Nucleotide excision repair hotspots of UV-induced DNA damage in the human genome

Yuchao Jiang[1,2,3,*,#], Wentao Li[4,*], Laura A Lindsey-Boltz[4,*], Yuchen Yang[2], Yun Li[1,2,5], Aziz Sancar[3,4,#]

[1] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA.

[2] Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

[3] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA.

[4] Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

[5] Department of Computer Science, College of Arts and Sciences, University of North Carolina, Chapel Hill, NC 27599, USA.

∗ These authors contributed equally.

To whom correspondence should be addressed. Email: yuchaoj@email.unc.edu, aziz_sancar@med.unc.edu.

ABSTRACT

We recently developed high-throughput sequencing approaches, eXcision Repair sequencing (XR-seq) and Damage-seq, to generate genome-wide mapping of DNA excision repair and damage formation, respectively, with single-nucleotide resolution. Here, we used time-course XR-seq data to profile UV-induced excision repair dynamics, paired with Damage-seq data to quantify the overall induced DNA damage. We identified genome-wide repair hotspots exhibiting high-level nucleotide excision repair immediately after UV irradiation. We show that such repair hotspots do not result from hypersensitivity to DNA damage, and are thus not damage hotspots. We find that the earliest repair occurs preferentially in promoters and enhancers from open-chromatin regions. The repair hotspots are also significantly enriched for frequently interacting regions and super-enhancers, both of which are themselves hotspots for local chromatin interactions. Further interrogation of chromatin organization to include DNA replication timing allows us to conclude that early-repair hotspots are enriched for early-replication domains. Collectively, we report genome-wide early-repair hotspots of UV-induced damage, in association with chromatin states and epigenetic compartmentalization of the human genome.

INTRODUCTION

UV in sunlight is a known mutagen and causative agent of skin cancer that induces DNA lesions as cyclobutane pyrimidine dimers (CPDs) and pyrimidine-pyrimidone (6-4) photoproducts [(6-4)PPs] (Hodis et al. 2012; Martincorena et al. 2015). In humans, both damage types are solely repaired by nucleotide excision repair (excision repair), which removes DNA lesions by dual incisions bracketing modified bases and fills and seals the resulting gap by DNA synthesis and ligation (Reardon and Sancar 2005). Excision repair consists of two pathways, global repair and transcription-coupled repair (Mellon et al. 1987; Hanawalt and Spivak 2008), that differ primarily in the damage recognition step (Kusakabe et al. 2019). For UV-induced DNA damage, CPD repair is highly associated with transcription, specifically with the transcribed strand, while (6-4)PP repair is distributed uniformly distributed throughput the genome (Hu et al. 2015).

Recently, we developed a high-throughput approach, excision repair sequencing (XR-seq), to isolate the oligonucleotides excised by excision repair and subject them to next-generation sequencing (Hu et al. 2015). This method has allowed for genome-wide mapping of UV-induced excision repair with single-nucleotide resolution in human (Hu et al. 2015), Lemur (Akkose et al. 2020), mouse (Yang et al. 2019), *Drosophila melanogaster* (Deger et al. 2019), *Saccharomyces cerevisiae* (Li et al. 2018), *Escherichia coli* (Adebali et al. 2017), and *Arabidopsis thaliana* (Oztas et al. 2018). We further developed another next-generation sequencing method, Damage-seq (Hu et al. 2016), to generate genome-wide mapping of UV damage formation with single-nucleotide resolution (Hu et al. 2017).

Using the combined UV damage maps and repair maps, it has been shown that the induced DNA damage is uniformly distributed throughout the genome and that the overall effect of damage in the genome is primarily driven, not by damage formation, but by repair efficiency (Hu et al. 2016; Hu et al. 2017). Indeed, different repair efficiencies across the genome have been previously reported, which are impacted by multiple factors, including transcription, chromatin state and structure, regulatory protein binding to DNA, and posttranscriptional modification of histones (Hu et al. 2015; Adar et al. 2016; Mao et al. 2016; Poulos et al. 2016; Sabarinathan et al. 2016; Garcia-Nieto et al. 2017; Hu et al. 2017; Mao et al. 2018; Yimit et al. 2019). Existing studies, however, have been focused

on profiling repair dynamics over a time course – 4 h for (6-4)PP and 48 h for CPD. Genomic regions that harbor high-level early repair shortly after damage formation have not been studied, nor have their associated genomic and epigenomic characteristics been systematically explored.

Here, we performed XR-seq at as early as 1 min for (6-4)PP and 12 min for CPD following UV irradiation to identify such repair hotspots. We systematically characterized the identified hotspots using additional high-throughput sequencing data that measures DNA damage formation, DNase I hypersensitivity, histone modifications, 3D chromatin interactions, and DNA replication timing. These extensive links between chromatin states and organization, cell cycle, DNA damage, and excision repair at specific genomic sites facilitate better understanding of mutagenesis and carcinogenesis and have the potential to improve prevention strategies and develop novel therapeutics for human diseases such as cancer.

RESULTS

Genome-wide profiling of DNA excision pair kinetics through ordered high-throughput experiments

In this work, we present an experimental and analytical framework where we systematically assay DNA excision repair of UV-induced DNA damage over a time course. Figure 1A gives an outline of the experimental design. Specifically, we adopted XR-seq (Hu et al. 2015) to measure repair of (6-4)PPs at 1 min, 2 min, 5 min, 20 min, 1 h, 2 h, and 4 h and CPDs at 12 min in normal human skin fibroblasts (NHF1) after 20J/m^2 UV treatment. The (6-4)PP XR-seq experiments from 1 min to 4 h were all performed with two biological replicates, with greater than 0.99 correlation coefficients between each pair of replicates (Figure S1). Quality control procedures show that all samples have good data quality with expected read length distributions and enrichments of TT and/or TC dinucleotides at the damage sites (Figure S2). To investigate the interplay between DNA damage and repair, we additionally adopted Damage-seq data (Hu et al. 2017) to profile the two types of UV-induced damage at 0 timepoint after UV irradiation. Refer to the Methods section for details on library preparation, sequencing, and bioinformatic analysis and Table S1 for detailed sample information.

The bulk repair pathways and kinetics of these two types of DNA damage are known to be quite different (Reardon and Sancar 2003; Hu et al. 2013). The global repair efficiently removes (6-4)PPs, with most of the damage excised within 4 h of UV treatment. We provide additional empirical evidence that the repair levels of (6-4)PP from the transcribed strand (TS) and the non-transcribed strand (NTS) are on par across all timepoints (Figure S3A-B). In contrast, CPDs are recognized primarily in a transcription-coupled manner and their complete repair requires 12 to 48 h, depending on the UV dose. In our experiment, however, given the sampling timepoint as early as 12 min for CPDs, transcription-coupled repair has not contributed much, if any at all, to the repair of CPDs (Figure S3C). This is of key importance so that our analysis is unbiasedly focused on global repair of (6-4)PP at any timepoints and of CPD at the early 12 min timepoint, without transcription as a confounder. Refer to the Discussion section for more details.

In this study, we focus on characterizing global repair kinetics and identifying early-repair hotspots. *In vivo* excision assay by gel autoradiography at 0 min and 2 min reveals that the primary excision products repair can be detected as early as 2 min after UV treatment (Figure 1B). Using genome-wide repair data by XR-seq, we performed principal component analysis (PCA) on the top 2,000 highly variable genes to generate a low-dimensional representation of the (6-4)PP repair data (Figure 1C). Since transcription-coupled repair does not contribute to the repair of (6-4)PPs, the PCA plots do not differ between the TS and NTS repair, except for reverted signs of the eigenvectors. Importantly, a repair trajectory on the principal component space can be reconstructed, which lines up well with the timepoints, suggesting differential repair kinetics over the time course (Figure 1C).

Identification of early-repair hotspots in normal human fibroblasts

We developed a computational framework to identify genome-wide early-repair hotspots using time-course XR-seq data. Refer to the Methods section for details. Briefly, we segmented the genome into consecutive bins of 50 bp long and identified bins that showed significantly higher number of reads at earlier or later timepoints (Figure S4). We define such genomic bins as repair hotspots and coldspots, respectively. Figure 1D shows the distributions of read counts per genomic bin across all (6-4)PP samples and

111 we note enrichment of both early repair at 1 min and late repair at 4 h, corresponding to
112 repair hotspots and coldspots, respectively. For repair of (6-4)PPs, we identified on the
113 genome-wide scale 175 and 156 repair hotspots from the plus and minus strand (Table
114 S2) and 48 and 57 repair coldspots from the plus and minus strand (Table S3),
115 respectively. For repair of CPDs, we focused on repair hotspots at an early timepoint to
116 negate the global impact of transcription, and identified 99 and 93 repair hotspots from
117 the plus and minus strands, respectively (Table S4). We could not search for intrinsic
118 CPD coldspots, because 30 min after UV irradiation transcription-coupled repair becomes
119 an important contributor to the repair profile.

120 The repair hotspots and coldspots reported are scattered across the entire human
121 genome (Figure S5). XR-seq signals from examples of repair hotspots and coldspots are
122 separated by strand and plotted across all timepoints in Figure 2. We also include
123 epigenomic signals by DNase-seq, ChIP-seq from ENCODE (ENCODE Project
124 Consortium 2012), and Damage-seq signals at 0 min after UV treatment (Hu et al. 2017).
125 Specifically, the XR-seq signals from an example (6-4)PP repair hotspot decrease
126 dramatically from 1 min to 20 min and can be barely seen at 1 h (Figure 2A). In contrast,
127 the XR-seq signals at a late-repair coldspot shown in Figure 2B increases over the time
128 course and peaks at 4 h. Another representative CPD repair hotspot at 12 min is shown
129 in Figure 2C.

130 We next performed sequence context analysis using all reads that are mapped to
131 the repair hotspots and coldspots, respectively. We trimmed the reads to be 15 bp long
132 centering at the damage sites and calculated strand-specific nucleotide frequencies in
133 repair hotspots, coldspots, and randomly picked spots. Interestingly, we identified an
134 enrichment of cytosine in the flanking regions of the damage/repair sites for both repair
135 hotspots and coldspots (Figure S6). Motif analysis by the MEME suite (Bailey et al. 2009)
136 confirmed such enrichment of cytosine adjacent to the repair sites, which are themselves
137 enriched with canonical sequences of CTCA for (6-4)PP and TT for CPD (Table S5).
138 While it has been previously shown that active transcription factor binding sites have
139 decreased levels of excision repair (Sabarinathan et al. 2016), here we report a
140 preference for cytosine at both early and late timepoints, which have increased levels of

141 repair compared to null regions. While an interesting observation, the underlying
142 biological mechanisms need further investigation.
143
144 **Early-repair hotspots are not hypersensitive to DNA damage**
145 A recent study assayed DNA damage by next-generation sequencing and reported a total
146 of 153 hyper-hotspots acquiring CPDs much more frequently than the genomic average
147 in primary human fibroblasts (Premi et al. 2019). Of these damage hotspots, 83 are from
148 the plus strand and 74 are from the minus strand, each having at least five recurrent
149 sequence reads (Premi et al. 2019). To investigate whether the identified repair hotspots
150 simply result from increased levels of DNA damage, we intersected the reported damage
151 hotspots from Premi et al. (2019) with the 99 and 93 CPD repair hotspots from the plus
152 and minus strand that we identified. We found that none of our identified repair hotspots
153 overlapped with the reported damage hotspots.
154 To further confirm this seemingly striking result, we analyzed genome-wide CPD
155 DNA damage data generated by our previously developed Damage-seq protocol (Hu et
156 al. 2017), to quantify damage levels at 0 timepoint after UV irradiation with single-
157 nucleotide resolution. After stringent quality control procedures (refer to the Methods
158 section for details), we identified 91 and 78 CPD damage hotspots from the plus and
159 minus strand, respectively, each having at least ten mapped reads (Table S6). Notably,
160 these Damage-seq hotspots are shown to be enriched for heterochromatin and repressed
161 regions (Figure S7), which is concordant with previous reports (Han et al. 2016; Hauer
162 and Gasser 2017). Importantly, none of the CPD damage hotspots, identified from this
163 parallel Damage-seq platform, overlap with the repair hotspots.
164 In addition, we compared the Damage-seq signals for (6-4)PP and CPD at our
165 identified hotspots and coldspots against those from randomly sampled regions along the
166 genome. Regions corresponding to the repair hotspots, coldspots, and random spots are
167 extended at both ends for 20bp and 500bp, respectively, to account for the different
168 sampling densities between Damage-seq and XR-seq. Our results, shown in Figure 3,
169 suggest that there is no significant difference in the damage levels between the three
170 repair categories (hotspot, coldspot, and random spot). The zoom-in and zoom-out views
171 of three example repair hotspots and coldspots in Figure 2 also suggest that the Damage-

172 seq reads are uniformly distributed in the flanking regions. Previous results have
173 demonstrated that the UV-induced DNA damage is indeed virtually uniform across the
174 entire human genome, while repair is affected by chromatin states, transcription factor
175 binding, etc., in a manner dependent on the type of DNA damage (Hu et al. 2017). While
176 we note that the shallow depth of coverage of Damage-seq can be a limiting factor (refer
177 to the Discussion section for details), our results validate our conclusion that the identified
178 repair hotspots are not damage hotspots.
179

180 **Early-repair hotspots are enriched for promoters and enhancers in open chromatin**
181 **regions**

182 The packaging of DNA into chromatin can hinder the access of repair proteins and affect
183 the repair efficiency (Adar et al. 2016; Yimit et al. 2019) and specific histone modifications
184 are associated with different functional and cytological chromatin states. To investigate
185 the relationship between the identified repair hotspots/coldspots and the epigenomic
186 markers, we used publicly available DNase-seq data and histone modification ChIP-seq
187 data for an adult human fibroblast cell line (NHLF) from ENCODE (ENCODE Project
188 Consortium 2012). Our results suggest that the earliest repair occurs preferentially in
189 active and open-chromatin regions. Specifically, chromatin accessibility by DNase-seq is
190 significantly higher in repair hotspots than in random genomic regions, and it is
191 significantly lower in repair coldspots (Figure 4). For histone modifications, repair hotspots
192 have significantly higher ChIP-seq signals compared to random spots for activation
193 markers, including H3K4me1, H3K4me3, and H3K27ac (Figure 4).

194 We adopted the segmented chromatin states by chromHMM (Ernst and Kellis 2012)
195 to annotate the repair hotspots for (6-4)PP and CPD and found a significant enrichment
196 of active promoters and enhancers (Figure 5A), which are characterized by nucleosome
197 loss and open chromatin. Genome-wide repair kinetics inferred across timepoints confirm
198 that enhancers are repaired at earlier timepoints and that repressed/heterochromatin
199 regions are repaired at later timepoints (Figure S8). Additionally, we performed genome-
200 wide annotations for CpG islands and genic/intergenic regions. Our results suggest that
201 the repair hotspots are enriched for inter CpG regions (i.e., depleted for CpG islands)
202 (Figure 5B) and intronic regions (Figure 5C), both of which are enriched for AT and

203 presumably result in more repair. Notably, the number of identified repair coldspots is too
204 low to generate reproducible annotation results between the replicates.
205
206 **Early-repair hotspots are enriched for frequently interacting regions (FIREs) and**
207 **super-enhancers**
208 With the genome-wide enrichment of active promoters and enhancers in open chromatin
209 regions of the detected repair hotspots, we hypothesize that the 3D genome structure
210 contributes to the observed differential repair kinetics between different regions of the
211 genome. To test this hypothesis, we sought to interrogate the publicly available Hi-C data
212 of human fibroblast cell line IMR90 (Dixon et al. 2012; Jin et al. 2013). Specifically, after
213 quality control procedures and data normalization, we profiled frequently interacting
214 regions (FIREs) using FIREcaller (Crowley et al. 2019). After overlapping the repair
215 hotspots and coldspots with the called FIREs (Table S7A), we found that a significantly
216 higher proportion of repair hotspots overlap with FIREs – 23.16% and 11.76% for (6-4)PP
217 and CPD, respectively – compared to a genome average of 6.93% based on the profiled
218 FIREs (Figure 6A). Conversely, the overlapping proportion of (6-4)PP repair coldspots is
219 3.23%, significantly lower than the genome average (Figure 6A).
220 FIREs have been previously reported to be enriched for super-enhancers (Schmitt
221 et al. 2016). We have demonstrated that the repair hotspots are enriched for both FIREs
222 and enhancers. We also observed that, across many cases, multiple enhancers that
223 overlapped with the repair hotspots were from the same genomic regions (Figure S9). As
224 such, we hypothesize that the repair hotspots are also enriched for super-enhancers and
225 thus additionally adopted a list of previously annotated super-enhancers in the human
226 fibroblasts (Roadmap Epigenomics Consortium, 2015) (Table S7B). We found that,
227 compared to a genome-wide average of 2.05%, the early-repair hotspots are indeed
228 enriched for super-enhancers (5.14% and 4.69% for (6-4)PP and CPD repair hotspots,
229 respectively), while none of the repair coldspots overlap with super-enhancers (0% for (6-
230 4)PP repair coldspot) (Figure 6B).
231 In addition, we called significant interactions based on the Hi-C contact matrix
232 using the Fit-Hi-C method (Ay et al. 2014) (Table S7C) and showed that early-repair
233 hotspots also overlap with a significantly higher number of significant interactions (Figure

234 6C). Refer to the Methods section for details on data analysis. The overlapping
235 information of the called repair hotspots and coldspots with the profiled FIREs, super-
236 enhancers, and significant chromatin interactions are included in Table S8. Figure 6D
237 illustrates the loop interactions of two identified repair hotspots. Notably, these two
238 hotspots also overlap with both FIREs and super-enhancers. These results collectively
239 provide a global picture of genetic regulation of repair kinetics via 3D genome structures.
240

241 **Early-repair hotspots are enriched for early-replication domains**

242 We next set out to test the relationship between the repair kinetics and replication timing.
243 We used the genome-wide Repli-Seq data of the human fibroblast cell line IMR90, which
244 maps high-resolution DNA replication patterns with respect to both cell-cycle time and
245 genomic position (Hansen et al. 2010). A recent re-analysis of the data using a deep-
246 learning model (Liu et al. 2016) segmented the genome into different replication domains
247 including the early-replication domain (ERD), late-replication domain (LRD), up-transition
248 zone (UTZ), and down-transition zone (DTZ). Refer to Table S9 for genome-wide
249 segmentation results. When overlapping the identified repair hotspots with the segmented
250 replication domains, we find that early-repair hotspots are also significantly enriched for
251 early-replication domains (Figure 7). While replication time has been shown to be
252 correlated with chromatin accessibility (Hansen et al. 2010), here we have demonstrated
253 a potential cell-cycle effect on DNA excision repair. That is, regions that are duplicated
254 early also tend to be repaired early and this has been shown to result in mutagenesis
255 asymmetries (Adar et al. 2016) – we further elaborate such implications in the Discussion
256 section.
257

258 **Profiling gene-level repair kinetics in normal human fibroblasts**

259 So far, our analysis has been conducted on the whole-genome level. Identifying dynamic
260 changes at individual gene levels can provide important insights about genes and
261 pathways that exhibit significant changes in repair dynamics (Li et al. 2019). We
262 calculated strand-specific reads per kilobase per million reads (RPKM) for all the genes
263 in the human genome. RPKMs across nine genes associated with circadian rhythm (e.g.,
264 *CLOCK*, *BMAL1*, *PER1*, *CRY1*) and excision repair (e.g., *XPA*, *XPC, CSA, CSB*) present

265 different repair dynamics over the time course (Figure S10). We adopted Trendy (Bacher
266 et al. 2018) to perform segmented regression analysis (Table S10) on the ordered XR-
267 seq data. We focused on significant genes with adjusted R^2 greater than 0.8 from the
268 regression analysis and identified 3,017 significant genes with fitted trend as "down,"
269 3,496 significant genes with fitted trend as "up-down," 1,024 significant genes with fitted
270 trend as "up," and 1,452 significant genes with fitted trend as "down-up" (Figure S11). The
271 number of genes that show the "down" trend is two-fold higher than the number of genes
272 that show the "up" trend, which is consistent with our previous observation – there are
273 more repair hotspots than coldspots. For these four groups of significant genes, we further
274 performed gene ontology (GO) enrichment analysis using PANTHER (Thomas et al.
275 2003). For genes with the "down" and "up-down" trend, GO biological processes are
276 enriched for a large number of biological regulations, including regulation of metabolic
277 process, gene expression, developmental process, cell cycle, and cell death, etc. (Table
278 S11A-B). On the other hand, there are no obvious GO enrichment terms for the "up" or
279 "down-up" genes (Table S11C-D).
280

281 **DISCUSSION**

282 We adopted time-course XR-seq data to detect genome-wide repair hotspots and
283 coldspots and integrated additional omics data from various next-generation sequencing
284 platforms to investigate the relationship between DNA damage, excision repair,
285 epigenomic markers, 3D genome organization, and replication timing. To our best
286 knowledge, this is the first demonstration at high-resolution of the connection between
287 repair kinetics and chromatin organization. The identified repair hotspots can serve as
288 additional functional genomic features in studies of the human genome at
289 multidimensional level. We believe the framework and multi-omics data we present here
290 can be useful to scientists both in the fields of chromatin dynamics and those interested
291 in determinants of repair rates in human genome. Our approach is applicable to studying
292 the response to other types of DNA damage including that induced by drugs used in
293 chemotherapy and thus has the potential for informing new therapeutic strategies.
294 Using paired XR-seq and Damage-seq data, we provided additional empirical
295 evidence that the damage levels in the regions of the repair hotspots were not enriched,

296 in concordance with previous reports (Hu et al. 2016; Hu et al. 2017). Therefore, we
297 attributed the observed differences in repair patterns to heterogeneous repair efficiencies
298 and not to damage formation. It is, however, noteworthy that the depth of coverage by
299 Damage-seq is an order of magnitude lower than that by XR-seq. Specifically, the
300 genome-wide average total number of reads between the two replicates for (6-4)PP
301 Damage-seq is 20.5 million, while that average for (6-4)PP XR-seq is 181.8 million (Table
302 S1). As such, it is possible that we do not have enough sequencing depth to detect
303 Damage-seq hotspots with high sensitivity. The fact that a good proportion of Damage-
304 seq reads get mapped to heterochromatin regions further complicates analysis and
305 lowers detection power. This issue also persists for the recently developed adductSeq
306 and FreqSeq, where a Poisson model was adopted with a genomic average rate of 0.07
307 reads per pyrimidine dinucleotide in fibroblasts (Premi et al. 2019). A Poisson distribution
308 with a mean parameter under the null much less than one could potentially suffer from
309 overdispersion and inflated false positive rates. Targeted Damage-seq offers a solution
310 for both discovery and validation. However, as with targeted DNA sequencing, the data
311 can be extremely noisy due to targeting, amplification, and sequencing biases and
312 artifacts (Jiang et al. 2018).

313 Using the genome-wide profile of DNA replication timing, we demonstrate that the
314 early-repair hotspots also tend to reside in early-replication domains. The inevitable effect
315 of replication on gene dosage and copy number could also have regulatory consequences
316 (Rhind and Gilbert 2013). Genes/regions that are duplicated early will be present at twice
317 the copy number of late replicating domains for most of the duration of S phase,
318 increasing both the DNA amount (as sister chromatids) and the transcriptional output.
319 The ability to perform excision repair for a greater fraction of the cell cycle has been shown
320 to result in lower mutation rates in the early-replication domains (Stamatoyannopoulos et
321 al. 2009; Lang and Murray 2011).

322 Granted, UV-induced DNA damage and excision repair have been linked with
323 mutagenesis and carcinogenesis. Recent studies have shown that mutation hotspots
324 exhibit strong increases in CPD formation efficacy (Elliott et al. 2018) and that excision
325 repair is attenuated in transcription factor binding sites, leading to elevated somatic
326 mutation rates (Sabarinathan et al. 2016). In terms of repair kinetics, Adar et al. (2016)

327 showed that globally late-repaired regions are associated with a higher level of cancer-
328 linked mutations. To further investigate this in our study context, we quantified the rates
329 of somatic point mutations associated with melanoma (Hodis et al. 2012) in the repair
330 hotspots and coldspots. However, due to the low number of hotspots and coldspots, as
331 well as their small sizes, we found very few overlapped with mutations and were thus
332 extremely underpowered to test for the differences in mutation rates between different
333 repair categories.

334 While the repair of (6-4)PP is predominantly carried out by global repair,
335 transcription-coupled repair plays an essential role in removing the more abundant and
336 less helix-distorting CPD adducts. In this study, we have intentionally focused on profiling
337 global repair hotspots for both (6-4)PP and CPD at early timepoints without the effect of
338 transcription-coupled repair. Genes exhibit both transcriptional dynamics over a time
339 course and biological fluctuations due to transcriptional bursting (Jiang et al. 2017) at the
340 same timepoint. This confounds and complicates the analysis. To disentangle the effects
341 of global repair and transcriptional-coupled repair in identifying repair hotspots is an
342 unsolved yet challenging problem.
343

344 **METHODS**

345 **Experimental methods**

346 ***Cell culture and UV irradiation***

347 Human NHF1 cells were obtained from W.K. Kaufmann (University of North Carolina,
348 Chapel Hill) (Heffernan et al. 2002) and cultured in Dulbecco's Modified Eagle Medium
349 (DMEM) with 10% FBS at 37 °C in a 5% CO_2 humidified chamber. For (6-4)PP XR-seq
350 at 1 min and 2 min timepoints, UV irradiation was performed as previously described (Hu
351 et al. 2015; Li et al. 2017). Briefly, the 80% confluent NHF1 cells were irradiated for 20 s
352 under a 250 nm UV lamp (1 $J/m^2/s$) after removing the culture medium and incubated at
353 37 °C in a 5% CO_2 humidified chamber for 1 min or 2 min with 37 °C culture medium.
354 Then, the cells were put on ice immediately and washed one time with ice-cold PBS
355 before being harvested by a cell scarper in 10 ml ice-cold PBS. In each replicate of (6-
356 4)PP XR-seq experiment, 50 and 30 petri dishes (150mm x 15mm) containing NHF1 cells
357 were used at 1 min and 2 min timepoints respectively. Cell culture, UV treatment, and

358 library preparation for (6-4)PP XR-seq at 5 min, 20 min, 1 h, 2 h, and 4 h, and CPD XR-
359 seq at 12 min were performed in previous studies (Hu et al. 2015; Adar et al. 2016; Deger
360 et al. 2019). For *in vivo* excision assay, UV irradiation was performed as aforementioned,
361 and 10 and 5 petri dishes (150mm x 15mm) containing NHF1 cells were used at 0 min
362 and 2 min timepoints respectively.
363
364 ***Excision assay***
365 The *in vivo* excision assay was performed as described (Hu et al. 2013; Hu et al. 2019).
366 Following UV irradiation, the excision products were isolated by gentle cell lysis and
367 nonchromatin fraction separation, and purified by TFIIH immunoprecipitation. The purified
368 excision products were then 3' radiolabeled by terminal deoxynucleotidyl transferase and
369 [α-^{32}P]-3'-dATP, and resolved in a 10% denaturing acrylamide gel. Ten and five petri
370 dishes (150mm x 15mm) of NHF1 cells were used at 0 min and 2 min, respectively.
371
372 ***XR-seq library preparation and sequencing***
373 XR-seq libraries were prepared as described in the previous protocol (Hu et al. 2019).
374 Briefly, the excision products were isolated by TFIIH immunoprecipitation following gentle
375 cell lysis and non-chromatin fraction separation, and ligated with adaptors. The ligated
376 excision products were then further purified by immunoprecipitation with anti-(6-4)PP
377 antibody and repaired by (6-4)PP photolyase before the library amplification by PCR.
378 Libraries were sequenced on an Illumina HiSeq 4000 platform.
379
380 **Data collection**
381 (6-4)PP XR-seq data at 5 min, 20 min, 1 h, 2 h, and 4 h were downloaded from the Gene
382 Expression Omnibus (GEO) with accession numbers GSE67941 (Hu et al. 2015) and
383 GSE76391 (Adar et al. 2016). CPD XR-seq data at 12 min were downloaded from GEO
384 with accession number GSE138846 (Deger et al. 2019). CPD and (6-4)PP Damage-seq
385 data of NHF1 were downloaded from with accession number GSE98025 (Hu et al. 2017).
386 Hyper-hotspots for UV-induced CPD damage in primary human fibroblasts were
387 downloaded from Premi et al. (2019). NHDF H3K4me1 (ENCODE Data Coordination
388 Center accession number ENCSR000ARV), H3K4me3 (accession number

389 ENCSR000DPR), H3K27ac (accession number ENCSR000APN), H3K27me3
390 (accession number ENCSR000APO), H3K9me3 (accession number ENCSR000ARX),
391 and DNase-seq (accession number ENCSR000EMP) data were downloaded from the
392 ENCODE portal (ENCODE Project Consortium 2012). NHLF chromatin state
393 segmentation results by chromHMM were downloaded from UCSC accession number
394 wgEncodeEH000792 (Ernst and Kellis 2012). Hi-C data of IMR90 were downloaded from
395 GEO with accession number GSE43070 (Jin et al. 2013) and from
396 https://bioconductor.org/packages/HiCDataHumanIMR90/ (Dixon et al. 2012). The list of
397 annotated super-enhancers in IMR90 was downloaded from the Roadmap Epigenomics
398 Consortium (Roadmap Epigenomics Consortium, 2015). Genomic categories of
399 replication timing from Repli-Seq data of IMR90 (Hansen et al. 2010) were downloaded
400 from GSE53984 (Liu et al. 2016).
401
402 **Bioinformatic and statistical analysis**
403 ***XR-seq bioinformatic pre-processing***
404 For XR-seq, cutadapt (Martin 2011) was used to trim reads with adaptor sequence
405 TGGAATTCTCGGGTGCCAAGGAACTCCAGTNNNNNNACGATCTCGTATGCCGTCT
406 TCTGCTTG at the 3' end and discard untrimmed reads. BWA (Li and Durbin 2009) was
407 used for alignment of single-end short reads. Unmapped reads and reads that map to
408 multiple locations with the same alignment quality were removed using Samtools (Li et al.
409 2009). Post-alignment filtering steps were adopted using Rsamtools
410 (http://bioconductor.org/packages/Rsamtools/). Specifically, if multiple reads share the
411 same 5' and 3' end coordinates, we keep only one to perform deduplication. We also only
412 keep reads that have mapping quality greater than 20 and are of lengths 21 bp to 31 bp.
413
414 ***Gene-level quantification of excision repair***
415 Reads from the TS and NTS strands were separated using known gene annotations for
416 hg19 by ENSEMBL. We use RPKM for within-sample normalization for the XR-seq data.
417 To perform gene-level quantification and downstream analysis including segmented
418 regression, we adopted a stringent quality control procedure and only retained genes that:
419 (i) had at least 10 TT or TC dinucleotides from either TS or NTS; (ii) were less than 300Kb;

420 and (iii) had at least 10 reads in total across all XR-seq samples. In addition, we took the
421 ratio of the reads from the TS and the NTS [TS/(TS+NTS)] to remove biases and artifact
422 that are shared between the two DNA stands, i.e., library size, gene length, and other
423 gene-specific biases, such as sequencing bias and antibody pulldown efficiency, etc. The
424 ratio is bound between 0 and 1 and sheds light upon how transcription-coupled repair and
425 global repair interplay (Figure S3).
426
427 ***Computational approach to identify repair hotspots and coldspots***
428 We started by segmenting the human reference genome into consecutive bins of 50 bp
429 long. We then calculated the observed depth of coverage per bin by XR-seq, separating
430 the plus-strand reads (+) and the minus-strand reads (-). To mitigate the effect of library
431 size/sequencing depth, we downsampled the reads in each sample to 7.7 million without
432 replacement. Given such data input, we identified strong repair hotspots in
433 pericentromeric regions, which were collapsed repeats annotated as unique sequences
434 in the reference genome (e.g., ribosomal DNA (Yang et al. 2019)). It is important to
435 exclude artifacts as stringently as possible and thus we undertook an additional quality
436 control step. "Blacklist" bins, including segmental duplication regions
437 (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab),
438 gaps in reference assembly from telomere, centromere, and/or heterochromatin regions
439 (https://gist.github.com/leipzig/6123703), and repeating elements by RepeatMasker
440 (https://genome.ucsc.edu/cgi-bin/hgTrackUi?g=rmsk) are masked in downstream
441 analysis. To identify repair hotspots and coldspot, we set a threshold on the number of
442 read counts per genomic bin in the 1 min and 4 h samples. Specifically, to identify (6-
443 4)PP repair hotspots, we require at least 15 reads mapped in both replicates at 1min and
444 at most 5 reads mapped in both replicates at 4 h. The read count threshold is relaxed for
445 the identification of coldspots, which have a smaller number compared to the hotspots.
446 For CPD repair, to avoid complicatedness due to transcription-coupled repair at later
447 timepoints, we focused on CPD repair hotspots only.
448

449 ***Hi-C data analysis***

450 We adopted the Hi-C data of human fibroblast cell line IMR90 (Dixon et al. 2012; Jin et
451 al. 2013) to investigate the relationship between identified repair hotspots and the 3D
452 genome structure. We took the raw contact matrix with 40 kb resolution as input and
453 detected FIREs, which play important roles in transcriptional regulations, across the entire
454 genome using FIREcaller (Crowley et al. 2019). To further investigate whether these
455 repair hotspots are involved in functional chromatin looping between regulatory elements
456 and their target genes, we adopted the Fit-Hi-C approach (Ay et al. 2014) to identify long-
457 range chromatin interactions on all 40 kb bin pairs within a maximal 3 MB region. The
458 interactions with p-value < 2.31e-11 were considered as statistically significant (Giusti-
459 Rodríguez et al. 2019).

460

461 **DATA AND CODE AVAILABILITY**

462 The data reported in this paper have been deposited in GEO with accession number
463 GSE148303. Scripts used in this paper are available at
464 https://github.com/yuchaojiang/damage_repair.

465

466 **COMPETING INTEREST STATEMENT**

467 The authors declare no competing interests.

468

469 **ACKNOWLEDGEMENTS**

470 This work was supported by NIH Grants R35 GM118102 (to A.S.), R01 ES027255 (to
471 A.S.), P01 CA142538 (to Y.J.), UL1 TR002489 (to Y.J.), K99 ES030015 (to W.L.), and a
472 pilot award from the UNC Computational Medicine Program (to Y.J.). We thank the
473 Sancar Lab members and Drs. Sheera Adar, Ming Hu, and Jeremy Simon for useful
474 comments and feedback.

475

476 **AUTHOR CONTRIBUTIONS**

477 AS envisioned and initiated the study, while LW and LAL-B performed the experiments.
478 YJ devised the analytical framework and all authors executed the data analysis. YJ wrote

479 the manuscript, with contributions from WL on experimental methods. The manuscript
480 was further edited and approved by all authors.
481

482 **REFERENCES**

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

485 Adebali O, Chiou YY, Hu J, Sancar A, Selby CP. 2017. Genome-wide transcription-coupled repair
486 in Escherichia coli is mediated by the Mfd translocase. *Proc Natl Acad Sci U S A* **114**: E2116-
487 E2125.

488 Akkose U, Kaya O, Lindsey-Boltz L, Karagoz Z, Brown AD, Larsen PA, Yoder AD, Sancar A, Adebali
489 O. 2020. Comparative analyses of two primate species diverged by more than 60 million
490 years show different rates but similar distribution of genome-wide UV repair events.
491 *bioRxiv*.

492 Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory
493 chromatin contacts. *Genome Res* **24**: 999-1011.

494 Bacher R, Leng N, Chu LF, Ni Z, Thomson JA, Kendziorski C, Stewart R. 2018. Trendy: segmented
495 regression analysis of expression dynamics in high-throughput ordered profiling
496 experiments. *BMC Bioinformatics* **19**: 380.

497 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009.
498 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

499 Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*
500 **489**: 57-74.

501 Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A,
502 Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human
503 epigenomes. *Nature* **518**: 317-330.

504 Crowley C, Yang Y, Qiu Y, Hu B, Won H, Ren B, Hu M, Li Y. 2019. FIREcaller: an R package for
505 detecting frequently interacting regions from Hi-C data. *bioRxiv*: 619288.

506 Deger N, Yang Y, Lindsey-Boltz LA, Sancar A, Selby CP. 2019. Drosophila, which lacks canonical
507 transcription-coupled repair proteins, performs transcription-coupled repair. *J Biol Chem*
508 **294**: 18092-18098.

509 Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in
510 mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376-
511 380.

512 Elliott K, Bostrom M, Filges S, Lindberg M, Van den Eynden J, Stahlberg A, Clausen AR, Larsson E.
513 2018. Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter
514 mutation hotspots in UV-exposed cancers. *PLoS Genet* **14**: e1007849.

515 Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization.
516 *Nat Methods* **9**: 215-216.

517 Garcia-Nieto PE, Schwartz EK, King DA, Paulsen J, Collas P, Herrera RE, Morrison AJ. 2017.
518 Carcinogen susceptibility is regulated by genome architecture and predicts cancer
519 mutagenesis. *EMBO J* **36**: 2829-2843.

520 Giusti-Rodríguez P, Lu L, Yang Y, Crowley CA, Liu X, Juric I, Martin JS, Abnousi A, Allred SC,
521 Ancalade N. 2019. Using three-dimensional regulatory chromatin interactions from adult
522 and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits.
523 *bioRxiv*: 406330.

524 Han C, Srivastava AK, Cui T, Wang QE, Wani AA. 2016. Differential DNA lesion formation and
525 repair in heterochromatin and euchromatin. *Carcinogenesis* **37**: 129-138.

526 Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and
527 surprises. *Nat Rev Mol Cell Biol* **9**: 958-970.

528 Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler
529 SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread
530 plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**: 139-144.

531 Hauer MH, Gasser SM. 2017. Chromatin and nucleosome dynamics in DNA damage and repair.
532 *Genes Dev* **31**: 2204-2221.

533 Heffernan TP, Simpson DA, Frank AR, Heinloth AN, Paules RS, Cordeiro-Stone M, Kaufmann WK.
534 2002. An ATR- and Chk1-dependent S checkpoint inhibits replicon initiation following
535 UVC-induced DNA damage. *Mol Cell Biol* **22**: 8552-8561.

536 Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L,
537 Place C et al. 2012. A landscape of driver mutations in melanoma. *Cell* **150**: 251-263.

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

541 Hu J, Adebali O, Adar S, Sancar A. 2017. Dynamic maps of UV damage formation and repair for
542 the human genome. *Proc Natl Acad Sci U S A* **114**: 6758-6763.

543 Hu J, Choi JH, Gaddameedhi S, Kemp MG, Reardon JT, Sancar A. 2013. Nucleotide excision repair
544 in human cells: fate of the excised oligonucleotide carrying DNA damage in vivo. *J Biol*
545 *Chem* **288**: 20918-20926.

546 Hu J, Li W, Adebali O, Yang Y, Oztas O, Selby CP, Sancar A. 2019. Genome-wide mapping of
547 nucleotide excision repair with XR-seq. *Nat Protoc* **14**: 248-282.

548 Hu J, Lieb JD, Sancar A, Adar S. 2016. Cisplatin DNA damage and repair maps of the human
549 genome at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **113**: 11507-11512.

550 Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. 2018. CODEX2: full-
551 spectrum copy number variation detection by high-throughput DNA sequencing. *Genome*
552 *Biol* **19**: 202.

553 Jiang Y, Zhang NR, Li M. 2017. SCALE: modeling allele-specific gene expression by single-cell RNA
554 sequencing. *Genome Biol* **18**: 74.

555 Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. 2013. A high-
556 resolution map of the three-dimensional chromatin interactome in human cells. *Nature*
557 **503**: 290-294.

558 Kusakabe M, Onishi Y, Tada H, Kurihara F, Kusao K, Furukawa M, Iwai S, Yokoi M, Sakai W,
559 Sugasawa K. 2019. Mechanism and regulation of DNA damage recognition in nucleotide
560 excision repair. *Genes Environ* **41**: 2.

561 Lang GI, Murray AW. 2011. Mutation rates across budding yeast chromosome VI are correlated
562 with replication timing. *Genome Biol Evol* **3**: 799-811.

563 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
564 *Bioinformatics* **25**: 1754-1760.

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

568 Li W, Adebali O, Yang Y, Selby CP, Sancar A. 2018. Single-nucleotide resolution dynamic repair
569 maps of UV damage in Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A* **115**:
570 E3408-E3415.

571 Li W, Hu J, Adebali O, Adar S, Yang Y, Chiou YY, Sancar A. 2017. Human genome-wide repair map
572 of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc Natl*
573 *Acad Sci U S A* **114**: 6752-6757.

574 Li W, Liu W, Kakoki A, Wang R, Adebali O, Jiang Y, Sancar A. 2019. Nucleotide excision repair
575 capacity increases during differentiation of human embryonic carcinoma cells into
576 neurons and muscle cells. *J Biol Chem* **294**: 5914-5922.

577 Liu F, Ren C, Li H, Zhou P, Bo X, Shu W. 2016. De novo identification of replication-timing domains
578 in the human genome by deep learning. *Bioinformatics* **32**: 641-649.

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

582 Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. 2016. Chromosomal landscape of UV damage
583 formation and repair at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **113**: 9057-
584 9062.

585 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
586 *EMBnet journal* **17**: 10-12.

587 Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A,
588 Alexandrov LB, Tubio JM et al. 2015. Tumor evolution. High burden and pervasive positive
589 selection of somatic mutations in normal human skin. *Science* **348**: 880-886.

590 Mellon I, Spivak G, Hanawalt PC. 1987. Selective removal of transcription-blocking DNA damage
591 from the transcribed strand of the mammalian DHFR gene. *Cell* **51**: 241-249.

592 Oztas O, Selby CP, Sancar A, Adebali O. 2018. Genome-wide excision repair in Arabidopsis is
593 coupled to transcription and reflects circadian gene expression patterns. *Nat Commun* **9**:
594 1503.

595 Poulos RC, Thoms JAI, Guan YF, Unnikrishnan A, Pimanda JE, Wong JWH. 2016. Functional
596 Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide
597 Excision Repair across the Motif. *Cell reports* **17**: 2865-2872.

598 Premi S, Han L, Mehta S, Knight J, Zhao D, Palmatier MA, Kornacker K, Brash DE. 2019. Genomic
599 sites hypersensitive to ultraviolet radiation. *Proc Natl Acad Sci U S A* **116**: 24196-24205.

600 Reardon JT, Sancar A. 2003. Recognition and repair of the cyclobutane thymine dimer, a major
601 cause of skin cancers, by the human excision nuclease. *Genes Dev* **17**: 2539-2551.

602 Reardon JT, Sancar A. 2005. Nucleotide excision repair. *Prog Nucleic Acid Res Mol Biol* **79**: 183-
603 235.

604 Rhind N, Gilbert DM. 2013. DNA replication timing. *Cold Spring Harb Perspect Biol* **5**: a010132.

605 Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. 2016. Nucleotide
606 excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**: 264-
607 267.

608 Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL et al. 2016. A Compendium
609 of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell*
610 *Rep* **17**: 2042-2059.

611 Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009.
612 Human mutation rate associated with DNA replication timing. *Nat Genet* **41**: 393-395.

613 Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A,
614 Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by
615 function. *Genome Res* **13**: 2129-2141.

616 Yang Y, Hu J, Selby CP, Li W, Yimit A, Jiang Y, Sancar A. 2019. Single-nucleotide resolution analysis
617 of nucleotide excision repair of ribosomal DNA in humans and mice. *J Biol Chem* **294**: 210-
618 217.

619 Yimit A, Adebali O, Sancar A, Jiang Y. 2019. Differential damage and repair of DNA-adducts
620 induced by anti-cancer drug cisplatin across mouse organs. *Nat Commun* **10**: 309.

621

622 **FIGURE & TABLE LEGENDS**

623 **Figure 1. Excision repair kinetics and early-repair hotspots in normal human**
624 **fibroblasts.** (A) Experimental design to measure DNA damage and excision repair of the
625 UV-induced (6-4)PP and CPD across different timepoints. (B) Detection of excision

626 products at 0 min and 2 min timepoints *in vivo*. Following UV irradiation, the excised
627 oligonucleotides were purified by TFIIH immunoprecipitation, radiolabeled, and resolved
628 in a 10% sequencing gel. DNA excision products of (6-4)PP can be detected as early as
629 2 min upon damage induction. (C) Principal component analysis of genome-wide excision
630 repair as measured by XR-seq shows repair kinetics across different timepoints. Between
631 1 min and 5 min, excised oligonucleotides were not degraded and thus XR-seq measured
632 cumulative repair. (D) Each row is a sample and each column is a specific total number
633 of reads per genomic bin. The color in the heatmap corresponds to the log counts of the
634 number of bins with specific read depths. Early-repair hotspots exist in samples collected
635 at early timepoints.
636
637 **Figure 2. Distribution of DNA damage, repair, and epigenomic markers at identified**
638 **repair hotspots and coldspots.** XR-seq and Damage-seq data are shown for both
639 strands, marked with + and -. Epigenetic data from ChIP-seq of histone modifications and
640 DNase-seq are plotted on the same scale for cross comparison. (A) A (6-4)PP repair
641 hotspot from chr2. (B) A (6-4)PP repair coldspot from chr1. (C) A CPD repair hotspot from
642 chr1. All examples are from intronic gene regions overlapping annotated enhancers.
643 Zoomed-in view of canonical sequences are overlaid in the bottom, with the
644 damage/repair sites shown in dashed boxes.
645
646 **Figure 3. Repair hotspots are not damage hotspots.** There is no enriched DNA
647 damage, as measured by Damage-seq, at the identified repair hotspots for (A) (6-4)PP
648 and (B) CPD. Regions corresponding to the different repair categories are extended at
649 both ends for 20bp and 500bp, respectively, to account for the different sampling densities
650 between Damage-seq and XR-seq.
651
652 **Figure 4. Genome-wide repair hotspots are associated with epigenomic markers.**
653 Chromatin accessibility (DNase I hypersensitivity) is higher for repair hotspots and lower
654 for coldspots. Repair hotspots are also characterized by higher ChIP-seq signals for
655 H3K4me1, H3K4me3, and H3K27ac, markers for gene activations.
656

657 **Figure 5. Genome-wide repair hotspots are enriched for enhancers.** Annotations for
658 (A) chromatin states, (B) CpG islands, and (C) genic/intergenic regions are shown for
659 repair hotspots for (6-4)PP (top) and CPD (bottom), separated by strands. Repair
660 hotspots are enriched for enhancers and promoters, which are in open-chromatin regions.
661 Repair hotspots are also enriched in inter-CpG and intronic regions, both of which are AT
662 rich.
663
664 **Figure 6. Genome-wide repair hotspots overlap with FIREs and super-enhancers**
665 **identified by Hi-C.** FIREs and super-enhancers were identified and annotated using Hi-
666 C data from human fibroblasts. Repair hotspots overlap with (A) FIREs and (B) super-
667 enhancers with significantly higher proportions compared to the genome-wide averages.
668 The solid density curves are generated from bootstrapping different regions along the
669 genome as the null case; the dashed vertical lines are the observed proportions for the
670 repair hotspots and coldspots. (C) Repair hotspots have significantly higher number of
671 significant interactions, identified by Hi-C. (D) Two examples of (6-4)PP repair hotspots
672 (chr8:13224201-13224300 and chr14:61994601-61994700) that overlap with both FIREs
673 and super-enhancers, and loop to different regions of the genome. Hi-C data has low
674 resolution and the significant interactions are drawn from the center of each bin, which
675 does not exactly overlap with the identified hotspot shown in red.
676
677 **Figure 7. Genome-wide repair hotspots are enriched for early-replication domains.**
678 Replication timing domains were identified using Repli-Seq data. Compared to genome-
679 wide average (random), there is a significantly higher proportion of the early-repair
680 hotspots located in the early-replication domains.

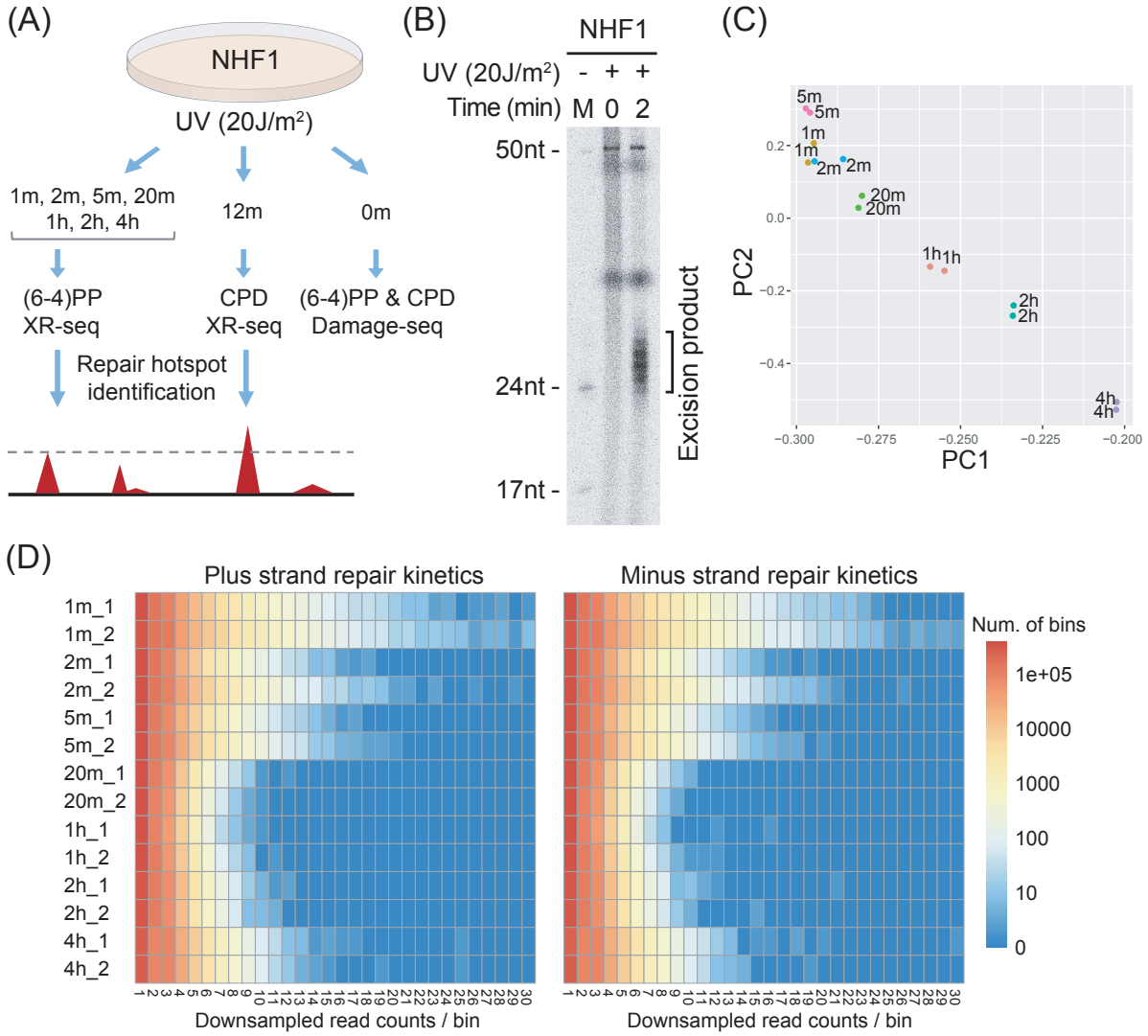
(A)
NHF1
UV (20J/m²)
1m, 2m, 5m, 20m
1h, 2h, 4h
12m
0m
(6-4)PP
XR-seq
CPD
XR-seq
(6-4)PP & CPD
Damage-seq
Repair hotspot
identification
(B)
NHF1
UV (20J/m²) - + +
Time (min) M 0 2
50nt -
24nt -
17nt -
Excision product
(C)
PC2
PC1
5m
1m
2m
20m
1h
2h
4h
(D)
Plus strand repair kinetics
Minus strand repair kinetics
1m_1
1m_2
2m_1
2m_2
5m_1
5m_2
20m_1
20m_2
1h_1
1h_2
2h_1
2h_2
4h_1
4h_2
Downsampled read counts / bin
Num. of bins
1e+05
10000
1000
100
10
0

Figure 1

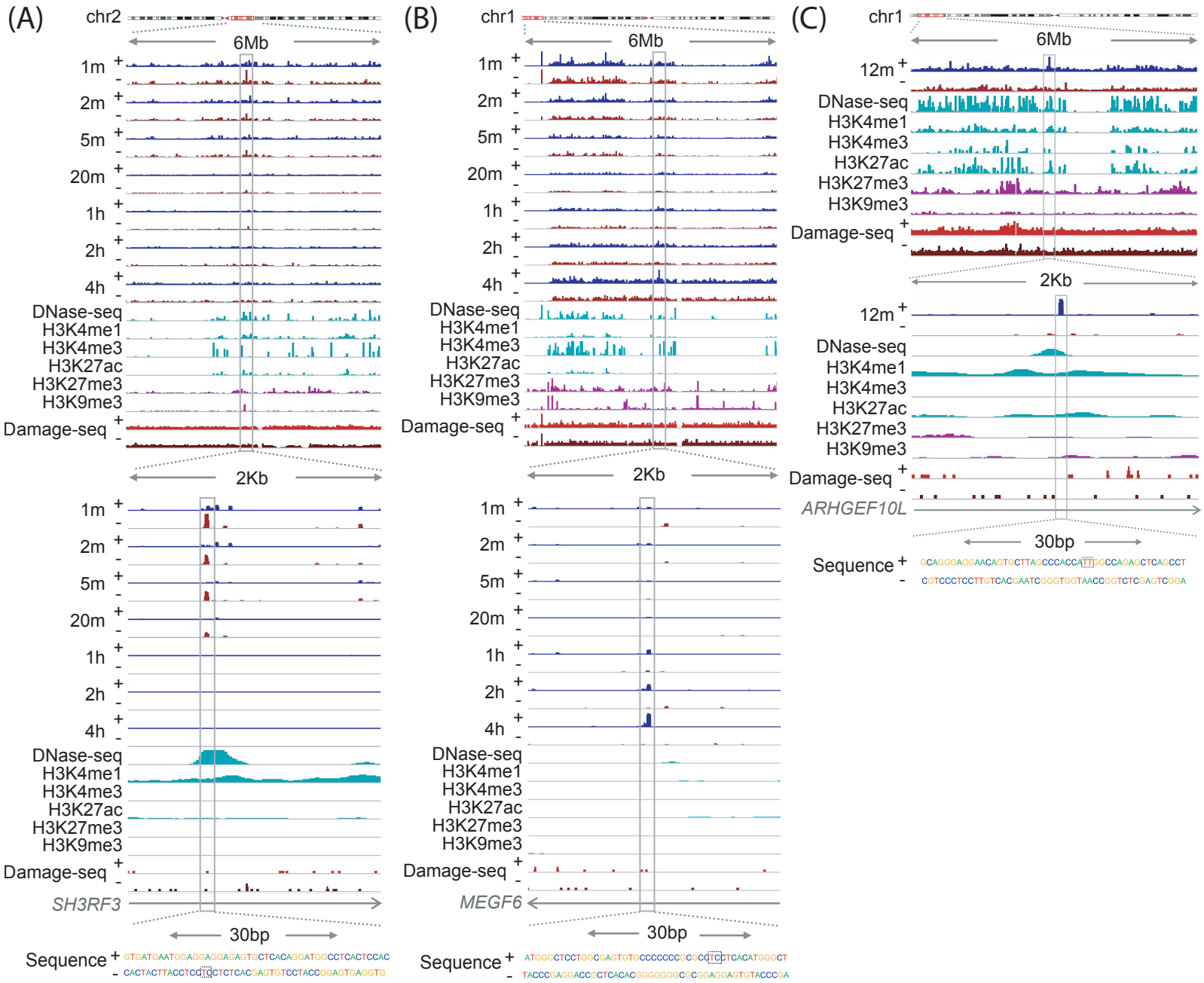
(A)
chr2
6Mb
1m
2m
5m
20m
1h
2h
4h
DNase-seq
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
Damage-seq
2Kb
SH3RF3
30bp
Sequence
(B)
chr1
MEGF6
(C)
12m
ARHGEF10L

Figure 2

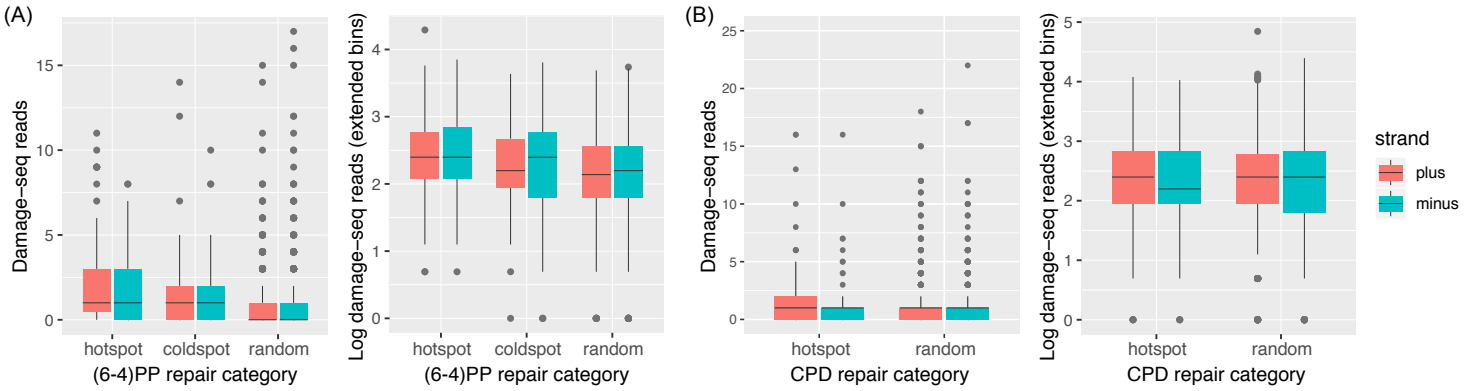
(A)
Damage−seq reads
hotspot
coldspot
random
(6-4)PP repair category
Log damage−seq reads (extended bins)
hotspot
coldspot
random
(6-4)PP repair category
(B)
Damage−seq reads
hotspot
random
CPD repair category
Log damage−seq reads (extended bins)
hotspot
random
CPD repair category
strand
plus
minus

Figure 3

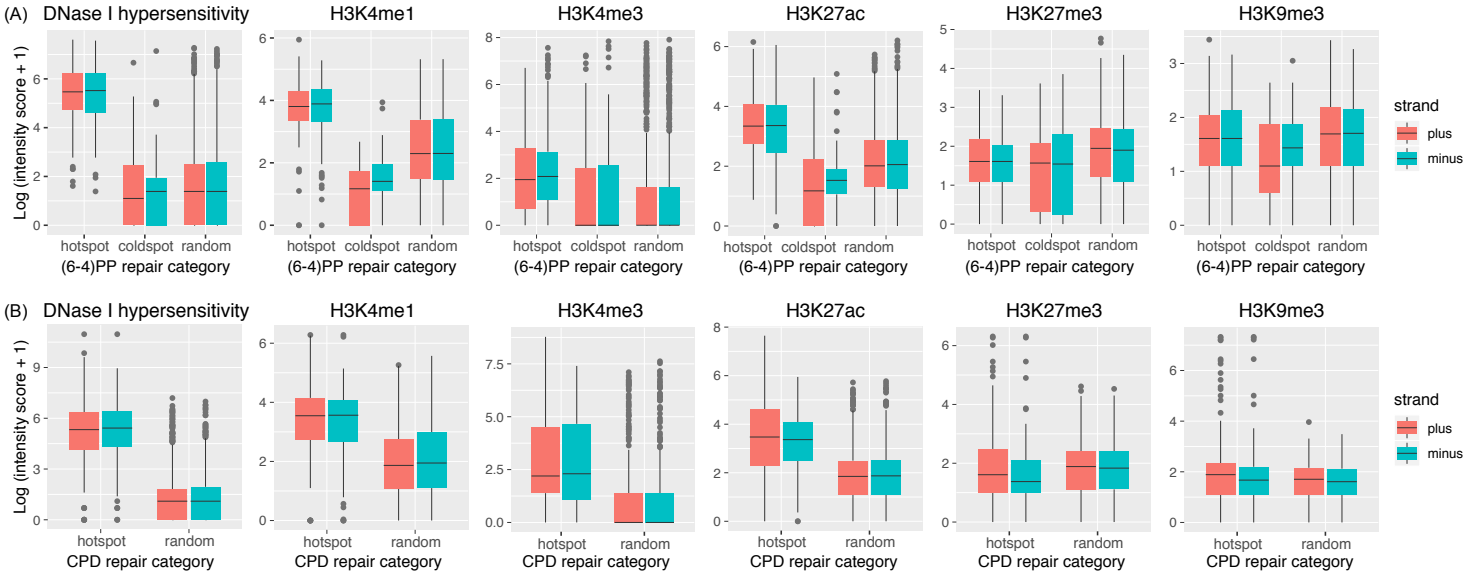
(A)
DNase I hypersensitivity
H3K4me1
H3K4me3
H3K27ac
H3K27me3
H3K9me3
Log (intensity score + 1)
hotspot
coldspot
random
(6-4)PP repair category
strand
plus
minus
(B)
CPD repair category

Figure 4

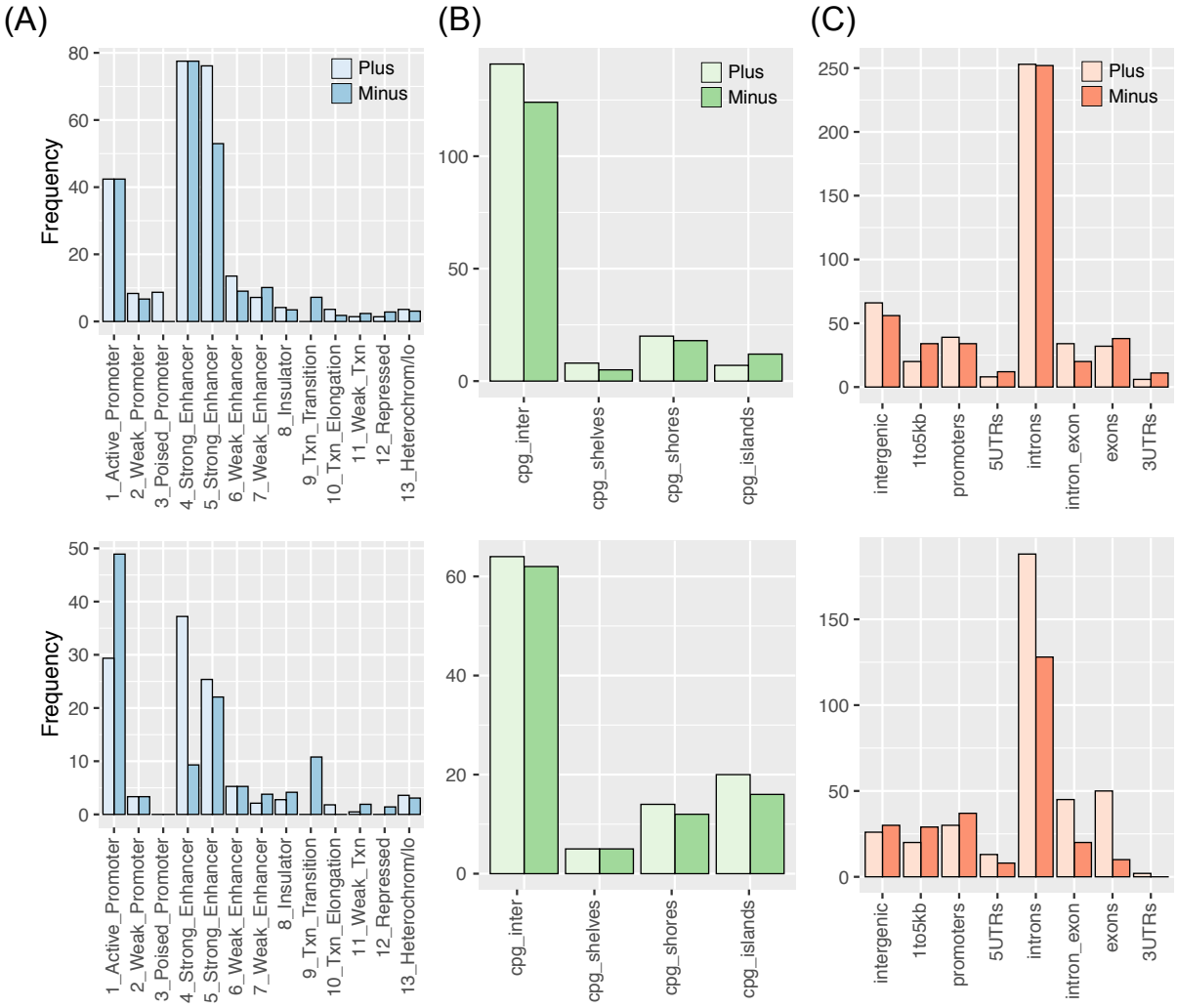
(A)
(B)
(C)
Frequency
Plus
Minus
1_Active_Promoter
2_Weak_Promoter
3_Poised_Promoter
4_Strong_Enhancer
5_Strong_Enhancer
6_Weak_Enhancer
7_Weak_Enhancer
8_Insulator
9_Txn_Transition
10_Txn_Elongation
11_Weak_Txn
12_Repressed
13_Heterochrom/lo
cpg_inter
cpg_shelves
cpg_shores
cpg_islands
intergenic
1to5kb
promoters
5UTRs
introns
intron_exon
exons
3UTRs

Figure 5

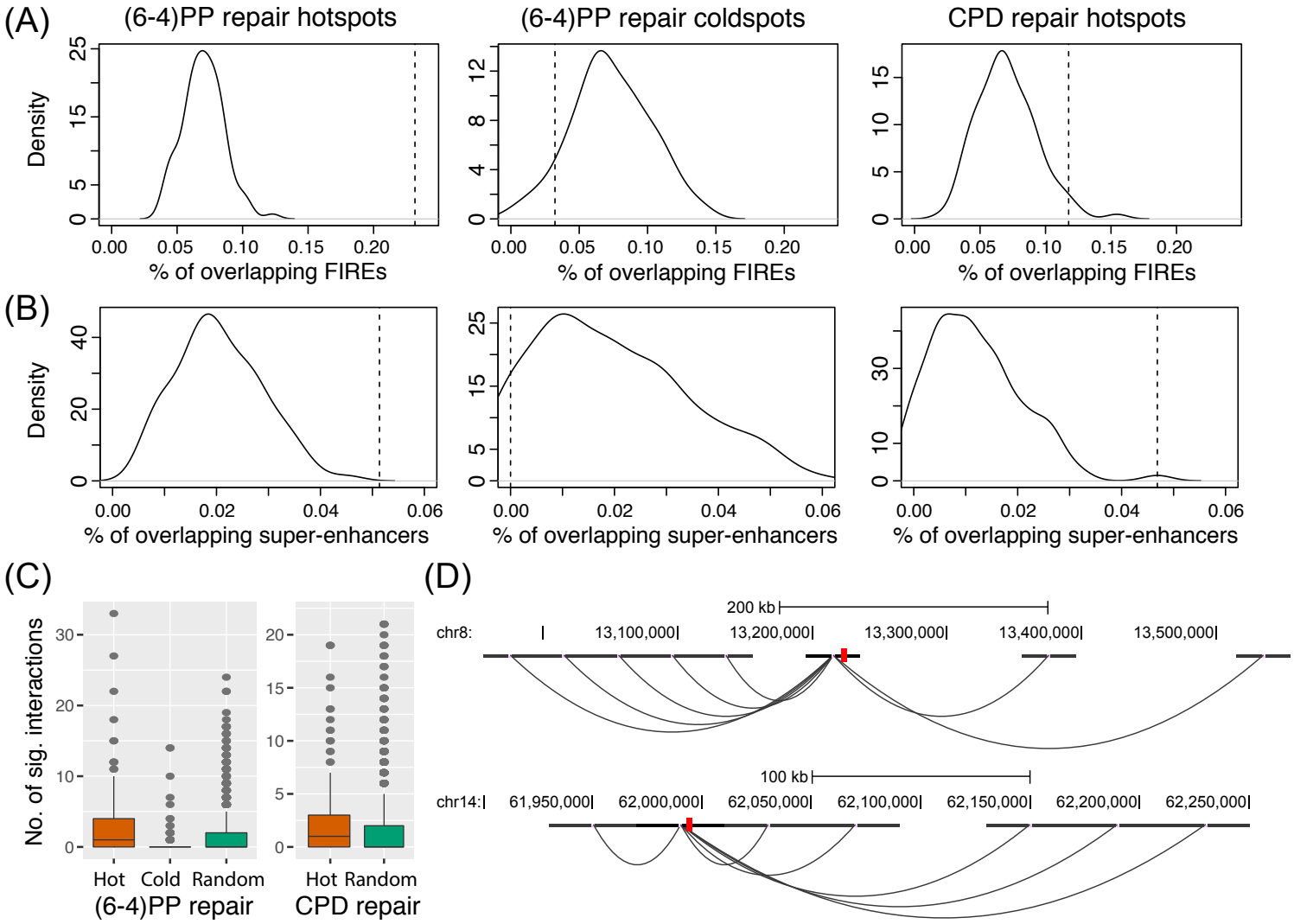
(A)
(6-4)PP repair hotspots
(6-4)PP repair coldspots
CPD repair hotspots
Density
% of overlapping FIREs
(B)
Density
% of overlapping super-enhancers
(C)
No. of sig. interactions
Hot
Cold
Random
(6-4)PP repair
Hot
Random
CPD repair
(D)
200 kb
chr8:
13,100,000
13,200,000
13,300,000
13,400,000
13,500,000
100 kb
chr14:
61,950,000
62,000,000
62,050,000
62,100,000
62,150,000
62,200,000
62,250,000

Figure 6

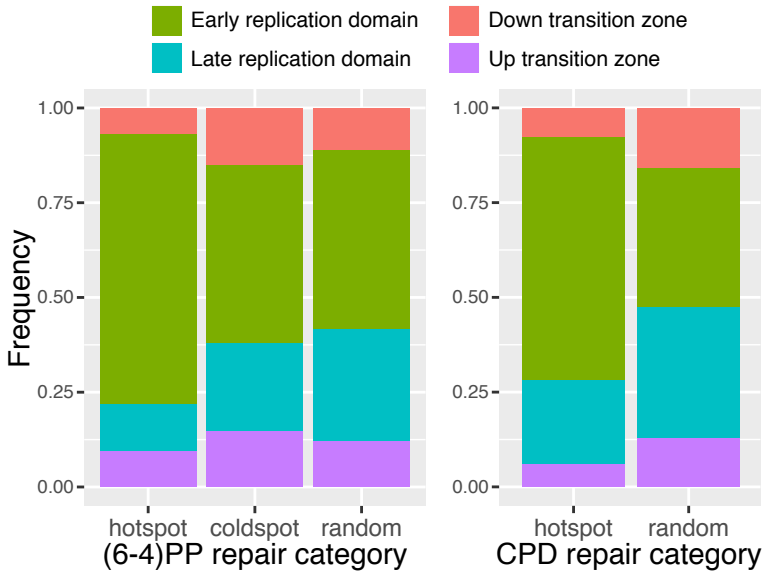
Early replication domain
Late replication domain
Down transition zone
Up transition zone
Frequency
1.00
0.75
0.50
0.25
0.00
hotspot
coldspot
random
(6-4)PP repair category
hotspot
random
CPD repair category

Figure 7