The landscape and consequences of transcription stress

Osama Hidmi[1], Sara Oster Flayshman[1], Diala Shatleh[1], Jonathan Monin[1] and Rami I. Aqeilan[1,2,]

[1]The Concern Foundation Laboratories, The Lautenberg Center for Immunology and Cancer Research, Department of Immunology and Cancer Research-IMRIC, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel
[2]Lead contact

ramiaq@mail.huji.ac.il (R.I.A.)

Abstract

Cancer is characterized by uncontrolled proliferation and hypertranscription of oncogenes, leading to transcription stress, a major source of DNA double-strand breaks (DSBs) that threaten genomic stability. However, the landscape and consequences of transcription stress remain incompletely understood. Here, using genome-wide maps of DSBs identified by sBLISS alongside transcription stress markers, we reveal that transcription stress regions are highly enriched at the super-enhancer (SE) regulatory landscape. Transcriptional induction through SEs markedly amplifies DSB accumulation compared to non-SE-driven transcription. γH2AX mapping revealed selective enrichment at transcription stress sites, with only a subset of DSB-enriched genes showing strong γH2AX marking. Notably, genes with high DSBs and strong γH2AX signals exhibited significantly higher DSB turnover and repair than γH2AX-low genes, correlating with increased vulnerability to mutagenesis. These findings identify SE-driven transcription as a major determinant of the transcription stress landscape in cancer and highlight a key mechanism threatening the genomic stability of oncogenes.

Introduction

The most fundamental hallmark of cancer cells is their ability to sustain proliferation, which they acquire through genetic alterations during neoplastic transformation[1]. Simultaneously, dysregulation of the transcriptional program emerges to accommodate the cell's increased demand for cellular factors essential for proliferation, growth, and survival[2,3]. Consequently, genes encoding these factors, i.e., oncogenes, are hypertranscribed to fulfill the needs of cancer cells[4].

A well-established mechanism by which cancer cells upregulate oncogenes involves the formation of oncogenic super-enhancers (SEs)[5–12]. In normal cells, SEs drive the expression of cell-type specific genes that encode master transcription factors (MTFs), which shape cellular identity and the transcriptional program[2]. In cancer cells, SEs drive the expression of cancer-type specific oncogenes and oncogenic MTFs that dictate their transcriptional programs and support their proliferation[2,12]. SEs can become dysregulated in cancer through various mechanisms: 1) overexpression of enhancer transcription factors renders SEs much more effective at driving the expression of their associated genes; 2) translocations can relocate SEs near an inactive oncogene in the corresponding normal cell, and 3) focal amplifications of SEs can lead to the increased expression of target genes[12]. These mechanisms contribute to the activation of proto-oncogenes by SEs, which drive the transcription of these genes to an exceptionally high level termed "hypertranscription."

Transcription is increasingly viewed as a source of DNA damage[13–18]. The most damaging type of DNA lesions produced during transcription is DNA double-strand breaks (DSBs)[19]. DSBs resulting from transcription are primarily facilitated by the formation of pathogenic co-transcriptional R-loops and the aberrant activity of topoisomerase 1 (TOP1)[13,20]. R-loops tend to form more frequently at highly expressed genes, and their formation causes TOP1 to become trapped in its cleavage complex form (TOP1cc)[13].

Although transcription is a highly regulated process, it can occasionally become abortive and result in DSBs. These transcriptional aberrations occur more frequently as transcription levels increase, in a phenomenon known as "transcription stress"[21]. Characterizing the "hotspots" of transcriptional stress is crucial because transcription-related DSBs can lead to cell death, genomic instability, and mutations[22,23]. Moreover, since cancer cells upregulate genes vital for their proliferation and survival, a deeper understanding of transcriptional stress and transcription-

58 associated DSBs linked to these oncogenes can aid in exploiting their vulnerabilities to target
59 cancer cells more effectively. Nevertheless, while transcription stress has been increasingly
60 studied[24–26], its landscape and consequences, particularly in cancer, remain incompletely
61 characterized.
62 DSBs are impediments to transcription[27,28] and must, therefore, be efficiently repaired to sustain
63 continuous high transcriptional levels. However, it is not understood how regions of transcription
64 stress are equipped to handle these deleterious DSBs.
65 In this study, we characterize the landscape of transcription stress in cancer cells and reveal
66 how these regions are structured to promote the rapid and efficient repair of transcriptional DSBs
67 that sustain hypertranscription.
68
69 **Results**
70
71 **Super-enhancer-regulated oncogenes are major transcription stress sites.**
72 Pathogenic R-loops preferentially form at highly expressed genes[29], facilitating aberrant TOP1
73 activity and TOP1cc trapping[13]. The high TOP1 occupancy, TOP1cc trapping, and co-
74 transcriptional R-loops all contribute to the formation of transcription-associated DSBs [20,30].
75 Therefore, we reasoned that these features, along with genome-wide maps of endogenous
76 DSBs (hereafter referred to as "Breakome"), could serve as markers of transcriptional stress. To
77 determine transcription stress sites in the genome, we analyzed sBLISS and ChIP-seq data for
78 the breast cancer cell line MCF7, and we applied a Hidden Markov Model (HMM)-based
79 approach to identify regions in the genome with co-occurrence across the four datasets (sBLISS,
80 DRIP-seq, TOP1cc, and TOP1 ChIP-seq). As expected, the majority of these regions were genic
81 (Figure S1A), while the rest were enriched mainly at intergenic super-enhancer (SEs) (Figure
82 S1B and C). To investigate which genes serve as transcription stress sites, we scored each
83 gene based on TOP1, TOP1cc, R-loops, and DSBs scores (see Methods). This scoring yielded
84 a shortlist of genes, many of which have well-established oncogenic functions (Figure 1A). Upon
85 closer analysis of this list, we noted that many of the genes were regulated by SEs in MCF7
86 cells, including *CCND1*, *NEAT1*, *miR21*, and *MYC* (SE-regulated genes were downloaded from
87 SEdb 2.0[31]) (Figure 1A). To statistically validate this observation, we systematically assessed
88 the representation of multiple gene sets among transcription stress sites by performing gene set

89 enrichment analysis (Figure 1B; see Methods). SE-regulated genes were the most significantly
90 enriched category, with SE-regulated oncogenes exhibiting the highest enrichment (~122-fold
91 over expectation under a random model, $P = 1.85 \times 10^{-15}$; hypergeometric test). Similar results
92 were obtained when quantifying the overlap between each gene set and the co-marked regions
93 identified by HMM (Figure S1D). Notably, the enrichment of SE-regulated genes was
94 substantially higher than that of typical-enhancer (TE)-regulated genes or highly expressed
95 genes lacking enhancer regulation (Figure 1B), despite having similar expression levels (Figure
96 S1E). To control for factors that might confound this analysis, we performed multivariate logistic
97 regression using SE-status, expression level, GC content, and gene length as predictors for
98 transcription stress at genes. The results of this model demonstrate that SE-status was the most
99 statistically significant independent predictor of transcription stress (Figure S1F), indicating
100 specificity to genes regulated by SEs.
101 DSBs at these genes coincide with TOP1cc and R-loops (Figure 1C). Analysis of sBLISS data
102 for MCF7 cells depleted for TOP1 and R-loops showed that most of these genes have
103 significantly fewer breaks upon TOP1 and R-loops depletion (Figure S2A). Additionally, DSB are
104 enriched at SE-regulated genes in both cycling and G1-synchronized cells (Figure S2B-D),
105 validating their transcriptional origin and independence from replication.
106 The relationship between transcription and DSBs was previously established[13–15]. However, our
107 results suggest that this association may be more pronounced in SE-regulated genes. This
108 prompted us to further investigate the discrepancies in the relationship between transcription
109 and DSBs in SE-regulated versus non-SE-regulated genes. Interestingly, the strong correlation
110 between gene expression and DSBs is observed only in SE-regulated genes and not in others
111 (Figure S2E-G). Moreover, highly expressed genes enriched with DSBs showed significantly
112 greater enrichment in TOP1, TOP1cc, and R-loops than similarly expressed genes lacking DSBs
113 (Figure S3A-E). Additionally, SE-regulated genes were significantly more prevalent among
114 highly expressed genes enriched with DSBs (Figure S3F). These findings indicate that the
115 association between transcription and DSBs is primarily attributed to SE-regulated genes.
116 Since SE activity is significantly more potent in cancer cells than in normal cells, we next aimed
117 to test whether the enrichment of DSBs at SE-regulated genes is a phenomenon specific to
118 cancer. To investigate this, we mapped the Breakome in 10 different cell lines, which included
119 three normal cell lines, one pre-malignant cell line, and six cancer cell lines. We found that SE-

120 regulated genes are more enriched with DSBs in cancer cells than in their normal counterparts
121 (Figure 1D), demonstrating the specificity of transcriptional stress to active SE-regulated
122 oncogenes in cancer cells.
123
124 **The super-enhancer landscape shapes transcription stress sites in cancer.**
125 The landscape of SEs is cell-type specific, enhancing the transcription of different genes based
126 on cell identity[12]. If SEs cause transcription stress in their target genes, one would predict that
127 various cancer cell types exhibit distinct transcription stress sites reflective of their SE
128 landscapes. To test this, we mapped the breakome of multiple cancer cell lines, focusing on
129 each cell line's break-prone SE-regulated genes. Interestingly, SE-regulated genes were
130 enriched in DSBs, predominantly in the cell type where they are regulated by an SE (Figure 2A).
131 This pattern was also apparent when comparing cell types with very similar characteristics
132 (MCF7 and T47D, both of which are ER+ luminal A breast cancer cell types), yet they exhibit
133 differing SE landscapes, suggesting that transcription stress sites correspond to the SE
134 landscape in cancer.
135 If SE activity indeed causes the accumulation of DSBs at transcription stress sites, then inhibiting
136 its activity would abolish SE-mediated DSB accumulation, while conversely, activating it would
137 enhance DSB accumulation at target genes. To test this hypothesis, we employed two
138 strategies. First, we inhibited SE activity in MCF7 and HeLa cells through two approaches: using
139 the BRD4 degron dBET6 and the CDK9 inhibitor (CDK9i) NVP-2, both of which have been shown
140 to lead to SE collapse and repress SE-mediated transcription[32,33]. Examining the breakome after
141 a 4-hour incubation with these inhibitors, which effectively suppressed SE-regulated
142 transcriptional activity as measured by spike-in real-time quantitative PCR (RT qPCR) (Figure
143 S2G-H), revealed that SE inhibition significantly reduced DSBs at SE-regulated genes in both
144 cell lines (Figure 2B). Furthermore, SE inhibition using either dBET6 or CDK9i eliminated the
145 cell-type-specific effect of the SE landscape on the breakome (Figure 2C). Secondly, we utilized
146 the fact that MCF7 cells are estrogen receptor-positive and that many of their SEs are estrogen-
147 responsive, and analyzed the breakome of these cells before and after estrogen addition. We
148 specifically focused on genes regulated by estrogen receptor-occupied SE-regulated genes
149 (ERSE-genes), which were previously characterized[34]. Consistently, ERSE induction
150 significantly increased DSBs in target genes but not random genes (Figure 2 D-E), indicating

151 that SE activity promotes the accumulation of DSBs at SE-regulated genes. Overall, these data
152 demonstrate that the landscape of SE activity in cancer cells shapes the landscape of
153 transcription stress in a cell-type and cancer-type-specific manner.
154

155 **Specificity of transcriptional stress at super-enhancer-regulated genes**

156 While multiple studies have demonstrated the formation of DSBs as a consequence of
157 transcription, independent of SE regulation [35–37], our results imply that this phenomenon is
158 markedly amplified when transcription is driven by SEs (Figure 1B; Figure S1E–F). To further
159 support this observation, we employed two complementary approaches. First, we examined
160 predictors for the genomic locations of transcription stress sites. Using SE-regulated genes,
161 typical enhancer (TE)-regulated genes, and generally highly expressed genes as predictors,
162 multivariate LASSO regression revealed that SE regulation was the strongest predictor of
163 transcription stress sites (odds ratio = 12.61; Figure 3A), even though the highly expressed gene
164 group exhibited the highest transcriptional levels overall (Figure 3B). Notably, SE regulation was
165 also the best predictor for genomic regions enriched in DSBs (Figure 3C).

166 Second, to experimentally assess whether transcriptional activation through SEs promotes
167 greater transcription stress and DSB accumulation, we conducted a deeper analysis of the
168 MCF7 breakome before and after estradiol treatment. Estrogen activates transcription via both
169 promoter-proximal binding and estrogen-responsive SEs[38], enabling us to compare the impacts
170 of these two mechanisms. We grouped genes by their level of transcriptional upregulation
171 (based on GRO-seq fold change) and further subdivided each group into SE-regulated and non-
172 SE-regulated categories. Consistently, SE-regulated genes accumulated significantly more
173 DSBs than their non-SE counterparts. This effect was especially prominent in the highest
174 induction group (fold change > 10; Figure 3D). A similar pattern was observed upon CDK9
175 inhibition, where SE-regulated genes exhibited a significantly greater reduction in DSBs
176 compared to non-SE-regulated genes with matched transcriptional downregulation (measured
177 by RNA-seq; Figure 3E).

178 Together, these results demonstrate that SE-mediated transcription is a major determinant of
179 transcription stress sites and results in increased DSB accumulation compared to non-SE
180 transcriptional activity, even at equivalent expression levels.[35]

181

182 **γH2AX marks transcription stress sites**

183 Since DSBs are transcription-blocking lesions[28], transcription stress sites must be rapidly and
184 efficiently repaired to maintain high expression levels. γH2AX is a marker for DDR activation and
185 is among the first responders to DSBs, as it is phosphorylated within a few minutes after DSB
186 formation[39]. To further explore the role of DDR activation in the signaling of DSB repair at
187 transcription stress sites, we conducted ChIP-seq for γH2AX on untreated MCF7 cells and
188 analyzed high-quality available γH2AX ChIP-seq data for HeLa and IMR90 cell lines.

189 In both MCF7 and HeLa cells, γH2AX signal is predominantly localized to gene bodies, while in
190 IMR90 cells, the distribution is more balanced between genic and intergenic regions (Figure
191 S4A). This pattern is consistent with a greater burden of transcription-associated stress in cancer
192 cells compared to non-transformed cells. Moreover, γH2AX is distributed across the gene bodies
193 of top targets (Figure S4B-D). Ranking genes based on their γH2AX enrichment revealed that
194 most exhibit relatively low γH2AX levels, whereas a subset displays high to very high γH2AX
195 density (Figure 4A). This finding indicates a non-uniform distribution of γH2AX, with preferential
196 enrichment at specific genes. Interestingly, all transcription stress sites are among the top 8%
197 of γH2AX targets (Figure 4A-B). This enrichment in the top γH2AX genes was considerably less
198 for the general high DSB genes (Figure 4C). Additionally, top γH2AX targets are enriched in SE-
199 regulated genes across each respective cell line (P= $2.3e^{-41}$, $2.9e^{-39}$, $8.5e^{-8}$, for MCF7, HeLa,
200 and IMR90 respectively; hypergeometric test) (Figure 4D-F).

201 The observed γH2AX enrichment across the three cell lines was consistent for several
202 oncogenes, including *miR21, NEAT1, MALAT1,* and *ACTG1* (Figure 4G). Interestingly, many
203 prominent γH2AX targets were unique to specific cell lines, most with established activity in their
204 respective cell types (Figure 4H). For example, two top γH2AX targets in MCF7, *TFF1* and
205 *GATA3*, are well-known oncogenic transcription factors primarily active in breast cancer and
206 were exclusively enriched in γH2AX in MCF7 cells. *KRT7* and *CCAT1*, which are overexpressed
207 and linked to poor prognosis in cervical cancer, are enriched in γH2AX only in HeLa cells.
208 Consequently, we questioned whether these variations in γH2AX enrichment stem from cell-
209 type-specific SEs. To investigate this, we compared γH2AX enrichment of top targets that are
210 regulated by a cell-type specific SE in each respective cell line. Indeed, SE-regulated γH2AX

211 targets are predominantly enriched in the cell line where they are regulated by an SE (Figure 4
212 I), underscoring that γH2AX marks cell type-specific transcriptional stress sites.
213
214 **γH2AX levels vary significantly across genes enriched with high endogenous DSBs**.
215 Next, we examined the correlation between DSBs and γH2AX in genes. Interestingly, analyzing
216 the correlation between each gene's breaks and its γH2AX density revealed significant variability
217 of γH2AX among genes with high DSBs (Figure 4 J-K and Figure S5A-E). Some genes enriched
218 with breaks are marked by γH2AX, while others are not in both MCF7 and HeLa cell lines (Figure
219 S5F-G). This discrepancy in γH2AX enrichment at various DSB-enriched genes might suggest
220 that regions enriched with endogenous DSBs in the genome are signaled with varying efficiency
221 and potentially repaired unevenly, depending on their context.
222
223 **γH2AX preferentially marks endogenous DSBs of highly active chromatin**
224 Although genes exhibit a moderate correlation between DSB density and γH2AX density
225 (Spearman's correlation = 0.57), non-linear regression analysis using Generalized Additive
226 Models (GAM) indicates that the transcription stress score explains a greater proportion of the
227 variance in γH2AX levels (R^2 = 0.73) than any individual component (Figure 5A and Figure S6).
228 To further investigate the underlying factors contributing to variability in γH2AX at high DSBs
229 genes, we defined two groups of genes with similar break density but different γH2AX levels
230 (Figure 4J): $DSBs^{high}$ $\gamma H2AX^{high}$ genes and $DSBs^{high}$ $\gamma H2AX^{low}$ genes. $DSBs^{high}$ $\gamma H2AX^{high}$ genes
231 are more expressed and enriched with transcription stress markers (TOP1, TOP1cc, and R-
232 loops) (Figure S7A-E). Furthermore, 64 of the 65 transcriptional stress sites were among
233 $DSBs^{high}$ $\gamma H2AX^{high}$ genes (Figure 4J and Figure 5B). This group also included all SE-regulated
234 transcription stress sites and is significantly enriched with active SE-regulated genes compared
235 to $DSBs^{high}$ $\gamma H2AX^{low}$ genes, which lack active SE-regulated genes and transcriptional stress
236 hotspots (Figure 5C and Figure S7F).
237 Chromatin context was previously shown to impact the phosphorylation of H2AX, specifically,
238 heterochromatin was shown to be refractory to H2AX phosphorylation[40]. Therefore, we
239 wondered whether the observed $DSBs^{high}$ $\gamma H2AX^{low}$ genes are close to, or within,
240 heterochromatic regions. To gain further insights into whether γH2AX preferentially marks

241 endogenous DSBs based on the chromatin context, we defined two groups of individual DSB
242 sites across the genome according to their γH2AX signal (Figure 5D and Figure S5G) and
243 mapped them to the different chromatin states of MCF7 and HeLa. γH2AX^{high} DSBs were biased
244 toward active chromatin states, while γH2AX^{low} DSBs were biased toward silent regions (Figure
245 5E and Figure S5H). These results made us question whether the differential γH2AX marking at
246 DSB-enriched genes was due to their chromatin context. Indeed, the DSBs^{high} γH2AX^{low} group,
247 but not the DSBs^{high} γH2AX^{high} group, is enriched with DSB-enriched genes that overlap with
248 heterochromatin and repressed chromatin states (Figure 5F and Figure S5I). Additionally, an
249 analysis of the chromatin state contexts of DSBs^{high} γH2AX^{high} and DSBs^{high} γH2AX^{low} in both
250 MCF7 and HeLa showed that DSBs^{high} γH2AX^{high} genes primarily overlap with chromatin regions
251 associated with strong transcriptional activity. Conversely, DSBs^{high} γH2AX^{low} has a higher
252 overlap with heterochromatin, repressed, and weak transcription chromatin states (Figure 5G
253 and Figure S5J).
254 Our results illustrate that DSBs alone do not fully account for the variation in γH2AX across genes
255 (Figure 5A). Conducting GAM on genes categorized by their chromatin context revealed that
256 DSBs primarily account for variation in γH2AX in genes that overlap with strong enhancer regions
257 and least in genes that overlap with repressed or heterochromatin regions (Figure 5H). This
258 suggests that breaks and γH2AX are most closely associated with genes spanning strong
259 enhancer elements.
260 Collectively, these results indicate that γH2AX preferentially marks endogenous DSBs in active
261 chromatin compared to inactive chromatin. This differential marking may influence variations in
262 the repair efficiency of distinct DSBs based on their chromatin context and γH2AX signaling.
263
264 **Genes marked with γH2AX exhibit high DSB turnover and repair efficiency**
265 sBLISS, like other methods used to map DSBs, captures a snapshot of the DSB landscape at
266 the moment of fixation[41–45]. Therefore, a gene that shows high levels of DSBs could experience
267 frequent DSB generation and repair cycles (high DSB turnover) or infrequent DSB generation
268 and repair cycles (low DSB turnover). To estimate the rate of endogenous DSB repair across
269 the genome, we mapped DSBs in HeLa cells after disrupting the primary DSB repair pathway,
270 non-homologous end-joining (NHEJ)[45], by using siRNAs that target XRCC4. Regions that are

271 frequently repaired should accumulate excessive DSBs due to repair impairment. In contrast,
272 regions with inherently low repair rates should show minimal DSB accumulation (Figure 6A). To
273 objectively determine which genes accumulate DSBs following XRCC4 knockdown (Figure 6B),
274 we employed the DESeq2 algorithm[46] on DSB count matrices, identifying 429 genes with
275 significant DSB accumulation upon XRCC4 KD (Figure 6C). Stratifying these genes by their
276 baseline DSB density revealed a trend in which groups with higher baseline DSB density showed
277 progressively greater impact following XRCC4 KD (Figure 6D), supporting the notion that DSB-
278 prone genes are repaired more effectively. Enrichment analysis indicated a strong enrichment
279 of DSBs^{high} γH2AX^{high} genes (Figure 6E), along with the top DSB-enriched and active SE-
280 regulated genes. In contrast, DSBs^{high} γH2AX^{low} genes and the bottom DSB-enriched and non-
281 SE-regulated genes were not significantly enriched. Directly examining the effects of repair
282 impairment demonstrated that NHEJ impairment leads to the most pronounced accumulation of
283 DSBs at DSBs^{high} γH2AX^{high} and SE-regulated genes, compared to DSBs^{high} γH2AX^{low} and non-
284 SE genes (Figure 6F). To investigate further how γH2AX patterns influence the repair and
285 turnover of endogenous DSBs, we used the accumulation of DSBs following repair impairment
286 as a proxy for repair efficiency (Figure 6G). Within DSBs^{high} γH2AX^{high} genes, repair efficiency
287 markedly increases with higher DSB levels; however, within DSBs^{high} γH2AX^{low} genes, this
288 relationship of repair efficiency is significantly weaker (Figure 6H). These findings suggest two
289 distinct types of endogenous DSBs: γH2AX-marked DSBs, which exhibit high turnover,
290 and poorly γH2AX-marked DSBs, which exhibit low turnover.
291
292 **Efficient repair at transcriptional stress sites is crucial for SE-regulated oncogene**
293 **expression.**
294 High γH2AX marking and repair efficiency at transcription stress sites under physiological
295 conditions suggests that these regions rely on DSB repair to sustain their elevated transcriptional
296 levels. To test this, we treated MCF7 cells with an ATM inhibitor (ATMi), which impairs DSBs
297 repair, and conducted sBLISS and spike-in RT qPCR. Indeed, ATM inhibition led to an
298 accumulation of DSBs at genes, biased toward active SE-regulated genes (Figure S8A-B).
299 Measuring mRNA levels of the top γH2AX marked SE-regulated genes indicated a significant
300 decrease in their expression levels (Figure S8C). To determine whether this reduction in

301 expression was due to the inhibition of the ATM axis or the accumulation of DSBs, we knocked
302 down XRCC4 under the same conditions as figure 5 and performed spike-in RT qPCR.
303 Interestingly, the top γH2AX marked SE-regulated genes, which exhibited DSB accumulation,
304 significantly decreased in expression (Figure S8D). These results indicate that the failure to
305 repair DSBs, rather than merely inhibiting DDR signaling, impedes the expression of SE-
306 regulated oncogenes.
307
308 **Efficiently repaired transcriptional DSBs cluster in the 3D genome.**
309 Induced DSBs cluster in repair foci[47,48], especially when introduced within active genes[48]. We
310 hypothesized that endogenous transcription-associated DSBs might also cluster within the 3D
311 genome as a mechanism to enhance repair. To investigate this, we analyzed high-quality Hi-C
312 data generated from MCF7 cells[49].
313 To quantify the clustering of genes associated with break density, we introduced two Hi-C
314 metrics that measure how a group of genes with a common characteristic (e.g., break density)
315 is clustered (Figure 7A): intra-group interactions, which represent the sum of genomic interaction
316 scores between genes within the same group, and inter-group interactions, the sum of genomic
317 interaction scores between genes from different groups. A clustered group of genes is expected
318 to exhibit higher intra-group interactions and lower inter-group interactions compared to a non-
319 clustered group (Figure 7A). To investigate whether DSB-enriched genes exhibit clustering,
320 genomic bins were arranged according to the break densities of their constituent genes and
321 categorized into nine groups based on break density. Indeed, bins containing DSB-enriched
322 genes demonstrate higher intra-group interactions and lower inter-group interactions than bins
323 with low DSB genes (Figure 7B), indicating the clustering of high DSB genes.
324 The percentage of strong interactions among $DSBs^{high}$ $\gamma H2AX^{high}$ genes is significantly higher
325 than that among $DSBs^{high}$ $\gamma H2AX^{low}$ genes (Figure 7C and Figure S9A). Additionally,
326 transcription stress sites exhibit a substantially higher percentage of strong interactions
327 compared to random sites with similar chromosomal distribution (Figure 7 D and Figure S9B),
328 suggesting that efficiently repaired transcription stress sites are clustered together. To avoid the
329 confounding effects of SE regulation on 3D interactions, we analysed Hi-C data from the
330 MCF10A cell line, which was conducted in parallel to MCF7[49]. MCF10A is a pre-cancerous cell
331 line that shares many common SE-regulated genes with MCF7 cells. However, these common

genes are significantly more enriched in DSBs within MCF7 cells (Figure 7E). This enabled a comparison of the same regions with the same SE-regulation status but differing DSB enrichment. Indeed, bins containing common SE-regulated genes, as opposed to random regions, showed significantly higher strong interchromosomal interactions in MCF7 cells than in MCF10A cells (Figure 7F-G), indicating that transcription stress sites that arise during carcinogenesis undergo increased clustering.

The repair of transcription stress sites is mutagenic.

Since transcription stress sites are more likely to be in close proximity (Figure 7), we suspected that recurrent DSBs at these sites could serve as mutational hotspots. To explore this, we downloaded and analysed mutational data from the COSMIC database[50], which contains extensive mutational data on human cancers. While mutations within coding regions are subject to selective pressures based on their impact on protein function and cellular fitness [51,52], non-coding mutations—particularly those in intronic regions—are less constrained by selection and can provide a more direct readout of a region's intrinsic susceptibility to mutation. To minimize the confounding effects of selection, we focused on non-coding intronic mutations in breast cancer and intersected them with our breakome data from the breast cancer cell line MCF7. Grouping genes by their DSB or γH2AX density revealed that genes with a higher break or γH2AX density exhibit greater intronic mutation density (Figure 8A-C). This association suggests that the repair of transcription-associated DSBs is inherently error-prone, leading to the accumulation of mutations over time.

To disentangle the effects of DSB density from repair and turnover, we modeled the relationship between γH2AX density and break density using nonlinear least squares (NLS) regression. The residuals from this model were then used to classify genes in each group into three subgroups: Under-responsive genes (negative residuals), Moderately responsive genes (residuals near zero), and Over-responsive genes (positive residuals). Notably, over-responsive genes showed the most increase in mutation density with increased DSBs (Figure 8D-F). This suggests that excessive γH2AX accumulation at DSB sites, which reflects recurrent DSB formation and turnover, and heightened repair activity, is error-prone. Furthermore, $DSBs^{high}$ $\gamma H2AX^{high}$ genes, but not $DSBs^{high}$ $\gamma H2AX^{low}$ genes, demonstrated intronic mutation enrichment, compared to

363 randomly selected genes (Figure 8G-I). Collectively, these results suggest that transcription
364 stress sites are prone to mutations, likely due to a combination of high DSB turnover, clustering,
365 and defective repair. These findings have important implications for understanding how
366 transcription-associated genome instability contributes to cancer evolution, as recurrent DSB
367 formation and error-prone repair may drive the accumulation of oncogenic mutations over time.
368
369
370 **Discussion**
371 Cellular processes–particularly those upregulated in cancer such as replication and
372 transcription–play a significant role in genomic instability[53]. Replication-dependent instability,
373 caused by replication stress[4,54], has been extensively studied, revealing its landscape[55,56],
374 causes, consequences[57], and potential for therapeutic targeting[58,59]. Transcription-replication-
375 dependent genomic instability[60], particularly transcription-replication conflicts (TRCs)[61], has also
376 been well characterized, with studies examining its genomic distribution, influencing factors, and
377 the impact of transcription-replication orientation[62]. Recently, attention has shifted toward
378 transcription-dependent, replication-independent genomic instability, namely, transcriptional
379 DSBs. While research has begun to uncover their causes and implications in disease[17,63],
380 several key aspects remain underexplored, including genomic hotspot patterns, repair
381 mechanisms, and their impact on cancer evolution. By utilizing multi-omic approaches, we
382 delineated the landscape of transcription stress and revealed that SEs shape the transcriptional
383 program and contribute to the formation of genomic instability hotspots in cancer.

384 SEs determine cell identity in normal cells[31]. In cancer cells, SEs enhance the transcription of
385 oncogenes to promote cell growth and proliferation [6,8,64,65]. Previous studies have established
386 that highly transcribed genes are prone to transcription-associated genome instability[13,15,20,30,66].
387 However, our data reveal that this relationship is significantly amplified in SE-regulated genes,
388 suggesting that transcriptional stress is not merely a function of high gene expression but is
389 modulated by SE activity. The enrichment of highly expressed SE-regulated genes among
390 transcription stress sites was over 30-fold higher than expected by chance, far exceeding that
391 of highly expressed TE-regulated genes or highly expressed genes lacking enhancer regulation.
392 This specificity suggests that SEs impose a unique transcriptional burden that predisposes their

393 target genes to DNA damage. The functional implications are significant, as many SE-regulated
394 genes encode critical oncogenes, including *CCND1*, *NEAT1*, *miR21*, and *MYC*, which play
395 essential roles in tumorigenesis.

396 The pronounced enrichment of DSBs at SE-regulated oncogenes was significantly greater in
397 cancer cell lines compared to their normal counterparts, indicating that SEs primarily drive
398 genomic instability in cancer cells. SEs have previously been recognized as hyper-mutated
399 regions that can alter the expression of target genes and contribute to carcinogenesis[6]. Our
400 findings enhance this understanding by demonstrating that SEs also dictate mutational
401 susceptibility at their target genes by inducing hyper-transcription.
402 Although sBLISS captures a snapshot of the DSB landscape at the time of fixation, we could
403 estimate the repair and turnover of DSBs across genes, revealing how transcription stress sites
404 particularly undergo high DSB turnover and repair efficiency. To accommodate this, these
405 regions experience constant activation of the DNA damage response (DDR) and H2AX
406 phosphorylation. Interestingly, not all DSB-enriched genes exhibit such efficient repair signaling,
407 suggesting two distinct types of DSB-bearing genes in the genome: genes that undergo frequent
408 breakage-repair cycles and are actively recognized by the repair machinery, and genes that,
409 despite being DSB-rich, experience infrequent breakage-repair cycles and remain persistent.
410 These persistent DSBs, found in regions with low expression, have a lower demand for repair.
411 Importantly, these observations pertain to physiological, endogenous DSBs under steady-state
412 conditions. In contrast, previous studies of induced DSBs, generated in large numbers by
413 exogenous agents, have reported different dynamics[40,67], likely reflecting a response to
414 unprogrammed and highly abundant damage. We speculate that these latter genes experience
415 breakage in a more programmed manner compared to genes with high turnover, which break as
416 a byproduct of transcription. The nature of these DSBs requires further investigation.
417 Additionally, we demonstrate that genes with efficient DSB repair signaling are more vulnerable
418 to mutations than genes with comparable DSB levels but less effective repair signaling,
419 emphasizing the importance of the repair process itself in mutagenesis at transcription stress
420 sites.
421

References

1. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Preprint, https://doi.org/10.1016/j.cell.2011.02.013 https://doi.org/10.1016/j.cell.2011.02.013.
2. Bradner, J.E., Hnisz, D., and Young, R.A. (2017). Transcriptional Addiction in Cancer. Preprint at Cell Press, https://doi.org/10.1016/j.cell.2016.12.013 https://doi.org/10.1016/j.cell.2016.12.013.
3. Lee, T.I., and Young, R.A. (2013). Transcriptional regulation and its misregulation in disease. Preprint at Elsevier B.V., https://doi.org/10.1016/j.cell.2013.02.014 https://doi.org/10.1016/j.cell.2013.02.014.
4. Bowry, A., Kelly, R.D.W., and Petermann, E. (2021). Hypertranscription and replication stress in cancer. Preprint at Cell Press, https://doi.org/10.1016/j.trecan.2021.04.006 https://doi.org/10.1016/j.trecan.2021.04.006.
5. Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Preprint at Nature Publishing Group, https://doi.org/10.1038/ng.3167 https://doi.org/10.1038/ng.3167.
6. Bal, E., Kumar, R., Hadigol, M., Holmes, A.B., Hilton, L.K., Loh, J.W., Dreval, K., Wong, J.C.H., Vlasevska, S., Corinaldesi, C., et al. (2022). Super-enhancer hypermutation alters oncogene expression in B cell lymphoma. Nature *607*, 808–815. https://doi.org/10.1038/s41586-022-04906-8.
7. Wang, X., Cairns, M.J., and Yan, J. (2019). Super-enhancers in transcriptional regulation and genome organization. Preprint at Oxford University Press, https://doi.org/10.1093/nar/gkz1038 https://doi.org/10.1093/nar/gkz1038.
8. Wang, M.T., Chen, Q.Y., Wang, S.J., Xie, H., Liu, J., Huang, R.X., Xiang, Y.F., Jiang, Y.Y., Tian, D.S., and Bian, E.B. (2023). Super-enhancers complexes zoom in transcription in cancer. Preprint at BioMed Central Ltd, https://doi.org/10.1186/s13046-023-02763-5 https://doi.org/10.1186/s13046-023-02763-5.
9. Bacabac, M., and Xu, W. (2023). Oncogenic super-enhancers in cancer: mechanisms and therapeutic targets. Preprint at Springer, https://doi.org/10.1007/s10555-023-10103-4 https://doi.org/10.1007/s10555-023-10103-4.
10. Hnisz, D., Schuijers, J., Lin, C.Y., Weintraub, A.S., Abraham, B.J., Lee, T.I., Bradner, J.E., and Young, R.A. (2015). Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Mol Cell *58*, 362–370. https://doi.org/10.1016/j.molcel.2015.02.014.
11. Morton, A.R., Dogan-Artun, N., Faber, Z.J., MacLeod, G., Bartels, C.F., Piazza, M.S., Allan, K.C., Mack, S.C., Wang, X., Gimple, R.C., et al. (2019). Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. Cell *179*, 1330-1341.e13. https://doi.org/10.1016/j.cell.2019.10.039.
12. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). XSuper-enhancers in the control of cell identity and disease. Cell *155*, 934. https://doi.org/10.1016/j.cell.2013.09.053.
13. Hidmi, O., Oster, S., Monin, J., and Aqeilan, R.I. (2024). TOP1 and R-loops facilitate transcriptional DSBs at hypertranscribed cancer driver genes. iScience *27*, 109082. https://doi.org/10.1016/j.isci.2024.109082.
14. Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E.Y., Tannahill, D., and Balasubramanian, S. (2016). DSBCapture: In situ capture and sequencing of DNA breaks. Nat Methods *13*, 855–857. https://doi.org/10.1038/nmeth.3960.

15. Oster, S., and Aqeilan, R.I. (2020). Programmed DNA Damage and Physiological DSBs: Mapping, Biological Significance and Perturbations in Disease States. Preprint at NLM (Medline), https://doi.org/10.3390/cells9081870 https://doi.org/10.3390/cells9081870.
16. Girasol, M.J., Krasilnikova, M., Marques, C.A., Damasceno, J.D., Lapsley, C., Lemgruber, L., Burchmore, R., Beraldi, D., Carruthers, R., Briggs, E.M., et al. (2023). RAD51-mediated R-loop formation acts to repair transcription-associated DNA breaks driving antigenic variation in Trypanosoma brucei. Proc Natl Acad Sci U S A *120*. https://doi.org/10.1073/pnas.2309306120.
17. Stork, C.T., Bocek, M., Crossley, M.P., Sollier, J., Sanz, L.A., Dé Ric Ché Din, F., Swigut, T., and Cimprich, K.A. (2016). Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. https://doi.org/10.7554/eLife.17548.001.
18. Sollier, J., Stork, C.T., García-Rubio, M.L., Paulsen, R.D., Aguilera, A., and Cimprich, K.A. (2014). Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability. Mol Cell *56*, 777–785. https://doi.org/10.1016/j.molcel.2014.10.020.
19. Halazonetis, T.D., Gorgoulis, V.G., and Bartek, J. An Oncogene-Induced DNA Damage Model for Cancer Development.
20. Cristini, A., Ricci, G., Britton, S., Salimbeni, S., Huang, S. yin N., Marinello, J., Calsou, P., Pommier, Y., Favre, G., Capranico, G., et al. (2019). Dual Processing of R-Loops and Topoisomerase I Induces Transcription-Dependent DNA Double-Strand Breaks. Cell Rep *28*, 3167-3181.e6. https://doi.org/10.1016/j.celrep.2019.08.041.
21. Milano, L., Gautam, A., and Caldecott, K.W. (2024). DNA damage and transcription stress. Preprint at Cell Press, https://doi.org/10.1016/j.molcel.2023.11.014 https://doi.org/10.1016/j.molcel.2023.11.014.
22. Min, J., Zhao, J., Zagelbaum, J., Lee, J., Takahashi, S., Cummings, P., Schooley, A., Dekker, J., Gottesman, M.E., Rabadan, R., et al. (2023). Mechanisms of insertions at a DNA double-strand break. Mol Cell *83*, 2434-2448.e7. https://doi.org/10.1016/j.molcel.2023.06.016.
23. Reijns, M.A.M., Parry, D.A., Williams, T.C., Nadeu, F., Hindshaw, R.L., Rios Szwed, D.O., Nicholson, M.D., Carroll, P., Boyle, S., Royo, R., et al. (2022). Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. Nature *602*, 623–631. https://doi.org/10.1038/s41586-022-04403-y.
24. Nakazawa, Y., Hara, Y., Oka, Y., Komine, O., van den Heuvel, D., Guo, C., Daigaku, Y., Isono, M., He, Y., Shimada, M., et al. (2020). Ubiquitination of DNA Damage-Stalled RNAPII Promotes Transcription-Coupled Repair. Cell *180*, 1228-1244.e24. https://doi.org/10.1016/j.cell.2020.02.010.
25. Gyenis, A., Chang, J., Demmers, J.J.P.G., Bruens, S.T., Barnhoorn, S., Brandt, R.M.C., Baar, M.P., Raseta, M., Derks, K.W.J., Hoeijmakers, J.H.J., et al. (2023). Genome-wide RNA polymerase stalling shapes the transcriptome during aging. Nat Genet *55*, 268–279. https://doi.org/10.1038/s41588-022-01279-6.
26. Saponaro, M., Kantidakis, T., Mitter, R., Kelly, G.P., Heron, M., Williams, H., Söding, J., Stewart, A., and Svejstrup, J.Q. (2014). RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. Cell *157*, 1037–1049. https://doi.org/10.1016/j.cell.2014.03.048.
27. Min, S., Ji, J.H., Heo, Y., and Cho, H. (2022). Transcriptional regulation and chromatin dynamics at DNA double-strand breaks. Preprint at Springer Nature,

514 https://doi.org/10.1038/s12276-022-00862-5 https://doi.org/10.1038/s12276-022-00862-
515 5.
516 28. Machour, F.E., and Ayoub, N. (2020). Transcriptional Regulation at DSBs: Mechanisms
517 and Consequences. Preprint at Elsevier Ltd, https://doi.org/10.1016/j.tig.2020.01.001
518 https://doi.org/10.1016/j.tig.2020.01.001.
519 29. Crossley, M.P., Bocek, M., and Cimprich, K.A. (2019). R-Loops as Cellular Regulators
520 and Genomic Threats. Preprint at Cell Press,
521 https://doi.org/10.1016/j.molcel.2019.01.024
522 https://doi.org/10.1016/j.molcel.2019.01.024.
523 30. Cristini, A., Park, J.H., Capranico, G., Legube, G., Favre, G., and Sordet, O. (2016).
524 DNA-PK triggers histone ubiquitination and signaling in response to DNA double-strand
525 breaks produced during the repair of transcription-blocking topoisomerase I lesions.
526 Nucleic Acids Res *44*, 1161–1178. https://doi.org/10.1093/nar/gkv1196.
527 31. Wang, Y., Song, C., Zhao, J., Zhang, Y., Zhao, X., Feng, C., Zhang, G., Zhu, J., Wang,
528 F., Qian, F., et al. (2023). SEdb 2.0: a comprehensive super-enhancer database of
529 human and mouse. Nucleic Acids Res *51*, D280–D290.
530 https://doi.org/10.1093/nar/gkac968.
531 32. Shu, S., Wu, H.J., Ge, J.Y., Zeid, R., Harris, I.S., Jovanović, B., Murphy, K., Wang, B.,
532 Qiu, X., Endress, J.E., et al. (2020). Synthetic Lethal and Resistance Interactions with
533 BET Bromodomain Inhibitors in Triple-Negative Breast Cancer. Mol Cell *78*, 1096-
534 1113.e8. https://doi.org/10.1016/j.molcel.2020.04.027.
535 33. Winter, G.E., Mayer, A., Buckley, D.L., Erb, M.A., Roderick, J.E., Vittori, S., Reyes, J.M.,
536 di Iulio, J., Souza, A., Ott, C.J., et al. (2017). BET Bromodomain Proteins Function as
537 Master Transcription Elongation Factors Independent of CDK9 Recruitment. Mol Cell *67*,
538 5-18.e19. https://doi.org/10.1016/j.molcel.2017.06.004.
539 34. Zheng, Z.Z., Xia, L., Hu, G.S., Liu, J.Y., Hu, Y.H., Chen, Y.J., Peng, J.Y., Zhang, W.J.,
540 and Liu, W. (2022). Super-enhancer-controlled positive feedback loop BRD4/ERα-RET-
541 ERα promotes ERα-positive breast cancer. Nucleic Acids Res *50*, 10230–10248.
542 https://doi.org/10.1093/nar/gkac778.
543 35. Geraud, M., Cristini, A., Salimbeni, S., Bery, N., Jouffret, V., Russo, M., Ajello, A.C.,
544 Fernandez Martinez, L., Marinello, J., Cordelier, P., et al. (2024). TDP1 mutation causing
545 SCAN1 neurodegenerative syndrome hampers the repair of transcriptional DNA double-
546 strand breaks. Cell Rep *43*. https://doi.org/10.1016/j.celrep.2024.114214.
547 36. Bunch, H., Lawney, B.P., Lin, Y.F., Asaithamby, A., Murshid, A., Wang, Y.E., Chen,
548 B.P.C., and Calderwood, S.K. (2015). Transcriptional elongation requires DNA break-
549 induced signalling. Nat Commun *6*. https://doi.org/10.1038/ncomms10191.
550 37. Madabhushi, R., Gao, F., Pfenning, A.R., Pan, L., Yamakawa, S., Seo, J., Rueda, R.,
551 Phan, T.X., Yamakawa, H., Pao, P.C., et al. (2015). Activity-Induced DNA Breaks
552 Govern the Expression of Neuronal Early-Response Genes. Cell *161*, 1592–1605.
553 https://doi.org/10.1016/j.cell.2015.05.032.
554 38. Carroll, J.S., Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky,
555 A.S., Keeton, E.K., Fertuck, K.C., Hall, G.F., et al. (2006). Genome-wide analysis of
556 estrogen receptor binding sites. Nat Genet *38*, 1289–1297.
557 https://doi.org/10.1038/ng1901.
558 39. γH2AX and cancer.

559 40. Kim, J.A., Kruhlak, M., Dotiwala, F., Nussenzweig, A., and Haber, J.E. (2007).
560 Heterochromatin is refractory to γ-H2AX modification in yeast and mammals. Journal of
561 Cell Biology *178*, 209–218. https://doi.org/10.1083/jcb.200612031.
562 41. Bouwman, B.A.M., Agostini, F., Garnerone, S., Petrosino, G., Gothe, H.J., Sayols, S.,
563 Moor, A.E., Itzkovitz, S., Bienko, M., Roukos, V., et al. (2020). Genome-wide detection of
564 DNA double-strand breaks by in-suspension BLISS. Nat Protoc *15*, 3894–3941.
565 https://doi.org/10.1038/s41596-020-0397-2.
566 42. Yan, W.X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M.W., Kallas, T.,
567 Custodio, J., Wernersson, E., Li, Y., Gao, L., et al. (2017). BLISS is a versatile and
568 quantitative method for genome-wide profiling of DNA double-strand breaks. Nat
569 Commun *8*. https://doi.org/10.1038/ncomms15058.
570 43. Canela, A., Sridharan, S., Sciascia, N., Tubbs, A., Meltzer, P., Sleckman, B.P., and
571 Nussenzweig, A. (2016). DNA Breaks and End Resection Measured Genome-wide by
572 End Sequencing. Mol Cell *63*, 898–911. https://doi.org/10.1016/j.molcel.2016.06.034.
573 44. Dobbs, F.M., van Eijk, P., Fellows, M.D., Loiacono, L., Nitsch, R., and Reed, S.H.
574 (2022). Precision digital mapping of endogenous and induced genomic DNA breaks by
575 INDUCE-seq. Nat Commun *13*. https://doi.org/10.1038/s41467-022-31702-9.
576 45. Wei, P.C., Chang, A.N., Kao, J., Du, Z., Meyers, R.M., Alt, F.W., and Schwer, B. (2016).
577 Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor
578 Cells. Cell *164*, 644–655. https://doi.org/10.1016/j.cell.2015.12.039.
579 46. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count
580 data. Genome Biol *11*. https://doi.org/10.1186/gb-2010-11-10-r106.
581 47. Arnould, C., Rocher, V., Saur, F., Bader, A.S., Muzzopappa, F., Collins, S., Lesage, E.,
582 Le Bozec, B., Puget, N., Clouaire, T., et al. (2023). Chromatin compartmentalization
583 regulates the response to DNA damage. Nature *623*, 183–192.
584 https://doi.org/10.1038/s41586-023-06635-y.
585 48. Aymard, F., Aguirrebengoa, M., Guillou, E., Javierre, B.M., Bugler, B., Arnould, C.,
586 Rocher, V., Iacovoni, J.S., Biernacka, A., Skrzypczak, M., et al. (2017). Genome-wide
587 mapping of long-range contacts unveils clustering of DNA double-strand breaks at
588 damaged active genes. Nat Struct Mol Biol *24*, 353–361.
589 https://doi.org/10.1038/nsmb.3387.
590 49. Barutcu, A.R., Lajoie, B.R., McCord, R.P., Tye, C.E., Hong, D., Messier, T.L., Browne,
591 G., van Wijnen, A.J., Lian, J.B., Stein, J.L., et al. (2015). Chromatin interaction analysis
592 reveals changes in small chromosome and telomere clustering between epithelial and
593 breast cancer cells. Genome Biol *16*. https://doi.org/10.1186/s13059-015-0768-0.
594 50. Sondka, Z., Dhir, N.B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey,
595 M., Ward, S., Wilding, J., Ahmed, M., et al. (2024). COSMIC: a curated database of
596 somatic variants and clinical data for cancer. Nucleic Acids Res *52*, D1210–D1217.
597 https://doi.org/10.1093/nar/gkad986.
598 51. Burrell, R.A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and
599 consequences of genetic heterogeneity in cancer evolution. Preprint,
600 https://doi.org/10.1038/nature12625 https://doi.org/10.1038/nature12625.
601 52. Yates, L.R., and Campbell, P.J. (2012). Evolution of the cancer genome. Preprint,
602 https://doi.org/10.1038/nrg3317 https://doi.org/10.1038/nrg3317.

603 53. Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability an
604 evolving hallmark of cancer. Preprint, https://doi.org/10.1038/nrm2858
605 https://doi.org/10.1038/nrm2858.
606 54. Hills, S.A., and Diffley, J.F.X. (2014). DNA replication and oncogene-induced replicative
607 stress. Preprint at Cell Press, https://doi.org/10.1016/j.cub.2014.04.012
608 https://doi.org/10.1016/j.cub.2014.04.012.
609 55. Kim, J., Sturgill, D., Sebastian, R., Khurana, S., Tran, A.D., Edwards, G.B., Kruswick, A.,
610 Burkett, S., Hosogane, E.K., Hannon, W.W., et al. (2018). Replication Stress Shapes a
611 Protective Chromatin Environment across Fragile Genomic Regions. Mol Cell *69*, 36-
612 47.e7. https://doi.org/10.1016/j.molcel.2017.11.021.
613 56. Gaillard, H., García-Muse, T., and Aguilera, A. (2015). Replication stress and cancer.
614 Preprint at Nature Publishing Group, https://doi.org/10.1038/nrc3916
615 https://doi.org/10.1038/nrc3916.
616 57. Zeman, M.K., and Cimprich, K.A. (2014). R E V I E W Causes and consequences of
617 replication stress.
618 58. Machour, F.E., R. Abu-Zhayia, E., Kamar, J., Barisaac, A.S., Simon, I., and Ayoub, N.
619 (2024). Harnessing DNA replication stress to target RBM10 deficiency in lung
620 adenocarcinoma. Nat Commun *15*. https://doi.org/10.1038/s41467-024-50882-0.
621 59. Cybulla, E., and Vindigni, A. (2023). Leveraging the replication stress response to
622 optimize cancer therapy. Preprint at Nature Research, https://doi.org/10.1038/s41568-
623 022-00518-6 https://doi.org/10.1038/s41568-022-00518-6.
624 60. Corazzi, L., Ionasz, V.S., Andrejev, S., Wang, L.C., Vouzas, A., Giaisi, M., Di Muzio, G.,
625 Ding, B., Marx, A.J.M., Henkenjohann, J., et al. (2024). Linear interaction between
626 replication and transcription shapes DNA break dynamics at recurrent DNA break
627 Clusters. Nat Commun *15*. https://doi.org/10.1038/s41467-024-47934-w.
628 61. Hamperl, S., and Cimprich, K.A. (2016). Conflict Resolution in the Genome: How
629 Transcription and Replication Make It Work. Preprint at Cell Press,
630 https://doi.org/10.1016/j.cell.2016.09.053 https://doi.org/10.1016/j.cell.2016.09.053.
631 62. Hamperl, S., Bocek, M.J., Saldivar, J.C., Swigut, T., and Cimprich, K.A. (2017).
632 Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates
633 Distinct DNA Damage Responses. Cell *170*, 774-786.e19.
634 https://doi.org/10.1016/j.cell.2017.07.043.
635 63. Geraud, M., Cristini, A., Salimbeni, S., Bery, N., Jouffret, V., Russo, M., Ajello, A.C.,
636 Fernandez Martinez, L., Marinello, J., Cordelier, P., et al. (2024). TDP1 mutation causing
637 SCAN1 neurodegenerative syndrome hampers the repair of transcriptional DNA double-
638 strand breaks. Cell Rep *43*. https://doi.org/10.1016/j.celrep.2024.114214.
639 64. Zhou, R.W., and Parsons, R.E. (2023). Etiology of super-enhancer reprogramming and
640 activation in cancer. Preprint at BioMed Central Ltd, https://doi.org/10.1186/s13072-023-
641 00502-w https://doi.org/10.1186/s13072-023-00502-w.
642 65. Huang, H., Hu, J., Maryam, A., Huang, Q., Zhang, Y., Ramakrishnan, S., Li, J., Ma, H.,
643 Ma, V.W.S., Cheuk, W., et al. (2021). Defining super-enhancer landscape in triple-
644 negative breast cancer by multiomic profiling. Nat Commun *12*.
645 https://doi.org/10.1038/s41467-021-22445-0.
646 66. Mosler, T., Conte, F., Longo, G.M.C., Mikicic, I., Kreim, N., Möckel, M.M., Petrosino, G.,
647 Flach, J., Barau, J., Luke, B., et al. (2021). R-loop proximity proteomics identifies a role

of DDX41 in transcription-associated genomic instability. Nat Commun *12*. https://doi.org/10.1038/s41467-021-27530-y.

67. Jakob, B., Splinter, J., Conrad, S., Voss, K.O., Zink, D., Durante, M., Löbrich, M., and Taucher-Scholz, G. (2011). DNA double-strand breaks in heterochromatin elicit fast repair protein recruitment, histone H2AX phosphorylation and relocation to euchromatin. Nucleic Acids Res *39*, 6489–6499. https://doi.org/10.1093/nar/gkr230.

68. Suehnholz, S.P., Nissan, M.H., Zhang, H., Kundra, R., Nandakumar, S., Lu, C., Carrero, S., Dhaneshwar, A., Fernandez, N., Xu, B.W., et al. (2024). Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. Cancer Discov *14*, 49–65. https://doi.org/10.1158/2159-8290.CD-23-0467.

69. Hidmi, O., Oster, S., Shatleh, D., Monin, J., and Aqeilan, R.I. (2024). Protocol for mapping physiological DSBs using in-suspension break labeling in situ and sequencing. STAR Protoc *5*. https://doi.org/10.1016/j.xpro.2024.103059.

70. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

71. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res *48*, D882–D889. https://doi.org/10.1093/nar/gkz1062.

72. Greer, Y.E., Hernandez, L., Fennell, E.M.J., Kundu, M., Voeller, D., Chari, R., Gilbert, S.F., Gilbert, T.S.K., Ratnayake, S., Tang, B., et al. (2022). Mitochondrial Matrix Protease ClpP Agonists Inhibit Cancer Stem Cell Function in Breast Cancer Cells by Disrupting Mitochondrial Homeostasis. Cancer Research Communications *2*, 1144–1161. https://doi.org/10.1158/2767-9764.CRC-22-0142.

676 **Method details**

677 **Cell culture and experimental conditions**

678 MCF7 (HTB-22) and T-47D (HTB-133) cells were grown in RPMI supplemented with 10%
679 (vol/vol) fetal bovine serum FBS (GIBCO), glutamine, and penicillin/streptomycin. HEK-293
680 (CRL-1573), MDA-MB-468 (HTB-132), MDA-MB-231 (HTB-26), and HeLa cells (CCL-2) were
681 grown in DMEM supplemented with 10% (vol/vol) FBS, glutamine, and penicillin/streptomycin.
682 A-4098 cells (HTB-44) were grown in. EMEM supplemented with 10% FBS, glutamine, and
683 penicillin/streptomycin. SH-SY5Y (CRL-2266) were grown in DMEM/F12 supplemented with
684 10% FBS, glutamine, and penicillin/streptomycin. HMLE cells were grown in Promocell
685 mammary epithelial cell basal media (C-21010) with added supplements (c-93110) whereas
686 MCF10A were grown on DMEM/F12 supplemented with 5% Horse serum, 20 ng/ml EGF, 0.5
687 mg/ml Hydrocortisone, 100 ng/ml Cholera toxin, 10 mg/ml Insulin and Pen/Strep. MCF-7 cells
688 were synchronized in the G1 cell cycle stage by culturing in serum-free media for 72 hours
689 followed by replacing the media with 5% FBS media for 5 hours. Cells were grown at 37°C under
690 a humidified atmosphere with 5% CO_2. Cells were routinely authenticated by STR profiling, and
691 tested for mycoplasma, and cell aliquots from early passages were used.

692 Cells were treated with DMSO (Sigma-Aldrich), dBET6 (Sigma-Aldrich, SML2638), or CDK9i
693 (NVP-2) (MCE, Cat# Y-12214A) 100 nM for 4 hours. Cells were then treated with DMSO or ATMi
694 (20 uM, KU-55933, Sigma) for 5 hours.

695

696 **Transcription stress score**

697 Our sBLISS, DRIP-seq, TOP1cc, and TOP1 ChIP-seq data used for transcription stress score
698 calculation are already published[13] and are available at GEO: GSE241309. For each dataset,
699 the signal density per gene was first converted to an empirical p-value based on its rank within
700 the distribution of all genes. Higher signal values correspond to lower p-values. These p-values
701 were then transformed into Z-scores using the standard normal inverse cumulative distribution
702 function:

703 $$Z_k = \Phi^{-1}(1 - p_k)$$

704 where p_k is the empirical p-value for dataset k, and Φ^{-1} denotes the inverse of the standard
705 normal cumulative distribution function.

706 The individual Z-scores from all four datasets were combined per gene using the Liptak
707 weighted Z-test (unweighted version):

708 $$TSS_{raw} = \frac{Z_1 + Z_2 + Z_3 + Z_4}{\sqrt{4}}$$

709 For better representation, the combined score was further min-max normalized and then
710 stretched by a power transformation to emphasize differences:
711
712 $$TSS_{final} = TSS_{scaled}^4$$

713 Genes not among the top 1000 in one or more parameters were excluded.
714

715 **Identification of Transcription Stress-Associated Domains using Multivariate HMM**

716 To identify genomic regions simultaneously enriched for markers of transcription stress, we
717 applied a multivariate hidden Markov model (HMM) using the *mhsmm* R package. The hg38
718 genome was divided into 1 kb non-overlapping bins, and mean signals for TOP1, TOP1cc, R-
719 loops, and DSBs were calculated per bin. Signal intensities were capped at the 0.5th and 99.5th
720 percentiles to minimize the influence of outliers. A two-state multivariate HMM was initialized
721 with uniform state priors and a transition matrix favouring self-transition (0.95 probability).
722 Emissions were modelled using multivariate Gaussian distributions with means offset by ±0.5
723 standard deviations and shared covariance. The model was trained by maximum likelihood
724 estimation, and Viterbi decoding was used to assign each bin to a state. The state with the
725 highest average signal across all datasets was designated as the "co-marked" state. To refine
726 the identified regions, we selected bins showing high signal (above the 95th percentile) in at
727 least three of the four datasets, merged adjacent bins separated by ≤1 kb, and filtered to retain
728 regions ≥3 kb. [13]
729

730 **Gene sets compilation and enrichment analysis**

731 Gene sets tested for enrichment in this manuscript were defined as follows: SE-regulated genes
732 for each cell line were downloaded from SEdb2.0[31]. A list of annotated oncogenes was
733 downloaded from OncoKB[68]. Highly expressed genes are defined as the top 1000 expressed
734 genes. TE-regulated genes are defined as genes that overlap enhancer regions and are not

735 annotated by an SE. Non-E genes are genes that do not overlap any enhancer element.
736 enrichment was calculated by dividing the number of observed overlapping genes by the
737 theoretical number of genes predicted under a random model. One-tailed (upper-tailed for values
738 > 1, and lower-tailed for values < 1) hypergeometric test using the phyper function in R to
739 determine the statistical significance of the observed enrichment.
740
741 **In-suspension break labelling *in situ* and sequencing (sBLISS)**
742 sBLISS was conducted as previously described[41,69]. In summary, 10^6 cells were fixed in 2%
743 paraformaldehyde in 10% FCS/PBS for 10 minutes at room temperature. The fixation was
744 quenched with 125 mM glycine for 5 minutes at room temperature, followed by another 5 minutes
745 on ice and two washes in ice-cold PBS. Cells were lysed for 60 minutes on ice, and their nuclei
746 were permeabilized for 60 minutes at 37°C. Next, nuclei were rinsed twice with CutSmart Buffer
747 containing 0.1% Triton X-100 (CS/TX100), and double-strand break (DSB) ends were blunted in
748 situ using NEB's Quick Blunting Kit for 60 minutes at room temperature. The blunted nuclei were
749 then washed twice with 1x CS/TX100 before in situ ligation of the sBLISS adapters to the DSB
750 ends. Adaptor ligation was carried out with T4 DNA Ligase for 20-24 hours at 16°C, with BSA
751 and ATP added. Following ligation, the nuclei underwent two washes with 1x CS/TX100, and
752 genomic DNA was extracted using Proteinase K at 55°C for 14-18 hours while shaking at 800
753 rpm. Proteinase K was then heat-inactivated for 10 minutes at 95°C, followed by extraction using
754 Phenol:Chloroform:Isoamyl Alcohol, Chloroform, and ethanol precipitation. The purified DNA
755 was sonicated in 100 µL of ultra-pure water using Covaris M220 for 60 seconds. Sonicated
756 samples were concentrated with AMPure XP beads (Beckman Coulter), and fragment sizes
757 were evaluated using a BioAnalyzer 2100 (Agilent Technologies), targeting a range of 300 bp to
758 800 bp with a peak around 400-600 bp. The sonicated DNA was then in vitro transcribed using
759 the MEGAscript T7 Kit for 14 hours at 37°C. After RNA purification and ligation of the 3'-Illumina
760 adaptors, the RNA underwent reverse transcription. The final library indexing and amplification
761 step was performed with NEBNext® Ultra™ II Q5® Master Mix.
762 sBLISS fastq files are initially de-multiplexed using sample barcodes. Quality control is
763 performed with trim galore to eliminate residual adapters, trim reads to a base quality of at least
764 20, and filter out short reads smaller than 20 bp. Initial and final sample qualities are assessed
765 with fastqc. Quality-processed fastq files are aligned to the GRCh38 assembly with hisat2, then

766 sorted and indexed using samtools. The resulting bam files are de-duplicated with umi-tools,
767 utilizing genomic coordinates and Unique Molecular Identifiers (UMI). Custom Python and R
768 scripts are employed to identify read start positions and convert bam files to bigwig format for
769 downstream analysis. An additional custom R script is used to discard a blacklist of positions,
770 primarily within centromeres.

771 **RNA-seq and GRO-seq data processing**

772 Bulk RNA-seq processed data for untreated MCF7, HeLa, and T-47D cell lines were obtained
773 from ENCODE[70,71], while data for MDA-MB-436 and MDA-MB-468 were downloaded from GEO
774 (GSE212143)[72]. For MCF7, the following datasets were used: ENCFF967AOT, ENCFF328CKZ,
775 ENCFF921PJP, ENCFF930TQB, ENCFF030IUO, ENCFF878CJF, ENCFF057ITQ,
776 ENCFF317XCR, ENCFF985KJE, and ENCFF456OYZ. For HeLa, we used ENCFF750TQI,
777 ENCFF622TLZ, ENCFF689BFU, ENCFF151QTZ, ENCFF262EJQ, ENCFF878CJF,
778 ENCFF964OAC, ENCFF346JDS, and ENCFF638MLA. For T-47D, we used ENCFF234TQB,
779 ENCFF412DJX, ENCFF475SHD, and ENCFF846PLI. TPM values were averaged across
780 replicates for each cell line to determine the expression rank of each gene.
781 RNA-seq data for MCF7 after CDK9i treatment was downloaded from GSE129012 (4h
782 timepoint). And GRO-seq data for MCF7 after E2 treatment was downloaded from GSE27463
783 (40 min timepoint).
784

785 **RT-qPCR and Spike-in RT-qPCR**

786 Total RNA was isolated utilizing TRI reagent (Biolab), following the manufacturer's guidelines for
787 the phenol/chloroform extraction technique. cDNA was synthesized from 1 µg of RNA using the
788 QScript cDNA synthesis kit (Quantabio). The SYBR Green PCR Master Mix (Applied
789 Biosystems) was employed for qRT-PCR. All assays were conducted in triplicate. Experiments
790 were performed to assess the levels of cDNA using primers targeting: Chicken RPL4 (FW,
791 GAGTGACTACAACCTGCCGA; REV, TTGGCGTATGGGTTCAGCTT), SIK1 (FW,
792 CTCCGGGTGGGTTTTTACGAC; REV, CTGCGTTTTGGTGACTCGATG), ID3 (FW,
793 GAGAGGCACTCAGCTTAGCC; REV, TCCTTTTGTCGTTGGAGATGAC), XBP1 (FW,
794 CCCTCCAGAACATCTCCCCAT; REV, ACATGACTGGGTCCAAGTTGT), CCND1 (FW,
795 GCTGCGAAGTGGAAACCATC; REV, CCTCCTTCTGCACACATTTGAA), HES1 (FW,
796 CCTGTCATCCCCGTCTACAC; REV, CACATGGAGTCCGCCGTAA), FOS (FW,

797 TCCCATCGGTCCACTAGGTTT; REV, AGGGCTGCACTGAGTTCTTTG), JUNB (FW,
798 ACGACTCATACACAGCTACGG; REV, GCTCGGTTTCAGGAGTTTGTAGT), TFF1 (FW,
799 CCCTCCCAGTGTGCAAATAAG; REV, GAACGGTGTCGTCGAAACAG), UBC (FW,
800 ATTTGGGTCGCGGTTCTTG; REV, TGCCTTGACATTCTCGATGGT), HPRT (FW,
801 TGACACTGGCAAAACAATGCA; REV, GGTCCTTTTCACCAGCAAGCT), MYC (FW,
802 ATGCCCCTCAACGTGAACTTC; REV, CGCAACATAGGATGGAGAGCA), XRCC4 (FW,
803 ATGTTGGTGAACTGAGAAAAGC; REV, GCAATGGTGTCCAAGCAATAAC), GREB1 (FW,
804 TGGTCCGTAATGCACAAGGG; REV, CTGCGTTTAGTGAGGGGTGA), GATA3 (FW,
805 GCCTCTGCTTCATGGATCCC; REV, CACACTCCCTGCCTTCTGTG), ACTB (FW,
806 TTTTGGCTATACCCTACTGGCA; REV, CTGCACAGTCGTCAGCATATC).
807 In RT-qPCR experiments, Ct values were normalized to HPRT levels. For spike-in RT-qPCR,
808 chicken RNA was added in proportion to the number of cells in each sample, and Ct values were
809 normalized to chicken RPL4.
810
811 **γH2AX ChIP-seq**
812 MCF7 cells (~10^6) were crosslinked with 1% formaldehyde (methanol free, Thermo Scientific
813 28906) for 10 min at room temperature and quenched with glycine, 125 mM final concentration.
814 Fixed cells were washed twice in PBS and incubated in sonication buffer (0.5M NaCl, 0.5%SDS
815 RIPA buffer containing PMSF, protease and phosphatase inhibitors) for 30min on ice. Cells were
816 sonicated using Covaris M220 for 10min to produce chromatin fragments of ~200-300 bp. The
817 sheared chromatin was centrifuged 10min at maximum speed. From the supernatant, 50ul were
818 saved as input DNA and the rest was diluted in RIPA buffer without NaCl and SDS. The
819 chromatin was immunoprecipitated by incubation with 5 µg of gamma-H2AX (phospho S139;
820 Abcam, AB2893) antibody pre mixed with 100ul Dynabeads Protein G (ThermoFisher, 10004D)
821 and incubated over-night at 4°C with rotation. Immunoprecipitates were washed twice with RIPA
822 buffer + 150mM NaCl, twice with RIPA buffer + 300 mM NaCl, and twice with TE buffer. The
823 chromatin was eluted from the beads with 200ul of direct elution buffer (10mM Tris pH 8, 0.3M
824 NaCl, 5mM EDTA, 0.5% SDS) and incubated overnight at 65°C to reverse the cross-linking.
825 Samples were treated with RNAse for 1h at 37 °C and with proteinase K at 55 °C for 2 h. DNA
826 was cleaned up by QIAquick PCR purification column (Qiagen), according to the manufacturer's

827 instructions, and eluted in 30 µl of elution buffer. The ChIPed and the Input DNA were used for
828 real-time PCR or to prepare libraries by Truseq (Ilumina) and sequenced in Nextseq (Illumina).
829 ChIPed and input fastq files are quality controlled with trim galore to remove residual adapters,
830 trim reads to base quality of at least 20 and filter out short reads with size smaller than 20 bp.
831 Initial and final sample qualities are evaluated with fastqc. The quality processed fastq files are
832 aligned to GRCh38 genome with hisat2 followed by sorting and indexing with samtools. The
833 utility bamCompare of deepTools is applied on the bam files to create a peak bigwig file. This
834 operation is set to ignore duplicates, discard a blacklist of regions, normalize by RPKM, and
835 compare by ratio in tiles of 50 bp. γH2AX ChIP-seq data for HeLa and IMR90 were downloaded
836 from GEO: GSM4564685 and GEO: GSM2067927, respectively.
837
838 **Statistical modelling and residual-based classification**
839 For multivariate logistic regression, genes were categorized based on whether they overlapped
840 transcription stress sites identified by Hidden Markov Modeling (HMM). Logistic regression was
841 performed using the glm function in R (family = binomial).
842 Multivariate Least Absolute Shrinkage and Selection Operator (LASSO) regression was
843 conducted using the glmnet package in R, with 10-fold cross-validation (cv.glmnet) to select the
844 optimal regularization parameter (λ). The model was trained to classify genes based on
845 transcription stress site overlap or DSB enrichment. Odds ratios were derived from the final
846 model coefficients to assess the relative predictive power of each feature.
847 Generalized additive model (GAM) was applied to model the relationship between each variable
848 and γH2AX across all genes in (Figure 5A), and for genes classified into their overlapping
849 chromatin states in (Figure 5H). The model was constructed using the gam function in
850 the mgcv package in R.
851 To dissect the effects of absolute DSBs from DSB turnover on mutation frequency, we modelled
852 the relationship between break density and transcription stress by nonlinear least squares (NLS)
853 regression using the following equation:
854

$$gH2AX\ density = a \times break\ density^{b}$$

855 The model was fitted using the nls function in R, with initial parameter estimates
856 for “a” and “b” set to 0.1. After fitting the model, residuals were extracted and used to classify

genes into three response categories based on predefined thresholds: genes with residuals > 0 were classified as 'Over-responsive,' genes with residuals < 0 as 'Under-responsive' and genes between -0.3 and 0.3 as 'Moderately responsive'.

Classifying individual DSB sites based on γH2AX

Each break site was assigned a γH2AX coverage level (γH2AX rank). To ensure a comparable distribution of break scores (i.e., the number of DSBs at the same site) between the two groups, break sites were classified into four categories: those with a break score of 1 (the large majority), those with a break score between 2 and 4, those with a break score between 5 and 10, and those with a break score between 11 and 50. Within each of the four categories, 1% of the top and bottom γH2AX-ranked breaks are selected. All categories are then re-combined so that the top set of breaks and the bottom set of breaks exhibit a similar break score distribution. Mapping to chromatin states is then computed and plotted for these sets.

Chromatin states

Chromatin states bed files were downloaded from ENCODE[70,71]. For MCF7, ENCFF631HCR file was used. For HeLa, ENCFF555HXM file was used.

Transient transfection

Transient transfection of siXRCC4 (Dharmacon; SMARTpool siRNA, Catalog ID: L-004494-00-0005) and siSc (Dharmacon; D-00181010) was carried out using Lipofectamine (ThermoFisher). On the day of transfection, cells were seeded to achieve a confluency of 60-70%. The transfection solution was prepared by mixing 1.5 ml of serum and antibiotic-free RPMI with 30 µL of Lipofectamine and incubating for 5 minutes before adding a mixture of 1.5 ml RPMI and 0.6 nmoles of siRNA, followed by a 20-minute incubation at room temperature. This mixture was then added to cells in a 10 cm plate cultured in antibiotic and serum-free media. The media was replaced with fully supplemented media after 5-6 hours, and the cells were incubated in the incubator for 48 hours.

HiC data analysis

886 Normalized Hi-C contact matrices for MCF7 and MCF10A cells were obtained from GEO
887 (GSE66733).
888 Genomic bins were intersected with all genes, and each bin was assigned to a single gene. Bins
889 overlapping multiple genes were excluded from the analysis.
890 In inter- and intra-group metrics calculation, genes were stratified into nine groups based on their
891 break density, excluding genes with zero breaks. Intra-group and inter-group interaction scores
892 were computed as follows: Intra-group interactions: For each gene, interaction scores with all
893 other genes within the same group were summed and normalized by the total number of genes
894 in that group. Inter-group interactions: For each gene, interaction scores with all genes outside
895 its group were summed and normalized by the total number of genes in all other groups. Genes,
896 along with their interaction scores, were ordered based on break density. The data was
897 visualized using loess smoothing.
898
899 **COSMIC mutations data analysis**
900 The file Cosmic_NonCodingVariants_Tsv_v101_GRCh37.tar, containing all noncoding variants,
901 was downloaded from COSMIC. This file was first filtered for breast cancer samples and
902 separated based on the type of mutation into SNVs, deletions, and insertions mutation files.
903 These files were then converted into genomic ranges objects and filtered to contain only
904 mutations within introns. To calculate mutation density, mutations per gene were counted and
905 normalized by the sum of intron length.
906
907 **Acknowledgments**
908 We sincerely thank the members of the Aqeilan lab for their valuable discussions and insights.
909 We are deeply grateful to Dr. Sheera Adar and Dr. Yotam Drier for their thoughtful guidance and
910 support as members of the PhD committee of O.H. We also appreciate the invaluable assistance
911 of Dr. Abed Nasereddin and Dr. Idit Shiff from the Core Research Facility of the Hebrew
912 University-Hadassah Medical School. Additionally, we extend our sincere gratitude to Prof.
913 Yinon Ben Neriah for generously providing dBET6 and CDK9i. This study was supported by
914 grants from the Israel Science Foundation (ISF) [No. 1056/21]. We also acknowledge the
915 support of the Carole and Andrew Harper Diversity Scholarship Program to O.H. and D.S,

Science Training Encouraging Peace (STEP) Fellowship to D.S. and S.O. and the VATAT PhD scholarship to O.H.

Author contributions

O.H. and R.I.A. conceived and designed the study. O.H. and J.M. performed computational analysis. S.O.F performed sBLISS on HMLE, MCF10A, MCF7, T47D, MDA-MB-436 and MDA-MB-468 cell lines. D.S. performed sBLISS on MCF7 cells synchronized in G1, SH-SY5Y cells, and NSCs. O.H. performed all other experiments. O.H. and R.I.A wrote the manuscript.

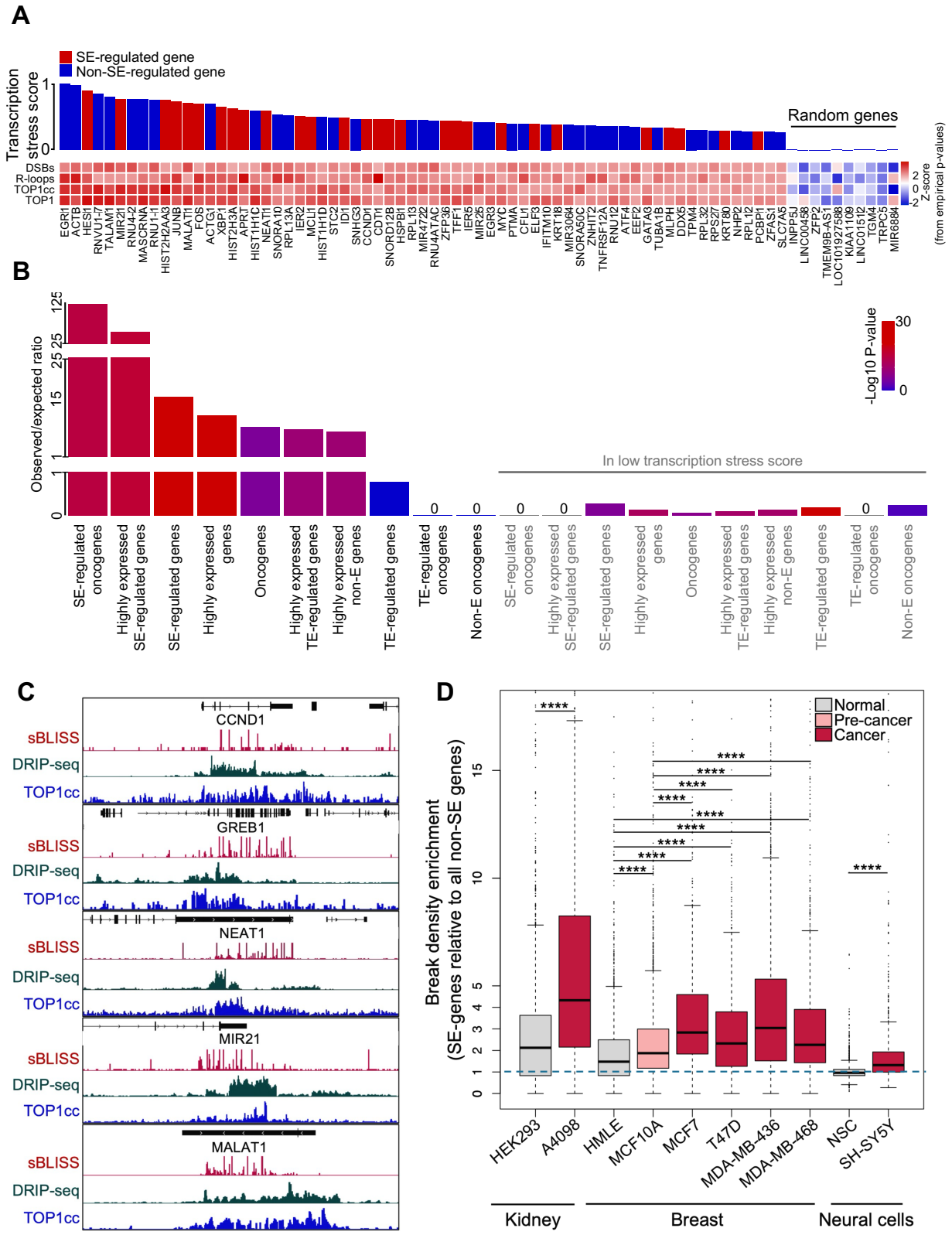


Figure 1: Super-enhancer-regulated oncogenes are major transcription stress sites.

931 **A,** Transcription stress sites ranked by their scores based on four transcription stress
932 parameters: TOP1, R-loops, TOP1cc, and DSBs in MCF7 cells. **B,** The ratio of observed to
933 expected numbers of genes in transcription stress sites. P-values were calculated using
934 hypergeometric testing (upper-tailed for enrichment in the top score and lower-tailed for the
935 bottom score). **C,** Genome browser snapshots depict DSBs, TOP1cc, and R-loops at selected
936 transcription stress sites. **D,** A boxplot illustrates DSBs enrichment at the SE-regulated genes
937 over other genes in each cell line (downloaded from SEdb$^{2.0}$ [31]). Enrichment was calculated by
938 dividing the break density of each SE-regulated gene by the median break density of non-SE-
939 regulated genes and the average values for replicates were plotted. In the boxplot, the center
940 line represents the median; the box limits indicate the upper and lower quartiles; the whiskers
941 denote 1.5X the interquartile range, while the dots represent outliers. P-values were computed
942 using the Mann-Whitney U test, ****$p < 0.0001$.

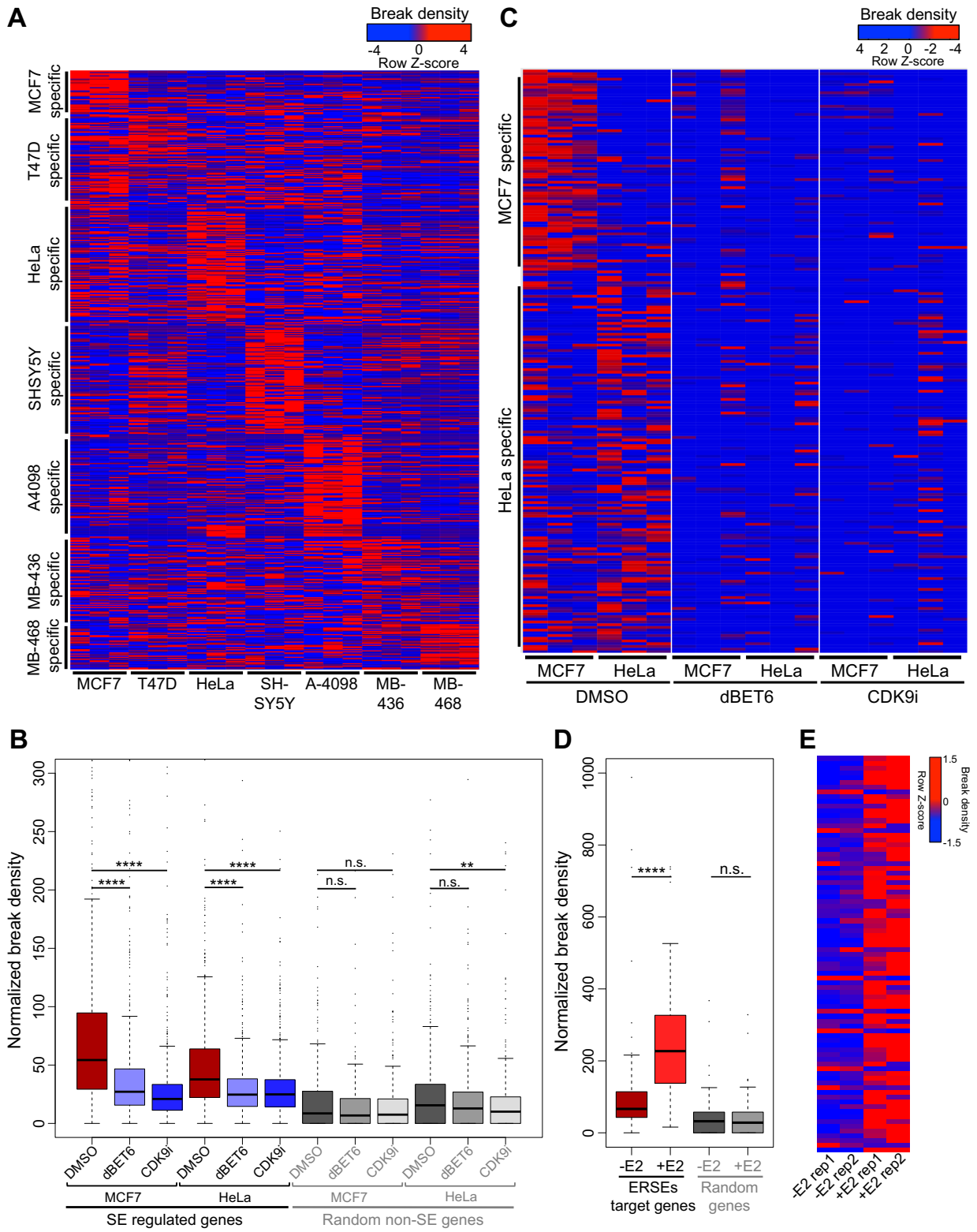


Figure 2: Super-enhancer landscape shapes transcription stress sites in cancer. **A**, Heatmap showing the normalized break density row Z-score for each cell line's unique SE-regulated genes. For each cell line, SE-regulated genes were identified, common SE-regulated

genes from other cell lines were excluded, and low to moderately expressed genes as well as genes with low break counts were filtered out. Each column corresponds to a biological replicate. Columns were normalized by median scaling. **B**, Boxplot illustrating the normalized break density of cells treated with DMSO (control), dBET6, or CDK9i (100nM for 4 hours each) for SE-regulated genes in each cell line or for randomly selected genes. values are averaged for three biological replicates. **C**, Heatmap depicting the normalized break density row Z-score for each cell line's unique SE-regulated genes following treatment with DMSO (control), dBET6, or CDK9i. Each column corresponds to a biological replicate. Each sample was normalized by downsizing to 270000 random breaks. **D**, Boxplot demonstrating the normalized break density of cells treated with vehicle (control) or estradiol (E2, 100nM for 1 hour) for significantly affected ($p < 0.05$) estrogen receptor-occupied SE (ERSE) target genes or for randomly selected genes. **E**, Heatmap representing the normalized break density row Z-score of significantly affected ERSE target genes. Boxplot features include: center line representing median; box limits denoting upper and lower quartiles; whiskers extending to 1.5x the interquartile range; and dots indicating outliers. P-values were computed using the Mann-Whitney U test, with n.s. indicating $p > 0.05$, ** indicating $p < 0.01$, and **** indicating $p < 0.0001$.

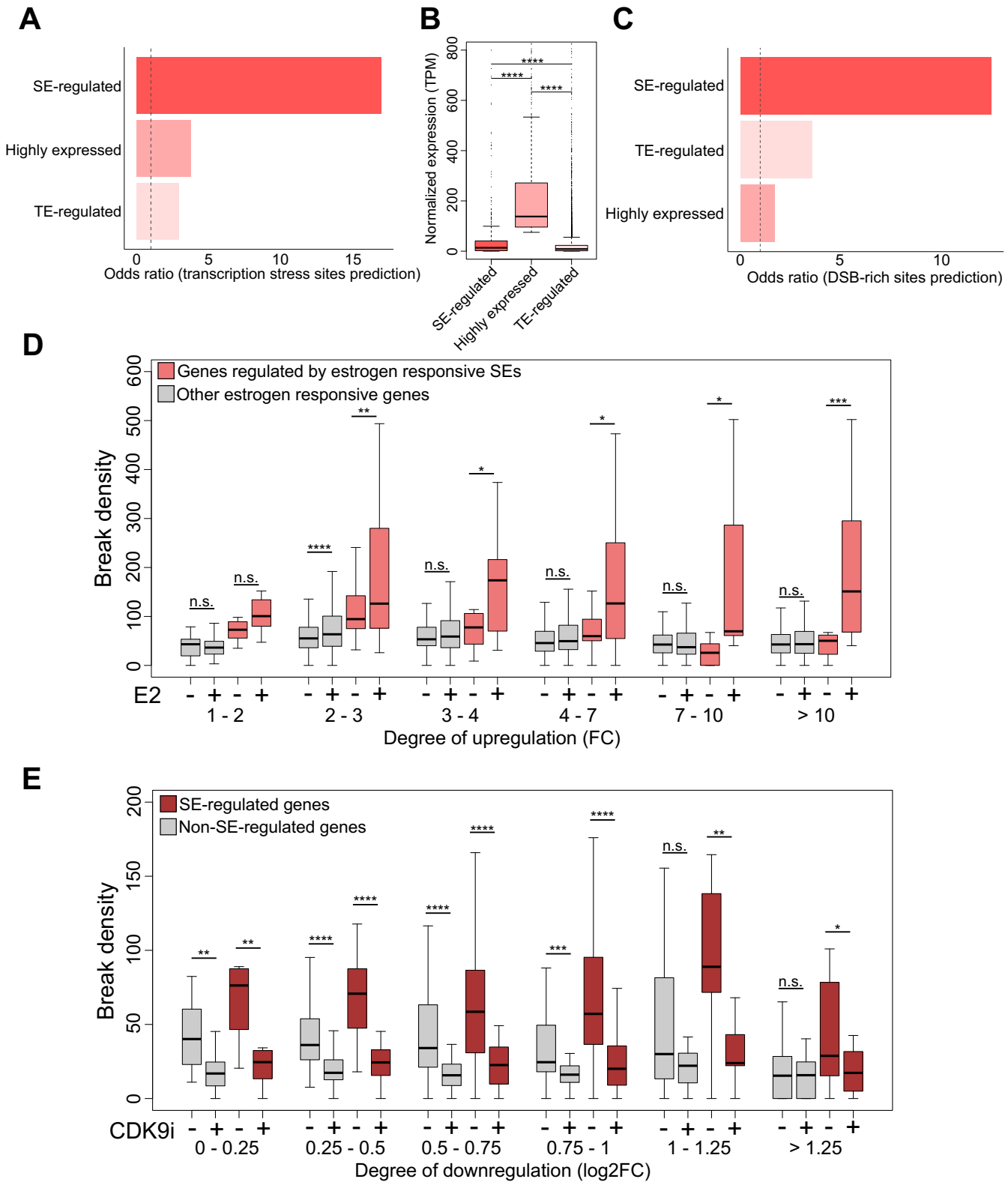


Figure 3. Specificity of transcription stress at super-enhancer (SE)–regulated genes. **A,** Multivariate LASSO regression identifying predictors of transcription stress site locations in 100kb genomic bins. Odds ratios are shown. **B,** Boxplot showing normalized expression (TPM)

967 of SE-regulated, highly expressed, and TE-regulated genes. Despite lower expression. **C,**
968 LASSO regression for prediction of DSB-rich genomic sites. the genome was binned in 10kb
969 bins, and the top 5% bind in break density were assigned as DSB-rich. **D,** Boxplot showing the
970 break density in estrogen-responsive genes upon E2 treatment, grouped by degree of
971 transcriptional upregulation (fold change, FC). Non-SE genes were randomly downsampled to
972 match SE gene numbers for comparable P-values. **E,** Boxplot showing the break density in
973 estrogen-responsive genes following CDK9 inhibition (CDK9i), grouped by degree of
974 transcriptional downregulation (log2 fold change). Non-SE genes were randomly downsampled
975 to match SE gene numbers for comparable P-values. Boxplot features include: center line
976 representing the median; box limits denoting upper and lower quartiles; whiskers extending to
977 1.5x the interquartile range; and dots indicating outliers. P-values in boxplots were calculated
978 using the Mann-Whitney U test in B, and Wilcoxon paired test in D and E. n.s. $p > 0.05$, *$p < 0.05$,
979 **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

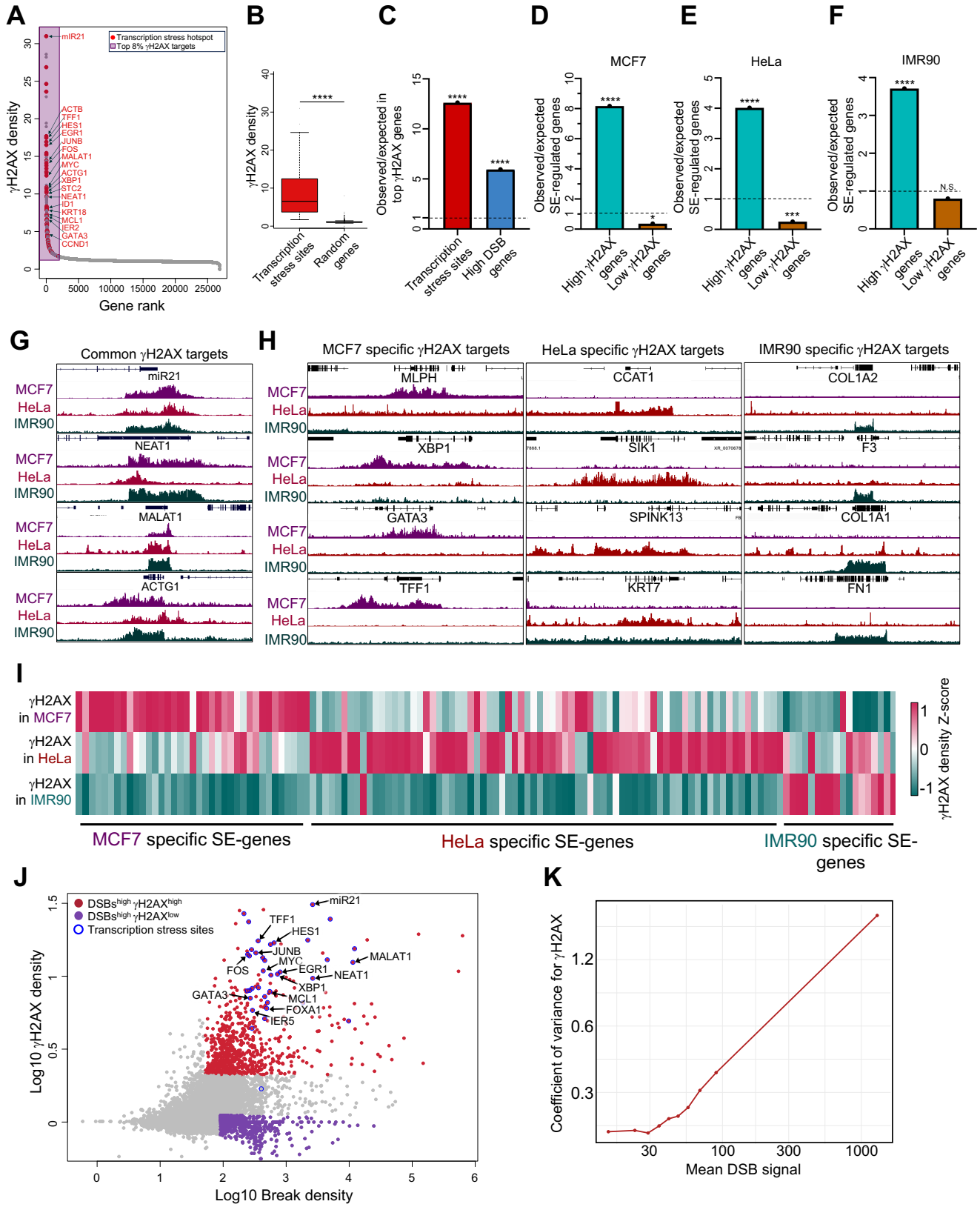
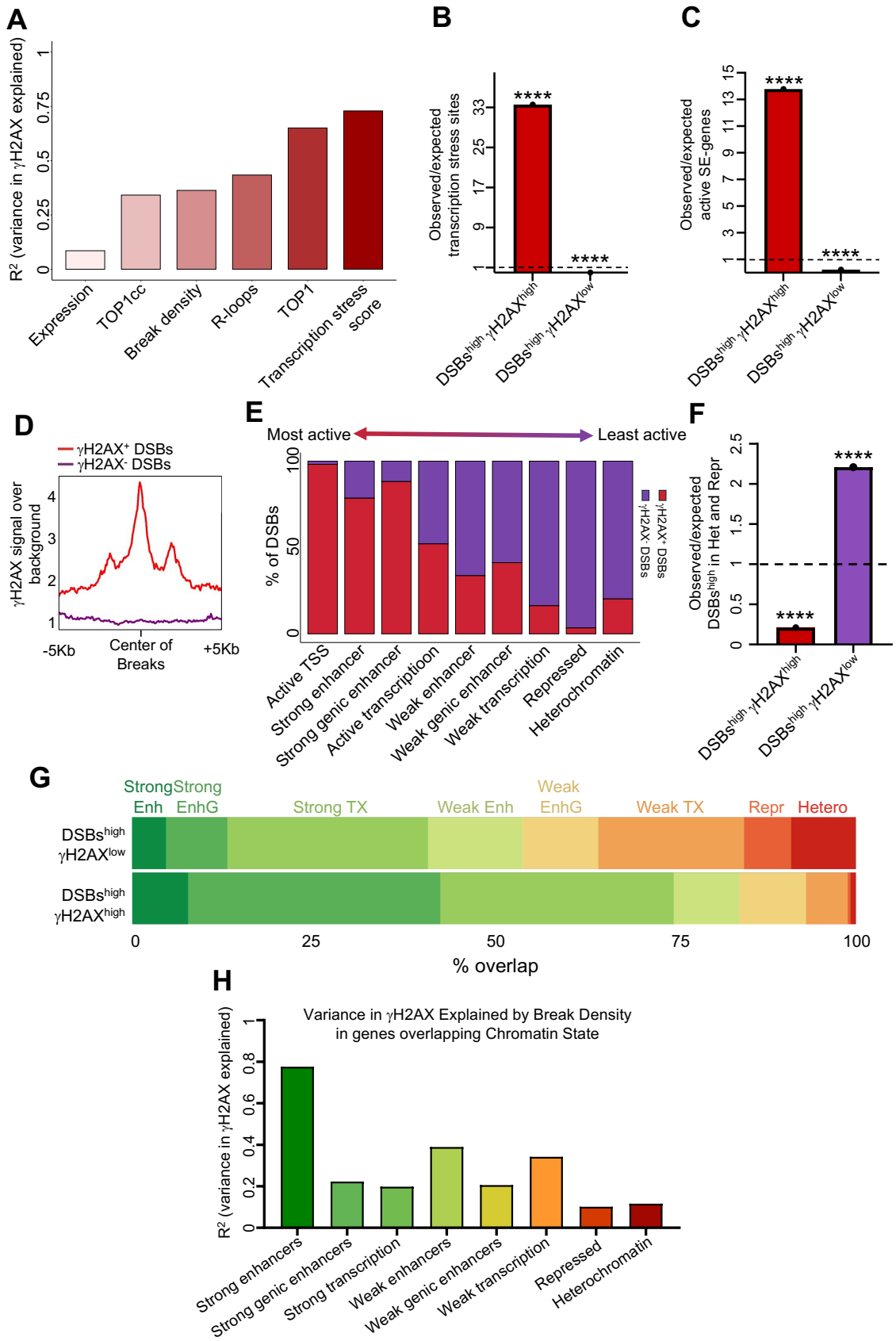


Figure 3: γH2AX marks transcription stress sites.

982 **A**, Ranked plot of genes in MCF7 cells based on γH2AX density. **B**, Boxplot showing γH2AX
983 density for transcription stress sites and randomly selected genes. **C**, The ratio of the observed
984 to the expected number of transcription stress sites or top 1000 break-prone genes in MCF7
985 among the top γH2AX enriched genes. **D**-**F**, The ratio of observed to expected numbers of SE-
986 regulated genes in the top and bottom 300 genes enriched in γH2AX for each cell line. P-values
987 were determined using the hypergeometric test: n.s. $p > 0.05$, *$p < 0.05$, ***$p < 0.001$, ****$p <$
988 0.0001. **G,** Genome browser snapshots of representative common γH2AX targets across MCF7,
989 HeLa, and IMR90 cell lines. **H**, Genome browser snapshots of representative unique γH2AX
990 targets in MCF7, HeLa, and IMR90 cell lines. **I**, Heatmap displaying the γH2AX density column
991 Z-score for each cell line's unique SE-regulated genes. For each cell line, SE-regulated genes
992 were identified, common SE-regulated genes with other cell lines, and genes with low γH2AX
993 counts were filtered out. **J**, Scatter plot showing the correlation between break density and
994 γH2AX density for each gene. **K,** Coefficient of variation (CV) of γH2AX density as a function of
995 mean DNA double-strand break (DSB) signal. Genes were binned into deciles based on their
996 DSB signal intensity (log-transformed), and for each bin, the mean γH2AX density, standard
997 deviation, and CV (standard deviation divided by mean) were calculated. Boxplot features
998 include: center line representing the median; box limits denoting upper and lower quartiles;
999 whiskers extending to 1.5x the interquartile range; and dots indicating outliers. P-values in
1000 boxplots were calculated using the Mann-Whitney U test, ****$p < 0.0001$.


A
R² (variance in γH2AX explained)
0
0.25
0.5
0.75
1
Expression
TOP1cc
Break density
R-loops
TOP1
Transcription stress score
B
Observed/expected transcription stress sites
1
9
17
25
33
****
****
DSBs^high γH2AX^high
DSBs^high γH2AX^low
C
Observed/expected active SE-genes
1
3
5
7
9
11
13
15
****
****
DSBs^high γH2AX^high
DSBs^high γH2AX^low
D
γH2AX⁺ DSBs
γH2AX⁻ DSBs
γH2AX signal over background
1
2
3
4
-5Kb
Center of Breaks
+5Kb
E
Most active
Least active
% of DSBs
0
50
100
γH2AX⁺ DSBs
γH2AX⁻ DSBs
Active TSS
Strong enhancer
Strong genic enhancer
Active transcriptioon
Weak enhancer
Weak genic enhancer
Weak transcription
Repressed
Heterochromatin
F
Observed/expected DSBs^high in Het and Repr
0
0.5
1
1.5
2
2.5
****
****
DSBs^high γH2AX^high
DSBs^high γH2AX^low
G
Strong Enh
Strong EnhG
Strong TX
Weak Enh
Weak EnhG
Weak TX
Repr
Hetero
DSBs^high γH2AX^low
DSBs^high γH2AX^high
0
25
50
75
100
% overlap
H
Variance in γH2AX Explained by Break Density in genes overlapping Chromatin State
R² (variance in γH2AX explained)
0
0.2
0.4
0.6
0.8
1
Strong enhancers
Strong genic enhancers
Strong transcription
Weak enhancers
Weak genic enhancers
Weak transcription
Repressed
Heterochromatin

1001

1002 **Figure 4: γH2AX preferentially marks endogenous DSBs of active chromatin**.
1003 **A**, Barplot showing R^2 values for the different variables tested regarding their effect on γH2AX
1004 variance across all genes. **B**-**C**. The ratio of observed to expected number of transcription stress
1005 sites (**B**) and active SE-regulated genes (**C**). **D**, γH2AX signal at the two defined sets of DSBs.
1006 **E**. The percentage of DSBs enriched with γH2AX ($\gamma H2AX^{high}$ DSBs) or lacking γH2AX ($\gamma H2AX^{low}$
1007 DSBs) in each chromatin state. Chromatin states were ordered based on what is typically known
1008 about their transcriptional activity. **F**, The ratio of observed to expected numbers of high DSB
1009 genes overlapping heterochromatin and repressed regions in both gene groups in MCF7 cells.
1010 **G**. The percentage of overlaps between each gene group and chromatin states in MCF7. For
1011 each group, the number of overlaps was counted for each chromatin state and then divided by
1012 the total number of overlaps. **H**. A barplot showing R^2 values for the effect of DSBs on γH2AX
1013 variance for genes overlapping different chromatin states. P-values were determined using the
1014 hypergeometric test: ****$p < 0.0001$.

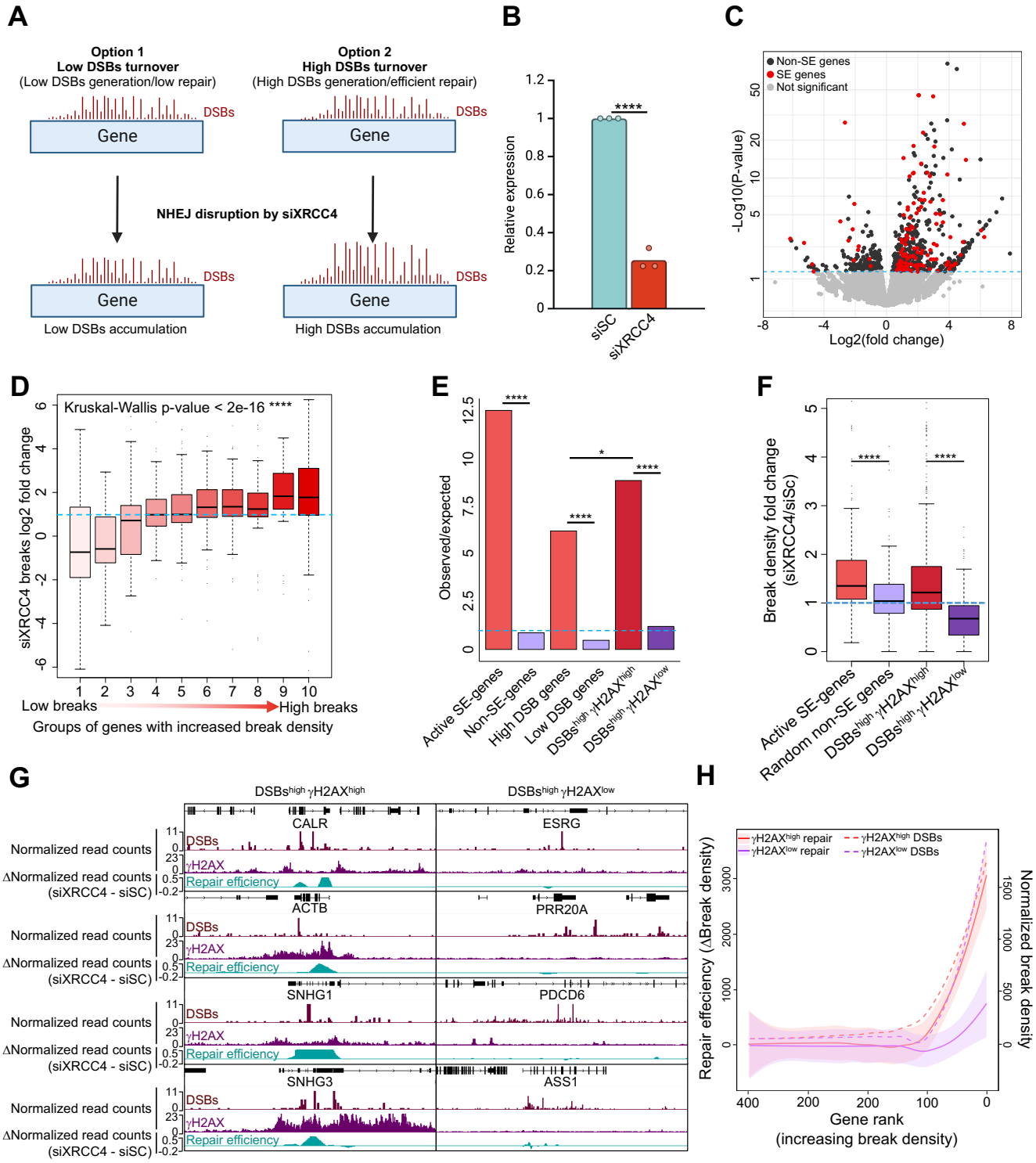


Figure 5: Genes marked with γH2AX show high DSB turnover and repair efficiency. A. Schematic representation of the rationale behind the experiment to estimate the rate of endogenous DSB repair and turnover. **B**. Validation of XRCC4 knockdown via RT qPCR, each dot represents the mean of three technical replicate of each biological replicate. **C**. Volcano plot depicting the results of the differential expression analysis of DSBs following XRCC4 KD,

assessed using the DESeq2 algorithm. The experiment was performed on two biological replicates per condition. **D**. Boxplot illustrating the log2FC of break density upon XRCC4 KD for significantly affected genes stratified according to their baseline break density into ten groups. **E**. The ratio of observed to expected number of genes in our defined gene sets. P-values were calculated using Fischer exact test. **F**. Boxplot of break density fold change. P-values were calculated using the Mann-Whitney U test. **G**. Genome browser snapshots of DSBs^{high} γH2AX^{high} and DSBs^{high} γH2AX^{low} gene examples, showing DSBs, γH2AX, and repair efficiency. Repair efficiency was estimated by pooling replicates, tiling the genome into 100bp tiles, subtracting the break score of each tile in the siXRCC4 sample from siSc (control), and smoothing the segments through sliding window averaging. **H**. Line plot for repair efficiency (log2FC after XRCC4 KD) for DSBs^{high} γH2AX^{high} versus DSBs^{high} γH2AX^{low} genes ordered by their break density. Lines were generated using LOESS smoothing. Boxplot: center line indicates median; box limits represent upper and lower quartiles; whiskers extend to 1.5x interquartile range; dots indicate outliers. *$p < 0.05$, ****$p < 0.0001$.

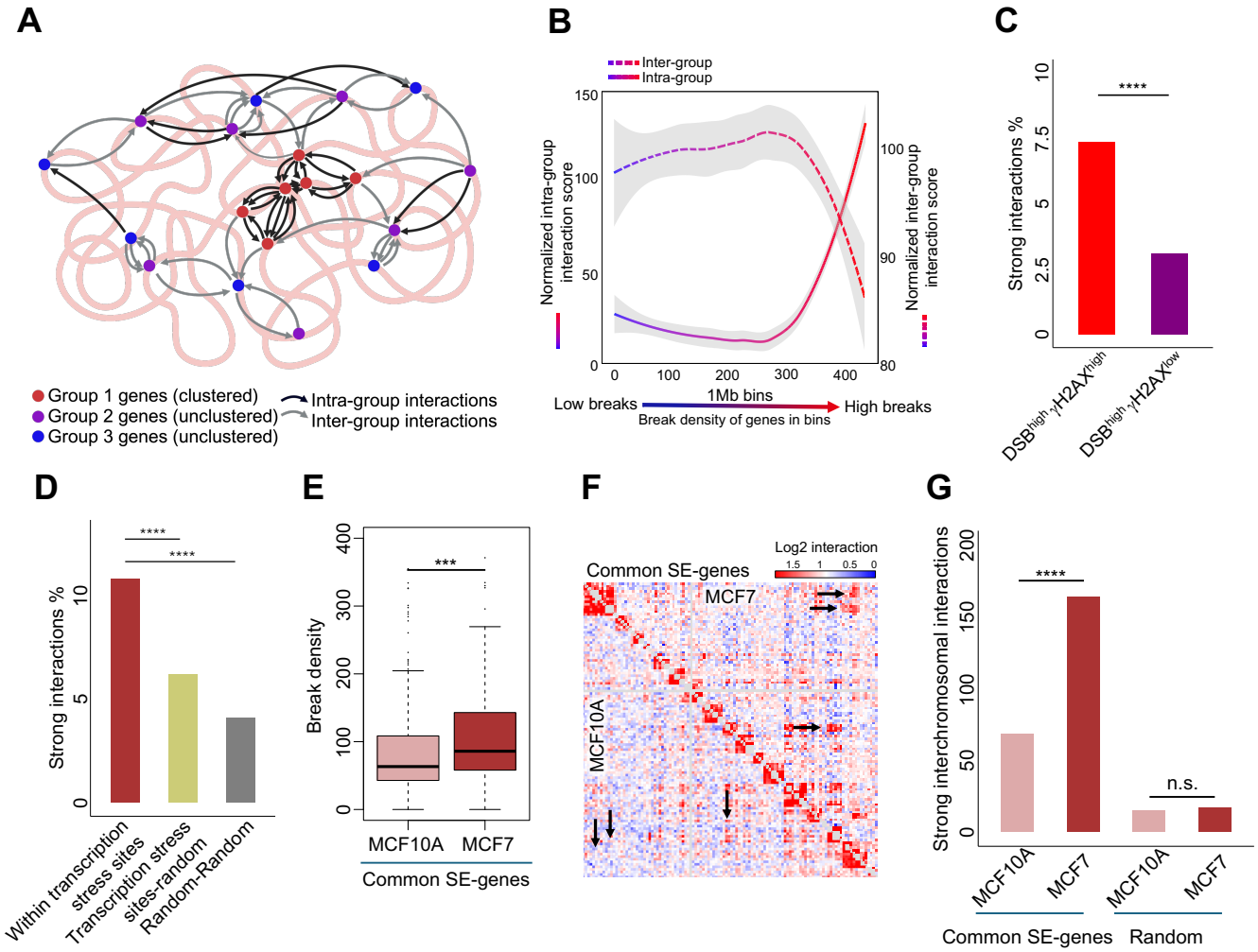
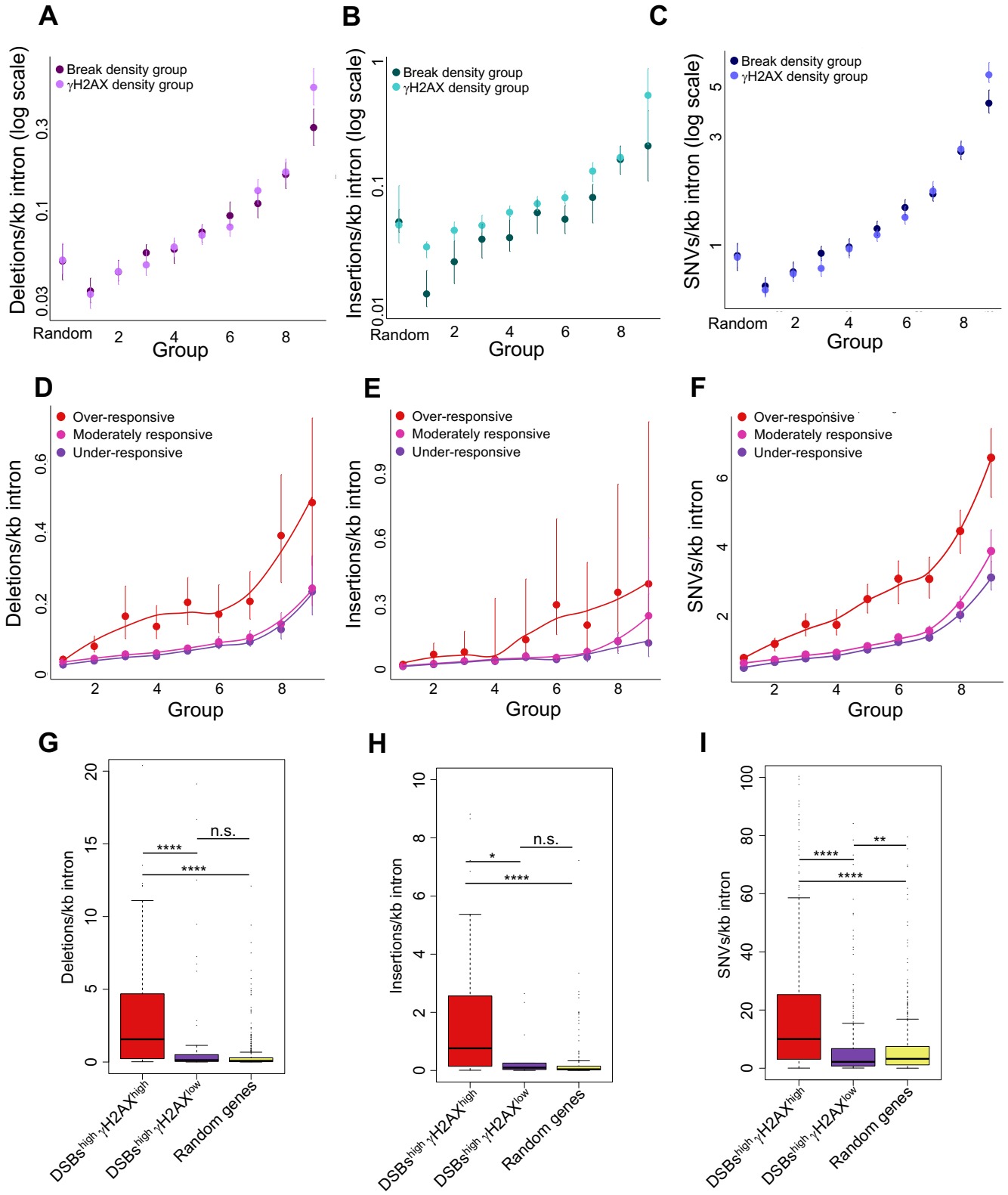
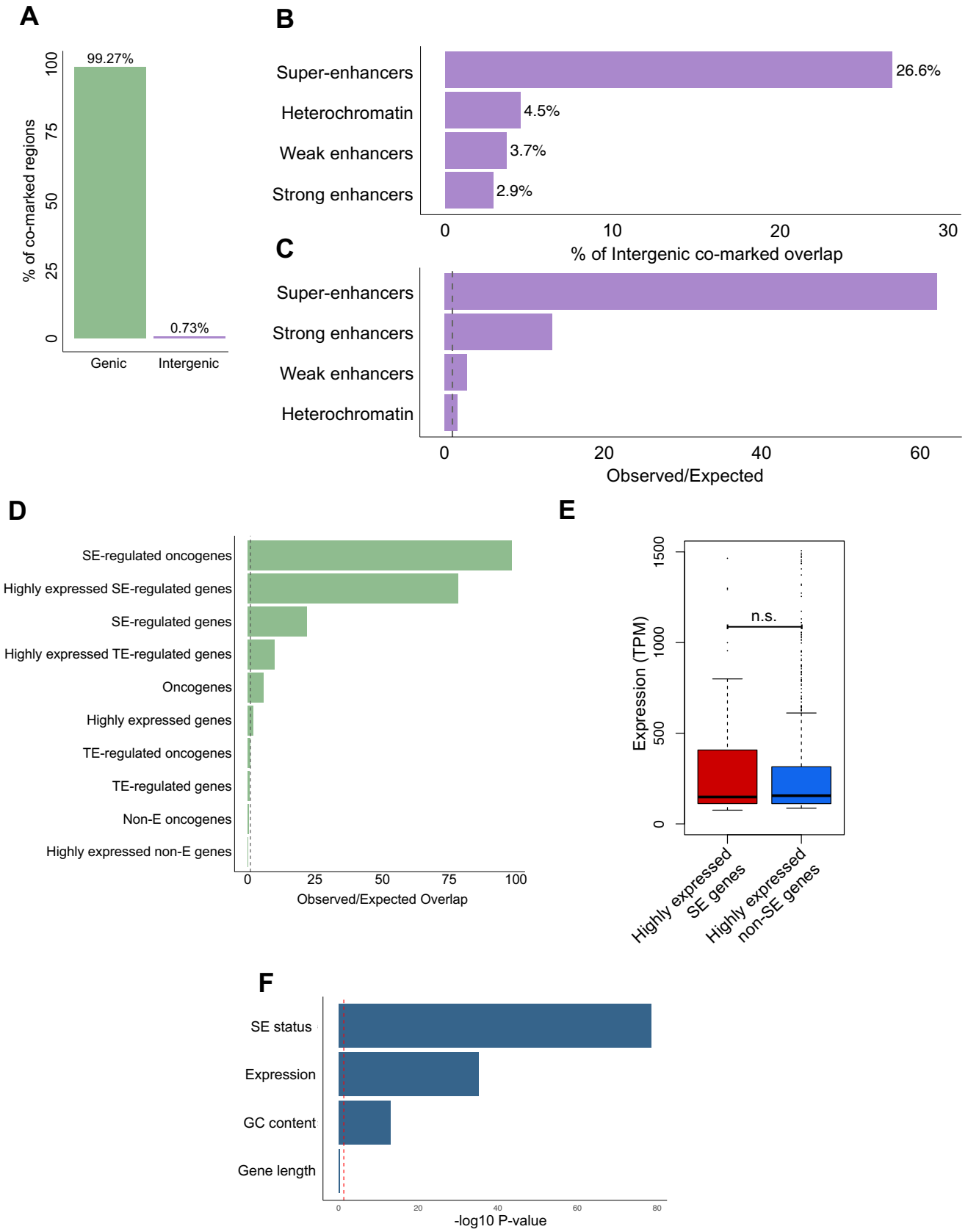


Figure 7: Efficiently repaired transcriptional DSBs cluster in the 3D genome.
A. Schematic illustrating inter- and intra-group interactions and their use to determine the clustering of genes with a common characteristic (high DSB density in this instance). **B**. Line plot depicting inter-group (dashed line) and intra-group (solid line) genomic bins arranged according to the break density of their overlapping genes. Lines were created using LOESS smoothing, with shaded areas representing the 95% confidence intervals. **C**. Percentage of strong interactions within genomic bins containing DSBs^{high} γH2AX^{high} compared to bins with DSBs^{high} γH2AX^{low} genes. Strong interactions are defined as a log2 interaction score > 2. The percentage is calculated by dividing the number of strong interactions by the total number of interactions. P-values were calculated by Chi-squared test. **D**. Percentage of strong interactions calculated as in (**C**). **E**. Boxplot reflecting break density of common SE-regulated genes between MCF7 and MCF10A cell lines. P-values were calculated by Mann-Whitney U test. **F**. Heatmap showing Log2 interaction scores for genomic bins containing common SE-regulated genes from both MCF7 and MCF10A cell lines. Arrows indicate distinct sites of variation in interactions. **G**. Count of interchromosomal strong interactions shown in the heatmap from (**F**). Random bins serve as a reference. Boxplot: center line indicates median; box limits represent upper and lower quartiles; whiskers extend to 1.5x interquartile range; dots indicate outliers. P-values were

calculated using chi-square test in (**c-d**) and (**g**), and Wilcoxon paired test in (e). ***p < 0.001, ****p < 0.0001.



1063 **Figure 7: transcriptional DSBs with high turnover are especially susceptible to mutations.**
1064 **A**-**C**. Intronic deletion density (**A**), insertion density (**B**), and SNV density (**C**) for genes
1065 categorized by their break or γH2AX density (excluding zero break genes). Random genes are
1066 included for reference. **D**-**F**, Rate of intronic mutations across genes categorized by their break
1067 density and further divided into three groups based on their γH2AX response to DSBs. **G-I**.
1068 Boxplot of intronic deletion density (**G**), insertion density (**H**), and SNV density (**I**) for $DSBs^{high}$
1069 $\gamma H2AX^{high}$ vs. $DSBs^{high}$ $\gamma H2AX^{low}$ vs. randomly selected genes.


A
% of co-marked regions
0
25
50
75
100
99.27%
0.73%
Genic
Intergenic
B
Super-enhancers
Heterochromatin
Weak enhancers
Strong enhancers
26.6%
4.5%
3.7%
2.9%
0
10
20
30
% of Intergenic co-marked overlap
C
Super-enhancers
Strong enhancers
Weak enhancers
Heterochromatin
0
20
40
60
Observed/Expected
D
SE-regulated oncogenes
Highly expressed SE-regulated genes
SE-regulated genes
Highly expressed TE-regulated genes
Oncogenes
Highly expressed genes
TE-regulated oncogenes
TE-regulated genes
Non-E oncogenes
Highly expressed non-E genes
0
25
50
75
100
Observed/Expected Overlap
E
Expression (TPM)
0
500
1000
1500
n.s.
Highly expressed SE genes
Highly expressed non-SE genes
F
SE status
Expression
GC content
Gene length
0
20
40
60
80
-log10 P-value

1070

1071 **Figure S1. Characterization of transcription stress sites defined by HMM analysis.**
1072 **A,** Distribution of transcription stress sites identified by HMM across genic and intergenic
1073 regions. Genic regions were defined as gene coordinates extended by ±10 kb flanking regions.
1074 **B,** Proportion of intergenic transcription stress sites identified by HMM overlapping key
1075 chromatin states and super-enhancers. **C,** Enrichment analysis showing observed over
1076 expected overlap of intergenic transcription stress sites with chromatin states and super-
1077 enhancers. **D,** Enrichment analysis of transcription stress sites within different gene categories.
1078 **E,** Boxplot displaying expression levels of noted genes. **F,** Logistic regression analysis testing
1079 the predictive power of SE status, gene expression, GC content, and gene length for the
1080 presence of transcription stress sites within genes.

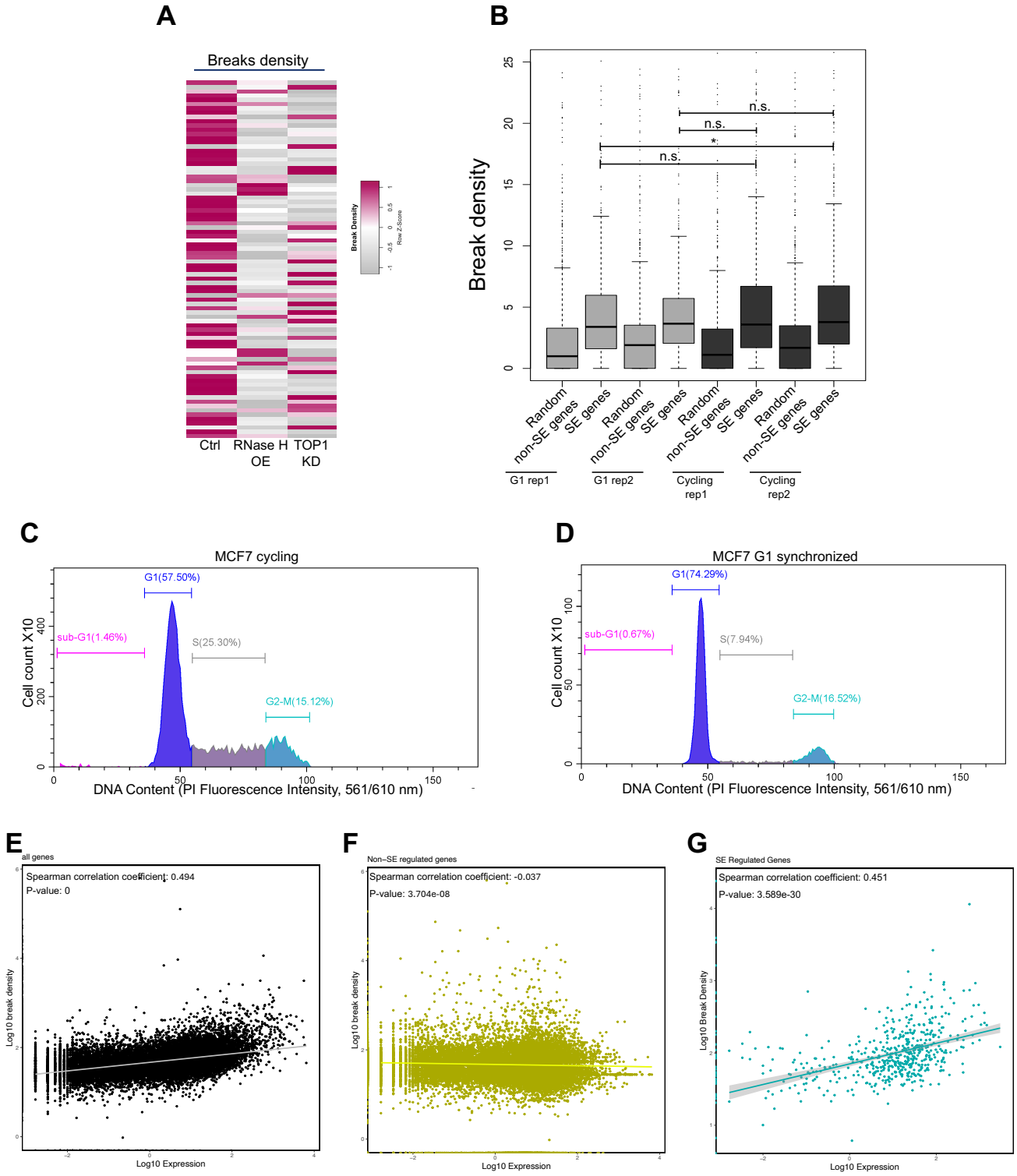


Figure S2

A, Heatmap illustrating break density of SE-regulated genes following RNase H overexpression and TOP1 KD. **B**, Boxplot showing break density of random genes or SE-regulated genes in MCF7 cells synchronized in G1 phase or cycling cells. **C–D,** Flow cytometry analysis of DNA content in MCF7 cells. (C) DNA content profile of asynchronously cycling MCF7 cells. (D) DNA

content profile of MCF7 cells synchronized in G1 phase, demonstrating enrichment of the population in G1. Cells were stained with propidium iodide (PI) and analyzed by flow cytometry. **E-G**, Scatter plot demonstrating the correlation between expression and break density for all genes (**E**), non-SE-regulated genes (**F**), and SE-regulated genes (**G**). Boxplot: center line indicates median; box limits represent upper and lower quartiles; whiskers extend to 1.5x interquartile range; dots indicate outliers. P-value was calculated using Mann Whitney U test. n.s. indicating $p > 0.05$, * indicating $p < 0.05$.

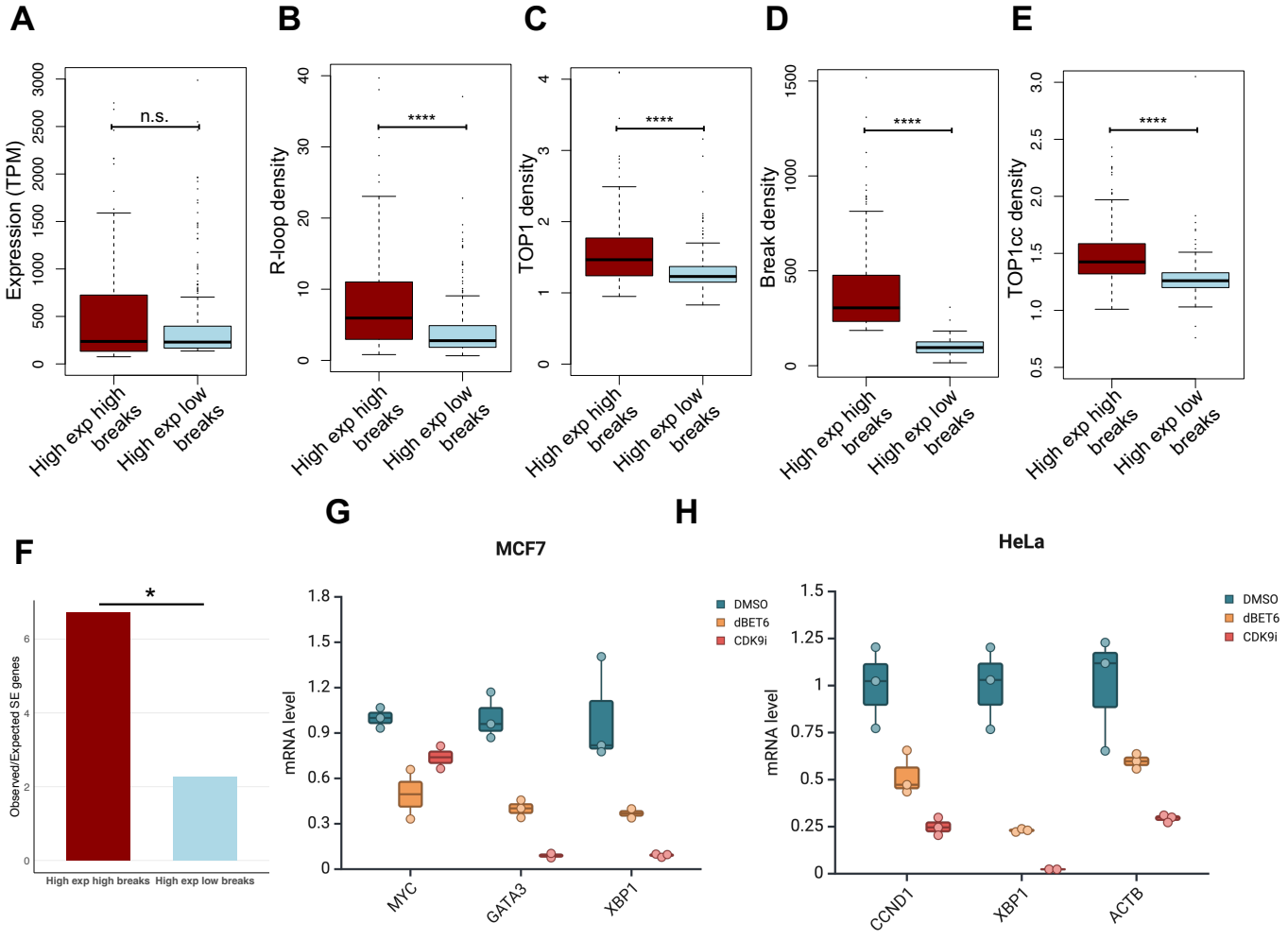


Figure S3

A-**E**, Boxplot showing expression levels (**A**), R-loop density (**B**), TOP1 density (**C**), break density (**D**), and TOP1cc density (**E**) for high expression high breaks genes versus high expression low breaks genes. **F**, observed to expected ratio of SE genes in high expression high breaks versus high expression low breaks genes. **g**-**h**, boxplot illustrating mRNA levels of sample SE-regulated genes in MCF7 (**G**) and HeLa (**H**) cells after dBET6 or CDK9i (100nM for 4 hours). Boxplot: center line indicates median; box limits represent upper and lower quartiles; whiskers extend to 1.5x interquartile range; dots indicate outliers. P-value was calculated using Mann Whitney U test. n.s. indicating p > 0.05, **** indicating p < 0.0001.

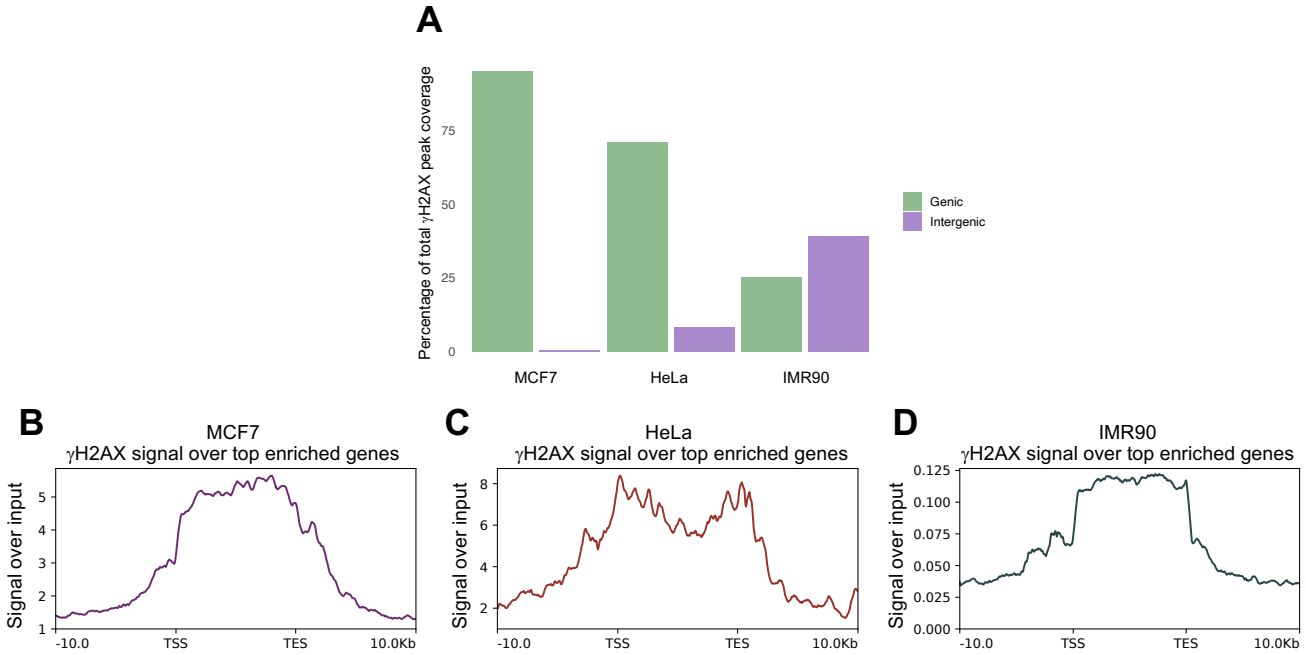


Figure S4. Distribution and enrichment of γH2AX signal at genic regions across different cell lines. A, Percentage of total γH2AX peak coverage occurring within genic and intergenic regions in MCF7, HeLa, and IMR90 cells. Genic regions were defined as gene coordinates extended by ±10 kb flanking regions. (**B**-**D**) γH2AX signal over input across the gene body and flanking regions (±10 kb) for the top enriched genes in (**B**) MCF7, (**C**) HeLa, and (**D**) IMR90 cells. TSS: transcription start site; TES: transcription end site.

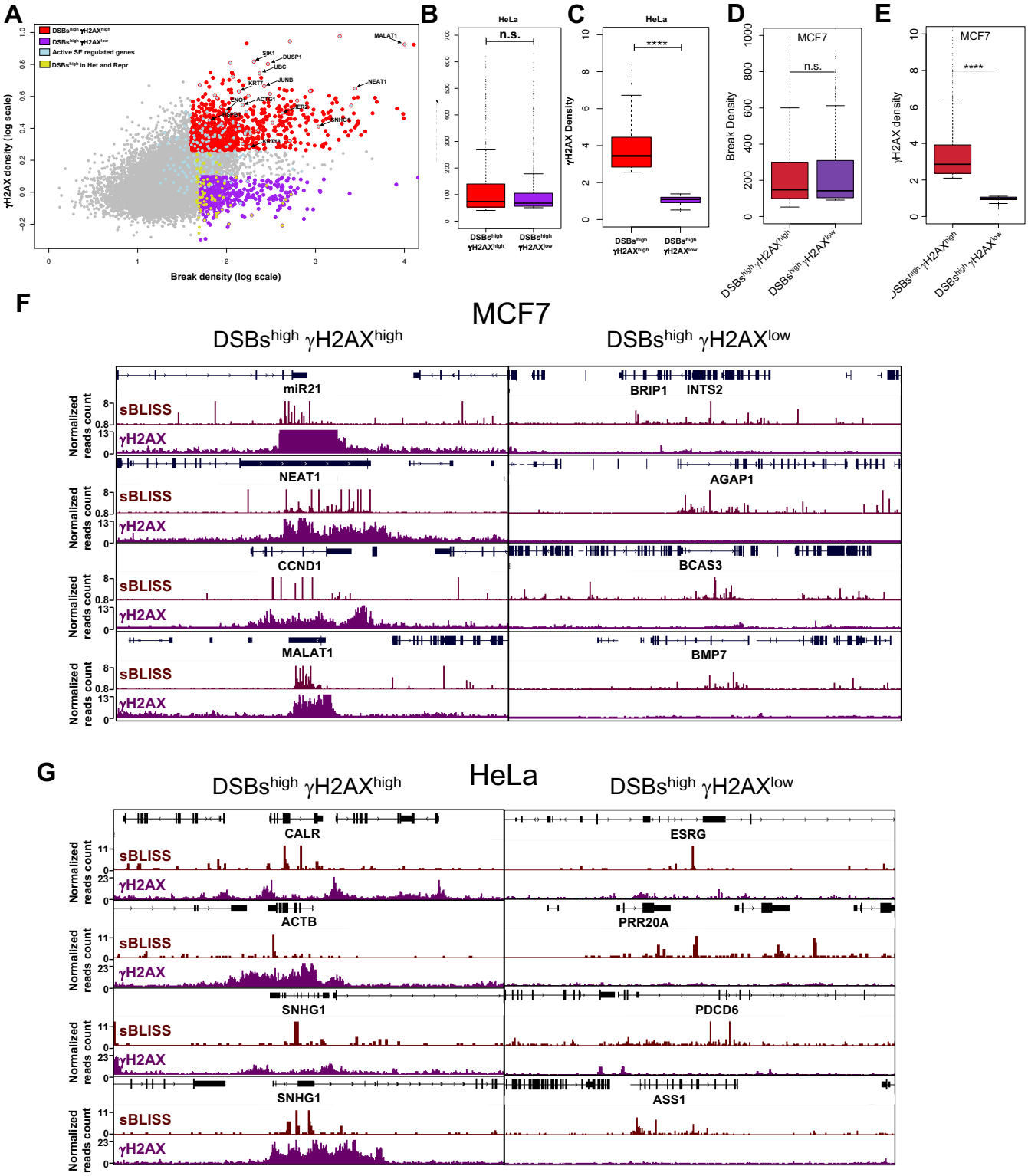


Figure S5

A, Scatter plot showing the correlation between break density and γH2AX density for each gene in HeLa cells. **B-E**, boxplots showing DSBs (B and D) and γH2AX density (C and E) for the two groups of genes defined in (a) and Figure 3F. **D-E,** Genome browser snapshots of representative

examples of DSBs^{high} γH2AX^{high} and DSBs^{high} γH2AX^{low} genes in MCF7 (D), and HeLa (E) cell lines. Boxplot: center line indicates median; box limits represent upper and lower quartiles; whiskers extend to 1.5x interquartile range; dots indicate outliers. P-values were calculated using Mann Whitney U test. n.s. indicating $p > 0.05$, **** indicating $p < 0.0001$.

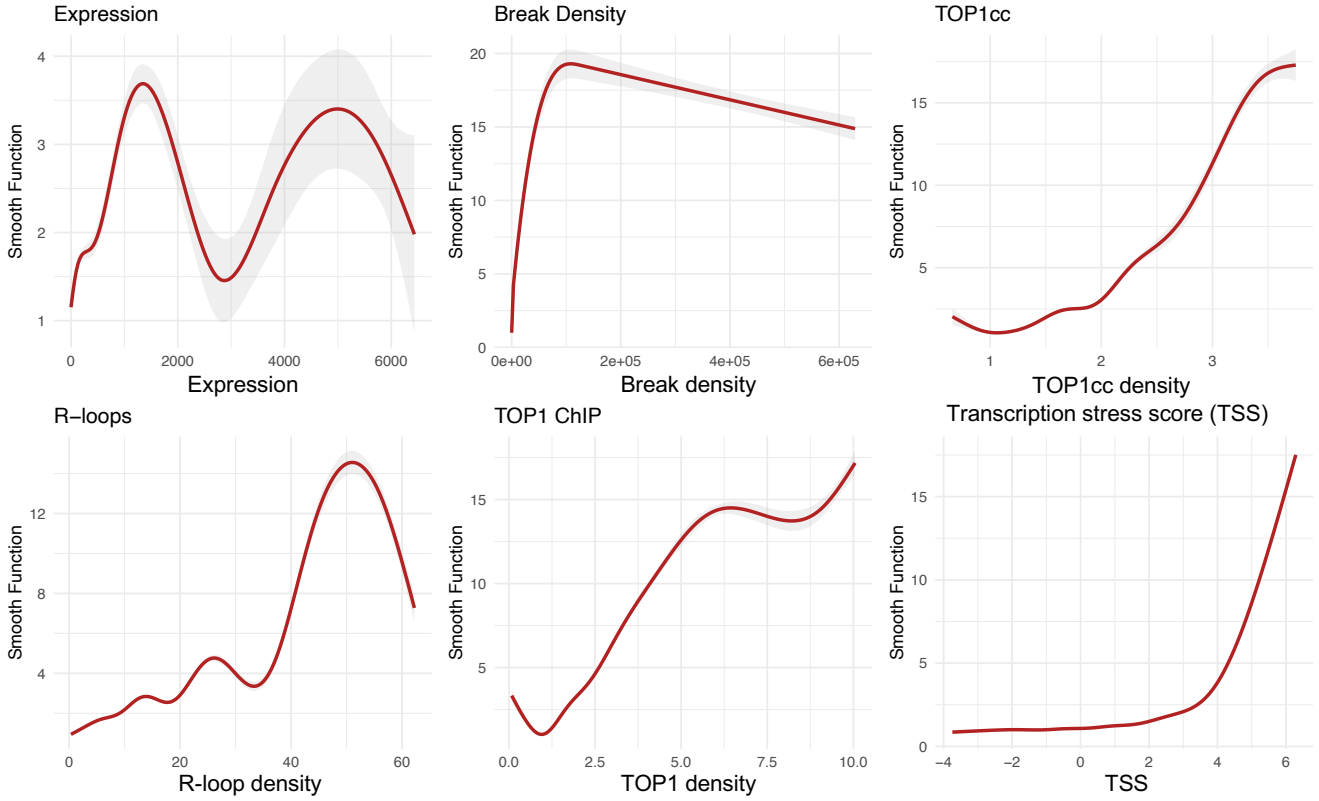


Figure S6

Smooth terms from Generalized Additive Models (GAMs) showing the effect of transcription-related predictors on γH2AX density. The y-axis shows the partial effect of each predictor (centered smooth function, s(x)) on γH2AX levels after accounting for non-linear relationships. The red line represents the fitted smooth function, and the shaded gray band indicates the 95% confidence interval.

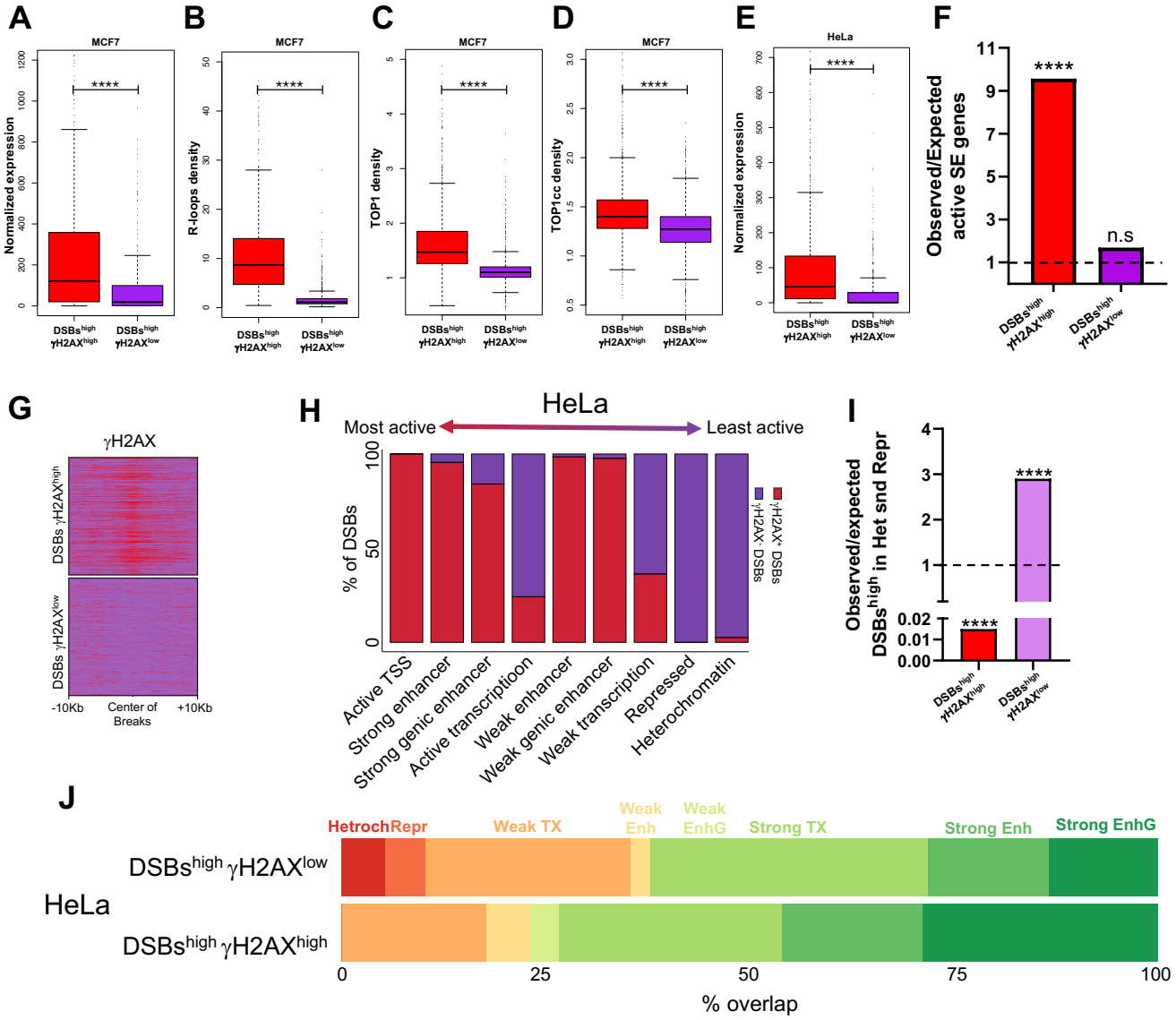


Figure S7

A-D, Boxplot showing expression levels (**A**), R-loop density (**B**), TOP1 density (**C**), TOP1cc density (**D**) for DSBs^{high} γH2AX^{high} and DSBs^{high} γH2AX^{low} genes in MCF7 cells. **E**, boxplot showing expression levels for DSBs^{high} γH2AX^{high} and DSBs^{high} γH2AX^{low} genes in HeLa cells. **F**, observed to expected ratio of active SE genes in DSBs^{high} γH2AX^{high} in HeLa cell line. **G**, heatmap showing the enrichment of γH2AX signal at the two defined sets of DSBs. **H**, The percentage of DSBs enriched with γH2AX (γH2AX^{high} DSBs) or lacking γH2AX (γH2AX^{low} DSBs) in each chromatin state of HeLa cell line. Chromatin states were ordered based on what is typically known about their transcriptional activity. **I**, The ratio of observed to expected numbers of high DSB genes overlapping heterochromatin and repressed regions in both gene groups in HeLa cells. **J**, The percentage of overlaps between each gene group and chromatin states in HeLa. For each group, the number of overlaps was counted for each chromatin state and then divided by the total number of overlaps.

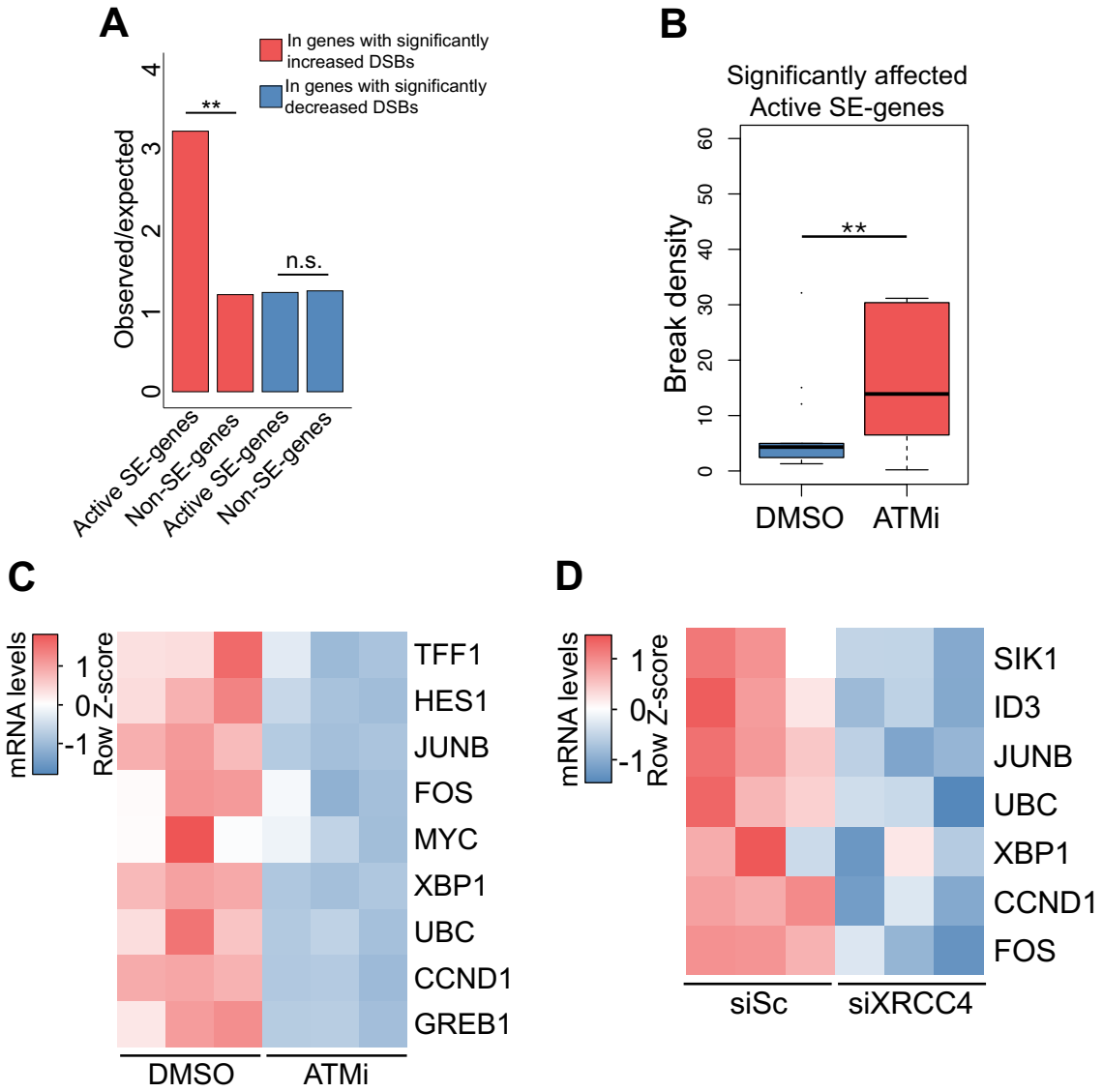


Figure S8: Efficient repair at transcription stress sites is required for SE-regulated oncogene expression

A, Ratio of observed to expected active SE-regulated genes and non-SE-regulated genes in genes with significantly increased vs decreased DSBs upon ATMi treatment. P-values were calculated using chi-square test. **B**, break density of significantly affected SE-regulated genes. C, heatmap showing the change of mRNA levels as measured by qPCR upon ATMi treatment of top γH2AX-enriched SE-regulated genes in MCF7 cells. P-values were calculated by Wilcoxon paired test. **D**, heatmap showing the change of mRNA levels as measured by qPCR upon XRCC4 KD of top γH2AX-enriched SE-regulated genes in HeLa cells. n.s. indicating $p > 0.05$, ** indicating $p < 0.01$.

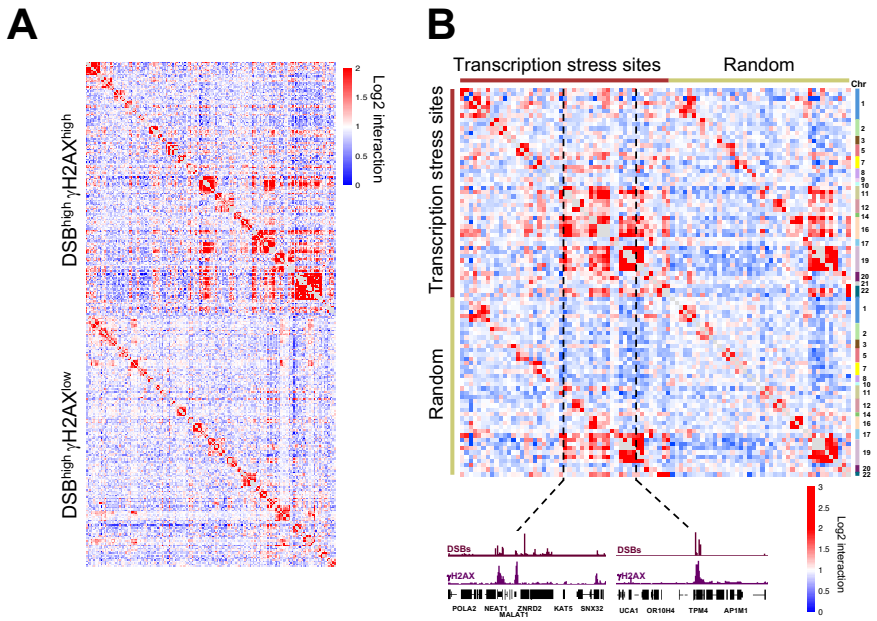


Figure S9

A, Heatmap showing Log2 interaction scores for genomic bins containing DSBs^{high} γH2AX^{high} (top) and DSBs^{high} γH2AX^{low}. (bottom). **B**, Heatmap showing Log2 interaction scores for genomic bins containing transcription stress sites or randomly selected bind with similar chromosomal distribution.