacco smoking, genetic susceptibility, and lung cancer risk. *Adv Clin Chem* 2017;**81**:231–77. https://doi.org/10.1016/bs.acc.2017.01.006
2. Phillips DH, Venitt S. DNA and protein adducts in human tissues resulting from exposure to tobacco smoke. *Intl J Cancer* 2012;**131**:2733–53. https://doi.org/10.1002/ijc.27827
3. Ma B, Stepanov I, Hecht SS. Recent studies on DNA adducts resulting from Human exposure to tobacco smoke. *Toxics* 2019;**7**:16. https://doi.org/10.3390/toxics7010016
4. Leffler S, Pulkrabek P, Grunberger D *et al.* Template activity of calf thymus DNA modified by a dihydrodiol epoxide derivative of benzo[a]pyrene. *Biochemistry* 1977;**16**:3133–36. https://doi.org/10.1021/bi00633a015
5. Nath ST, Lee MS, Romano LJ. Effect of carcinogenic adducts on transcription by T7 RNA polymerase. *Nucleic Acids Res* 1987;**15**:4257–71. https://doi.org/10.1093/nar/15.10.4257
6. Merav M, Bitensky EM, Heilbrun EE *et al.* Gene architecture is a determinant of the transcriptional response to bulky DNA damages. *Life Sci Alliance* 2024;**7**:e202302328. https://doi.org/10.26508/lsa.202302328
7. Marien K, Mathews K, van Holde K *et al.* Replication blocks and sequence interaction specificities in the codon 12 region of the c-ha-ras proto-oncogene induced by four carcinogens in vitro. *J Biol Chem* 1989;**264**:13226–32. https://doi.org/10.1016/S0021-9258(18)51618-5
8. Moore P, Strauss BS. Sites of inhibition of in vitro DNA synthesis in carcinogen- and UV-treated phi X174 DNA. *Nature* 1979;**278**:664–66. https://doi.org/10.1038/278664a0
9. Hassanain H, Tseitline D, Hacohen T *et al.* A practical site-specific method for the detection of bulky DNA damages. *J Mol Biol* 2024;**436**:168450. https://doi.org/10.1016/j.jmb.2024.168450
10. Gibbons DL, Byers LA, Kurie JM. Smoking, p53 mutation, and lung cancer. *Mol Cancer Res* 2014;**12**:3–13. https://doi.org/10.1158/1541-7786.MCR-13-0539
11. Alexandrov LB, Ju YS, Haase K *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* 2016;**354**:618–22. https://doi.org/10.1126/science.aag0299
12. Nik-Zainal S, Kucab JE, Morganella S *et al.* The genome as a record of environmental exposure. *Mutagenesis* 2015;**30**:763–70.
13. Kucab JE, Zou X, Morganella S *et al.* A compendium of mutational signatures of environmental agents. *Cell* 2019;**177**:821–836. https://doi.org/10.1016/j.cell.2019.03.001
14. Sancar A. Mechanisms of DNA repair by photolyase and excision nuclease (Nobel Lecture). *Angew Chem Int Ed* 2016;**55**:8502–27. https://doi.org/10.1002/anie.201601524
15. Spivak G. Nucleotide excision repair in humans. *DNA Repair (Amst)* 2015;**36**:13–8. https://doi.org/10.1016/j.dnarep.2015.09.003
16. Cleaver JE, Lam ET, Revet I. Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. *Nat Rev Genet* 2009;**10**:756–68. https://doi.org/10.1038/nrg2663
17. Kiyohara C, Takayama K, Nakanishi Y. Lung cancer risk and genetic polymorphisms in DNA repair pathways: a meta-analysis. *J Nucleic Acids* 2010;**2010**:701760. https://doi.org/10.4061/2010/701760
18. Popanda O, Schattenberg T, Phong CT *et al.* Specific combinations of DNA repair gene variants and increased risk for non-small cell lung cancer. *Carcinogenesis* 2004;**25**:2433–41. https://doi.org/10.1093/carcin/bgh264
19. Li W, Li K, Zhao L *et al.* DNA repair pathway genes and lung cancer susceptibility: a meta-analysis. *Gene* 2014;**538**:361–65. https://doi.org/10.1016/j.gene.2013.12.028
20. Wang M, Liu H, Liu Z *et al.* Genetic variant in DNA repair gene GTF2H4 is associated with lung cancer risk: a large-scale analysis of six published GWAS datasets in the TRICL consortium. *Carcinogenesis* 2016;**37**:888–96. https://doi.org/10.1093/carcin/bgw070
21. Cohen Y, Adar S. Novel insights into bulky DNA damage formation and nucleotide excision repair from high-resolution genomics. *DNA Repair (Amst)* 2023;**130**:103549. https://doi.org/10.1016/j.dnarep.2023.103549
22. Hu J, Adar S, Selby CP *et al.* Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev* 2015;**29**:948–60.
23. Hu J, Adebali O, Adar S *et al.* Dynamic maps of UV damage formation and repair for the human genome. *Proc Natl Acad Sci USA* 2017;**114**:6758–63. https://doi.org/10.1073/pnas.1706522114
24. Mao P, Brown AJ, Esaki S *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat Commun* 2018;**9**:2626. https://doi.org/10.1038/s41467-018-05064-0
25. Mao P, Smerdon MJ, Roberts SA *et al.* Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc Natl Acad Sci USA* 2016;**113**:9057–62. https://doi.org/10.1073/pnas.1606667113
26. Teng Y, Bennett M, Evans KE *et al.* A novel method for the genome-wide high resolution analysis of DNA damage. *Nucleic Acids Res* 2011;**39**:e10. https://doi.org/10.1093/nar/gkq1036
27. Heilbrun EE, Merav M, Adar S. Exons and introns exhibit transcriptional strand asymmetry of dinucleotide distribution, damage formation and DNA repair. *NAR Genom Bioinform* 2021;**3**:lqab020. https://doi.org/10.1093/nargab/lqab020
28. Brown AJ, Mao P, Smerdon MJ *et al.* Nucleosome positions establish an extended mutation signature in melanoma. *PLoS Genet* 2018;**14**:e1007823. https://doi.org/10.1371/journal.pgen.1007823
29. Pich O, Muinos F, Sabarinathan R *et al.* Somatic and germline mutation periodicity follow the orientation of the DNA Minor groove around nucleosomes. *Cell* 2018;**175**:1074–1087. https://doi.org/10.1016/j.cell.2018.10.004
30. Elliott K, Singh VK, Bostrom M *et al.* Base-resolution UV footprinting by sequencing reveals distinctive damage signatures for DNA-binding proteins. *Nat Commun* 2023;**14**:2701. https://doi.org/10.1038/s41467-023-38266-2
31. Selvam K, Sivapragasam S, Poon GMK *et al.* Detecting recurrent passenger mutations in melanoma by targeted UV damage sequencing. *Nat Commun* 2023;**14**:2702. https://doi.org/10.1038/s41467-023-38265-3
32. Bohm KA, Morledge-Hampton B, Stevison S *et al.* Genome-wide maps of rare and atypical UV photoproducts reveal distinct patterns of damage formation and mutagenesis in yeast chromatin.

Proc Natl Acad Sci USA 2023;**120**:e2216907120. https://doi.org/10.1073/pnas.2216907120

33. Yancoskie MN, Khaleghi R, Gururajan A *et al.* ASH1L guards cis-regulatory elements against cyclobutane pyrimidine dimer induction. *Nucleic Acids Res* 2024;**52**:8254–70. https://doi.org/10.1093/nar/gkae517
34. Elliott K, Bostrom M, Filges S *et al.* Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLoS Genet* 2018;**14**:e1007849. https://doi.org/10.1371/journal.pgen.1007849
35. Premi S, Han L, Mehta S *et al.* Genomic sites hypersensitive to ultraviolet radiation. *Proc Natl Acad Sci USA* 2019;**116**:24196–205. https://doi.org/10.1073/pnas.1907860116
36. Adar S, Hu J, Lieb JD *et al.* Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc Natl Acad Sci USA* 2016;**113**:E2124–2133. https://doi.org/10.1073/pnas.1603388113
37. Denissenko MF, Pao A, Tang M *et al.* Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* 1996;**274**:430–32. https://doi.org/10.1126/science.274.5286.430
38. Thrall BD, Mann DB, Smerdon MJ *et al.* Nucleosome structure modulates benzo[a]pyrenediol epoxide adduct formation. *Biochemistry* 1994;**33**:2210–16. https://doi.org/10.1021/bi00174a030
39. Denissenko MF, Chen JX, Tang MS *et al.* Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proc Natl Acad Sci USA* 1997;**94**:3893–98. https://doi.org/10.1073/pnas.94.8.3893
40. Weisenberger DJ, Romano LJ. Cytosine methylation in a CpG sequence leads to enhanced reactivity with benzo[a]pyrene diol epoxide that correlates with a conformational change. *J Biol Chem* 1999;**274**:23948–55. https://doi.org/10.1074/jbc.274.34.23948
41. Chen JX, Zheng Y, West M *et al.* Carcinogens preferentially bind at methylated CpG in the p53 mutational hot spots. *Cancer Res* 1998;**58**:2070–75.
42. Tretyakova N, Matter B, Jones R *et al.* Formation of benzo[a]pyrene diol epoxide-DNA adducts at specific guanines within K-ras and p53 gene sequences: stable isotope-labeling mass spectrometry approach. *Biochemistry* 2002;**41**:9535–44. https://doi.org/10.1021/bi025540i
43. Tretyakova N, Guza R, Matter B. Endogenous cytosine methylation and the formation of carcinogen carcinogen-DNA adducts. *Nucleic Acids Symp Ser* 2008;49–50. https://doi.org/10.1093/nass/nrn025
44. Hu W, Feng Z, Tang MS. Preferential carcinogen-DNA adduct formation at codons 12 and 14 in the human K-ras gene and their possible mechanisms. *Biochemistry* 2003;**42**:10012–23. https://doi.org/10.1021/bi034631s
45. Tang MS, Zheng JB, Denissenko MF *et al.* 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* 1999;**20**:1085–89. https://doi.org/10.1093/carcin/20.6.1085
46. Yoon JH, Smith LE, Feng Z *et al.* Methylated CpG dinucleotides are the preferential targets for G-to-T transversion mutations induced by benzo[a]pyrene diol epoxide in mammalian cells: similarities with the p53 mutation spectrum in smoking-associated lung cancers. *Cancer Res* 2001;**61**:7110–17.
47. Dong H, Bonala RR, Suzuki N *et al.* Mutagenic potential of benzo[a]pyrene-derived DNA adducts positioned in codon 273 of the human P53 gene. *Biochemistry* 2004;**43**:15922–28. https://doi.org/10.1021/bi0482194
48. Li W, Hu J, Adebali O *et al.* Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc Natl Acad Sci USA* 2017;**114**:6752–57. https://doi.org/10.1073/pnas.1706021114
49. Hu J, Lieb JD, Sancar A *et al.* Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc Natl Acad Sci USA* 2016;**113**:11507–11512. https://doi.org/10.1073/pnas.1614430113
50. Jiang Y, Mingard C, Huber SM *et al.* Quantification and mapping of alkylation in the human genome reveal single nucleotide resolution precursors of mutational signatures. *ACS Cent Sci* 2023;**9**:362–72. https://doi.org/10.1021/acscentsci.2c01100
51. Maslov AY, Makhortov S, Sun S *et al.* Single-molecule, quantitative detection of low-abundance somatic mutations by high-throughput sequencing. *Sci Adv* 2022;**8**:eabm3259. https://doi.org/10.1126/sciadv.abm3259
52. Zhu Y, Tan Y, Li L *et al.* Genome-wide mapping of protein–DNA damage interaction by PADD-seq. *Nucleic Acids Res* 2023;**51**:e32. https://doi.org/10.1093/nar/gkad008
53. McKenna A, Hanna M, Banks E *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303. https://doi.org/10.1101/gr.107524.110
54. Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools.. *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008
55. Blokzijl F, Janssen R, van Boxtel R *et al.* MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* 2018;**10**:33. https://doi.org/10.1186/s13073-018-0539-0
56. Vierstra J, Lazar J, Sandstrom R *et al.* Global reference mapping of human transcription factor footprints. *Nature* 2020;**583**:729–36. https://doi.org/10.1038/s41586-020-2528-x
57. Korhonen J, Martinmaki P, Pizzi C *et al.* MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* 2009;**25**:3181–82. https://doi.org/10.1093/bioinformatics/btp554
58. Guilbaud G, Murat P, Wilkes HS *et al.* Determination of human DNA replication origin position and efficiency reveals principles of initiation zone organisation. *Nucleic Acids Res* 2022;**50**:7436–50. https://doi.org/10.1093/nar/gkac555
59. Yaacov A, Vardi O, Blumenfeld B *et al.* Cancer mutational processes vary in their association with replication timing and chromatin accessibility. *Cancer Res* 2021;**81**:6106–16. https://doi.org/10.1158/0008-5472.CAN-21-2039
60. Godambe AV, Patil GP. Some characterizations involving additivity and infinite divisibility and their applications to Poisson mixtures and Poisson sums. In: Patil GP, Kotz S, Ord JK (eds.), *A Modern Course on Statistical Distributions in Scientific Work*. *NATO Advanced Study Institutes Series*. Dordrecht: Springer Netherlands, 1975, 339–51. https://doi.org/10.1007/978-94-010-1848-7_31
61. Cerami E, Gao J, Dogrusoz U *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4. https://doi.org/10.1158/2159-8290.CD-12-0095
62. Zhao H, Sun Z, Wang J *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2014;**30**:1006–7. https://doi.org/10.1093/bioinformatics/btt730
63. Chen T, Guestrin C. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: Association for Computing Machinery, 2016, 785–94.
64. Encode Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247
65. Angeloni A, Bogdanovic O. Sequence determinants, function, and evolution of CpG islands. *Biochem Soc Trans* 2021;**49**:1109–19. https://doi.org/10.1042/BST20200695
66. Bock C, Paulsen M, Tierling S *et al.* CpG island methylation in human lymphocytes is highly correlated with DNA sequence,

repeats, and predicted DNA structure. *PLoS Genet* 2006;**2**:e26. https://doi.org/10.1371/journal.pgen.0020026

67. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. https://doi.org/10.1007/BF00994018
68. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2000, 1–30.
69. Breiman L. Random forests. *Machine Learning* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324
70. Friedberg EC, Walker GC, Siede W *et al. DNA Repair and Mutagenesis*. ASM Press, 2006.
71. Mao P, Smerdon MJ, Roberts SA *et al.* Asymmetric repair of UV damage in nucleosomes imposes a DNA strand polarity on somatic mutations in skin cancer. *Genome Res* 2020;**30**:12–21. https://doi.org/10.1101/gr.253146.119
72. Szeliga J, Dipple A. DNA adduct formation by polycyclic aromatic hydrocarbon dihydrodiol epoxides. *Chem Res Toxicol* 1998;**11**:1–11. https://doi.org/10.1021/tx970142f
73. Guza R, Kotandeniya D, Murphy K *et al.* Influence of C-5 substituted cytosine and related nucleoside analogs on the formation of benzo[a]pyrene diol epoxide-dG adducts at CG base pairs of DNA. *Nucleic Acids Res* 2011;**39**:3988–4006. https://doi.org/10.1093/nar/gkq1341
74. Wang C, Li T, Wang Z *et al.* Quantitative study of stereospecific binding of monoclonal antibody to anti-benzo (a)pyrene diol epoxide-N (2)-dG adducts by capillary electrophoresis immunoassay. *J Chromatogr A* 2010;**1217**:2254–61. https://doi.org/10.1016/j.chroma.2010.02.024
75. Hsu TM, Liu TM, Amin S *et al.* Determination of stereospecificity of benzo[a]pyrene diolepoxide-DNA antisera with site-specifically modified oligonucleotides. *Carcinogenesis* 1995;**16**:2263–65. https://doi.org/10.1093/carcin/16.9.2263
76. Geacintov NE, Shahbaz M, Ibanez V *et al.* Base-sequence dependence of noncovalent complex formation and reactivity of benzo[a]pyrene diol epoxide with polynucleotides. *Biochemistry* 1988;**27**:8380–87. https://doi.org/10.1021/bi00422a013
77. Muheim R, Buterin T, Colgate KC *et al.* Modulation of human nucleotide excision repair by 5-methylcytosines. *Biochemistry* 2003;**42**:3247–54. https://doi.org/10.1021/bi0268504
78. He Z, Zhang R, Chen S *et al.* FLT1 hypermethylation is involved in polycyclic aromatic hydrocarbons-induced cell transformation. *Environ Pollut* 2019;**252**:607–15. https://doi.org/10.1016/j.envpol.2019.05.137
79. Pfeifer GP, Grunberger D, Drahovsky D. Impaired enzymatic methylation of BPDE-modified DNA. *Carcinogenesis* 1984;**5**:931–35. https://doi.org/10.1093/carcin/5.7.931
80. Wilson VL, Jones PA. Inhibition of DNA methylation by chemical carcinogens in vitro. *Cell* 1983;**32**:239–46. https://doi.org/10.1016/0092-8674(83)90514-7
81. Wilson VL, Jones PA. Chemical carcinogen-mediated decreases in DNA 5-methylcytosine content of BALB/3T3 cells. *Carcinogenesis* 1984;**5**:1027–31. https://doi.org/10.1093/carcin/5.8.1027
82. Yang P, Ma J, Zhang B *et al.* CpG site-specific hypermethylation of p16INK4alpha in peripheral blood lymphocytes of PAH-exposed workers. *Cancer Epidemiol Biomarkers Prev* 2012;**21**:182–90. https://doi.org/10.1158/1055-9965.EPI-11-0784
83. Ye F, Xu XC. Benzo[a]pyrene diol epoxide suppresses retinoic acid receptor-beta2 expression by recruiting DNA (cytosine-5-)-methyltransferase 3A. *Mol Cancer* 2010;**9**:93. https://doi.org/10.1186/1476-4598-9-93
84. Mann DB, Springer DL, Smerdon MJ. DNA damage can alter the stability of nucleosomes: effects are dependent on damage type. *Proc Natl Acad Sci USA* 1997;**94**:2215–20. https://doi.org/10.1073/pnas.94.6.2215
85. Sivapragasam S, Stark B, Albrecht AV *et al.* CTCF binding modulates UV damage formation to promote mutation hot spots in melanoma. *EMBO J* 2021;**40**:e107795. https://doi.org/10.15252/embj.2021107795
86. Frigola J, Sabarinathan R, Gonzalez-Perez A *et al.* Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. *Nucleic Acids Res* 2021;**49**:891–901. https://doi.org/10.1093/nar/gkaa1219
87. Perera D, Poulos RC, Shah A *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* 2016;**532**:259–63. https://doi.org/10.1038/nature17437
88. Sabarinathan R, Mularoni L, Deu-Pons J *et al.* Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* 2016;**532**:264–67. https://doi.org/10.1038/nature17661
89. Johnson DG, Coleman A, Powell KL *et al.* High-affinity binding of the cell cycle-regulated transcription factors E2F1 and E2F4 to benzo[a]pyrene diol epoxide-DNA adducts. *Mol Carcinog* 1997;**20**:216–23. https://doi.org/10.1002/(SICI)1098-2744(199710)20:2%3c216::AID-MC8%3e3.0.CO;2-K
90. MacLeod MC, Powell KL, Tran N. Binding of the transcription factor, Sp1, to non-target sites in DNA modified by benzo[a]pyrene diol epoxide. *Carcinogenesis* 1995;**16**:975–83. https://doi.org/10.1093/carcin/16.5.975
91. Garcia-Nieto PE, Schwartz EK, King DA *et al.* Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J* 2017;**36**:2829–43. https://doi.org/10.15252/embj.201796717
92. Akkose U, Adebali O. The interplay of 3D genome organization with UV-induced DNA damage and repair. *J Biol Chem* 2023;**299**:104679. https://doi.org/10.1016/j.jbc.2023.104679
93. Shen H, Spitz MR, Qiao Y *et al.* Smoking, DNA repair capacity and risk of nonsmall cell lung cancer. *Intl J Cancer* 2003;**107**:84–8. https://doi.org/10.1002/ijc.11346

Received: October 30, 2024. **Revised:** January 10, 2025. **Editorial Decision:** January 13, 2025. **Accepted:** January 20, 2025