Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene

Wentao Li1, Jinchuan Hu2, Ogun Adebalia3, Sheera Adara3,4, Yanyan Yang5, Yi-Ying Chiou5, and Aziz Sancar1,1

*Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC 27599; and Department of Microbiology and Molecular Genetics, The Hebrew University Hadassah Medical School, Jerusalem 91120, Israel

Contributed by Aziz Sancar, May 11, 2017 (sent for review April 11, 2017; reviewed by Núria López-Bigas and Gerd P. Pfeifer)

Benzo[a]pyrene (BaP), a polycyclic aromatic hydrocarbon, is the major cause of lung cancer. BaP forms covalent DNA adducts after metabolic activation and induces mutations. We have developed a method for capturing oligonucleotides carrying bulky base adducts, including UV-induced cyclobutane pyrimidine dimers (CPDs) and BaP diol epoxide-deoxyguanosine (BPDE-dG), which are removed from the genome by nucleotide excision repair. The isolated oligonucleotides are ligated to adaptors, and after damage-specific immunoprecipitation, the adaptor-ligated oligonucleotides are converted to dsDNA with an appropriate translesion DNA synthesis (TLS) polymerase, followed by PCR amplification and next-generation sequencing (NGS) to generate genome-wide repair maps. We have termed this method translesion excision repair-sequencing (tXR-seq). In contrast to our previously described XR-seq method, tXR-seq does not depend on repair/removal of the damage in the excised oligonucleotides, and thus it is applicable to essentially all DNA damages processed by nucleotide excision repair. Here we present the repair maps for CPDs and BPDE-dG adducts generated by tXR-Seq for the human genome. In addition, we report the sequence specificity of BPDE-dG excision repair using tXR-seq.

Results

tXR-seq of CPD and BPDE-dG. To perform tXR-seq, the human lymphocyte cell line GM12878, which has been extensively characterized by the ENCODE project (14), was either exposed to UV (20 J/m² at 254 nm) or treated with (±)-anti-BPDE (2 μM). After allowing 4 h or 1 h for repair, respectively, cells were lysed gently, and low molecular weight DNA, consisting mostly of excised 26–27 mers, was separated from chromosomal high molecular weight DNA by centrifugation.

The basic features of sequencing library generation for tXR-seq are shown in Fig. 1. In brief, the excised oligomers are immunoprecipitated with either anti-CPD antibody or anti-XPB and anti-p62 antibodies, because the excision product is released in a relatively stable complex with TFIH, which also protects the excision product from degradation by nonspecific nucleases. This is followed by extraction of the excised oligomers and ligation to adaptors and a second immunoprecipitation of the sequences to the genome. Here we present the application of this approach to the mapping of CPD repair and BPDE-dG repair in the human genome.

Significance

Benzo[a]pyrene (BaP) is a widespread potent carcinogen found in food, coal tar, cigarette smoke, and industrial smoke. Cigarette smoking is the leading cause of lung cancer, and the mutagenesis in smoking-associated lung cancer is determined by multiple factors, including nucleotide excision repair. We have developed a general method for genome-wide mapping of nucleotide excision repair at single-nucleotide resolution and applied it to generate repair maps of UV- and BaP-induced DNA damage in human. Results show a novel sequence specificity of BaP diol epoxide-deoxyguanosine repair. This general method can be used to study repair of all types of DNA damage that undergo nucleotide excision repair.

Author contributions: W.L., J.H., and A.S. designed research; W.L. performed research; W.L., J.H., Q.A., S.A., Y.Y., Y.-Y.C., and A.S. analyzed data; and W.L. and A.S. wrote the paper.

Reviewers: N.L.-B., Institute for Research in Biomedicine Barcelona; and G.P.P., Van Andel Research Institute of the City of Hope.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo/ (accession no. GSE97675).

1To whom correspondence should be addressed. Email: aziz_sancar@med.unc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1706021114/-/DCSupplemental.

PNAS | June 27, 2017 | vol. 114 | no. 26 | 6752–6757

www.pnas.org/cgi/doi/10.1073/pnas.1706021114
Li et al.

**Sequence Specificity of BPDE-dG Repair.** Several studies have been conducted on the effects of nearest-neighbor and distant-neighbor sequences on the repair rate of BPDE-dG. The data have been interpreted within the context of structural deformity and duplex backbone flexibility caused by the adduct in various sequence contexts, and some general rules have been derived (24–27). Although these pioneering studies have been quite useful, some of the rules were derived from the excision rates of the BPDE-dG adduct from in vitro experiments with human cell-free extracts and oligonucleotide duplexes 135 bp in length (28).

The tXR-seq method, which produces at least ~20 million reads in each biological replicate, provides a model-independent means for analyzing the effects of nearest and distant (one base flanking the nearest bases) sequence neighbors on BPDE-dG excision efficiencies. To secure the vast majority of the excised oligonucleotides used in this analysis containing BPDE-dG at the same position, we classified the sequencing reads from tXR-seq based on read lengths. An analysis of single-nucleotide frequencies along reads 26–29 nt from both tXR-seq and conventional XR-seq for CPD (Fig. S1A and B). In both experiments, there was strong TT enrichment at the fixed position of 6 nt from the 3′ end, further confirming the validity of tXR-seq.

**Length Distribution and Nucleotide Frequency Analysis of Excision Products.** The PCR products, approximately 145 bp long, can be seen only in the presence of purified excised oligonucleotides containing either CPD or BPDE-dG, because polymerases can bypass CPD and BPDE-dG damage, respectively, in an error-free manner (15–17). This is followed by PCR amplification with index primers and NGS on the Illumina HiSeq 2500 platform. Anti-BPDE antibody. Human DNA polymerases  and  are used for primer extension in the presence of purified excised oligonucleotides containing either CPD or BPDE-dG, because polymerases and can bypass CPD and BPDE-dG damage, respectively, in an error-free manner (15–17). This is followed by PCR amplification with index primers and NGS on the Illumina HiSeq 2500 platform.

*Fig. 1.* Schematic of sequencing library construction strategy for tXR-seq. The red asterisk represents a DNA lesion processed by nucleotide excision repair. Excised oligomers are precipitated with TFIH antibodies, extracted, and ligated to adaptors. Then, the adaptor-containing oligomers are precipitated with anti-lesion antibody, the primer with index is annealed and extended by an appropriate TLS DNA polymerase (pink), and the extension product is amplified by PCR to obtain the dsDNA library for NGS.

by CPD photolyase. In both CPD tXR-seq and BPDE-dG tXR-seq cases, oligomers in the size range of 24–28 nt predominate, with a median size of 26 nt, in agreement with the well-established dual incision mode of the human nucleotide excision repair system (18–20). Importantly, the length distribution of the oligomers obtained by tXR-seq for CPD is in remarkable agreement with that obtained by conventional CPD XR-seq, affirming the validity of tXR-seq.

We further analyzed the base sequence distributions along the excised CPD oligomers obtained by tXR-seq and conventional XR-seq (Fig. 2C). As expected, for CPD oligomers isolated by both conventional XR-seq and tXR-seq methods, positions 19–21 from the 5′ end were enriched for thymines (Ts), in agreement with the dual incision pattern of the human excision nuclease system (21, 22). In contrast, for the BPDE-dG oligomers, guanines (Gs) are enriched at positions 19–22, consistent with BPDE-dG monoadducts at one of these positions (Fig. 2D). As a control, when randomly chosen oligomers (10 million) of 26 nt from the human genome were analyzed for sequence composition, all four nucleotides were uniformly distributed throughout the length of the oligomers, with ~60% AT and 40% GC at all positions, consistent with Chargaff’s rules (23) (Fig. 2D). We also analyzed the frequencies of di-pyrimidines (TT, TC, CT, and CC) at each position along the excised oligomers of 26–29 nt from both tXR-seq and conventional XR-seq for CPD (Fig. S1A and B). In both experiments, there was strong TT enrichment at the fixed position of 6 nt from the 3′ end, further confirming the validity of tXR-seq.
positions (Fig. 3 B and C and Table 1). The 3′ nearest-neighbor base also followed the same rule, except for AG*, which had an A>C>T order. For G*A and G*T, the 5′ nearest-neighbor bases with the second-highest excision frequency were A and T, respectively. We also selected the specific G at positions 19, 20, and 21 for 26, 27, and 28 oligomers, respectively, and analyzed the sequence context effect in the same way (Fig. S3 and Table S1). The results suggest that the excision frequency was similar and also generally followed the C>T>A rule. Finally, these conclusions must be qualified, because they are based on the assumption of no sequence effect on adduct formation. In fact, BPDE preferentially binds at methylated CpG sites over any other NpG sites (29). In view of these reported sequence preferences for damage formation, genome-wide BPDE-dG damage distribution must be experimentally demonstrated by Damage-seq (10) for a definitive generalization of the sequence effect on repair.

Effect of Transcription on BPDE-dG Repair. Transcription stimulates repair on the transcribed strand (TS) (30). As a general rule, this effect is more apparent in highly transcribed genes and for lesions that are poorly recognized by the core nucleotide excision repair machinery. To explore the effect of transcription on BPDE-dG repair, we first generated both CPD and BPDE-dG repair maps and integrated the total RNA-seq signal across the entire human genome using tXR-seq reads and ENCODE RNA-seq data from GM12878 cells (Fig. S4). We chose the TP53 gene to analyze the effect of transcription on BPDE-dG repair for several reasons. First, it is mutated as a driver both in UV-induced and cigarette smoke-induced cancers. Second, it is transcribed at moderately high levels in many cell lines, including the GM12878 cell line used in this study. Finally, previous high-resolution repair studies by ligation-mediated PCR (LM-PCR) on short fragments of TP53 may be used as references for tXR-seq data (31). Fig. 4 shows the transcription and repair maps of human chromosome 17, which carries the TP53 gene, at resolutions ranging from megabase to single base or dinucleotide in TP53. This figure also shows tXR-seq data for both CPD and BPDE-dG repair in GM12878 cells, as well as two excision products that can be unambiguously assigned to specific TT and G damages, respectively (Fig. 4). As is apparent from the repair data at all resolutions, the TS for both TP53 and the adjacent WRAP53 gene transcribed in opposite direction were repaired at higher efficiency than the nontranscribed strand (NTS). In addition, for
both damage types, there appeared to be hotspots for repair, although deeper sequencing is needed to be able to comment on the cause and significance of these repair hotspots.

In addition, we analyzed CPD and BPDE-dG repair around the transcription start sites (TSS) and transcription end sites (TES) in NHF1 and GM12878 cell lines, respectively (Fig. 5). The trend was in general agreement with the data obtained from the NHF1 cell line for the repair of CPD and (6-4)PP as measured by the conventional XR-seq method. As expected, all of the repair profiles had repair peaks around TSS, in agreement with the documented high RNA polymerase II density and nascent RNA levels in this region (13, 32–34). In addition, as we reported previously (13), a low level of enriched excision repair near TSS was also seen in CB1 cell line, which can be linked to the open chromatin structure close to TSS. In fact, the repair ratio at 1 h for BPDE-dG fell somewhere between that for CPD and (6-4)PP, indicating that BPDE-dG is recognized by the core excision repair complex at an affinity between that of (6-4)PP and CPD (Fig. 5A).

Effect of Chromatin States on Repair. Based on DNA sequence elements and histone posttranslational modifications, 15 chromatin states have been defined (35). In previous studies, we found that although these states did not affect cisplatin-induced damage formation, they did affect repair efficiencies, particularly in active promoters, enhancers, and transcribed gene bodies (10, 13). We analyzed all 15 chromatin states for CPD, BPDE-dG, and (6-4)PP repair (Fig. 5B). Again, the general pattern observed in tXR-seq was similar to the patterns seen in CPD and (6-4)PP repair obtained by conventional XR-seq. Interestingly, for both CPD and BPDE-dG repair at 1 h after UV treatment, both exhibited high repair levels over active chromatin states but had distinct repair levels over TS and NTS around TSS and TES (Fig. 5A and B). For BPDE-dG repair, there was only a minor increase on TS compared with NTS, whereas there was much higher repair on TS than on NTS in CPD repair. This finding further confirms the unique BPDE-dG repair characteristics.

The foregoing findings not only show that BPDE-dG as a bulky adduct is processed by the cell in a manner similar to other bulky DNA lesions, but also show that tXR-seq, as a simpler and more versatile method, can be used for damages not suitable for analysis by the XR-seq method.

Discussion
The recently developed XR-seq method has been quite useful for mapping nucleotide excision repair of UV and cisplatin and oxaliplatin damage. However, the method in its original form is applicable only to damages that can be either enzymatically or chemically reversed. Here we have used TLS polymerases to circumvent this obstacle and made the method more widely applicable. In doing so, we have also generated a human genome applicable only to damages that can be either enzymatically or chemically reversed. Here we have used TLS polymerases to circumvent this obstacle and made the method more widely applicable. In doing so, we have also generated a human genome

Table 1. Sequence specificity of BPDE-dG repair

<table>
<thead>
<tr>
<th>5’ distant</th>
<th>XG*</th>
<th>3’ nearest</th>
<th>5’ nearest</th>
<th>G*X</th>
<th>3’ distant</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&gt;T-A</td>
<td>CG*</td>
<td>C&gt;T&gt;A</td>
<td>C&gt;A&gt;T</td>
<td>G*C</td>
<td>C&gt;T&gt;A</td>
</tr>
<tr>
<td>C&gt;T-A</td>
<td>AG*</td>
<td>A&gt;C&gt;T</td>
<td>C&gt;A&gt;T</td>
<td>G*A</td>
<td>C&gt;T&gt;A</td>
</tr>
<tr>
<td>C&gt;T-A</td>
<td>TG*</td>
<td>C&gt;T&gt;A</td>
<td>C&gt;T&gt;A</td>
<td>G’T</td>
<td>T&gt;A-T</td>
</tr>
</tbody>
</table>

G* represents the BPDE-dG lesion, and X represents the possible A, C, and T bases.

and that these cancer-causing mutagenic adducts are in the nontranscribed strand and are repaired at slower rates than other BPDE-dG adducts in the same strand in TP53 gene, in the context of G* TC (157), CG* G (248), and CG* T (273) (31). The sequencing depth in this tXR-seq was insufficient to detect repair at these sites and thus make a meaningful comparison with repair at other sites to confirm these findings. With new improvements in NGS technology, we expect that deeper sequencing will be affordable and will allow us to address the issue of whether carcinogenic mutations at these sites are isolated with high damage frequency combined with poor repair efficiency to give rise to lung cancers.

We note that although the DNA polymerases η and θ that we used in this study are not applicable to all lesions processed by nucleotide excision repair. However, many TLS polymerases have been identified and characterized in recent years (36, 37); thus, for virtually any bulky DNA lesion, it is possible to use an appropriate TLS polymerase or a combination of two TLS polymerases to accomplish translesion synthesis and construct genomic DNA repair maps. In conclusion, the tXR-seq method...
opens the door to genome-wide mapping repair of all types of DNA damages processed by nucleotide excision repair at single-nucleotide resolution.

Nonetheless, we note that although tXR-seq is a significant improvement over XR-seq, further technical and computational optimizations are needed for rigorous and quantitative analyses of all damages processed by nucleotide excision repair. The translesion synthesis efficiency and accuracy of the TLS polymerases used need to be taken into consideration, as does the neighboring sequence effect on TLS. In the analysis of BPDE-dG adds, the elimination of sequences of reads containing Gs three bases upstream and four bases downstream of the added G residue precludes the analysis of neighboring Gs on excision efficiency of the added G. Although these limitations are important, we believe that we will be able to address them in the near future. The translesion synthesis efficiency and neighboring sequence effects for most TLS polymerases are known, and we expect to determine these factors for the TLS polymerases that have not yet been analyzed. Regarding the neighboring G effects on BPDE-dG adds, we are developing computational tools to predict with precision the adducted G even when there are neighboring Gs, and thus we will be able to analyze the effect of all sequence contexts on the repair of BPDE-dG damage and the damage induced by other agents, such as N-acetoxy-2-acetylaminofluorene, that specifically attack G residues.

Materials and Methods

Antibodies and TLS DNA Polymerases. The following antibodies were used in this study: anti-mouse IgG (sc-2025), anti-rabbit IgG (sc-2027), anti-XPB (sc-293), and anti-p62 (sc-292) from Santa Cruz Biotechnology; rabbit anti-mouse IgG (ab46540) from Abcam; anti-CPD from Kamiya Biomedical; and mouse IgG (ab46540) from Abcam; anti-BPDE (clone 8E11) from Trevigen. TLS DNA polymerases used need to be taken into consideration, as does the neighboring sequence effect on TLS. The figures represent data from merged two biological replicates.

UV Irradiation and BPDE Treatment. GM12878 cells were grown to ~80% confluency before UV irradiation or (+)-anti-BPDE (MRIGlobal) treatment. The UV (20 J/m² at 254 nm) irradiation was performed as described previously (6), except that 20 mL of GM12878 cells were irradiated in RPMI medium 1640 (no phenol red) and 2 mM glutamine at 37 °C in a 5% CO₂ humidified chamber.

Cell Line and Culture Conditions. Human GM12878 cells were cultured in RPMI medium 1640 (no phenol red) with 15% FBS and 2 mM glutamine at 37 °C in a 5% CO₂ humidified chamber.

tXR-seq Library Preparation and Sequencing. After UV or BPDE treatment, GM12878 cells were incubated for 4 h and 1 h respectively. At appropriate time points, cells were collected by centrifugation and lysed in ice-cold Buffer A (25 mM HEPES pH 7.9, 100 mM KCl, 12 mM MgCl₂, 0.5 mM EDTA, 2 mM DTT, 12.5% glycerol, 0.5% Nonidet P-40) for 10 min, followed by 40 strokes in a glass hand homogenizer. The chromatin fraction was pelleted by centrifugation at 16,800 × g for 30 min at 4 °C, and the supernatant was subjected to TFIIF immunoprecipitation by adding 2 μg of anti-XBP, 1 μg of anti-p62, and 200 μg of RNase A (Sigma-Aldrich; R4642) per 20 mL of original cell culture.

After 3 h, rotation at 4 °C, 15 μL of recombinant protein A/G Plus-agarose (Santa Cruz Biotechnology; sc-2000) and 20 mL of normal cell culture was added, and the mixture was gently rotated overnight at 4 °C. The excised oligonucleotides were eluted from the recombinant protein A/G Plus-agarose beads by 100 μL of Buffer C (10 mM Tris·Cl pH 7.5, 1 mM EDTA, and 1% SDS) for 15 min at 65 °C after two washes with Buffer A and Buffer B (25 mM HEPES pH 7.9, 100 mM KCl, 12 mM MgCl₂, 0.5 mM EDTA, 2 mM DTT, 12.5% glycerol, and 1% Nonidet P-40), respectively. The eluted excised oligonucleotides were extracted by phenol-chloroform, and the ethanol-precipitated excised oligos were incubated with 5 μL of RNase A/T1 mixture (Thermo Fisher Scientific; EN0551) for 1 h at 37 °C. Before adaption ligation, another round of phenol-chloroform extraction was applied, and a G50 filtration column (GE Healthcare) was used to further purify the excised oligonucleotides.

For one adaption ligation reaction, 1 μL of ‘S’ adaptor (20 pmol), 2 μL of ‘T’ adaptor (40 pmol), and 1.2 μL of 10 K hybridization buffer (20 mM Tris·HC, pH 8.5, 500 mM NaCl, and 0.2 mM EDTA) were added into the 5.8 μL of purified excised oligonucleotides, and the reaction mixture was incubated at 60 °C for 10 min, followed by a 5-min incubation at 16 °C. Then 4 μL of 5× ligation buffer, 1 μL of T4 DNA ligase (Thermo Fisher Scientific; 15224-01), 1 μL of 50% PEG8000 (New England BioLabs), and 4 μL of ddH₂O were added to the mixture, and the ligation reaction mixture was incubated overnight at 16 °C. The ‘S’ adaptor and ‘T’ adaptor were the same as described previously (6).

After phenol-chloroform extraction and ethanol precipitation, the excised oligonucleotides with adaptors were boiled for 5 min and then immediately placed in ice water. Either 2 μL of anti-CPD or 4 μL of anti-BPDE was pre-incubated with 5 μL of Protein G Dynabeads (Thermo Fisher Scientific; 1004D), 5 μL of anti-rabbit Dynabeads (Thermo Fisher Scientific; 11203D) and 4 μL of 50% rabbit anti-mouse IgG (Abcam; ab46540) for 2–3 h, followed by mixing with 100 μL of Reaction Buffer (20 mM Tris·Cl at pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, and 0.5% sodium deoxycholate) containing the denatured excised oligonucleotides with adaptors. The mixtures were rotated at 4 °C overnight. The subsequent washing, elution, phenol-chloroform extraction, and ethanol precipitation steps were performed as described previously (4).

In the primer extension step, human DNA polymerases η and κ were used to bypass CPD and BPDE-dG damage, respectively. The 10 μL of excised oligonucleotides with adaptors was mixed with 15 μL of 2× TLS polymerase buffer (50 mM potassium phosphate pH 7.0, 10 mM MgCl₂, 5 mM TTP, 200 μg/mL BSA, 200 μg/mL bovine serum albumin, and 4 μL of 10 μM RTP primer (where r represents different index sequences). The single-cycle primer extension procedure is as follows: initial denaturing at 98 °C for 3 min; ramping to 65 °C at 0.1 °C/min and holding for 10 min; ramping to 37 °C at 0.1 °C/min;
addition of either 2 μL of polymerase η or 1 μL of polymerase κ into the reaction with mixing; and incubation at 37 °C for 30 min. Then the primer extension product was purified by phenol-chloroform extraction and ethanol precipitation.

The dsDNA library was constructed by PCR amplification using Kapa HotStart ReadyMix with RP1 and RP1 primers, and purified from 10% native polyacrylamide gel as described previously (6). In CPD tXR-seq and BPDe-DG tXR-seq, the primer extension products were amplified for 11 and 14 PCR cycles, respectively. All tXR-seq sequencing libraries were sequenced on the Illumina HiSeq 2500 platform.

Read Processing and Genome Alignment. At least 22 million reads were obtained in tXR-seq. All adaptors in reads were trimmed using Trimmomatic, and duplicate reads were removed with the FASTX-Toolkit (38). Reads >50 mer are filtered before further analysis. The processed reads were aligned to the hg19 human female genome using bowtie with the following command options: -x -q –best -p 4.

Data Visualization. All sequencing data were obtained from two biological replicates of tXR-seq dsDNA libraries. To compare the repair signal, we normalized all count data by the sequencing depth and visualized them in the Integrative Genomics Viewer (40). Sequencing data for both CPD tXR-seq and (6-4)PP tXR-seq at 1 h in HNF1 cells were obtained from a previously published dataset (6). Genome Expression Omnibus (GEO) accession no. GSE67941. The raw data and bigwig tracks for tXR-seq in this study were deposited in the GEO database, https://www.ncbi.nlm.nih.gov/geo/ (accession no. GSE97675).

Ensemble Data. GM12878 long nonpolyA RNA-seq [University of California Santa Cruz (UCSC) genome browser accession no. wgEncodeEH000148], chromatin state datasets for GM12878 (accession no. wgEncodeEH000784), and NHLF (accession no. wgEncodeEH000792) were downloaded from the ENCODE portal (genome.ucsc.edu/ENCODE).

Length Distribution and Nucleotide Frequencies. We obtained the length distribution and nucleotide frequencies using custom scripts after adaptor trimming and removal of duplicate reads. To normalize nucleotide frequencies in excised oligonucleotides to human whole genome nucleotide frequencies, we used our previously published unmapped human genomic sequencing dataset (10) (GEO accession no. GSE82213) and selected size-specific genomic sequences at random for further analysis.

Repair Profiles of Strands and Chromatin State Analysis. The transcript coordinates from hg19 reference genome were downloaded from the UCSC genome browser. The transcripts with an expression score of ≥200 were taken into account. Among those transcripts, those with another transcript in the 6 kb downstream vicinity and those <15 kb were filtered out. Regions 5 kb upstream and 15 kb downstream from the transcription start and end sites were binned into 100-bp windows.

The dataset of chromatin states for GM12878 mapped on the hg19 reference genome was downloaded from ENCODE. Reads per kilobase per mapped reads (RPKM) values were calibrated and plotted. Two biological replicate sets were combined to generate the boxplots. Bedtools software was used to count the reads for each analysis. All quantitative data were analyzed and plotted using R or GraphPad Prism 6 software.

ACKNOWLEDGMENTS. This work was supported by NIH Grants GM118102 and ES027255.

Fig. S1. Dipyrimidine frequencies along the reads of 26–29 nt from CPD tXR-seq in GM12878 cells and conventional CPD XR-seq in NHF1 cells. The enrichment of TT frequency is at a fixed position of 6 nt from 3’ end in both CPD tXR-seq (A) and conventional CPD XR-seq (B). The fixed position in all reads of different lengths is highlighted in dashed line.
Fig. S2. Single-nucleotide frequencies for 26–29 oligomers obtained by BPDE-dG tXR-seq. The enrichment of Gs at the two positions of 6 nt from 3′ end is highlighted by the turquoise dashed line. The figures represent data from merged two biological replicates.

Fig. S3. Sequence specificity of BPDE-dG repair. (A) Same as in Fig. 3A, except specific Gs were chosen at positions 19, 20, and 21 for 26, 27, and 28 oligomers, respectively. (B and C) Same as in A, except that the trinucleotide frequencies were used for excised oligomers and random oligomers. The figures represent data from merged two biological replicates.
Fig. S4. Whole genome map of CPD and BPDE-dG repair generated using the tXR-seq method in GM12878 cells. ENCODE nonpolyA RNA-seq signals in blue, separated by strands, are plotted on top of the tXR-seq tracks.

Table S1. Sequence specificity of BPDE-dG repair

<table>
<thead>
<tr>
<th>5’ distant</th>
<th>XG*</th>
<th>3’ nearest</th>
<th>5’ nearest</th>
<th>G*X</th>
<th>3’ distant</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&gt;T&gt;A</td>
<td>CG*</td>
<td>C&gt;T&gt;A</td>
<td>C&gt;T&gt;A</td>
<td>G*C</td>
<td>A&gt;T&gt;C</td>
</tr>
<tr>
<td>C&gt;T&gt;A</td>
<td>TG*</td>
<td>T&gt;C&gt;A</td>
<td>C&gt;T&gt;A</td>
<td>G*T</td>
<td>A&gt;C&gt;T</td>
</tr>
<tr>
<td>C&gt;T&gt;A</td>
<td>AG*</td>
<td>C&gt;A&gt;T</td>
<td>C&gt;A&gt;T</td>
<td>G*A</td>
<td>C&gt;A&gt;T</td>
</tr>
</tbody>
</table>

G* represents the BPDE-dG lesion, and X represents the possible A, C, and T bases.