SUPPLEMENTAL METHODS

UV irradiation, yeast genomic DNA isolation, and CPD-seq library preparation

Yeast strains were grown in YPD (yeast extract-peptone-dextrose) to mid-log phase (OD_{600} ≈ 0.8). Yeast cells were collected and resuspended in sterile deionized H_2O and irradiated with 125 J/m^2 UVC light. Samples were taken before UV irradiation (i.e., No UV) and immediately after UV treatment (i.e., 0hr). The remaining cells were incubated in fresh YPD for repair for different times (i.e., 2hr). Cells in each sample were centrifuged and the cell pellet was stored at -80˚C. For genomic DNA isolation, yeast cells were suspended in lysis buffer [2% (vol/vol) Triton X-100, 1% SDS, 100 mM NaCl, 10 mM Tris-HCl, pH 8.0, 1 mM EDTA], PCI (phenol:chloroform:isopropanol=25:24:1), glass beads and vortexed vigorously (Mao et al. 2016). Purified genomic DNA was quantified and ~12 µg of DNA was used for each CPD-seq library preparation. CPD-seq libraries were barcoded and mixed at equal volume for multiplexed sequencing. The detailed CPD-seq protocol, including library preparation, sequencing reads alignment, and extraction of putative CPD sites, is described in our recent study (Mao et al. 2016), with the only major difference being that the 3' end of each sequencing read was trimmed by 1 bp, since in some cases this nucleotide corresponded to the A-tailed end of the sonicated DNA fragment. CPD-seq data for the wild-type (WT) 0hr, 1hr, and 2hr repair time points were originally published in (Mao et al. 2016).

Yeast CPD-seq data analysis

CPD-seq sequencing reads associated with lesions at dipyrimidine sequences (i.e., TT, TC, CT, and CC) were included for subsequent data analysis and processing to

generate wig files, which were used for subsequent analysis. This method in effect counts a CPD read for both positions in the dinucleotide damage site. Bin analysis of CPD repair along the TS and NTS of each yeast gene (e.g., Supplemental Figure S1C) was performed as previously described (Mao et al. 2017), except the transcription start site (TSS) and transcription end site (TES) of each yeast gene were obtained from a more recent study (Park et al. 2014). The TS and NTS of each gene were divided into 6 equal sized bins. Three additional bins (167 bp for each) were tabulated upstream of the TSS and downstream of the TES, respectively. We also corrected a small error in the code from our original study (Mao et al. 2016) in computing CPD repair along the NTS. The number of CPD reads in each bin was counted at each repair time point (e.g., 2hr) and divided by the CPD reads at 0hr. The resulting fraction of CPDs remaining (i.e., CPD-2hr/CPD-0hr) was calculated for each bin separately for the TS and NTS.

CPD repair was also analyzed at single nucleotide resolution in DNA regions neighboring the TSS of 5,205 yeast genes (Park et al. 2014), after excluding genes that overlap with the repetitive rDNA (i.e., YLR154C-G, YLR154C-H, YLR154W-C, YLR155C, YLR161W, YLR162W, YLR162W-A) and *CUP1* (i.e., YHR053C, YHR054C, YHR055C) loci. The number of CPD reads along each DNA strand (i.e., NTS and TS) was counted at positions ranging from 200 bp upstream and 640 bp downstream of the TSS. To account for differences in DNA sequence composition, the number of CPD reads at each position was normalized to the dipyrimidine (i.e., TT, TC, CT, and CC) frequency at each position to determine the number of CPD reads per dipyrimidine sequence (e.g., Figure 1A). If an individual nucleotide was part of two dipyrimidine sequences (e.g., T<u>C</u>T; underlined C makes up part of a T<u>C</u> and a <u>C</u>T dipyrimidine), then

it was counted as two dipyrimidine sequences. To analyze repair efficiency, the number of CPD reads associated with each position relative to the TSS was divided by the initial CPD read count at 0hr. The resulting fraction of remaining CPDs, which inversely correlates with DNA repair efficiency, is used to estimate repair efficiency at each nucleotide position. For comparison, we plotted the average nucleosome dyad coverage from a published MNase-seq nucleosome map (Weiner et al. 2015), or the average nucleosome dyad score for a published high-resolution chemical cleavage nucleosome map (Brogaard et al. 2012), relative to the TSS.

The gene cluster and treeview programs (Eisen et al. 1998; Saldanha 2004) were used to display the fraction of CPDs remaining for genes across the yeast genome. The fraction of CPDs remaining was calculated for 30 bp sliding window that moved at 15 bp steps from -500 to +640 nt from the TSS of 5,205 yeast genes (Park et al. 2014). For this gene plot analysis, CPD lesions were assigned half-integer positions between the two pyrimidines comprising the CPD lesions, and repair along both DNA strands (TS and NTS) were analyzed separately. A similar procedure was used to plot the normalized CPD levels (i.e., CPDs per dipyrimidine position) across all yeast genes. Genes were sorted by their published transcription rate (Holstege et al. 1998).

Calculating repair asymmetry in yeast nucleosomes

To analyze repair in the +1, +2, and +3 nucleosomes for each yeast gene, we obtained the dyad positions of yeast nucleosomes from a published MNase-seq map (Weiner et al. 2015), and used the nucleosome annotation from this study to extract +1, +2, and +3 nucleosome positions for each yeast gene. Only nucleosomes derived from genes with

a mapped TSS (Park et al. 2014), after excluding genes that overlap with the repetitive rDNA and *CUP1* loci, were analyzed. The number of CPD reads following 2hr repair was determined at each position relative to the nucleosome dyad axis (i.e., from -73 to +73 nt away from the dyad), and the fraction of CPDs remaining (F_{remain}) was calculated by dividing the number of CPD reads following 2hr repair (N_{2hr}) by the initial 0hr CPD read count (N_{0hr}).

$$F_{remain}(x) = N_{2hr}(x)/N_{0hr}(x)$$

Only CPD reads associated with lesions on the NTS were counted, and the NTS was oriented in the 5'->3' direction. CPD repair in +1 nucleosomes, identified by a high-resolution chemical cleavage method (Chereji et al. 2018), was also analyzed.

The percent asymmetry in the fraction of CPDs remaining at each nucleosome position (i.e., +1, +2, +3) was calculated by computing the difference in the fraction of CPDs remaining between equivalent positions on opposite sides of the nucleosome dyad (e.g., -50 and +50) and dividing by the average fraction of CPDs remaining at these two positions (see equation below). Symmetric repair across the nucleosome dyad should result in an average relative difference equal to zero, so a *t*-test was used to test this hypothesis for positions between -73 and +73 nt away from the dyad. The average relative difference (multiplied by 100) was reported as the percent asymmetry.

$$Asymmetry\ (\%) = 100 * \frac{1}{73}\sum_{x=1}^{73}\frac{F_{remain}(x) - F_{remain}(-x)}{(F_{remain}(x) + F_{remain}(-x)) * \frac{1}{2}}$$

Human nucleosome map

Nucleosome positions were called using a new algorithm that combines both DNase-seq (Degner et al. 2012; Zhong et al. 2016) and MNase-seq data (Gaffney et al. 2012). Nucleosome scores derived from DNase-seq data (Zhong et al. 2016) were used to identify potential nucleosome dyad positions using a greedy algorithm, as previously described (Brown et al. 2018). This DNase-seq data provides a very accurate measure of the rotational positioning of the nucleosome, but has more difficulty in determining the translational positioning. To address this issue, we used MNase-seq data (Gaffney et al. 2012) to better localize the position of the nucleosome dyad (i.e., mnase_mids_combined_147.wig from http://eqtl.uchicago.edu/nucleosomes/midpoints/). The algorithm compared the number of MNase-seq reads (147bp in length) associated with the potential dyad position identified by the greedy algorithm, as well as positions at 10bp intervals (in order to preserve the rotational setting) up to 20bp away in either direction (i.e., ±10bp or ±20 bp), and the position with the maximum number of MNase-seq reads was chosen as the dyad position. For increased flexibility, we also analyzed MNase-seq reads at positions ±1bp from each of these locations when finding the position with the maximum number of MNase-seq reads (i.e., at positions 0, ±1bp, ±9bp, ±10bp, ±11bp, ±19bp, ±20bp, and ±21bp from the potential dyad identified from the DNase-seq data). If none of these positions had at least one MNase-seq read, then the potential dyad position was discarded, and the nucleotide position with next highest nucleosome score was analyzed in a similar manner. If a nucleosome was called at one of these positions, then the neighboring 147 bp on each side of the identified dyad were excluded from the greedy algorithm. In order to call strongly positioned nucleosomes, only nucleosome scores greater than or equal to 10 were selected by the greedy

algorithm, and nucleosome positions overlapping with ENCODE blacklisted regions were excluded, as previously described (Brown et al. 2018).

We identified nucleosomes located within protein-coding genes by intersecting the list of nucleosome positions with gene coordinates from GENCODE (version 19, hg19; (Harrow et al. 2012)). The transcribed and non-transcribed strands for each intragenic nucleosome were determined based on the transcription orientation of its associated gene. In cases where more than one gene was associated with an individual nucleosome position, the nucleosome was excluded if the genes were transcribed in opposite directions, but was retained if the genes were transcribed in the same direction. Nucleosome positions not overlapping with any protein-coding genes were deemed intergenic nucleosomes.

For simplicity, most analysis used human genome version hg19, since the melanoma mutations were called using this genome version. However, using genome version GRCh38 (hg38) should not affect the conclusions.

Analysis of XR-seq, HS-Damage-seq, DNase-seq

Human XR-seq reads was obtained from (Adar et al. 2016), trimmed, and mapped to the hg19 genome using Bowtie 2 (Langmead and Salzberg 2012). The resulting SAM files were converted to BAM files using SAMtools (Li et al. 2009) and then converted to BED files using BEDTools (Quinlan and Hall 2010). Intact DNA fragments excised by NER typically range from 23 to 31 nt (Hu et al. 2015; Adar et al. 2016), so XR-seq reads outside this length range were excluded. Sequence analysis indicates that positions 5-6 and 6-7 nt from the 3' end of the CPD XR-seq read show the strongest enrichment in

dipyrimidine sequences ((Hu et al. 2015) and unpublished data), indicating the likely location of the CPD lesion in the XR-seq read. The dinucleotide frequencies at position 5-6 most resembled the expected pattern for UV lesions (e.g., frequency of TT > TC > CT > CC), so we assigned the putative CPD lesion to this location (i.e., 5-6 nt from 3' end) in each XR-seq read.

HS-Damage-seq sequencing reads associated with CPD lesions (Hu et al. 2017) were trimmed, aligned to the hg19 genome using Bowtie 2 (Langmead and Salzberg 2012). The resulting SAM file was converted to a BED file using SAMtools (Li et al. 2009) and BEDTools (Quinlan and Hall 2010). The putative CPD lesion site was inferred by extracting the dinucleotide sequence on the 5' end of the mapped read (on the same DNA strand). Only HS-Damage-seq reads associated with dipyrimidine sequences at the lesion site were included in subsequent analysis.

DNase-seq data (Degner et al. 2012) was downloaded from the authors' website (http://eqtl.uchicago.edu/dsQTL_data/MAPPED_READS/) and converted to BED format. Only DNase-seq reads from individuals used in the MNase-seq nucleosome mapping data (i.e., NA18507, NA18508, NA18516, NA18522, NA19193, NA19238, and NA19239) were included in the analysis. The 5' end of each DNase-seq read, which represents the DNase cleavage site, was analyzed. DNase-seq bed files were sorted, split into plus and minus strands, and converted to wig files using IGVtools (Robinson et al. 2011). Since the DNase-seq data was from genome version hg18, an hg18 version of the nucleosome map was used to analyze the DNase-seq data.

Melanoma and squamous cell carcinoma mutation data

Melanoma mutation data were derived from 183 unique tumor samples derived from the ICGC data portal (data release 20), and were processed as previously described (Mao et al. 2018). Only single nucleotide somatic mutations were used in subsequent analysis.

Somatic mutations from squamous cell carcinomas derived from eight individuals with proficient NER (WT) and five individuals with deficient GG-NER (i.e., individuals with $XPC^{-/-}$ germline mutations) were obtained from (Zheng et al. 2014). Only nucleotide variants displaying quality scores > 100 were used for analysis.

The vast majority of skin cancer mutations are UV signature mutations, which are typically caused by mutagenic bypass of UV photoproducts (i.e., CPDs and 6-4PPs). Since UV photoproducts form at dipyrimidine sequences, we assigned the mutation to the pyrimidine-containing strand (i.e., C or T nucleotide).

Calculating repair and mutational asymmetry in human nucleosomes

Repair asymmetry was analyzed for the NTS in intragenic nucleosomes (i.e., nucleosomes occurring in protein-coding gene sequences) and for both DNA strands in intergenic nucleosomes using the called nucleosome dyad positions. For intragenic nucleosomes (553,477 nucleosome positions), the density of XR-seq reads at each repair time point was determined at each position relative to the nucleosome dyad axis. The density of XR-seq reads was normalized using the density of 0hr HS-Damage-seq reads at each position in the nucleosome, which served as an estimate of the initial CPDs at each position in the nucleosome (i.e., see Fig. 3B). For intragenic nucleosomes, only XR-seq reads mapping to the NTS of the associated gene were analyzed, and the NTS was always oriented in the 5'-to-3' direction. For intergenic

nucleosomes (450,141 nucleosome positions), XR-seq reads mapping to both DNA strands were analyzed separately. We limited our analysis to the central 121 bp of the nucleosome (i.e., from -60 to +60 nt away from the dyad), as the XR-seq and HS-Damage-seq data (and melanoma mutations) were more variable near the nucleosomal DNA ends.

The percent asymmetry in repair activity for intragenic nucleosomes was calculated by computing the difference in the normalized XR-seq reads between equivalent positions on opposite sides of the nucleosome dyad (e.g., -50 and +50) and dividing by the average normalized XR-seq reads at these two positions, as described for the yeast repair asymmetry. Symmetric repair across the nucleosome dyad should result in an average relative difference equal to zero, so a *t*-test was used to test this hypothesis for positions between -60 and +60 nt away from the dyad. Since the normalized XR-seq reads measures repair activity, instead of the fraction of unrepaired damage, lower repair of the 3' side of the nucleosome dyad results in a negative asymmetry value for the normalized XR-seq data, but a positive asymmetry value for the fraction of unrepaired CPDs (i.e., CPD-seq data).

Similar analysis was performed to calculate the asymmetry in mutation density and enrichment in intragenic nucleosomes. Mutation density along the NTS in intragenic nucleosomes was normalized using the strand-specific expected mutation density. The percent asymmetry was calculated as described above for positions between -60 and +60 nt away from the dyad, and the significance of the asymmetry was determined using a *t*-test. A similar analysis was used to measure asymmetry in the distribution of

CPD lesions, after normalizing by the scaled UV-irradiated naked DNA control (Brown et al. 2018; Mao et al. 2018).

To analyze asymmetry specifically at 'out' positions in nucleosomes, we identified the highest level of mutation enrichment within 2 nt of each 'out' position (i.e., ±10, ±20, ±30, ±41, ±52, ±62 nt from the dyad axis). The average relative difference in the mutation enrichment peak at each pair of 'out' positions (e.g., -10 nt [5' side of dyad] versus +10 nt [3' side of the dyad]) was used to determine the percent asymmetry, as described above, and a *t*-test was used to determine the significance of this asymmetry. A similar method was used to analyze asymmetry specifically at 'out' positions for the normalized XR-seq reads, DNase-seq reads, and CPD enrichment.

Supplemental References

Adar S, Hu J, Lieb JD, Sancar A. 2016. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc Natl Acad Sci U S A* **113**: E2124-2133.

Brogaard K, Xi L, Wang JP, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**: 496-501.

Brown AJ, Mao P, Smerdon MJ, Wyrick JJ, Roberts SA. 2018. Nucleosome positions establish an extended mutation signature in melanoma. *PLoS Genet* **14**: e1007823.

Chereji RV, Ramachandran S, Bryson TD, Henikoff S. 2018. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol* **19**: 19.

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390-394.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863-14868.

Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK. 2012. Controls of nucleosome positioning in the human genome. *PLoS Genet* **8**: e1003036.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774.

Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**: 717-728.

Hu J, Adar S, Selby CP, Lieb JD, Sancar A. 2015. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev* **29**: 948-960.

Hu J, Adebali O, Adar S, Sancar A. 2017. Dynamic maps of UV damage formation and repair for the human genome. *Proc Natl Acad Sci U S A* doi:10.1073/pnas.1706522114.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Mao P, Brown AJ, Esaki S, Lockwood S, Poon GMK, Smerdon MJ, Roberts SA, Wyrick JJ. 2018. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat Commun* **9**: 2626.

Mao P, Brown AJ, Malc EP, Mieczkowski PA, Smerdon MJ, Roberts SA, Wyrick JJ. 2017. Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Res* **27**: 1674-1684.

Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. 2016. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **113**: 9057-9062.

Park D, Morris AR, Battenhouse A, Iyer VR. 2014. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Res* **42**: 3736-3749.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.

Saldanha AJ. 2004. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**: 3246-3248.

Weiner A, Hsieh TH, Appleboim A, Chen HV, Rahat A, Amit I, Rando OJ, Friedman N. 2015. High-resolution chromatin dynamics during a yeast stress response. *Mol Cell* **58**: 371-386.

Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, Collisson EA, Vemula SS, Naujokas A, Chiotti KE et al. 2014. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Rep* **9**: 1228-1234.

Zhong J, Luo K, Winter PS, Crawford GE, Iversen ES, Hartemink AJ. 2016. Mapping nucleosome positions using DNase-seq. *Genome Res* **26**: 351-364.