# R-loops and Topoisomerase 1 facilitate formation of transcriptional DSBs at gene bodies of hypertranscribed genes

Osama Hidmi and Rami I. Aqeilan*

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

*Corresponding author: ramiaq@mail.huji.ac.il

**Keywords**: R-loops, TOP1, γH2AX, DSBs, hypertranscription, oncogenes

**Conflict of Interest:** non

## Abstract

DNA double-stranded breaks (DSBs) are considered the most deleterious type of DNA damage. Although cytotoxic levels of DSBs occur mainly due to external factors, such as ionizing irradiation and UV radiation, DSBs can occur naturally under physiological conditions through cell's natural processes such as transcription and replication. However, how physiological DSBs are generated during transcription is still poorly understood. Here, we mapped DSBs using in-suspension break labeling *in situ* and sequencing (sBLISS) and re-analysis of ChIP-seq and DRIP-seq data, along with immunofluorescence staining for DNA damage markers in MCF-7 cells. Mapping of DSBs revealed their occurrence at highly expressed genes that are enriched with Topoisomerase 1 (TOP1), the main enzyme responsible for resolving positive and negative supercoiling, and with R-loops. Depleting R-loops and TOP1 decreased DSBs specifically at highly expressed and TOP1/R-loops-enriched genes implying that TOP1 and R-loops are the main causes of transcriptional DSBs. These findings propose a model in which DSBs in highly expressed genes, mainly oncogenes, are caused by resolving of R-loops and the catalytic activity of TOP1. Our data also reveal that these genes, despite being highly fragile, are covered by γ-H2AX suggesting efficient repair signaling. Our findings underscore the critical roles of TOP1 and R-loops in regulating DSBs in hypertranscribed oncogenes implicated in carcinogenesis.

## Introduction

Double-stranded breaks (DSBs) are considered to be the most deleterious type of DNA damage, and thus must be repaired to preserve genome integrity. Although cytotoxic levels of DSBs occur mainly by external factors such as ionizing radiation and UV radiation, DSBs can occur naturally under physiological conditions through the cell's natural processes such as transcription and replication (Wilhelm, Said and Naim, 2020).

Transcription can cause DSBs in replicating cells due to collision with the replication machinery, especially when transcription and replication have opposite directions(Mirkin and Mirkin, 2005; Helmrich, Ballarino and Tora, 2011; Hamperl *et al.*, 2017). This collision (also called transcription-replication conflict) happens when the replication machinery, mainly DNA polymerases, and the transcriptional machinery, mainly RNA polymerases, encounter each other in close proximity within the gene. Since cancer cells are highly active exhibiting hypertranscription and high replication levels, they are expected to exhibit a high occurrence of such conflicts (Helmrich, Ballarino and Tora, 2011; Hamperl *et al.*, 2017). This collision between transcription and replication was found to occur mainly at transcription termination site (TTS), and to be prevented by the activity of topoisomerase 1 (TOP1) (Promonet *et al.*, 2020). Elevated transcription was found to strongly correlate with DSBs (Lensing *et al.*, 2016). However, how transcription by itself causes DSBs is still under investigation.

During transcription elongation, the DNA unwinds around the transcribing RNA Pol II (Choder and Aloni, 1988), which increases the chance of newly transcribed RNA to anneal to its template (García-Muse and Aguilera, 2019), forming a three-stranded nucleotide structure composed of a DNA:RNA hybrid and a dislocated single-stranded DNA, known as an R-loop. Very short R-loops (<10 bp) are important for various cellular processes and are formed during transcription and removed afterward by the enzyme RNase H1 (Cerritelli and Crouch, 2009) and RNA-DNA helicases (Aguilera and Gómez-González, 2017). Longer R-loops on the other hand, are known to be harmful to genome integrity due to generating DSBs. The specific mechanism by which R-loops generate DSBs is still poorly understood. A proposed mechanism is that they expose

the displaced ssDNA to endonucleases (Sollier and Cimprich, 2015), making it more susceptible to single-stranded breaks (SSBs). Another observation is that R-loop structures can be recognized by specific endonucleases such as XPG, XPF, and FEN1 (Cristini *et al.*, 2019), which cleave the ssDNA hybridized to the RNA in the R-loop leaving a SSB. However, how these SSBs turn into DSBs is unknown.

Transcription elongation causes negative and positive supercoiling behind and ahead of the transcribing RNA PolII, respectively (Liu and Wang, 1987; Ma and Wang, 2016). This supercoiling has to be resolved to preserve efficient transcription, especially at highly transcribed genes (Ma and Wang, 2016). TOP1 is the main enzyme responsible for resolving positive and negative supercoiling; TOP1 cuts the DNA single-strandedly through its catalytic domain allowing the DNA to relax, followed by re-ligation of the cut strand (Pommier *et al.*, 2016, 2022). Optimally, TOP1 forms a transient Topoisomerase cleavage complex (TOP1cc) which cuts the DNA, and this transient complex is rapidly reversed and TOP1 is released, eventually leading to no persistent DNA damage. However, if TOP1 is trapped at its catalytic state, either by physiological or pathological causes, the ssDNA would remain cut, and TOP1cc that is bound to the DNA should be removed through the Tyrosyl-DNA Phosphodiesterase 1 (TDP1) excision pathway for the break to be repaired (Cristini *et al.*, 2016; Pommier *et al.*, 2022).

DSBs can be detected using various methods, including antibodies against markers of DNA damage like anti-γ-H2AX which is helpful to detect changes in DNA damage with different treatments and conditions, however, it fails to locate the site of these breaks in the genome. Using those antibodies for chromatin immunoprecipitation sequencing assays (ChIP-seq) can detect the location of those breaks (DeCaprio and Kohl, 2020), but because of the nature of γ-H2AX distribution, which can be as wide as several kilo-basepairs around the break, it is less efficient to detect the site of the break with high resolution. For these reasons, we used the newly established technique, break labeling *in situ* and sequencing (sBLISS) (Bouwman *et al.*, 2020), which detects DSBs in a single nucleotide resolution. sBLISS can detect the levels of DSBs across the genome, with high discrimination between DSBs occurring at proximal genomic locations and has

also been confirmed to detect physiological DSBs (Hazan *et al.*, 2019), which makes it a very powerful technique for studying transcriptional DSBs.

Here we used multiple techniques, including sBLISS, ChIP-seq, and DRIP-seq data to uncover a mechanism in which transcription, R-loops, and TOP1 altogether facilitate physiological DSBs formation at active regions and especially at highly transcribed genes.

# Results

## Physiological breaks are enriched at highly transcribed genes

In our previous work (Hazan *et al.*, 2019), we found physiological DSBs to be highly enriched at active regions such as promoters, insulators, and super-enhancers using the BLISS method. To further uncover the mechanism regulating physiological DSBs, we mapped DSBs distribution across the genome (breakome) using the newly developed sBLISS method (Bouwman *et al.*, 2020) in untreated MCF-7 breast cancer luminal cells. Comparing the observed breaks to those that are expected at different chromatin states, we found that the observed:expected ratio is higher at active promoters, strong enhancers, transcription elongation, and transcription transition chromatin states, as opposed to repressed regions, where the ratio was significantly lower (Figure 1A). The increased break enrichment relative to expected at active but not at repressed regions further confirms the involvement of transcription in the formation of DSBs in these regions.

To further test this, relative gene expression was measured using the RNA sequencing technique CEL-seq (Hashimshony *et al.*, 2012), and relative gene expression for each gene was correlated with its break density. Interestingly, relative gene expression was found to be positively correlated with break density (Figure 1B and C, Supplementary Figure 1A, B and D-K), indicating that higher expressed genes tend to have higher break density. Moreover, gene ontology analysis of the top 400 broken genes in MCF-7 showed enrichment in biological processes such as translation, gland morphogenesis, and response to estradiol, which are biological processes that are expected to be highly

active in breast epithelial tissue (Supplementary Figure 1C), indicating that active genes tend to exhibit high break densities.

To investigate these transcriptional DSBs, the top 1000 expressed and the lower 1000 expressed genes were selected from our CEL-seq, and for each gene, the ratio of observed to expected breaks was calculated (Figure 1D). As predicted, highly expressed genes displayed a higher ratio (~7 folds) compared to low expressed genes (<1) (Figure 1E). When examining break density distribution across the gene body of these genes (Figure 1F), we observed an elevated break density at the gene body of the highly transcribed genes compared to low transcribed genes, with two distinct peaks at transcription start site (TSS) and transcription termination site (TTS).

These data clearly suggest that physiological DSBs are more enriched at active regions, specifically at highly transcribed genes, and point towards transcription as a direct or indirect cause of physiological DSBs.

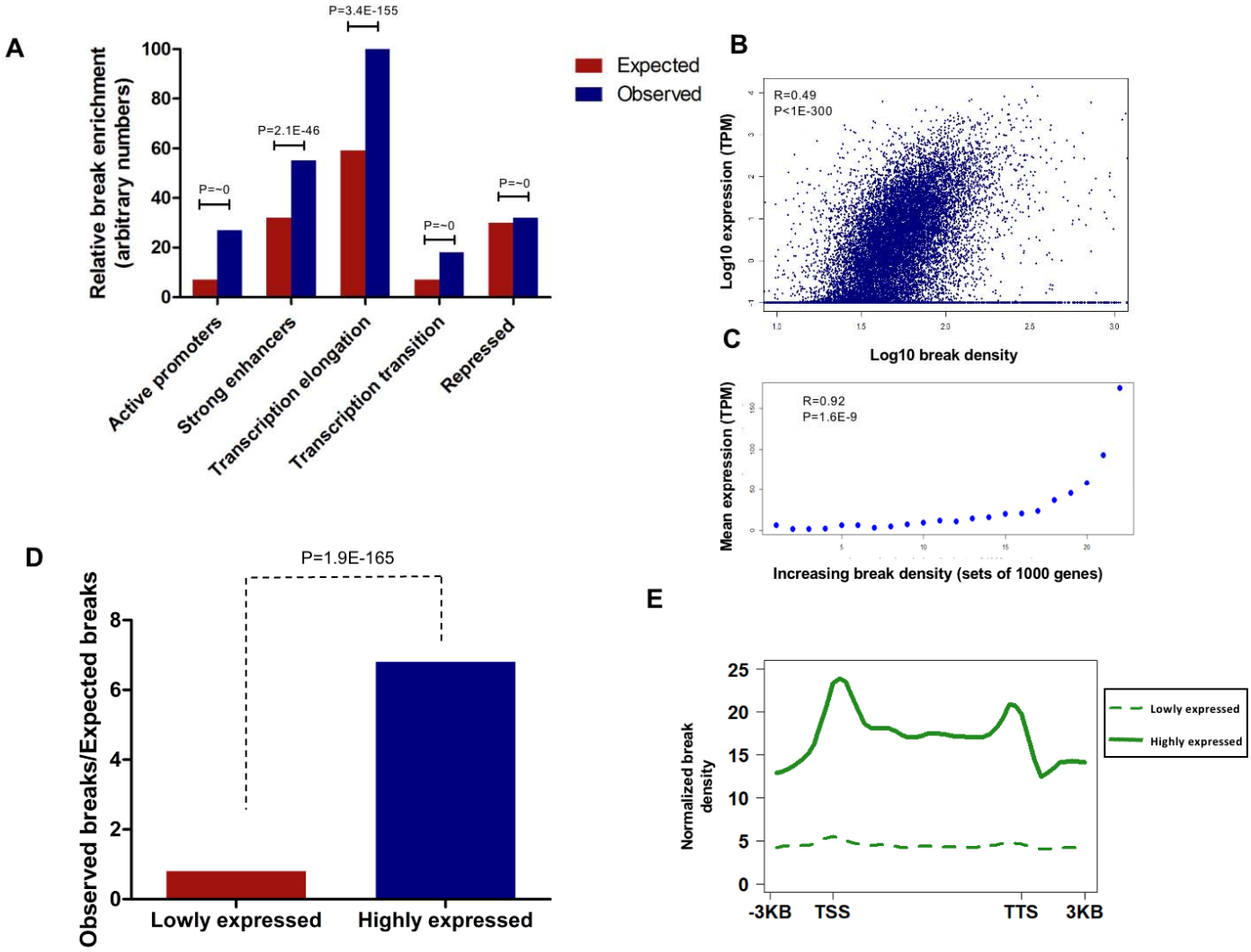


**Figure 1: Breaks are enriched at highly expressed genes**. **A**. The distribution of DSBs in breast cancer MCF7 cells along ChromHMM-defined chromatin states of HMEC. Bar height represents break enrichment relative to other chromatin states. **B**. A zoomed log scaled scatter plot showing a positive correlation between break density and expression levels; each dot represents a gene, and for each gene break density and expression were measured and normalized to gene size. **C**. Mean expression (TPM) positively correlates with break density, genes were grouped into 22 groups with increasing break enrichment, the 22nd group being the group with the highest mean break enrichment. **D**. Ratio of observed breaks vs expected breaks at highly and low expressed genes. **E.** Plot showing break density at gene body of the 1000 most transcribed genes (continuous line) and 1000 least transcribed genes (dashed line). sBLISS experiments were done in duplicates, and CEL-seq was done in quadruplicates.

## Transcriptional DSBs are associated with TOP1

Previously, TOP1 was found to overlap regulatory elements such as insulators, promoters, and enhancers, especially at those which exhibited high break density (Baranello *et al.*, 2016; Hazan *et al.*, 2019). To directly test the involvement of TOP1 in transcriptional DSBs formation, we correlated TOP1 levels and expression levels. Interestingly, the expression of moderately and highly expressed genes revealed a positive correlation with TOP1 levels (Figure 2A, B, Supplementary Figure 2A). Moreover, we found high enrichment of breaks at TOP1 peaks across the genome (Figure 2C). Furthermore, we compared the ratio of the observed breaks to the expected breaks between high and low TOP1 enriched genes and found a greater ratio in high TOP1 genes (~9 folds) compared to low TOP1 genes (Figure 2C). To further confirm the significance of TOP1, we knocked down TOP1, using siRNA (Supplementary Figure 2B), and examined the consequences on DNA breaks using immunofluorescence staining. As seen in Fig 2E and F, a significant decrease in γ-H2AX and 53BP1 foci per nucleus were observed in TOP1-depleted MCF7 cells compared to scramble siRNA control cells (Figure 2E, F). Altogether, these results indicate an important role of TOP1 in the formation of transcriptional DSBs.

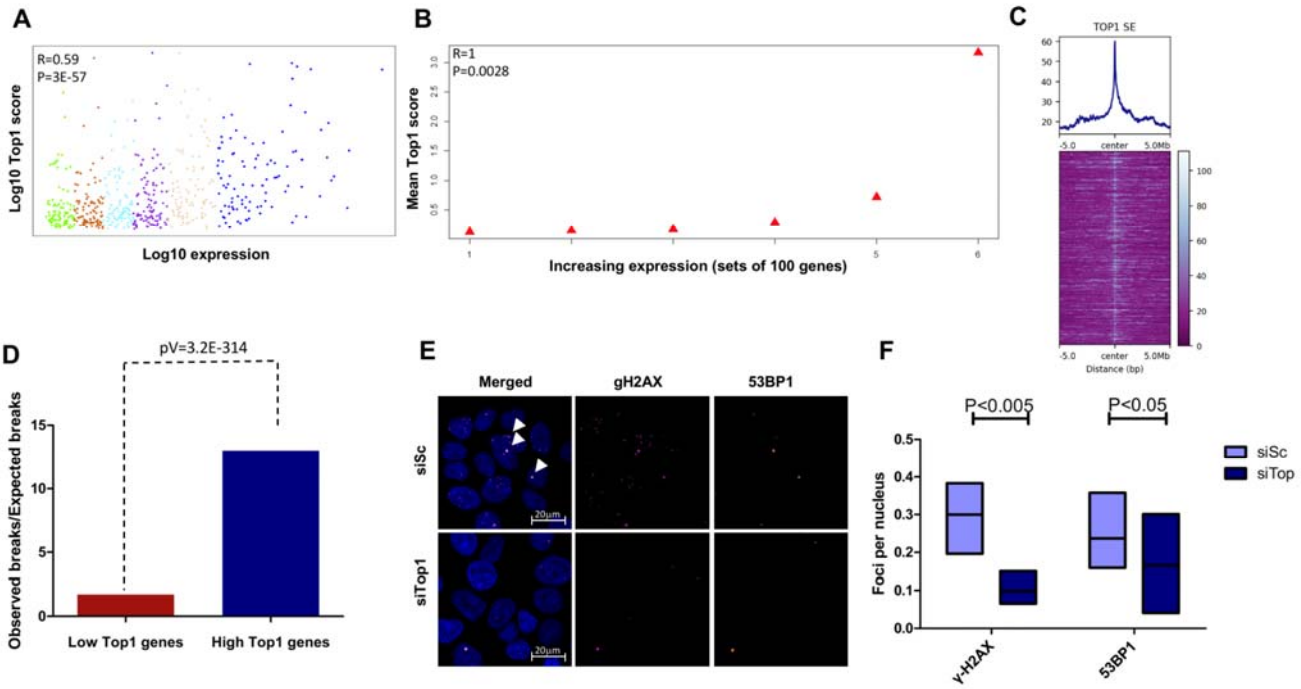



**Figure 2: Transcriptional DSBs are associated with TOP1**. **A.** A log scaled scatter plot showing positive correlation between expression and TOP1 levels for the top 50% expressed genes; each dot represents a gene, and each color represents a gene group used in **B**. **B.** Median of expression (TPM) correlates with TOP1 levels, for the top 50% expressed genes. **C.** Heatmap of break density at TOP1 binding regions across the genome. **D.** Ratio of observed breaks to expected breaks for low TOP1 and high TOP1 genes. **E.** Representative images of immunofluorescent staining of MCF-7 ctrl (siSc) and TOP1-depleted (siTOP1) MCF-7 cells using γ-H2AX and 53BP1 antibodies. **F.** Quantification of the number of γ-H2AX and 53BP1 foci per nucleus in MCF-7 ctrl and MCF-7 siTop (E). sBLISS and immunofluorescence experiments were done in duplicates.

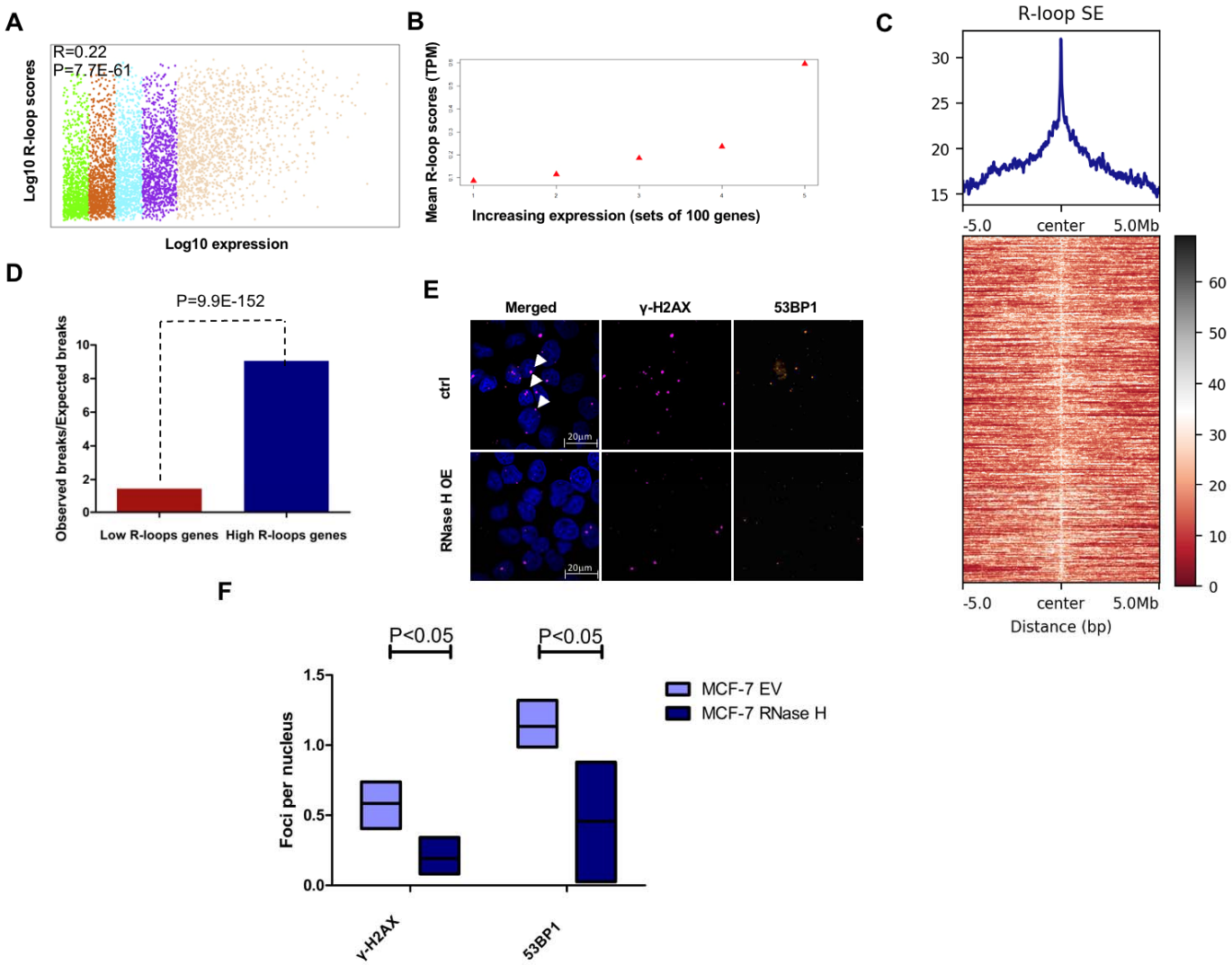
## Transcriptional DSBs are associated with R-loops

R-loops are known for their harmful effect on genome integrity, and their formation is known to be facilitated by the transcriptional machinery (Aguilera and García-Muse, 2012; García-Muse and Aguilera, 2019). Given the connection between TOP1 and R-loops, which was further validated by correlating TOP1 levels with R-loops levels (Supplementary Figure 2C), we next set to test the involvement of R-loops in the formation of transcriptional DSBs.

Correlating R-loops levels, using published DRIP-seq data done on untreated MCF-7 cells, and expression revealed a positive correlation in the top 50% expressed genes (Figure 3A and B, Supplementary Figure 2D). When mapping break enrichment at R-loops peaks, we observed elevated break enrichment at these peaks (Figure 3C). Furthermore, comparing observed to expected breaks ratio between high and low R-

loops genes revealed that genes enriched with R-loops have a higher ratio (~9 folds) compared to those deprived of R-loops (Figure 3D).

To test the effect of R-loops depletion on global DSBs, the enzyme RNase H1 which degrades the RNA moiety of R-loops (García-Muse and Aguilera, 2019) was overexpressed in MCF-7 cells, and immunofluorescence staining was performed for 53BP1 and γ-H2AX (Figure 3E, F). Interestingly, cells overexpressing RNase H1 had fewer γ-H2AX and 53BP1 foci per nucleus compared to control cells, indicating the




involvement of R-loops in the formation of physiological DSBs.

**Figure 3: Transcriptional DSBs are associated with R-loops**. **A**. A log scaled scatter plot showing positive correlation between expression and R-loops levels for the top 50% expressed genes, each dot represents a gene. DRIP data was extracted from GSE81851. **B.** Mean expression (TPM) correlates with R-loops levels for the top 50% expressed genes. **C.** Heatmap of break density at R-loops regions across the genome. **D.** Ratio of observed breaks to expected breaks for low R-loops and high R-loops genes. **E.**

Representative images of immunofluorescent staining of MCF-7 EV and MCF-7 overexpressing RNase H1 for γ-H2AX and 53BP1. **F.** Number of γ-H2AX and 53BP1 foci per nucleus, for MCF-7 EV and MCF-7 RNase H1. sBLISS and immunofluorescence experiments were done in duplicates.

## TOP1 knockdown and RNase H1 overexpression decrease break enrichment specifically at the gene body of highly transcribed genes

To investigate the contribution of TOP1 and R-loops in transcriptional DSBs formation, TOP1 was depleted in MCF-7 cells overexpressing RNase H1 or control cells (supplementary Figure 3A, B, and C), and changes in DSBs were mapped by sBLISS (Figure 4A). We first examined break enrichment at different chromatin states and found that regions associated with transcription elongation and transcription transition chromatin states, but not active promoters, had fewer breaks in the treated samples compared to the control (Figure 4B). To investigate the effect of RNase H1 overexpression and TOP1 knockdown on highly active regions, we looked specifically at the highest expressed genes selecting the top 1000 expressed genes from CEL-seq data that was done on those treated MCF7 cells. Remarkably, we found that highly transcribed genes exhibit a decrease in DSBs enrichment when TOP1 is depleted and/or RNase H1 is overexpressed (Figure 4C, Supplementary Figure 3D). In contrast, no significant effect was detected in low expressed genes (Figure 4C, Supplementary Figure 3D). In addition, genes enriched with R-loops displayed a significant decrease in break enrichment after RNase H1 overexpression and/or TOP1 knockdown, while no significant change was detected in genes deprived of R-loops (Figure 4D). Similarly, genes enriched with TOP1 exhibited a significant decrease in break density as compared to those which are deprived of TOP1 (Figure 4E). We next examined how the distribution of breaks changed across the gene body upon TOP1 depletion and/or RNase H1 overexpression and found that breaks density decreases in gene body and TTS, but not at TSS (Figure 4F and Supplementary Figure 3E). These findings are consistent with the observation that TOP1 is catalytically active at gene bodies but not at TSS (Baranello *et al.*, 2016).

Interestingly, both manipulations (RNase H1 OE and TOP1 KD) together did not give a complementary effect compared to each treatment separately. To further uncover this, we selected active genes that exhibited the most decrease in break density with each

treatment separately (calculated by the average of the two separate treatments) and examined break density of the three groups. Interestingly, although genes were selected based on the separate manipulations (SR or TE), break density was lowest with the combined manipulations (TR) at these genes (Supplementary Figure 3 F). Similar results were obtained when we examined breaks at overlapping genes between high R-loops/TOP1 and active genes (Supplementary Figure 3 H, J), indicating that both TOP1 and R-loops coordinate the induction of DSBs at these active genes.

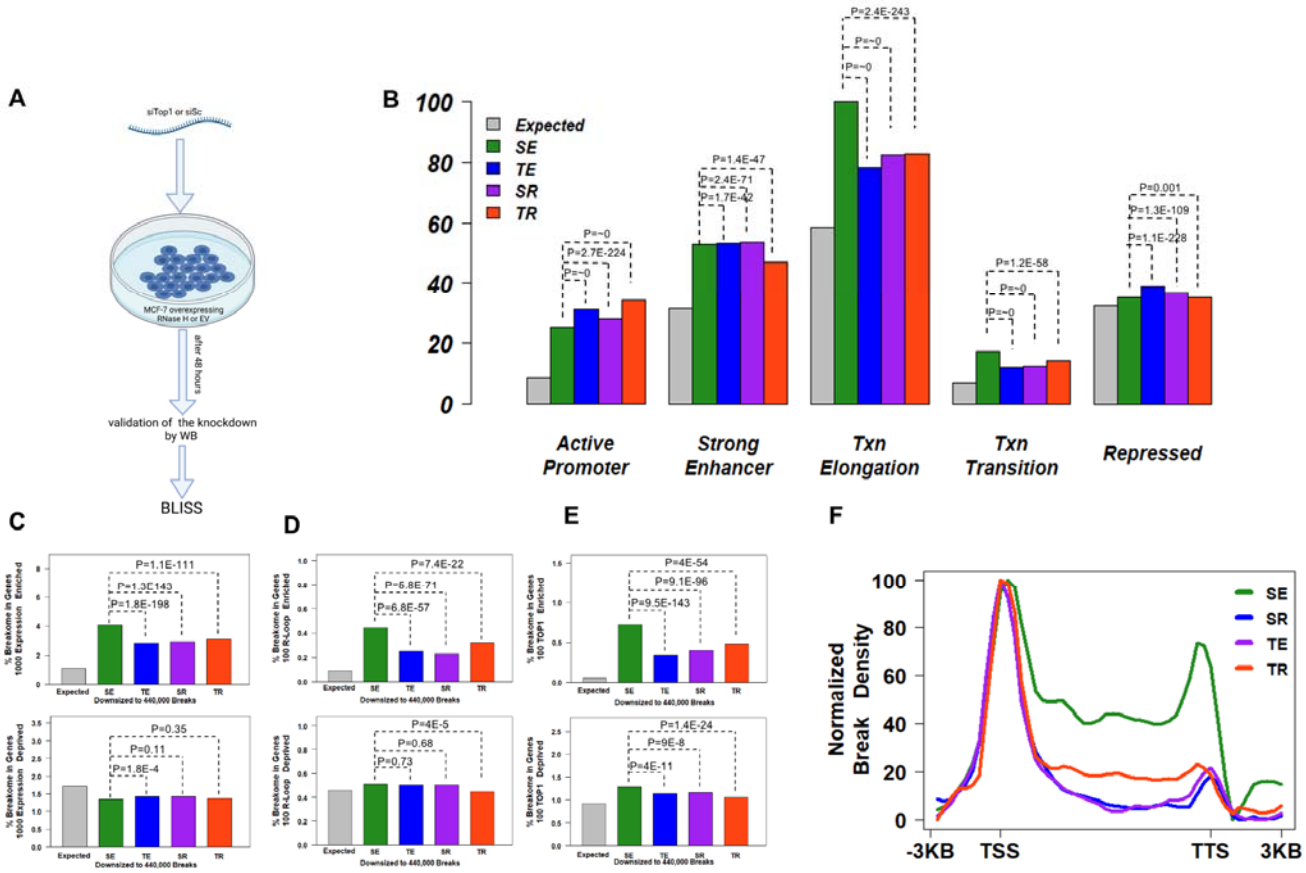



**Figure 4: RNase H1 overexpression and TOP1 knockdown decrease break enrichment at gene body of highly expressed genes. A.** Graphical illustration of the experimental workflow. **B**. The distribution of DSBs in breast cancer MCF7 cells along ChromHMM-defined chromatin states of HMEC in control cells (SE), cells knocked down for TOP1 (TE), cells overexpressing RNase H1 (SR), and cells knocked down for TOP1 and overexpressing RNase H1 (TR). Bar height is break enrichment relative to other chromatin states. **C**. Breakome percentage in 1000 most transcribed genes (top) and 1000 least transcribed genes (bottom) with the different treatments. **D**. Breakome percentage in 100 top R-loop enriched genes (top) and 100 top R-loop deprived genes (bottom) with the different treatments. **E**. Breakome percentage in 100 top TOP1 enriched genes (top) and 100 top TOP1 deprived genes (bottom) with the different treatments. **F**. Plot showing break density at gene body of the 1000 highest transcribed genes with the different treatments. sBLISS experiments were done in duplicates.

## Fragility of estradiol responsive genes is mediated by R-loops and TOP1

To test if an increase in transcription would increase global DNA damage, transcription in MCF-7 cells was induced by treatment with estradiol (E2) followed by immunofluorescence staining for γ-H2AX and 53BP1 (Figure 4A, B). As expected, we found that cells displayed more 53BP1 and γ-H2AX foci per nucleus when treated with E2 (increased transcription), and this increase is suppressed with RNase H1 overexpression or TOP1 knockdown, indicating that the increase in DNA damage upon E2 treatment is mediated by R-loops and TOP1. To further specify these results, we utilized sBLISS to map DSBs in control cells (NE/SN), cells treated with estradiol (EE/SE2), and cells overexpressing RNase H1 or knocked down for TOP1 and treated with estradiol (ER/TE2). We found that estrogen responsive genes, exhibiting significant increased expression upon E2 treatment (extracted from GSE27463) (Hah *et al.*, 2011), displayed increased break density as opposed to down-regulated genes (Supplementary Figure 4 A-D). Importantly, the majority of these genes had decreased break density upon depletion R-loops or TOP1 (ER/TE2) as compared to estradiol only, indicating that the increase in transcriptional DSBs upon E2 treatment is mediated by R-loops and TOP1.

Furthermore, selecting estrogen responsive genes which exhibited increased break density following estradiol induction showed strongly that the fragility of these genes was highly attenuated by the depletion of R-loops or TOP1 (Figure 5 E, F), further confirming the involvement of TOP1 and R-loops in transcriptional DSBs in general, and specifically estrogen associated DSBs. Moreover, disease related pathway analysis for these estrogen responsive genes using DisGenNET (Piñero *et al.*, 2021) revealed enrichment in tumor progression, neoplasm metastasis, breast carcinoma, and malignant neoplasm of breast (Supplementary Figure 4 E) implying that estrogen treatment up-regulates breast cancer-associated genes and this upregulation causes DSBs at those genes through R-loops and TOP1. Altogether, these data imply that an increase in transcription results in an increase in DSBs, and this transcriptional DSBs is associated with R-loops and TOP1.

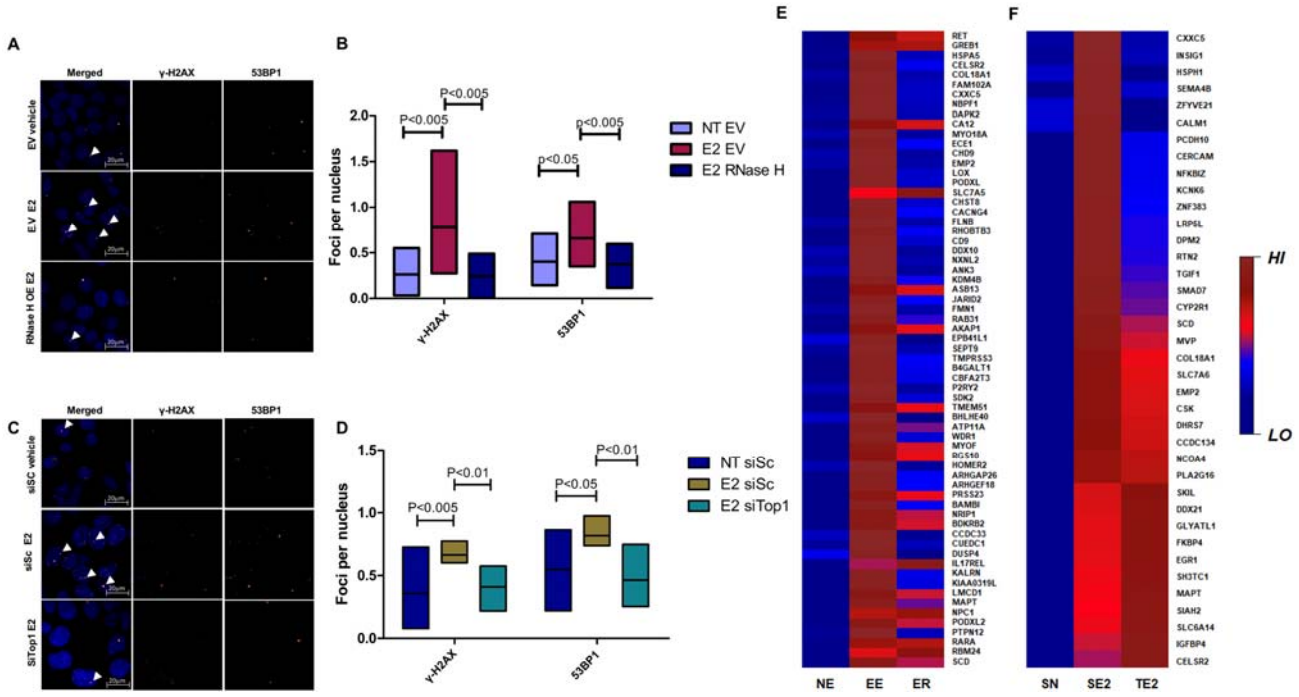



**Figure 5: Estradiol associated DSBs are mediated by R-loops and TOP1. A.** Representative images of immunofluorescent staining of MCF-7 EV non treated, MCF-7 EV treated with estradiol (E2), and MCF-7 cells overexpressing RNase H1 and treated with E2, for γ-H2AX and 53BP1. **B.** Quantification of number of γ-H2AX and p53BP1 foci per nucleus, for MCF-7 EV NT and MCF-7 EV treated with E2, and MCF-7 cells overexpressing RNase H1 treated with E2. **C.** Representative images of immunofluorescent staining of MCF-7 siSc non treated, MCF-7 siSc treated with estradiol (E2), and MCF-7 cells knocked down for TOP1 and treated with E2, for γ-H2AX and 53BP1. **D.** Quantification of number of γ-H2AX and p53BP1 foci per nucleus, for MCF-7 siSc NT and MCF-7 siSc treated with E2, and MCF-7 cells knocked down for TOP1 and treated with E2. **E.** Color coded heatmap showing the change in break density between control (NE), cells incubated with estradiol (EE), cells incubated with estradiol and overexpressing RNase H1 (ER), for estrogen responsive genes with the highest positive differential break density. **F.** Color coded heatmap showing the change in break density between control (SN), cells incubated with estradiol (SE2), cells knocked down for TOP1 (TN), for estrogen responsive genes with the highest positive differential break density. sBLISS and immunofluorescence experiments were done in duplicates.

## DSBs at highly transcribed genes might limit transcription

To make sure the decrease in breaks enrichment at highly expressed genes was not due to decreased transcriptional levels caused by the treatment itself, we examined the change in normalized expression counts of the top 1000 expressed genes with the different treatments using CEL-seq (Supplementary Figure 5 A, B). Surprisingly, cells overexpressing RNase H1 (SR) or depleted for TOP1 (TE), but not with both treatments (TR), had more genes with increased normalized expression counts indicating increased expression of the top 1000 expressed genes with treatments (62.6%, 67.7%,

and 47.3% for each treatment, respectively) (Supplementary Figure 5C-E) Furthermore, comparing gene sets sorted by expression in the control sample (same group of genes in all conditions), we found that the highest expressed group (group 1) has the largest percentage of genes with increased transcription upon treatments (Supplementary Figure 5 F). Additionally, comparing the normalized expression counts of these gene sets, we observed the highest expressed group to have the major increase in normalized expression counts (Supplementary Figure 5 G). However, when we examined the change of expression specifically at genes that exhibited the most decrease in break density the pattern seemed to be variable (Supplementary Figure 3 G, I, and K), indicating that the effect of TOP1 depletion and/or RNase H1 OE -and therefore decreased breaks- on expression to be different at different genes. These findings suggest that transcriptional DSBs, facilitated by TOP1 and R-loops, might be hampering transcription at highly expressed genes.

## TOP1 and R-loops are the main causes of DSBs at highly transcribed genes

To further investigate the contribution of TOP1 and R-loops in the formation of DSBs of highly expressed genes at physiological conditions, we compared different gene sets with similar expression levels in untreated MCF-7 cells. Interestingly, TOP1 and R-loops density in break-prone genes is greater than in break-deprived genes although both groups exhibited similar expression levels (Figure 6A, B). These results indicate that TOP1/R-loops density dictates the level of DSBs formation at actively transcribed genes. Moreover, break density in highly expressed genes that are deprived of R-loops/TOP1 is lower as compared to similarly expressed genes with high R-loops/TOP1 genes, and have slightly lower breaks as compared to low expressed genes (Figure 6C). These data clearly suggest that TOP1 and R-loops are contributing the major source of DSBs at highly expressed genes.

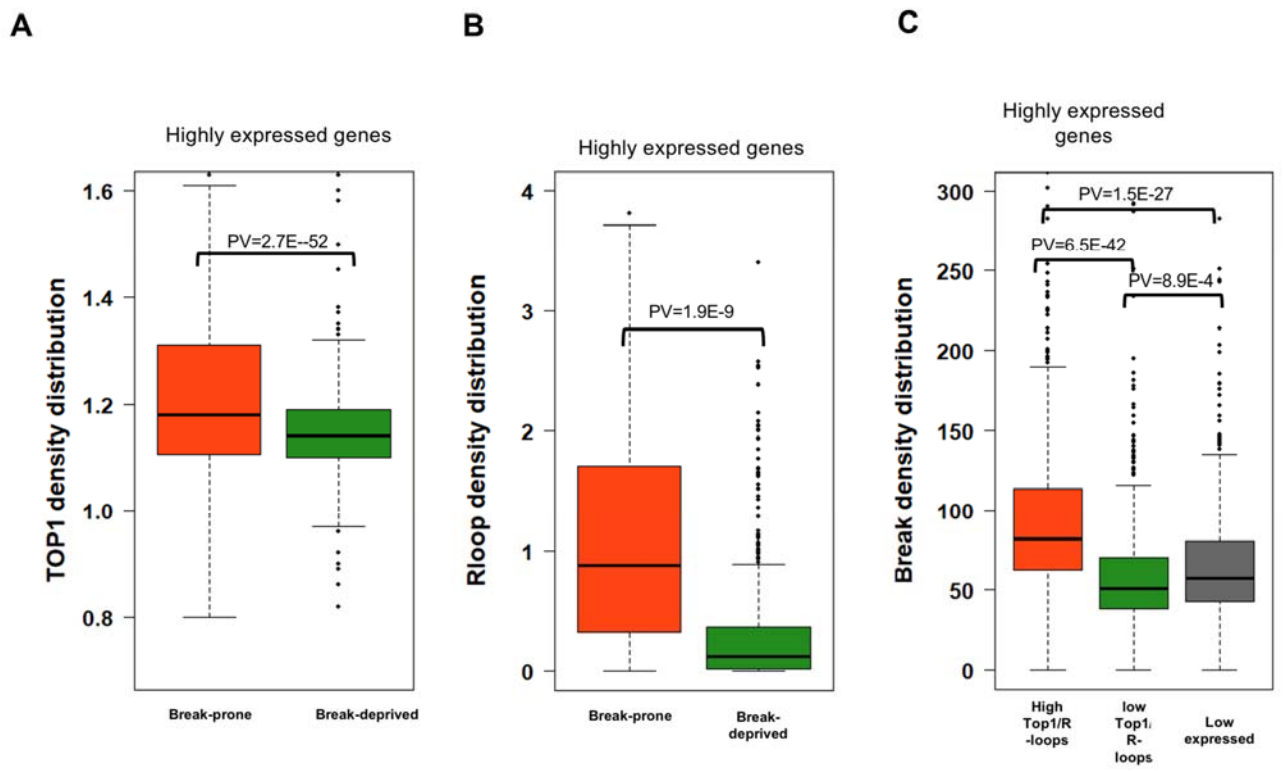



**Figure 6: TOP1 and R-loops are the main causes of transcriptional DSBs. A.** Box-plot showing TOP1 density of break-prone genes versus break-deprived genes with the expression of both groups being equated and high. **B.** Box-plot showing R-loops density of break-prone genes versus break-deprived genes with the expression of both groups being similar and high. **C.** Box-plot showing break density of high TOP1/R-loop/expression genes, low TOP1/R-loop/high expression genes, and low expressed genes. sBLISS experiments were done in duplicates.

## Transcriptional DSBs can cause mutations at cancer genes

Endogenous DSBs are a source of genomic instability and mutations that contributes to cancer (Tubbs and Nussenzweig, 2017). To investigate the involvement of transcriptional DSBs in cancer driver mutations, we examined break density change with each treatment, at well-established cancer driver mutated genes (Nik-Zainal et al., 2016; Hu et al., 2020). Interestingly, from the 97 genes that have breakome data, 77 genes (79%) exhibited a decrease in break density with TOP1 knockdown and/or RNase H overexpression (Figure 7A), with 62 genes (80%) having a decrease in both treatments. Sorting these genes according to their break enrichment shows that most of these genes exhibited a decrease in break density with RNase H1 overexpression and/or TOP1 knockdown, especially at those genes with the highest break density in the

control (Figure 7A). Moreover, these genes have a significant decrease in the total relative breaks with each treatment but not both treatments (Figure 7B). These results indicate that TOP1 and R-loops mediated transcriptional DSBs can be a cause of mutations at cancer genes.

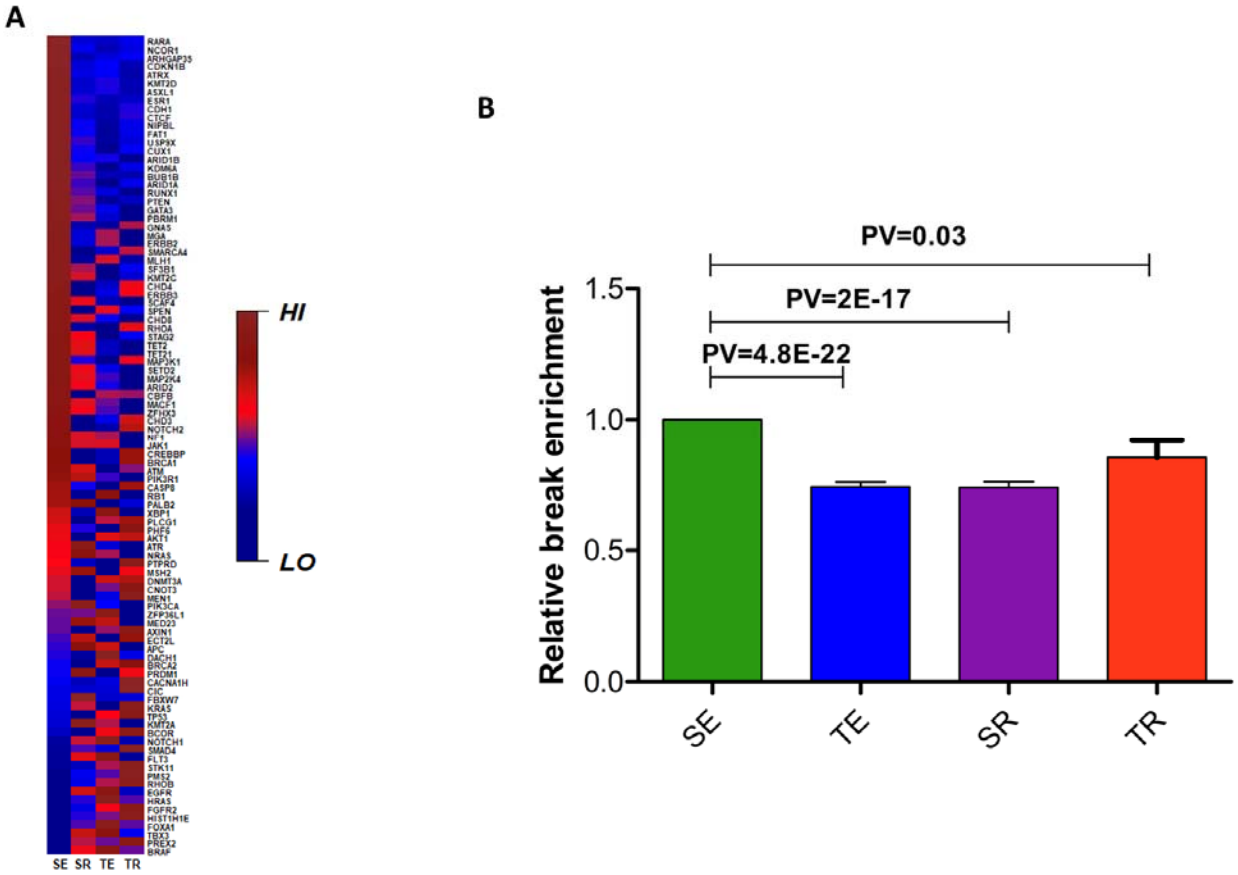



**Figure 7: RNase H1 overexpression and TOP1 knockdown decrease break enrichment at highly mutated genes in breast cancer. A.** Color coded heatmap showing the change in break density with treatments for genes with established driver mutations in breast cancer. **B.** The percentage of genes with decreased break density with treatments, in genes with known driver mutations in breast cancer. sBLISS experiments were done in duplicates.

## γ-H2AX is enriched at highly transcribed genes, mainly oncogenes

To investigate repair signaling of transcriptional DSBs, we performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) using a γ-H2AX antibody, known to mark DNA damage and repair sites (Mah, El-Osta and

Karagiannis, 2010), in MCF7 cells without exposing them to external DNA damage. The distribution of γ-H2AX highly overlapped with RNA polymerase II (Pol II) (Figure 8A) and correlated with transcription level and was enriched in highly expressed genes (Figure 8B). These findings imply that γ-H2AX may be required for high transcriptional activity, as previously suggested (Seo *et al.*, 2012). Next, we compared γ-H2AX coverage between different gene sets and found that break-prone genes display a much higher γ-H2AX enrichment as compared to break-deprived genes (Figure 8C), further validating the role of γ-H2AX in DNA damage repair. Moreover, γ-H2AX was only enriched at highly expressed genes (Figure 8D), consistent with DSBs data generated by sBLISS (Figure 1E and Supplementary Figure 3E).

To further uncover the dynamics of DSBs at well established oncogenes, we investigated break enrichment at known oncogenes that are highly active in MCF7 cells. Previously, systematic screening of genes required for tumor survival have defined a dependency score for each gene in ~500 cancer cell-lines, including MCF-7 (Tsherniak *et al.*, 2017). We found that many MCF7-dependent oncogenes like *ESR1, GATA3, NRAS, CCND1, ACTB, MYC, HES1, EGR1* and *TFF1* were very enriched with DSBs (top 5% broken genes) (Supplementary Table 1), which are significantly decreased after TOP1 KD and/or RNase H1 overexpression (Supplementary Figure 6A, B) with no significant change in normalized expression counts (Supplementary Figure 6 C). Notably, these genes were also very enriched with TOP1 (mostly top 1% of genes) (Supplementary table 2). Given the dependency of MCF-7 cells on these genes for growth and survival, and the fact that most of these genes are highly expressed (Supplementary Table 3), we hypothesized that these genes must indeed be actively repaired. To test this hypothesis, we looked at γ-H2AX enrichment at these oncogenes. Interestingly, we found that these MCF7-dependent oncogenes were indeed marked by high levels of γ-H2AX (Fig. 8E). A zoom-out view demonstrates the specificity of γ-H2AX enrichment around these genes as well as other well-characterized oncogenes like *XBP1*, *FOS* and *miR21* (Supplementary Figure 6 D).

To test whether the decreased breaks observed at these oncogenes are also accompanied by a decrease in γ-H2AX levels, we performed ChIP-qPCR for selected

genes that are most enriched with γ-H2AX (Figure 8F). Interestingly, *miR21*, *EGR1*, and *MYC all showed significant decrease in* γ-H2AX enrichment upon RNase H OE and TOP1 KD, with the highest decrease with combined manipulations (TR), consistent with our results using sBLISS (Supplementary Figure 3 F, H, and I). These results further indicate that transcriptional DSBs are caused by TOP1 and R-loops, and that their depletion decrease DSBs as detected by direct labeling with sBLISS, and ChIP-qPCR of the DNA damage marker γ-H2AX.

Moreover, gene ontology analysis of 400 genes with the highest score for γ-H2AX showed that most of these genes are involved in cellular functions like DNA repair, cell-cycle control, gene expression, translation, and response to stress (Supplementary Figure 6E) and cancer-associated signaling pathways including TP53, FOXO, HGFR, NFAT, IGF1, HIF1α, INFγ, EGFR, and RB (Supplementary Figure 6 F). These results further confirm that highly transcribed genes including oncogenes are covered by γ-H2AX marks, to facilitate efficient repair of transcriptional DSBs, and promote hypertranscription of oncogenes.

To determine how these transcriptional DSBs are being repaired, we analyzed the breakome of MCF-7 cells depleted for RAD51 (HR) or XRCC4 (NHEJ), two major proteins involved in DSB repair pathways. We found that while RAD51 depletion slightly increased break density at both TSS and gene bodies (Figure 8 G, H), XRCC4 depletion remarkably increased break density only at gene bodies of highly expressed genes but not at the TSS (Figure 8 I, J), consistent with previously published data showing the repair signaling of DSBs produced by transcription-blocking TOP1 lesions to be triggered by DNA-PK which recruits NHEJ factors (Cristini *et al.*, 2016). These results imply that transcriptional DSBs caused by TOP1 and R-loops at gene bodies might be repaired by the NHEJ pathway.

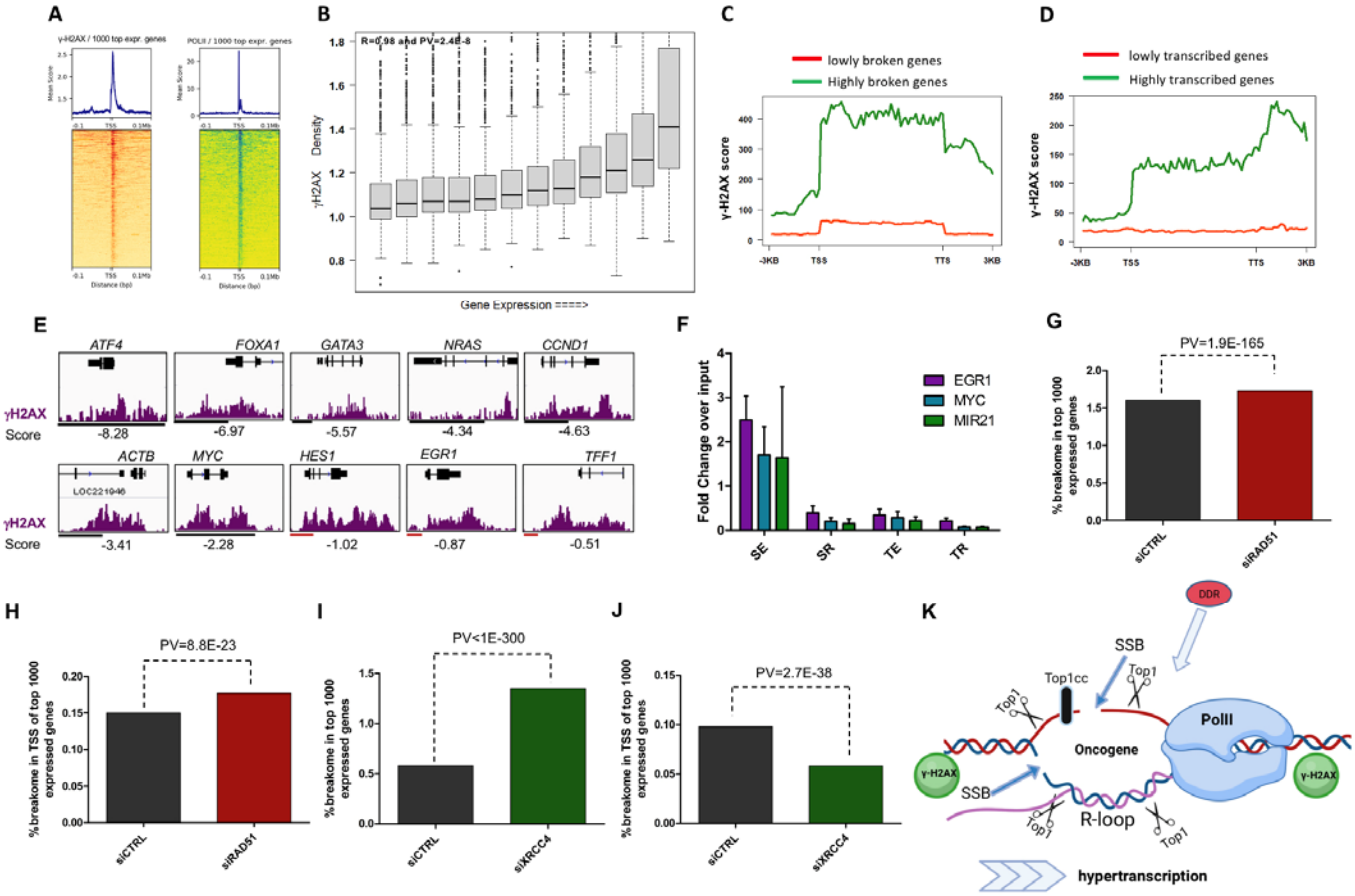



**Figure 8: Efficient repair of transcriptional DSBs is required to sustain the hypertranscription of oncogenes**. **A**. Heatmap of highly expressed genes showing overlap between γ-H2AX and RNA PolII. **B**. Positive correlation between γ-H2AX coverage and RNA levels. **C**. A plot showing γ-H2AX coverage across the gene body of break-prone genes and break-deprived genes. **D**. A plot showing γ-H2AX coverage across the gene body of highly transcribed genes and low transcribed genes. **E.** γ-H2AX coverage across breast cancer dependent oncogenes. **F.** ChIP-qPCR data for γ-H2AX at EGR1, MYC, and MIR21, in control cells (SE), cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR). **G.** %breakome in the gene body of the top 1000 expressed genes in MCF-7 cells knocked down for RAD51 and control cells. **H.** %breakome in the transcription start site (TSS) of the top 1000 expressed genes in MCF-7 cells knocked down for RAD51 and control cells. **I.** %breakome in the gene body of the top 1000 expressed genes in MCF-7 cells knocked down for XRCC4 and control cells. **J.** %breakome in the transcription start site (TSS) of the top 1000 expressed genes in MCF-7 cells knocked down for XRCC4 and control cells. **K.** Model of transcriptional DSBs formation at hyper-transcribed oncogenes and its repair. Oncogene hypertranscription by RNA PolII increases the negative supercoiling (-SC) behind the transcribing RNA PolII. This -SC increases the demand for the catalytic activity of TOP1 to resolve it, which eventually leads to the stabilization of TOP1cc and formation of a persistent SSB. The increased -SC caused by the hypertranscription also facilitate the formation of R-loops that can turn into a SSB, which together with the SSB caused by TOP1cc, could result in a DSB. To ensure efficient and sustained hypertranscription of oncogenes, oncogenes are enriched with the mark γ-H2AX to ensure strong DNA damage response (DDR) and efficient repair of transcriptional DSBs. The illustration was generated using BioRender. sBLISS and ChIP-seq experiments were done in duplicates.

## Discussion

Our data show that highly transcribed genes overlap both TOP1 and R-loops, and that R-loops and TOP1 genomic levels can predict the damage at these regions. Moreover, we showed that TOP1 knockdown or RNase H1 overexpression significantly decreases transcriptional DSBs at highly expressed genes, including oncogenes. These findings propose a model (Figure 8 K) in which DSBs in oncogenes are mainly caused by the accumulation of R-loops and the catalytic activity of TOP1. Our data also reveal that these oncogenes, despite being highly broken (fragile), are also highly expressed, and this high expression is mediated by efficient repair signaling as indicated by γ-H2AX accumulation and involvement of XRCC4.

### Transcriptional DSBs at TSS are different from gene body

Our data propose a model in which transcriptional DSBs at TSS are produced and processed differently from transcriptional DSBs at gene body. TOP1 and R-loops depletion reduced DSBs only at gene bodies of the highly expressed genes, and not at TSS, consistent with the finding that DSBs at TSS are produced by high-mobility group AT-hook 2 protein (HMGA2) which is required for transcription initiation (Dobersch *et al.*, 2021). This is also consistent with our previous observation that the pattern of DSBs is behaving differently at regulatory element including promoters, strong enhancers and insulators (Hazan *et al.*, 2019). Moreover, TOP1 and R-loop depletion slightly increased break density at promoters and enhancers, implicating that they might serve another function at these regions.

Our data also indicate that not only the origin of DSBs at promoters are different from gene body but also how these DSBs are processed, consisting with a recent report showing that promoters are prone to oxidative damage and are protected by NuMA (Ray *et al.*, 2022). These observations are proposing a model inwhich DNA damage and repair are subjected to different 'rules' at different regions across the gene.

### The coordination between TOP1 and R-loops

The mechanism by which TOP1 and R-loops cause DSBs has been previously explored (Cristini *et al.*, 2019). A mechanism was proposed in which TOP1 and R-loops cause

DSBs through the induction of two adjacent SSBs, one produced by TOP1cc processing, and the other by the activity of endonucleases which recognizes R-loops structures. In our data, although the general view of highly expressed genes points towards the previously proposed mechanism, a more focussed view on genes with the highest decreased breaks with the separate manipulations showed that these genes had the lowest break enrichment with both RNase H1 OE and TOP1 KD (Supplementary Figure 3 F, H, and J), which was also validated by γ-H2AX ChIP-qPCR (Figure 8F). These findings imply that R-loops and TOP1 cause DSBs independently of each other. A possible mechanism could be that R-loops and TOP1 themselves could accumulate SSBs at both strands leading to DSBs. Another posibility is that SSBs induced by TOP1 and R-loops could be adjacent to SSBs induced by other agents such as oxidative damage (Ray *et al.*, 2022).

## Estrogen responsiveness and DSBs

Previously, R-loops have been proposed to be the main cause of estrogen-induced DNA damage (Stork *et al.*, 2016), pointing towards a role of R-loops in transcriptional DSBs. Our sBLISS data on E2 responsive genes after TOP1 KD or RNase H1 OE shows induction of breaks at these genes after estradiol treatment, further validating a causing effect of transcription on DSBs formation. Furthermore, TOP1 KD or RNase H1 OE significantly attenuated the effect of estradiol, further validating the involvement of TOP1 and R-loops in transcriptional DSBs formation.

## The repair of transcriptional DSBs

Our results show that despite many of MCF7 dependent oncogenes are enriched in the break density list, they are also highly expressed, implying that these regions are repaired very efficiently. Our findings suggest that highly expressed genes including oncogenes, are covered by the histone mark γ-H2AX, indicating that DDR signaling is very active in these regions likely to ensure repair of the DSBs resulting from the hypertranscription of these genes. We also showed that depletion of XRCC4 increased DSBs at the gene body of highly expressed genes as compared to RAD51 knockdown,

indicating that the NHEJ repair pathway might be the main pathway for transcriptional DSBs repair at gene bodies, while HR might be more prominent at TSS.

## Acknowledgments

We are grateful for Ms. Sara Oster for help with sBLISS technique and Mr. Jonathon Monin for bioinformatic analysis. The Aqeilan's lab is funded by the European Research Council (ERC) [No. 682118] and Israel Science Foundation (ISF) [No. 1056/21].

## References

Aguilera, A. and García-Muse, T. (2012) ‘R Loops: From Transcription Byproducts to Threats to Genome Stability’, *Molecular Cell*, 46(2), pp. 115–124. doi: 10.1016/j.molcel.2012.04.009.

Aguilera, A. and Gómez-González, B. (2017) ‘DNA-RNA hybrids: The risks of DNA breakage during transcription’, *Nature Structural and Molecular Biology*. Nature Publishing Group, 24(5), pp. 439–443. doi: 10.1038/nsmb.3395.

Baranello, L. *et al.* (2016) ‘RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription’, *Cell*. Elsevier Inc., 165(2), pp. 357–371. doi: 10.1016/j.cell.2016.02.036.

Bouwman, B. A. M. *et al.* (2020) ‘Genome-wide detection of DNA double-strand breaks by in-suspension BLISS’, *Nature Protocols*. Springer US, 15(12), pp. 3894–3941. doi: 10.1038/s41596-020-0397-2.

Cerritelli, S. M. and Crouch, R. J. (2009) ‘Ribonuclease H: The enzymes in eukaryotes’, *FEBS Journal*, 276(6), pp. 1494–1505. doi: 10.1111/j.1742-4658.2009.06908.x.

Choder, M. and Aloni, Y. (1988) ‘RNA polymerase II allows unwinding and rewinding of the DNA and thus maintains a constant length of the transcription bubble’, *Journal of Biological Chemistry*. Â© 1988 ASBMB. Currently published by Elsevier Inc; originally published by American Society for Biochemistry and Molecular Biology., 263(26), pp. 12994–13002. doi: 10.1016/s0021-9258(18)37661-0.

Cristini, A. *et al.* (2016) ‘DNA-PK triggers histone ubiquitination and signaling in response to DNA double-strand breaks produced during the repair of transcription-blocking topoisomerase I lesions’, *Nucleic Acids Research*, 44(3), pp. 1161–1178. doi: 10.1093/nar/gkv1196.

Cristini, A. *et al.* (2019) ‘Dual Processing of R-Loops and Topoisomerase I Induces Transcription-Dependent DNA Double-Strand Breaks’, *Cell Reports*, 28(12), pp. 3167-3181.e6. doi: 10.1016/j.celrep.2019.08.041.

DeCaprio, J. and Kohl, T. O. (2020) ‘Immunoprecipitation’, *Cold Spring Harbor Protocols*, 2020(11), pp. 449–461. doi: 10.1101/pdb.top098509.
Dobersch, S. *et al.* (2021) ‘Positioning of nucleosomes containing γ-H2AX precedes active DNA demethylation and transcription initiation’, *Nature Communications*. Springer US, 12(1). doi: 10.1038/s41467-021-21227-y.
García-Muse, T. and Aguilera, A. (2019) ‘R Loops: From Physiological to Pathological Roles’, *Cell*, 179(3), pp. 604–618. doi: 10.1016/j.cell.2019.08.055.
Hah, N. *et al.* (2011) ‘A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells’, *Cell*. Elsevier Inc., 145(4), pp. 622–634. doi: 10.1016/j.cell.2011.03.042.
Hamperl, S. *et al.* (2017) ‘Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses’, *Cell*. Elsevier Inc., 170(4), pp. 774-786.e19. doi: 10.1016/j.cell.2017.07.043.
Hashimshony, T. *et al.* (2012) ‘CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification’, *Cell Reports*. The Authors, 2(3), pp. 666–673. doi: 10.1016/j.celrep.2012.08.003.
Hazan, I. *et al.* (2019) ‘Activation of Oncogenic Super-Enhancers Is Coupled with DNA Repair by RAD51’, *Cell Reports*. ElsevierCompany., 29(3), pp. 560-572.e4. doi: 10.1016/j.celrep.2019.09.001.
Helmrich, A., Ballarino, M. and Tora, L. (2011) ‘Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes’, *Molecular Cell*. Elsevier, 44(6), pp. 966–977. doi: 10.1016/j.molcel.2011.10.013.
Hu, Z. *et al.* (2020) ‘Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases’, *Nature Genetics*. Springer US, 52(7), pp. 701–708. doi: 10.1038/s41588-020-0628-z.
Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.
Lensing, S. V. *et al.* (2016) ‘DSBCapture: In situ capture and sequencing of DNA

breaks', *Nature Methods*, 13(10), pp. 855–857. doi: 10.1038/nmeth.3960.
Li, M. *et al.* (2015) 'RECQ5-dependent SUMOylation of DNA topoisomerase I prevents transcription-associated genome instability', *Nature Communications*. Nature Publishing Group, 6, pp. 1–13. doi: 10.1038/ncomms7720.
Liu, L. F. and Wang, J. C. (1987) 'Supercoiling of the DNA template during transcription.', *Proceedings of the National Academy of Sciences of the United States of America*, 84(20), pp. 7024–7027. doi: 10.1073/pnas.84.20.7024.
Ma, J. and Wang, M. D. (2016) 'DNA supercoiling during transcription', *Biophysical Reviews*. Biophysical Reviews, 8, pp. 75–87. doi: 10.1007/s12551-016-0215-9.
Mah, L. J., El-Osta, A. and Karagiannis, T. C. (2010) 'γh2AX: A sensitive molecular marker of DNA damage and repair', *Leukemia*. Nature Publishing Group, 24(4), pp. 679–686. doi: 10.1038/leu.2010.6.
Manzo, S. G. *et al.* (2018) 'DNA Topoisomerase I differentially modulates R-loops across the human genome', *Genome Biology*. Genome Biology, 19(1), pp. 1–18. doi: 10.1186/s13059-018-1478-1.
Mirkin, E. V. and Mirkin, S. M. (2005) 'Mechanisms of Transcription-Replication Collisions in Bacteria', *Molecular and Cellular Biology*, 25(3), pp. 888–895. doi: 10.1128/mcb.25.3.888-895.2005.
Nik-Zainal, S. *et al.* (2016) 'Landscape of somatic mutations in 560 breast cancer whole-genome sequences', *Nature*. Nature Publishing Group, 534(7605), pp. 47–54. doi: 10.1038/nature17676.
Ohle, C. *et al.* (2016) 'Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair', *Cell*, 167(4), pp. 1001-1013.e7. doi: 10.1016/j.cell.2016.10.001.
Piñero, J. *et al.* (2021) 'The DisGeNET cytoscape app: Exploring and visualizing disease genomics data', *Computational and Structural Biotechnology Journal*, 19, pp. 2960–2967. doi: 10.1016/j.csbj.2021.05.015.
Pommier, Y. *et al.* (2016) 'Roles of eukaryotic topoisomerases in transcription,

replication and genomic stability', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 17(11), pp. 703–721. doi: 10.1038/nrm.2016.111.
Pommier, Y. *et al.* (2022) 'Human topoisomerases and their roles in genome stability and organization', *Nature Reviews Molecular Cell Biology*. Springer US, 23(6), pp. 407–427. doi: 10.1038/s41580-022-00452-3.
Promonet, A. *et al.* (2020) 'Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites', *Nature Communications*. Springer US, 11(1), pp. 1–12. doi: 10.1038/s41467-020-17858-2.
Ramírez, F. *et al.* (2016) 'deepTools2: a next generation web server for deep-sequencing data analysis', *Nucleic Acids Research*, 44(W1), pp. W160–W165. doi: 10.1093/NAR/GKW257.
Ray, S. *et al.* (2022) 'A mechanism for oxidative damage repair at gene regulatory elements', *Nature*. Springer US, 609(7929), pp. 1038–1047. doi: 10.1038/s41586-022-05217-8.
Sagie, S. *et al.* (2017) 'Telomeres in ICF syndrome cells are vulnerable to DNA damage due to elevated DNA:RNA hybrids', *Nature Communications*, 8, pp. 1–12. doi: 10.1038/ncomms14015.
Seo, J. *et al.* (2012) 'Genome-wide profiles of H2AX and γ-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells', *Nucleic Acids Research*, 40(13), pp. 5965–5974. doi: 10.1093/nar/gks287.
Sollier, J. and Cimprich, K. A. (2015) 'Breaking bad: R-loops and genome integrity', *Trends in Cell Biology*. Elsevier Ltd, 25(9), pp. 514–522. doi: 10.1016/j.tcb.2015.05.003.
Stork, C. T. *et al.* (2016) 'Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage', *eLife*, 5(AUGUST), pp. 1–21. doi: 10.7554/eLife.17548.
Tsherniak, A. *et al.* (2017) 'Defining a Cancer Dependency Map', *Cell*. Elsevier Inc., 170(3), pp. 564-576.e16. doi: 10.1016/j.cell.2017.06.010.
Tubbs, A. and Nussenzweig, A. (2017) 'Endogenous DNA Damage as a Source

of Genomic Instability in Cancer', *Cell*, 168(4), pp. 644–656. doi: 10.1016/j.cell.2017.01.002.

Wilhelm, T., Said, M. and Naim, V. (2020) 'Dna replication stress and chromosomal instability: Dangerous liaisons', *Genes*, 11(6), pp. 1–35. doi: 10.3390/genes11060642.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9). doi: 10.1186/gb-2008-9-9-r137.

# Method details

## cell culture

MCF7 (HTB-22) were grown in RPMI supplemented with 10% (vol/vol) FBS (GIBCO), glutamine, and penicillin/streptomycin (Beit-Haemek). Cells were grown at 37°C under a humidified atmosphere with 5% $CO_2$. Cells were routinely authenticated by STR profiling, tested for mycoplasma, and cell aliquots from early passages were used.

## Transient transfection and plasmids

Transient transfection of siTOP1 (Dharmacon; SMARTpool siRNA, L-005278-00-0005) and siSc (Dharmacon; D-00181010) was achieved using lipofectamine (ThermoFisher). Cells were seeded to reach a confluency of 60-70% on the day of transfection. Transfection solution was prepared by mixing 1.5ml serum/antibiotic-free RPMI with 30µL lipofectamine and incubating for 5 minutes, before adding a mixture of 1.5ml RPMI and 0.6nmoles of siRNA followed by 20 minutes incubation at RT. The mixture was then added to the cells in a 10cm plate cultured in antibiotic/serum-free media. The media was replaced by fully supplemented media after 5-6 hours, and cells were incubated in the incubator for 48 hours.

For RNase H1 overexpression, RNase H1 plasmid and its empty control plasmid containing only GFP, a kind gift from Dr. Sara Selig (Molecular Medicine Laboratory, Rambam Health Care Campus and Rappaport Faculty of Medicine, Technion, Haifa 31096, Israel), were used. GFP expression was used as an indicative of RNase H1 overexpression. The plasmid was prepared as described in (Sagie *et al.*, 2017).

## Immunofluorescence assay

Cells were seeded on coverslips to a 50-70% confluency at the time of fixation. Fixation was done using 4% paraformaldehyde for 10 minutes, followed by three washes with 1X PBS. Cells were then permeabilized with 0.25% Triton X for 7 minutes followed by another three washes with 1X PBS. This was followed by blocking with 10% goat serum for 1 hour. Cells were then stained with anti-γ-H2AX antibody (abcam, ab26350) using 1:1000 dilution, and anti-53BP1 anti-body (abcam, ab36823) using 1:500 dilution, and incubated in a humidity chamber in the cold room overnight. After three washes with

PBS, cells were incubated with the secondary antibodies in a humidity chamber for one hour at room temperature, and then stained with Hoechst stain (1:2000 dilution in 1X PBS) followed by two washes with PBS. Coverslips were then mounted, and immunofluorescent pictures were taken using confocal microscopy.

## CEL-seq

Total RNA was extracted using TRI reagent (BioLabs) according to the manufacturer specifications. RNA concentration was measured using Qubit Flex Fluorometer (ThermoFisher Scientific). RNA sequencing libraries were prepared using the CEL-Seq2 protocol in the Technion genomic Center (TGC), as published by (Hashimshony *et al.*, 2012), with one modification; instead of single-cells as input, 2 ng purified RNA was taken as input for library preparation. The CEL-Seq2 libraries were analyzed for average fragment size using Agilent 2200 TapeStation (Agilent) and concentration was measured using Qubit Flex Fluorometer (ThermoFisher Scientific). The libraries were sequenced on the Illumina NextSeq 2000 sequencer (Illumina), 12 bases for read 1 and 88 bases for read 2. Demultiplexing was performed in two steps. First, Illumina demultiplexing was performed using bcl2fastq Illumina software with the following parameters: barcode-mismatches =1, minimum-trimmed-read-length = 0, and mask-short-adapter-reads = 0. Second, Cell-seq demultiplexing using the pipeline described in (Hashimshony *et al.*, 2012) was executed with the following parameters: min_bc_quality = 10, bc_length = 6, umi_length = 6, and cut_length = 88. RNA measurements, library preparation and sequencing were performed by the Technion Genomics Center, Technion, Israel.

## Western blot

Protein lysates were prepared by incubating cells in lysis buffer containing 50 mM Tris (pH7.5), 150 mM NaCl, 10% glycerol, 0.5% Nonidet P-40 (NP-40), with protease and phosphatase inhibitors (1:100). Samples were run on SDS-PAGE gel for 90-120 minutes and then blotted on nitrocellulose membrane using semi-dry blotting machine (Biorad). Two antibodies were used, Topoisomerase 1 antibody (Cat# ab3825) for

validation of siRNA knockdown, and GAPDH antibody (Calbiochem; CB1001), as a housekeeping gene antibody.

### Transcription induction using Estradiol

MCF-7 cells were hormonally starved by growing in RPMI media supplemented with 10% charcoal stripped fetal bovine serum for 48 hours. Transcription was induced by replacing the media with RPMI media supplemented with charcoal stripped FBS and 100nM β-Estradiol (Sigma-Aldrich) (pre-dissolved in ethanol), and incubated for 1 hour.

### In-suspension break labelling *in situ* and sequencing (sBLISS)

sBLISS was performed as described in (Bouwman *et al.*, 2020). Briefly, $10^6$ cells were fixed in 2% paraformaldhyde in 10%FCS/PBS for 10min at room temperature. Fixation was quenched with 125 mM glycine for 5min at RT, and another 5min on ice, followed by two washes in ice cold PBS. The cells were lysed for 60min on ice and the nuclei were permeabilized for 60min at 37°c. Then, nuclei were washed twice with CutSmart Buffer supplemented with 0.1% Triton X-100 (CS/TX100), and DSB ends were in situ blunted with NEB's Quick Blunting Kit for 60min at RT. Blunted nuclei were washed twice with 1x CS/TX100 before proceeding with in situ ligation of sBLISS adapters to the blunted DSB ends. Adaptor ligation was performed with T4 DNA Ligase for 20-24h at 16C and supplemented with BSA and ATP. After ligation, nuclei were washed twice with 1x CS/TX100 and genomic DNA was extracted with Proteinase K at 55°c for 14-18h while shaking at 800rpm. Afterward, Proteinase K was heat-inactivated for 10 min at 95°c, followed by extraction using Phenol:Chloroform:Isoamyl Alcohol, Chloroform, and ethanol precipitation. The purified DNA was sonicated in 100 µL ultra pure water using Covaris M220 for 60s. Sonicated samples were concentrated with AMPure XP beads (Beckman Coulter) and fragment sizes were assessed using a BioAnalyzer 2100 (Agilent Technologies) to range from 300bp to 800bp with a peak around 400-600bp. The sonicated DNA was in vitro transcribed using MEGAscript T7 Kit for 14 h at 37°c. After RNA purification and ligation of the 3'-Illumina adaptor, the RNA was reverse transcribed. The final step of library indexing and amplification was performed using NEBNext® Ultra™ II Q5® Master Mix.

## Chromatin immunoprecipitation (ChIP) and ChIP-seq

MCF7 cells (~$10^6$) were crosslinked with 1% formaldehyde (methanol free, Thermo Scientific 28906) for 10 min at room temperature and quenched with glycine, 125 mM final concentration. Fixed cells were washed twice in PBS and incubated in sonication buffer (0.5M NaCl, 0.5%SDS RIPA buffer containing PMSF, protease and phosphatase inhibitors) for 30min on ice. Cells were sonicated using Covaris M220 for 10min to produce chromatin fragments of ~200-300 bp. The sheared chromatin was centrifuged 10min at maximum speed. From the supernatant, 50 µl were saved as input DNA and the rest was diluted in RIPA buffer without NaCl and SDS. The chromatin was immunoprecipitated by incubation with 5 µg of gamma-H2AX (phospho S139; Abcam, AB2893) for 5 mg of TOP1 antibodies (ab3825, ab85038) antibody pre mixed with 100 µl Dynabeads Protein G (ThermoFisher, 10004D) and incubated over-night at 4°C with rotation. Immunoprecipitates were washed twice with RIPA buffer + 150mM NaCl, twice with RIPA buffer + 300 mM NaCl, and twice with TE buffer. The chromatin was eluted from the beads with 200 µl of direct elution buffer (10mM Tris pH 8, 0.3M NaCl, 5mM EDTA, 0.5% SDS) and incubated overnight at 65°C to reverse the cross-linking. Samples were treated with RNAse for 1h at 37 °C and with proteinase K at 55 °C for 2 h. DNA was cleaned up by QIAquick PCR purification column (Qiagen), according to the manufacturer's instructions, and eluted in 30 µl of elution buffer. The ChIPed and the Input DNA were used for real-time PCR or to prepare libraries by Truseq (Ilumina) and sequenced in Nextseq (illumine).

## Bioinformatic analysis and data processing

### sBLISS analysis

sBLISS fastq files are first de-multiplexed by sample barcodes. Quality control is applied with trim_galore to remove residual adapters, trim reads to base quality of at least 20 and filter out short reads with size smaller than 20 bp. Initial and final sample qualities are evaluated with fastqc. Quality processed fastq files are aligned to GRCh38 assembly with hisat2 and then sorted and indexed with samtools. Resulting bam files are de-duplicated with umi-tools employing genomic coordinates and Unique Molecular

Identifiers (UMI). Custom python and R scripts are applied to identify read start position and convert bam files to break bigwig format for downstream analysis. A custom R script is applied to discard a blacklist of positions (mainly within centromeres). For most analyses, breaks are aggregated over genomic tiles using R tileGenome utility.

### CEL-Seq analysis

Quality control is applied on CEL-Seq fastq files with trim_galore to remove residual adapters, polyA tails, trim reads to base quality of at least 20 and filter out short reads with size smaller than 20 bp. Initial and final sample qualities are evaluated with fastqc. Quality processed fastq files are then aligned to GRCh38 transcriptome using salmon aligner in it mapping mode while enabling the write_mapping key to create a sam file of transcripts. The sam file is deduplicated with a custom python script which scans for primary alignments only and deduplicates with respect to the transcript and Unique Molecular Identifiers (UMI). Non-duplicated primary alignments and all their associated secondary alignments are kept. Salmon aligner is executed again in its alignment mode, with sam deduplicated file as input resulting in a de-duplicated transcript count table.

### DRIP-Seq analysis

DRIP-Seq peak file is downloaded from GEO datasets - GSE81851. The peaks selected for analysis are for T0-Input after liftover to GRCh38.

### TOP1 ChIP-Seq analysis

MCF7 TOP1 ChIP-Seq treated and input fastq files are quality controlled with trim_galore to remove residual adapters, trim reads to base quality of at least 20 and filter out short reads with size smaller than 20 bp. Initial and final sample qualities are evaluated with fastqc. The quality processed fastq files are aligned to GRCh38 genome with hisat2 followed by sorting and indexing with samtools. The utility bamCompare of deepTools is applied on the bam files to create a peak bigwig file. This operation is set to ignore duplicates, discard a blacklist of regions, normalize by RPKM, and compare by ratio in tiles of 50 bp

## γ-H2AX ChIP-seq data processing

The sequenced files, containing 84bp single-ended reads, were trimmed for Illumina adapter and for base quality, keeping Phred score above 20, with trim galore utility of Babraham Bioinformatics. Reads were then aligned to the human genome (GRCh38) with Bowtie2 (v2.2.9) (Langmead and Salzberg, 2012) keeping the default end-to-end scoring coefficients. Coverage files, both bigwig and bedgraph, were created with the bamCompare utility of deeptools (Ramírez *et al.*, 2016) by comparing the treated BAM file with the control input file while ignoring duplicates. We have also created our own black list of regions, mainly consisting of chromosomal centromeres, and filtered the results accordingly.

Peak calling was performed with the MACS2 tool (Zhang *et al.*, 2008). Prior to applying peak calling, filtered bed files were created from the aligned SAM files (for both treated and control) using a custom script. The filter kept reads exhibiting high mapping quality of 42 (highest figure for for Bowtie2) and an average base quality of at least 32. For peak calling, the treated file was compared the control file while ignoring duplicates, setting fold limits at their default of 5-50 but lowering q-value to a stricter figure of 0.01.

## Algorithms used for plotting the data

Heatmaps were plotted using the following pipeline. First deepTools computeMatrix is applied with center reference-point. Next, an R script discards outlier spikes from the matrix. Finally, deepTools plotHeatmap is performed. In two-part plots, each part is defined as an independent region.

Regions of chromatin states are based on GSE57498 for HMEC cell line. For each of the regions, breaks are counted within that region and compared between treated samples or between treated sample and an expected count. The expected count is calculated as the proportion of breaks expected due to regions size relative to overall genome size. Plots are created with base R functions.

Figures showing break percentage at high expression/ R-loops/ TOP1 were generated using the following pipeline, Downsized breaks are counted across the gene body for all genes-of-interest (e.g., 1000 topmost expressed genes) and summed up for each of the treated samples. Downsizing is performed to reduce possible bias related to library

preparation. An expected break value is also calculated based on the relative occupancy of the genes-of-interest relative to overall genome size. Bar-plots are created with base R functions.

For plots showing breaks distribution across gene body, Each gene-of-interest is divided into 30 successive regions and breaks are counted separately in each region. In addition, for each gene, two flanks of size 3000 bp are subdivided into 10 regions and breaks are counted in the flanks. The plot shows break counts for several treated samples after normalization of each curve between 0 and 100.

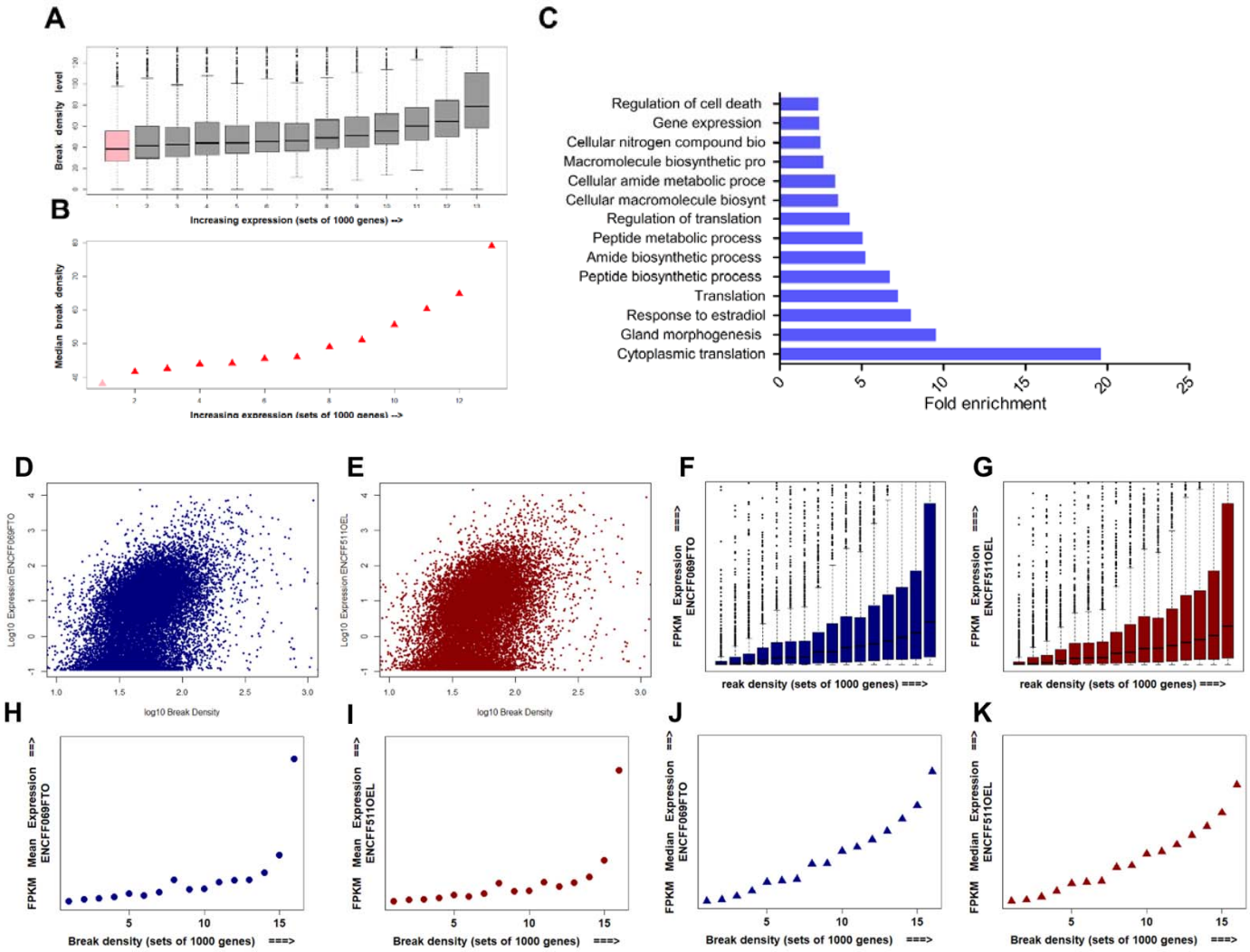
For color coded heatmaps, Break density is assigned to each gene-of-interest (cancer related genes) and plotted as a heatmap after ranking by SE sample. Heatmap is created with heatmap.2 function of R gplots package.

Figures with bars and boxplots, and log-log scatter plots are created using base R functions.

## Data availability

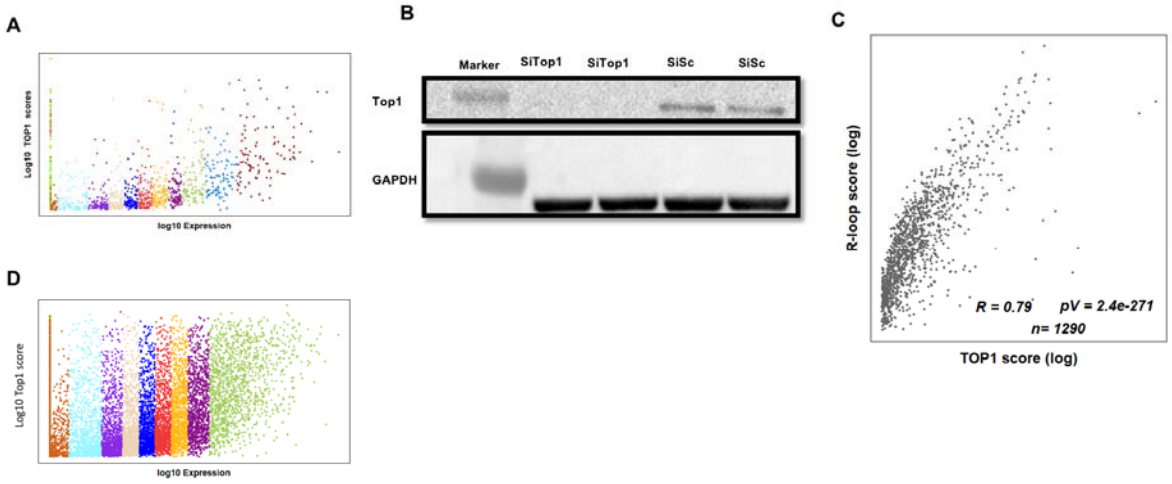
Raw data are available upon request.

# Supplementary Figures



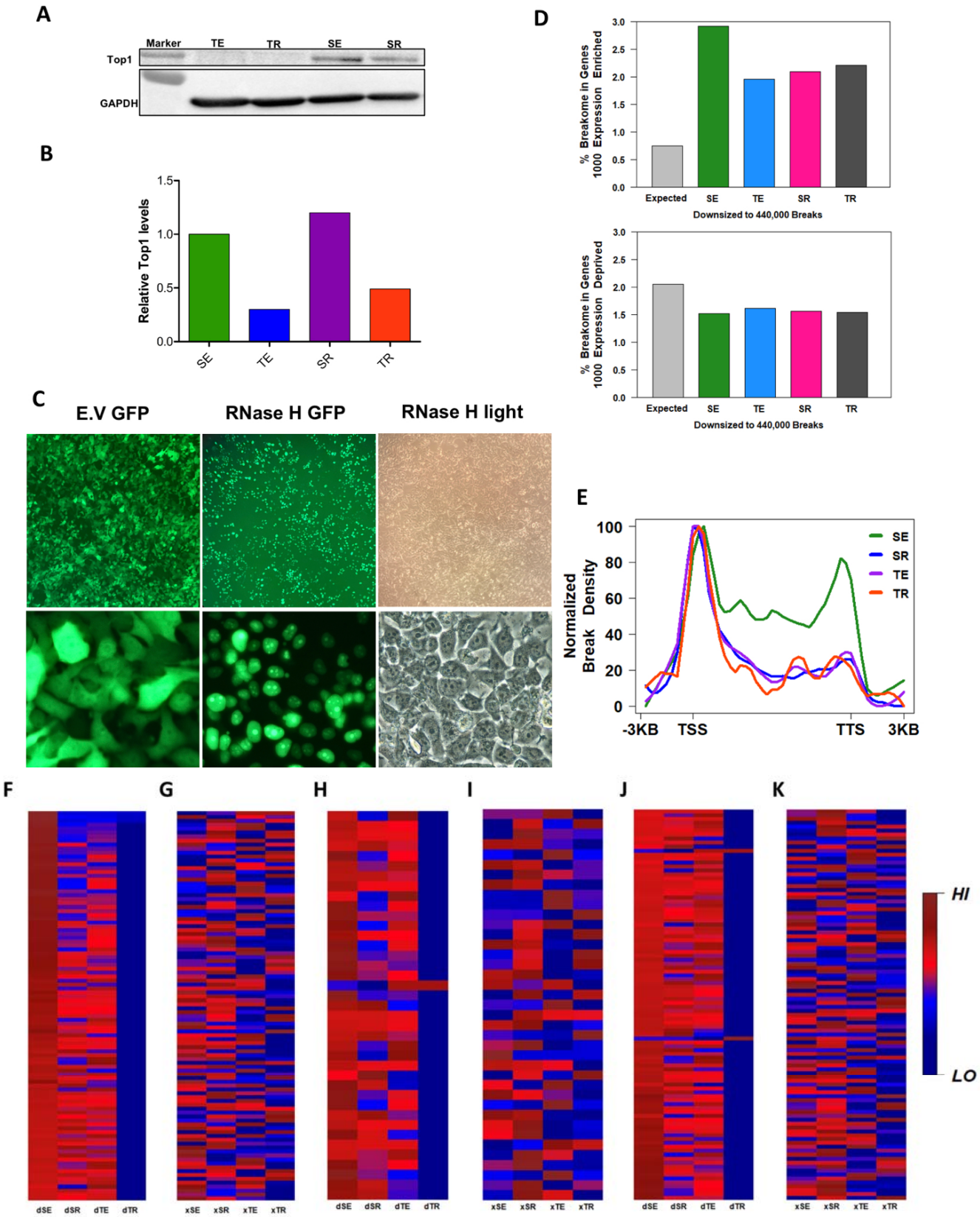

**Supplementary Figure 1: DSBs are enriched at highly expressed genes**. **A**. Boxplot showing the correlation between break density and expression level, genes were grouped into 13 groups with increasing expression, the 13th group being the group with the highest mean expression. **B.** Median of break density positively correlates with break density. A zoomed log scaled scatter plot showing the correlation between expression levels and TOP1 level, each dot represents a gene. **C**. Gene ontology analysis of the 400 top broken genes in MCF-7 cells. **D** and **E**. A zoomed log scaled scatter plot showing

a positive correlation between break density and expression levels; each dot represents a gene, and for each gene break density and expression were measured and normalized to gene size. **F** and **G**. Boxplot showing the correlation between break density and expression level. **H** and **I.** Mean expression (TPM) positively correlates with break density. **J** and **K.** Median expression (TPM) positively correlates with break density.



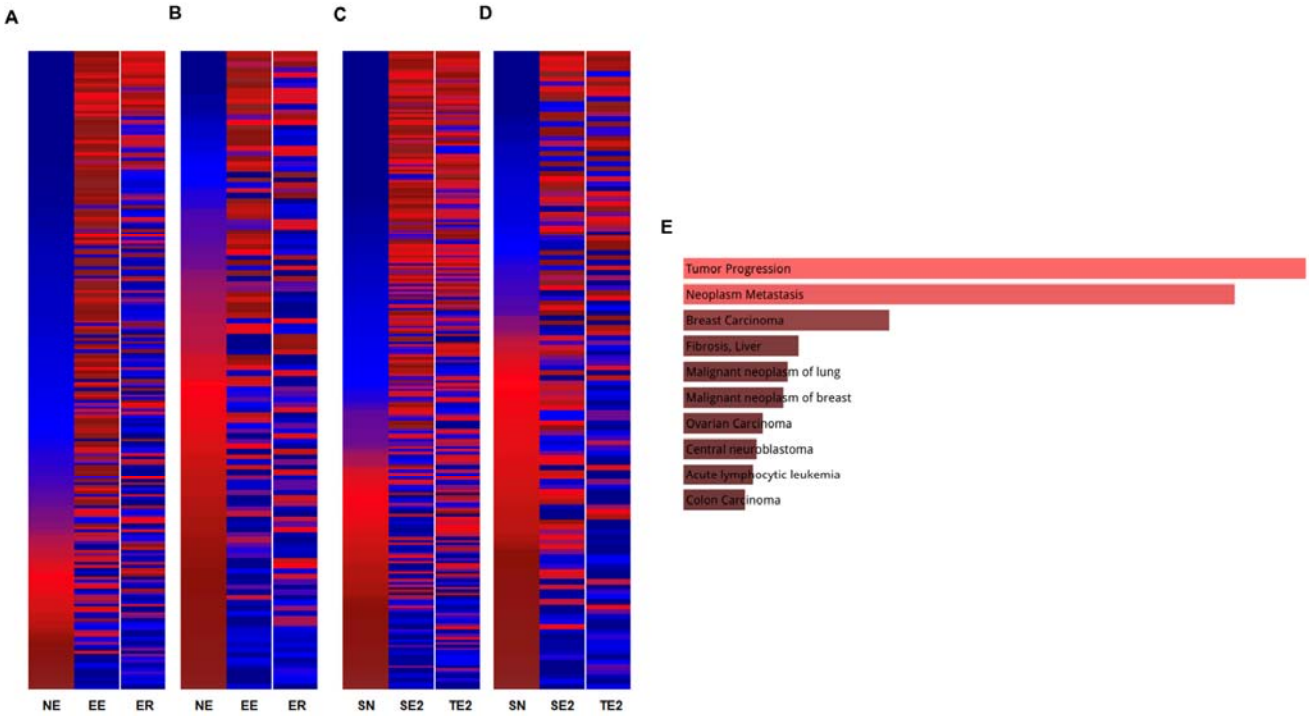

**Supplementary Figure 2: Transcriptional DSBs are associated with TOP1 and R-loops. A.** A log scaled scatter plot showing positive correlation between expression and TOP1 levels; each dot represents a gene, and each color represents a gene group. **B.** Validation of TOP1 knockdown by Western Blot. **C**. A log scaled scatter plot showing the correlation between R loops and TOP1 levels, each dot represents a gene. DRIP data was extracted from GSE81851. **D.** A log scaled scatter plot

showing positive correlation between expression and R-loop levels; each dot represents a gene, and each color represents a gene group.



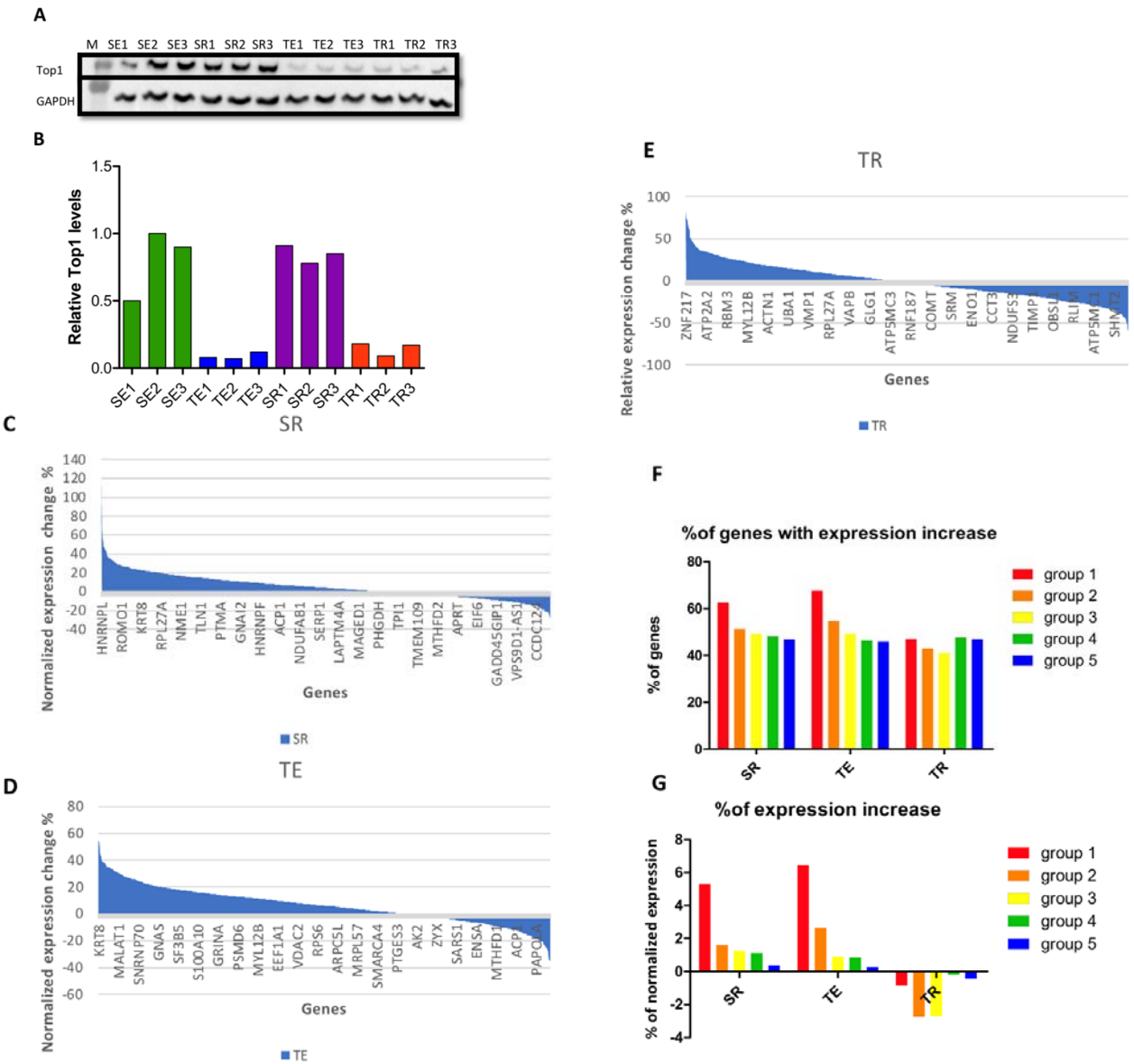

**Supplementary Figure 3: RNase H1 overexpression and TOP1 knockdown decrease break**

**enrichment at gene body of highly expressed genes. A**. Validation of TOP1 knockdown using western blotting. **B.** Quantification of relative TOP1 levels after TOP1 knockdown. **C.** GFP pictures for RNase H1-GFP overexpression. **D.** Breakome percentage in 1000 most transcribed genes (top) and 1000 least transcribed genes (bottom) with the different treatments, expression data was extracted from encode projects ENCFF069FTO_FPKM and ENCFF511OEL_FPKM for MCF7. **E.** Plot showing break density at gene body of the 1000 highest transcribed genes with the different treatments. P-values were calculated using chi-square test. **F.** Color coded heatmap showing the change in break density between control (SE) cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR). **G.** Color coded heatmap showing the change in normalized expression between control (SE) cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR). **H.** Color coded heatmap showing the change in break density between control (SE) cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR). **I.** Color coded heatmap showing the change in normalized expression between control (SE) cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR). **J.** Color coded heatmap showing the change in break density between control (SE) cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR). **K.** Color coded heatmap showing the change in normalized expression between control (SE) cells overexpressing RNase H (SR), cells knocked down for TOP1 (TE), and both manipulations (TR).

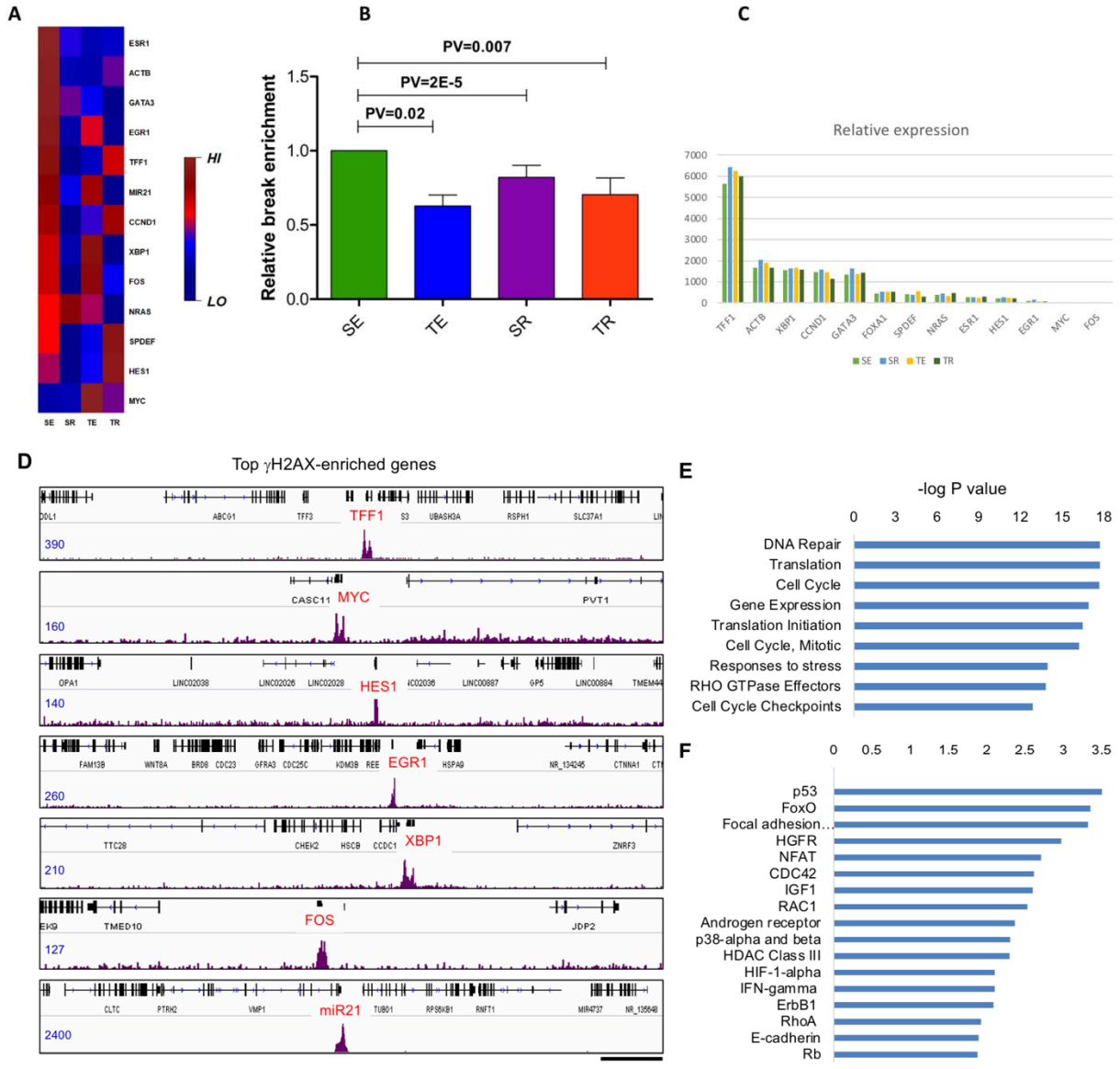



**Supplementary Figure 4: Estradiol associated DSBs are mediated by R-loops and TOP1. A.** and **B.** Color coded heatmap showing the change in break density between control (NE), cells incubated with estradiol (EE), cells incubated with estradiol and overexpressing RNase H1 (ER), for estrogen positive

responsive genes (E) and negative responsive genes (F). **C.** And **D.** Color coded heatmap showing the change in break density between control (SN), cells incubated with estradiol (SE2), cells knocked down for TOP1 (TN), for estrogen positive responsive genes (G) and estrogen negative responsive genes (H). **E.** Disease related pathways for estradiol responsive genes using DisGeNET database (Piñero *et al.*, 2021).




**Supplementary Figure 5: Highly expressed genes exhibited an increase in normalized expression counts after RNase H1 overexpression and TOP1 knockdown. A.** Validation for TOP1 knockdown using Western Blotting. **B.** Quantification of relative TOP1 levels after TOP1 knockdown. **C, D,** and **E**. Area chart showing the proportion of genes with an increase or decrease in the normalized expression counts with each treatment. **F**. Percentage of genes that exhibited an increase in their normalized expression counts with RNase H1 overexpression (SR), TOP1 knockdown (TE), or both treatments (TR), the top 5000 genes were sorted into 5 groups (1000 gene each) according to expression, the first group being the group with the top 1000 expressed genes. **G**. Percentage of normalized expression counts change with RNase H1 overexpression (SR), TOP1 knockdown (TE), or both treatments (TR), the top 5000 genes were sorted into 5 groups (1000 gene each) according to expression, the first group being the group with the top 1000 expressed genes.




**Supplementary Figure 6: efficient repair of transcriptional DSBs is required to sustain the hypertranscription of oncogenes**. **A**. Heatmap showing the change in break enrichment in MCF-7 dependent oncogenes. **B.** Relative break enrichment change of breast cancer dependent oncogenes with the different treatments. **C.** Normalized expression counts change of breast cancer dependent oncogenes with the different treatments. **D.** A zoom-ed out view of γ-H2AX coverage at breast cancer dependent oncogenes. **E** and **F.** Gene ontology analysis of 400 gene with the highest score of γ-H2AX enrichment.

| Gene | Break density rank | % |
|---|---|---|
| MIR21 | 90 | 0.368083105 |
| ACTB | 105 | 0.429430289 |
| XBP1 | 230 | 0.940656824 |
| HES1 | 262 | 1.071530817 |
| EGR1 | 285 | 1.165596499 |
| CCND1 | 313 | 1.280111243 |
| MYC | 356 | 1.455973171 |
| TFF1 | 484 | 1.979469142 |
| NRAS | 682 | 2.789251973 |
| GATA3 | 702 | 2.871048219 |
| FOS | 762 | 3.116436956 |
| ESR1 | 1395 | 5.705288127 |
| SPDEF | 1550 | 6.33920903 |

**Supplementary Table 1: Rank of breast cancer dependent genes in break density:** rank was extracted from MCF-7 breakome, and % of rank was calculated as follows: %=gene rank/all genes' number *100%. Total genes number is 24451.

| Oncogenes | TOP1 density rank | % |
|---|---|---|
| HES1 | 2 | 0.007440199 |
| EGR1 | 6 | 0.022320598 |
| MIR21 | 13 | 0.048361296 |
| XBP1 | 30 | 0.111602991 |
| TFF1 | 35 | 0.130203489 |
| ACTB | 40 | 0.148803988 |
| ESR1 | 45 | 0.167404486 |
| FOS | 71 | 0.264127079 |
| GATA3 | 85 | 0.316208474 |
| MYC | 104 | 0.386890369 |
| CCND1 | 192 | 0.714259142 |
| SPDEF | 4232 | 15.74346192 |
| NRAS | 7683 | 28.58152598 |

**Supplementary Table 2: Rank of breast cancer-dependent genes in TOP1 density:** rank was extracted from MCF-7 TOP1 ChIP-seq, and % of rank was calculated as follows: %=gene rank/all genes' number with >0 density*100%. Total genes number is 26881.

| Gene | Normalized expression counts rank | % |
|---|---|---|
| TFF1 | 1 | 0.006404509 |
| ACTB | 39 | 0.249775842 |
| XBP1 | 47 | 0.301011912 |
| CCND1 | 52 | 0.333034456 |
| GATA3 | 60 | 0.384270526 |
| FOXA1 | 584 | 3.740233124 |
| NRAS | 616 | 3.945177405 |
| SPDEF | 727 | 4.656077879 |
| ESR1 | 1423 | 9.113615986 |
| HES1 | 1706 | 10.92609197 |
| EGR1 | 2896 | 18.54745741 |
| MYC | 7969 | 51.03753042 |
| FOS | 9403 | 60.221596 |

**Supplementary Table 3: Rank of breast cancer-dependent genes in normalized expression counts:** rank was extracted from MCF-7 CEL-seq, and % of rank was calculated as follows: %=gene rank/all genes' number *100%. Total number of genes is 15614.