Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations

Gaetano Ivan Dellino[1,2,10*], Fernando Palluzzi[1,8,10], Andrea Maria Chiariello[3], Rossana Piccioni[1], Simona Bianco[3], Laura Furia[1], Giulia De Conti[1], Britta A. M. Bouwman[4], Giorgio Melloni[1,9], Davide Guido[5], Luciano Giacò[1], Lucilla Luzi[1], Davide Cittaro[6], Mario Faretta[1], Mario Nicodemi[3,7], Nicola Crosetto[4] and Pier Giuseppe Pelicci[1,2*]

It is not clear how spontaneous DNA double-strand breaks (DSBs) form and are processed in normal cells, and whether they predispose to cancer-associated translocations. We show that DSBs in normal mammary cells form upon release of paused RNA polymerase II (Pol II) at promoters, 5′ splice sites and active enhancers, and are processed by end-joining in the absence of a canonical DNA-damage response. Logistic and causal-association models showed that Pol II pausing at long genes is the main predictor and determinant of DSBs. Damaged introns with paused Pol II-pS5, TOP2B and XRCC4 are enriched in translocation breakpoints, and map at topologically associating domain boundary-flanking regions showing high interaction frequencies with distal loci. Thus, in unperturbed growth conditions, release of paused Pol II at specific loci and chromatin territories favors DSB formation, leading to chromosomal translocations.

Double-strand breaks (DSBs) are generated by external genotoxic agents or cell-intrinsic processes, including DNA replication, transcription or oxidative stress[1,2], and pose continuous challenges for the maintenance of cell viability and genome integrity. If unrepaired, in normal mammalian cells DSBs activate intracellular checkpoints involving the p53 tumor suppressor, leading to senescence or apoptosis[3–6]. Alternatively, misrepair of the DSB ends (following inappropriate processing or joining) may lead to mutations or chromosomal aberrations, increasing the risk of neoplastic transformation[7].

Much of our knowledge about this processing derives from conditional induction of DSBs at specific genomic sites using restriction enzymes, zinc-finger nucleases or CRISPR/Cas-based endonucleases[8–12]. DSB formation triggers a DNA-damage response (DDR) that involves recruitment of the MRE11–RAD50–NBS1 complex at DSBs, ataxia telangiectasia mutated (ATM)-dependent histone H2AX phosphorylation (γH2AX)[13] and recruitment of DSB-repair proteins, such as XRCC4 (non-homologous end-joining; NHEJ), PARP1 (alternative end-joining; alt-EJ) and RAD51 (homologous recombination). Homologous recombination occurs largely during S and G2 phases, when the undamaged sister template is available[14], allowing more faithful repair than NHEJ, which is active throughout the cell cycle.

Less is known about the formation and processing of endogenous DSBs. Emerging evidence suggests that DSBs can be generated by endogenous nucleases, such as topoisomerases, and accumulate at discrete genomic sites, including transcriptionally active genes or chromosome loop anchors[15]. Topoisomerases induce transient DSBs following transcriptional activation to resolve the DNA supercoiling that accumulates ahead of and behind the transcription machinery[16]. In normal and transformed cells, TOP2B-induced DNA breaks have been documented at enhancers or gene promoters following transcriptional activation by different signals[17–21], and they have been linked to chromosomal translocations[18,22].

A number of technologies allow genome-wide mapping of DSBs at base-pair (bp) resolution, either indirectly, through the identification of translocated DSBs[23], or directly, through sequencing DSB ends[24–26]. Application of these technologies to the analyses of normal cells grown in unperturbed conditions revealed the presence of hundreds to thousands of persistent DSBs at discrete genomic sites suggesting that, although normal cells are proficient for the activation of p53-mediated apoptosis or senescence[26,27], they tolerate persistent DSBs. In normal epidermal keratinocytes[26] and primary neural stem/progenitor cells[27], persistent DSBs at transcription start sites (TSSs) showed a tight association with high levels of transcription, suggesting that steady-state transcription also favors DSB formation. It remains unclear, however, how transcription induces DSBs at specific genomic regions in normal cells in unperturbed conditions, how they are processed and whether they predispose to cancer-associated translocations.

Results

Endogenous DSBs accumulate at promoters and active enhancers. Endogenous DSBs were mapped in the diploid mammary epithelium

[1]Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Milan, Italy. [2]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [3]Department of Physics, University of Naples Federico II, and INFN Complesso di Monte Sant'Angelo, Naples, Italy. [4]Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [5]Neurology, Public Health and Disability Unit, Foundation IRCCS Neurological Institute Carlo Besta, Milan, Italy. [6]Center for Translational Genomics and Bioinformatics, IRCCS San Raffaele Hospital, Milan, Italy. [7]Berlin Institute of Health, MDC-Berlin, Berlin, Germany. [8]Present address: Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy. [9]Present address: Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [10]These authors contributed equally: Gaetano Ivan Dellino, Fernando Palluzzi. *e-mail: gaetano.dellino@ieo.it; piergiuseppe.pelicci@ieo.it

cells MCF10A-AsiSIER[28], which express the 4-hydroxytamoxifen (4-OHT)-inducible AsiSI endonuclease. The 4-OHT induced formation of γH2AX foci, p53 activation and cell-cycle arrest (Supplementary Fig. 1). Under these conditions, BLISS (Breaks Labeling In Situ and Sequencing)[29] identified ~1% of the undigested AsiSI sites and ~70% of the AsiSI-digested sites, identified by chromatin immunoprecipitation sequencing (ChIP-seq) of DSB-sensing and -repair factors (NBS1, γH2AX, XRCC4 and RAD51; Fig. 1a and Supplementary Figs. 2 and 3). Thus, BLISS allows mapping of exogenously induced DSBs at the expected genomic locations.

In addition to the digested AsiSI sites, control and 4-OHT-treated cells showed ~8,100 DSB hotspots (Tier1 endogenous DSBs), largely overlapping (~70%) with the ~8,000 Tier1 DSBs from an independent experiment (Fig. 1b–f and Supplementary Table 1). Among the 3,975 DSBs (~50%) mapping within 3,046 genes ($P=0.98$; Fig. 1g,h), 634 mapped within 627 promoters (fragile promoters; $P=2.2\times10^{-40}$; Supplementary Table 2), with a bell-shaped DSB distribution around the position +700 bp from the TSS (Fig. 1i). H3K4me3, H3K4me1 and H3K27ac ChIP-seq identified an additional 48 DSBs at non-annotated TSSs (Fig. 1e,j and Supplementary Fig. 4a), and 818 DSBs within 799 active enhancers ($P=7.86\times10^{-104}$; Fig. 1d,f,k, Supplementary Fig. 4a and Supplementary Table 2). Thus, endogenous DSBs identified in proliferating DDR-proficient mammary epithelial cells accumulate preferentially at promoters and active enhancers.

Pol II is paused at the TSS of fragile promoters. We investigated whether endogenous DSBs at fragile promoters are associated with transcription. The observed-to-expected ratio of fragile promoters in genes with increasing transcription levels (measured by global run-on sequencing (GRO-seq): class 1–4 genes) showed over-representation of fragile promoters among moderately and highly transcribed genes (classes 3 and 4) and under-representation in the transcriptionally inactive class 1 genes (Supplementary Tables 3 and 4). However, ~86% of the moderately to highly transcribed genes ($n=9,843$ out of 11,387) did not contain Tier1 DSBs, and half of the DSB-free genes ($n=870$ out of 1,731) were moderately to highly transcribed, suggesting that although transcription might favor DSB formation, it is not sufficient.

We then analyzed ChIP-seq levels of the Ser 5-phosphorylated isoform of RNA polymerase II (Pol II-pS5), which associates with co-transcriptional capping and early transcriptional elongation[30]. Levels of Pol II-pS5 at TSSs correlated with BLISS signals (Supplementary Fig. 5) and increased progressively from class 1 to class 4 at both fragile and control promoters (Fig. 2a). Pol II-pS5 was markedly higher at fragile promoters (Supplementary Fig. 5) within each transcription class (Fig. 2a), despite the presence of comparable levels of transcription (Fig. 2b). To test whether this was due to Pol II pausing during early steps of transcriptional elongation, we calculated the Pol II pausing index for each gene by computing the ratio of promoter to gene body signals of either GRO-seq or Ser 2-phosphorylated Pol II (Pol II-pS2, which is associated with productive transcriptional elongation)[31]. Consistently, the Pol II pausing index at the fragile promoters of class 3 and 4 genes was higher than at controls (Fig. 2c,d). We could not measure the pausing index of class 1 and 2 genes, owing to the extremely low signal-to-noise ratio of the Pol II-pS2 ChIP-seq or GRO-seq profiles. However, Pol II-pS5 was also extremely high in the fragile promoters of the inactive class 1 genes, compared to same-class promoters (Fig. 2a), suggesting the presence of paused Pol II. These results demonstrate that paused Pol II is a common and unique feature of fragile promoters.

TOP2B accumulates specifically at fragile promoters. Efficient release of Pol II pausing requires transient DSB formation mediated by TOP2B, as observed upon transcriptional activation of inducible genes[19,20]. We investigated whether Pol II pausing correlates with recruitment of topoisomerases to fragile promoters during steady-state transcription. Treatment with TOP2B or TOP1 inhibitors (etoposide or camptothecin, respectively) showed recruitment of TOP2B, and to a lesser extent TOP1, specifically to fragile promoters (Fig. 3a,b) at sites of Pol II pausing and DSB accumulation. In control promoters TOP1 was not detectable, while TOP2B was extremely low in all transcriptional classes (Fig. 4a), consistent with the low levels of TOP2B found at promoters of inducible genes before transcriptional activation[20]. In all cases, levels of topoisomerases increased with increasing transcription (Fig. 4a) and correlated with levels of Pol II-pS5 (Fig. 4b and Supplementary Fig. 6). Thus, we found high and positively correlating levels of topoisomerases and Pol II-pS5 specifically at fragile promoters, suggesting that both Pol II pausing and topoisomerase activity contribute to endogenous DSB formation.

We next investigated whether DSB formation at fragile promoters is critical for transcription. Inhibition of TOP2B by etoposide ($t=1$ h) induced mild modifications of expression of genes with either control or fragile promoters, with more downregulated genes among those with fragile promoters ($P=1.36\times10^{-5}$; Supplementary Fig. 7). However, when RNA-seq analyses were restricted to the first exons, differences in transcriptional downregulation between fragile and control promoters increased markedly ($P=6.85\times10^{-9}$; Fig. 4c), suggesting that TOP2B is required for the early events of transcription elongation at genes with fragile promoters under steady-state conditions.

DSBs at fragile promoters recruit XRCC4 and PARP1. To investigate DDR and DNA repair at endogenous DSBs, we analyzed NBS1, γH2AX, XRCC4, PARP1 and RAD51 ChIP-seq signals across the four transcription classes. Fragile promoters showed very low (class 4) or no (classes 1–3) NBS1 signal, and no γH2AX enrichment, compared to input DNA (Figs. 3b and 4a). XRCC4 and PARP1, but not RAD51, were found at fragile promoters and showed gradual increase from poorly to highly transcribed genes, mirroring Pol II-pS5 and TOP2B signals (Figs. 3b and 4a). Thus, endogenous DSBs at fragile promoters, regardless of the transcription levels of the associated genes, are processed by proteins of the c-NHEJ or alt-EJ repair pathway and, unlike exogenously induced DSBs, do not elicit a canonical DDR.

TOP2B, XRCC4 and Pol II-pS5 interact in intact cells. Pol II and c-NHEJ proteins, including XRCC4, are part of the same multiprotein complex[32]. We investigated the proximity of Pol II-pS5, XRCC4 and TOP2B in intact nuclei, and their association with γH2AX, using the in situ proximity ligation assay (PLA; spatial resolution of 50–80 nm)[33]. PLA- and γH2AX-fluorescence images were analyzed by robotized microscopy and processed by Automated Image Cytometry[34]. We obtained a total of ~4×10^5 XRCC4–TOP2B PLA sites and ~1×10^6 γH2AX foci (Fig. 5a) by analyses of ~12,700 nuclei (Fig. 5b,c, Supplementary Fig. 8 and Supplementary Table 5). Strikingly, almost no increase in local intensity of the γH2AX signal was found at XRCC4–TOP2B PLA sites, compared to the rest of the nucleus (ratio of 1.1; Supplementary Fig. 8o). Consistently, only a minority of the XRCC4–TOP2B PLA signals mapped in close proximity (<400 nm) to γH2AX foci (Fig. 5d,e and Supplementary Table 5). Similar results were obtained for the TOP2B–Pol II-pS5 PLA sites (Supplementary Fig. 9 and Supplementary Table 5). Thus, transient interactions exist among TOP2B, XRCC4 and Pol II-pS5, despite differences in binding kinetics (Fig. 3a,b), and are consistent with accumulation of Pol II-pS5, TOP2B and XRCC4 at fragile promoters in the absence of detectable γH2AX enrichment, as observed by ChIP-seq.

To investigate whether the identified XRCC4–TOP2B interaction sites are associated with DNA replication or transcription, we performed concomitant analyses of DNA content, DNA

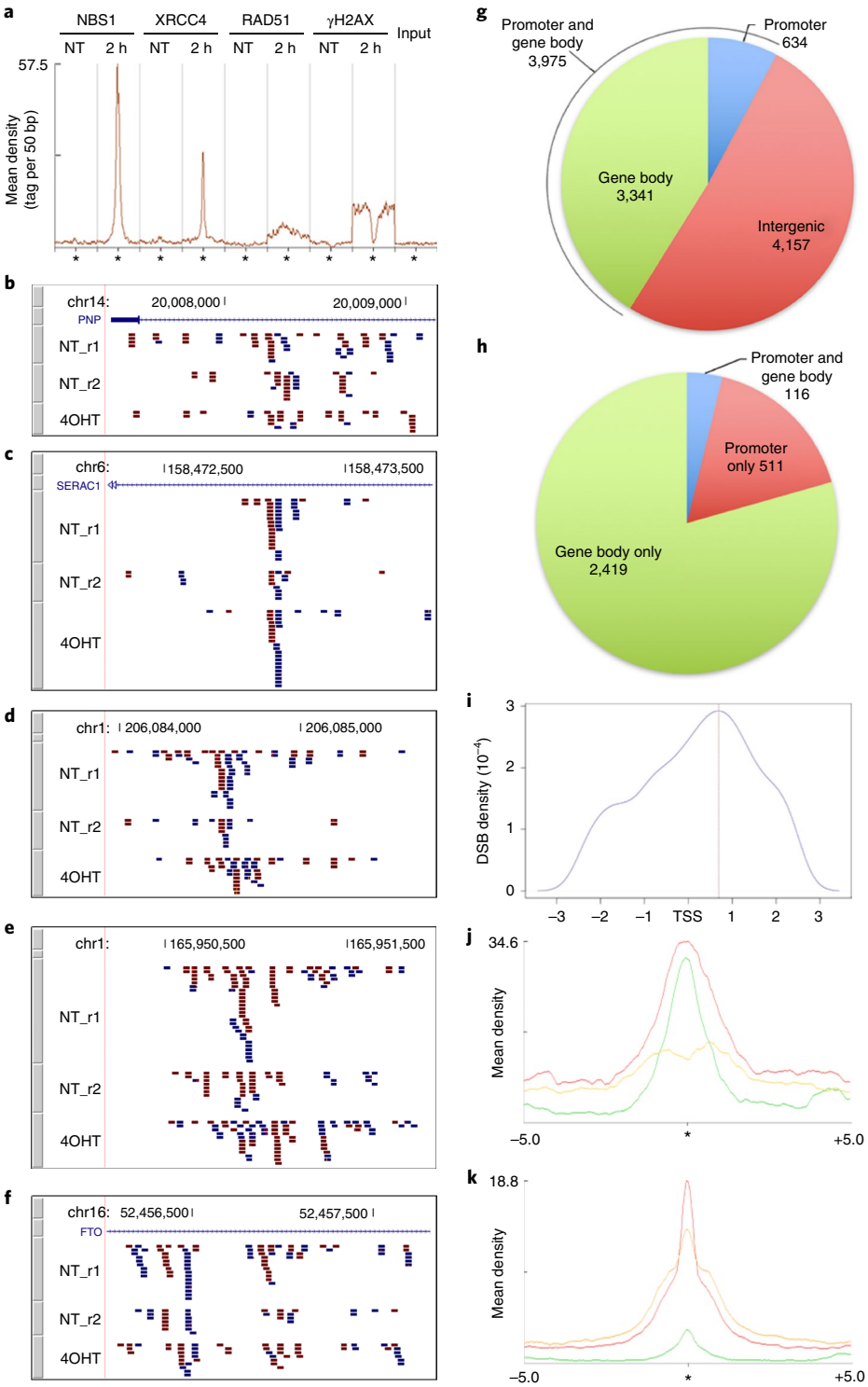

Fig. 1 | Genome-wide mapping of digested AsiSI sites and endogenous DSBs in diploid mammary epithelial cells. a, Mean-density profiles (±5 kb from the AsiSI sites, indicated with stars) of NBS1, XRCC4, RAD51 and γH2AX normalized ChIP-seq signals, and input DNA (signal intensity is indicated on the *y* axis) from a representative cluster of digested AsiSI sites, in untreated (NT) and 4-OHT-treated ($t = 2$ h) cells. **b**–**f**, Screenshots from the UCSC Genome Browser showing representative Tier1 endogenous DSBs (clusters of enriched BLISS reads, red and/or blue, over total and local background, showing ≥2-fold enrichment of NBS1, XRCC4 or RAD51 ChIP-seq signals over input DNA) from two biological replicates (r1 and r2) of untreated and 4-OHT-treated MCF10A-AsiSIER cells, mapping within the promoter (±2.5 kb from TSS) (**b**), gene body (**c**), intergenic region (**d**), non-annotated TSS (**e**) or active enhancer (**f**). **g**, Pie chart of the 8,132 endogenous Tier1 DSBs mapping within promoters, gene bodies or intergenic regions, as indicated. DSBs were enriched within promoters and enhancers, but not within genes (Supplementary Fig. 4b and Supplementary Table 2). **h**, Pie chart of the 3,046 damaged genes containing endogenous DSBs only within their promoter or gene body, or within both promoter and gene body, as indicated. **i**, DSB distribution around the TSS of fragile promoters. Summit (vertical red dashed line) at +695 bp from the TSS. Seventy per cent of promoter DSBs ($n = 441$ out of 634) mapped downstream of the TSS. **j**,**k**, Mean-density profiles (tag per 50 bp) of H3K4me3, H3K4me1 and H3K27ac ChIP-seq signals ±5 kb from the H3K4me3 peak summit (*) at non-annotated TSSs (**j**) or the H3K27ac peak summit (*) at damaged enhancers (**k**).

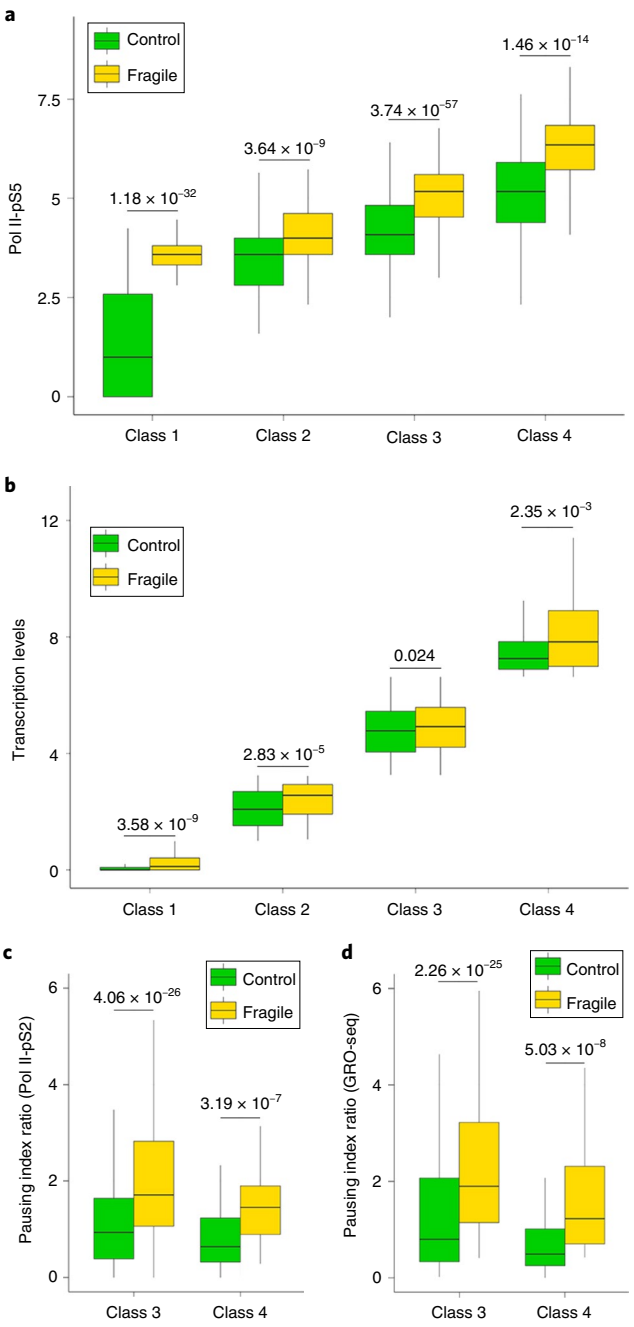

Fig. 2 | Pol II-pS5 accumulates at fragile promoters. a, Box plots showing levels of Pol II-pS5 measured by ChIP-seq at ±1 kb from the TSS of class 1–4 control (n = 417, 444, 757, 113, respectively) and fragile (n = 81, 99, 367, 80, respectively) promoters. P values (one-sided Wilcoxon rank-sum test) as indicated. **b**, Box plots showing transcription (GRO-seq) levels of genes associated with class 1–4 control and fragile promoters shown in panel **a**. P values (one-sided Wilcoxon rank-sum test) as indicated. **c**,**d**, Box plots showing the pausing index ratio, as measured by Pol II-pS2 ChIP-seq (**c**) or GRO-seq (**d**) of class 3 and 4 control (n = 870) and fragile (n = 447) promoters. P values (one-sided Wilcoxon rank-sum test) as indicated. In the box plots of panels **a**–**d**, the center line, box edges and whiskers indicate the median, upper and lower quartiles, and 1.5× interquartile range, respectively.

synthesis, transcription, γH2AX foci and XRCC4–TOP2B PLA sites (Supplementary Fig. 8). As reported[34], γH2AX foci were distributed across the cell cycle, with a marked prevalence in S phase (Fig. 5f and Supplementary Table 5). PLA sites were instead uniformly distributed (median of 23 and 32 sites per cell in G1 and S, respectively; Fig. 5f and Supplementary Table 5) with a low degree of proximity to γH2AX foci in all phases, particularly in G1 (Fig. 5f and Supplementary Table 5), suggesting that the vast majority of the XRCC4–TOP2B interactions form in the absence of DNA replication. Transcription, measured as mean fluorescence intensity per unit area, was homogeneously distributed across the cell cycle and in the nuclear space, including the volume occupied by the XRCC4–TOP2B PLA sites (ratio < 1.1 in all phases; Supplementary Fig. 8q), consistent with findings that c-NEHJ proteins form a multiprotein complex with Pol II (ref. [32]). Thus, the majority of the XRCC4–TOP2B interaction sites are not proximal to either γH2AX foci or nuclear sites with increased levels of local transcription, and are distributed across the cell cycle.

Pol II pausing, TOP levels and gene length predict DSBs. We then investigated mechanism(s) of Pol II-pS5 and topoisomerase enrichment at fragile promoters. Topoisomerases are required for resolution of the topological tension conferred by long transcripts[35,36] or two Pol II complexes at closely spaced promoters (bidirectional transcription)[37]. Bidirectional transcription was enriched at the fragile promoters ($P = 2.08 \times 10^{-4}$; Supplementary Fig. 10), and genes with fragile promoters were significantly longer than controls in all transcription classes (Fig. 5g). Genes with class 4 fragile promoters were the shortest, yet they were longer than their control counterparts (median length of ~5 and <3 kilobases (kb), respectively; Fig. 5g), suggesting that 3 kb is the critical threshold for DSB accumulation in human cells, as shown in yeast[35].

We then asked which factors are predictive of DSB formation at gene promoters (Supplementary Note) and found: (1) highly significant association of promoter DSBs with gene length, Pol II-pS5, TOP2B and TOP1 (using nested logistic models; Supplementary Table 6), (2) ~85% accuracy of DSB prediction by these four factors (using a random-forest classifier; Supplementary Table 7); (3) gene length as the main DSB predictor (using the mean decrease in Gini impurity index (MDG); Fig. 6a and Supplementary Table 8), consistently with the finding that etoposide-induced downregulation of transcription from the most expressed fragile promoters was much stronger for the first exons of the longest genes (Fig. 4c). Thus, gene length, Pol II-pS5 and topoisomerases together are distinguishing features of the vast majority of the DSB-containing promoters, and are sufficient to predict DSB occurrence.

Pol II-pS5 release is the main determinant of DSB formation. To investigate mechanism(s) of DSB formation, we modeled cause–effect relationships among the identified factors (Supplementary Fig. 11, Supplementary Tables 9–14 and Supplementary Note) and found that: (1) levels of Pol II-pS5 are causally associated with DSB accumulation and (2) this association is either direct or indirect, through topoisomerases (using a structural equation model; Fig. 6b). In both cases, however, the causal association of Pol II-pS5 with XRCC4 was consistently stronger than with PARP1 (Fig. 6b). Thus, Pol II-pS5 pausing is the main determinant of DSB formation at fragile promoters, either directly or through topoisomerases.

To investigate how Pol II pausing causes DSBs, we tested the hypothesis that their formation is necessary for the transition of paused Pol II into productive elongation. MCF10A cells were incubated with 5,6-dichloro-1-β-ribofuranosil benzimidazole (DRB), which inhibits the CDK9 kinase (responsible for Pol II Ser 2-phosphorylation) and prevents productive transcription[38,39], as observed after short DRB treatment (Supplementary Fig. 12a). Concurrently, Pol II-pS5 increased at most of the fragile promoters and at a fraction of controls (Fig. 6c,d), consistent with the inhibitory effect of DRB on transcriptional elongation and subsequent Pol II pausing (Supplementary Fig. 12b). DRB induced the concomitant loss, or

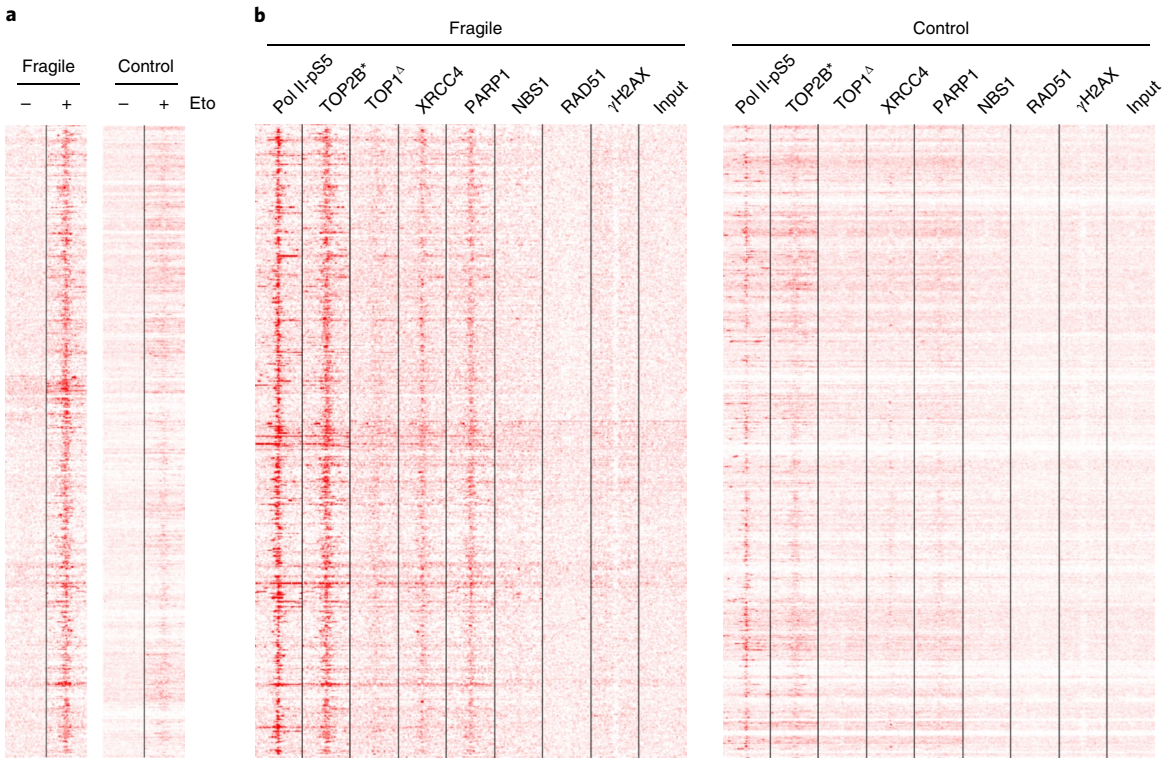

Fig. 3 | Effect of etoposide (Eto) treatment on TOP2B ChIP-seq signals, and ChIP-seq signals of Pol II-pS5 and DNA-repair factors around the TSSs of fragile and control promoters. Heat map of genomic distribution around the TSS (±5 kb) of: TOP2B-normalized ChIP-seq signals before (–) and after (+) etoposide administration (5 μM, 1 h) at the 627 fragile and 1,731 control promoters (**a**); Pol II-pS5, TOP2B, TOP1, XRCC4, PARP1, NBS1, RAD51 and γH2AX normalized ChIP-seq signals, and input DNA, at fragile and control promoters in MCF10A-AsiSIER cells as indicated (**b**). Asterisks, triangles: ChIP-seq assays performed with etoposide- or camptothecin-treated cells, respectively.

strong reduction, of XRCC4 at fragile promoters (Fig. 6c,d) and BLISS signal at both fragile promoters and associated gene bodies (Fig. 6e), thus demonstrating that Pol II pausing per se is not sufficient to induce DSBs. Similar results were obtained using flavopiridol, another CDK9 inhibitor (Supplementary Fig. 13). Following DRB removal, Pol II-pS5 signals at TSSs dropped to steady-state levels, with the concomitant full recovery of XRCC4 signals at fragile promoters (Fig. 6c,d and Supplementary Fig. 12c), showing that the release of Pol II pausing is directly involved in DSB formation. Weak XRCC4 signals appeared after DRB removal also at promoters of ~50% of highly transcribed control genes (class 2–4; $n = 631$ out of 1,313) compared to steady-state levels (Supplementary Fig. 12c). The same genes also showed a slight, yet significant increase of the Pol II pausing index following DRB administration (Supplementary Fig. 12d) suggesting that DSB formation generally occurs at all sites of Pol II release. Together, these data demonstrate that release at sites of Pol II pausing (either induced by DRB at control promoters or occurring physiologically at fragile promoters) is the main determinant of DSB formation.

Although at lower levels than at fragile promoters, Pol II-pS5 was detectable at fragile enhancers (DSB-positive; Fig. 7a) but not at control active enhancers (Supplementary Fig. 14). Enhancers are characterized by lower transcription levels than promoters and decreased paused Pol II stability[40]. However, fragile enhancers and fragile promoters showed similar responses to topoisomerase inhibitors and distribution of DNA-repair proteins, with the exception of weak γH2AX signals, which spread only a few kb from fragile enhancers (Fig. 7a,b). Notably, fragile enhancers showed similar responses to DRB administration/removal (Fig. 7c,d). Thus, Pol II pause/release and TOP2B contribute to formation of DSBs also at fragile enhancers.

Fragile introns have DSBs at 5′ splice sites or active enhancers. The vast majority of translocation breakpoints map within introns. All the MCF10A gene-associated DSBs mapped within introns (D^+ introns), including those associated with promoters, namely introns with their 5′ splice site within +2.5 kb from the TSS (promoter D^+ introns; $n = 554$ out of 2,143; Supplementary Tables 15 and 16). Thus, we investigated the association of the 5′ splice site of promoter D^+ introns with DSBs and Pol II-pS5/TOP2B/XRCC4. While the Pol II-pS5 signal was sharp and symmetrically distributed around the TSS of control promoters (approximately ±600 bp), at fragile promoters it was wider (approximately ±2 kb), slightly shifted downstream of the TSS and enriched at the 5′ splice site of promoter D^+ introns (Supplementary Fig. 15a,b). TOP2B showed similar profiles, while XRCC4 and DSBs extended further downstream, with ~45% of promoter D^+ introns showing DSBs at <2.5 kb from the 5′ splice site (Fig. 1i and Supplementary Figs. 15c,d and 16a). Unfortunately, the proximity of 5′ splice sites to TSSs did not allow unambiguous separation of 5′ splice site- and TSS-specific signals. However, when only the 5′ splice sites of distal-promoter D^+ introns were considered (introns with 5′ splice site > 0.6 kb from the TSS; $n = 95$ out of 554; Supplementary Table 16), Pol II-pS5, TOP2B and XRCC4 signals appeared as a second peak, clearly distinct from the TSS-associated one (Fig. 8a). DRB treatment induced marked accumulation of Pol II-pS5 and disappearance of XRCC4 at these 5′ splice sites, while drug removal decreased Pol II-pS5 and increased XRCC4 (Fig. 8a), as observed at fragile promoters and enhancers. Visual inspection confirmed the presence of a second peak of Pol II-pS5, TOP2B and XRCC4 at the 5′ splice site (Supplementary Fig. 16b). Thus, paused Pol II release at the 5′ splice site of promoter introns contributes to DSB formation within fragile promoters. Promoter D^+ introns are particularly long, suggesting

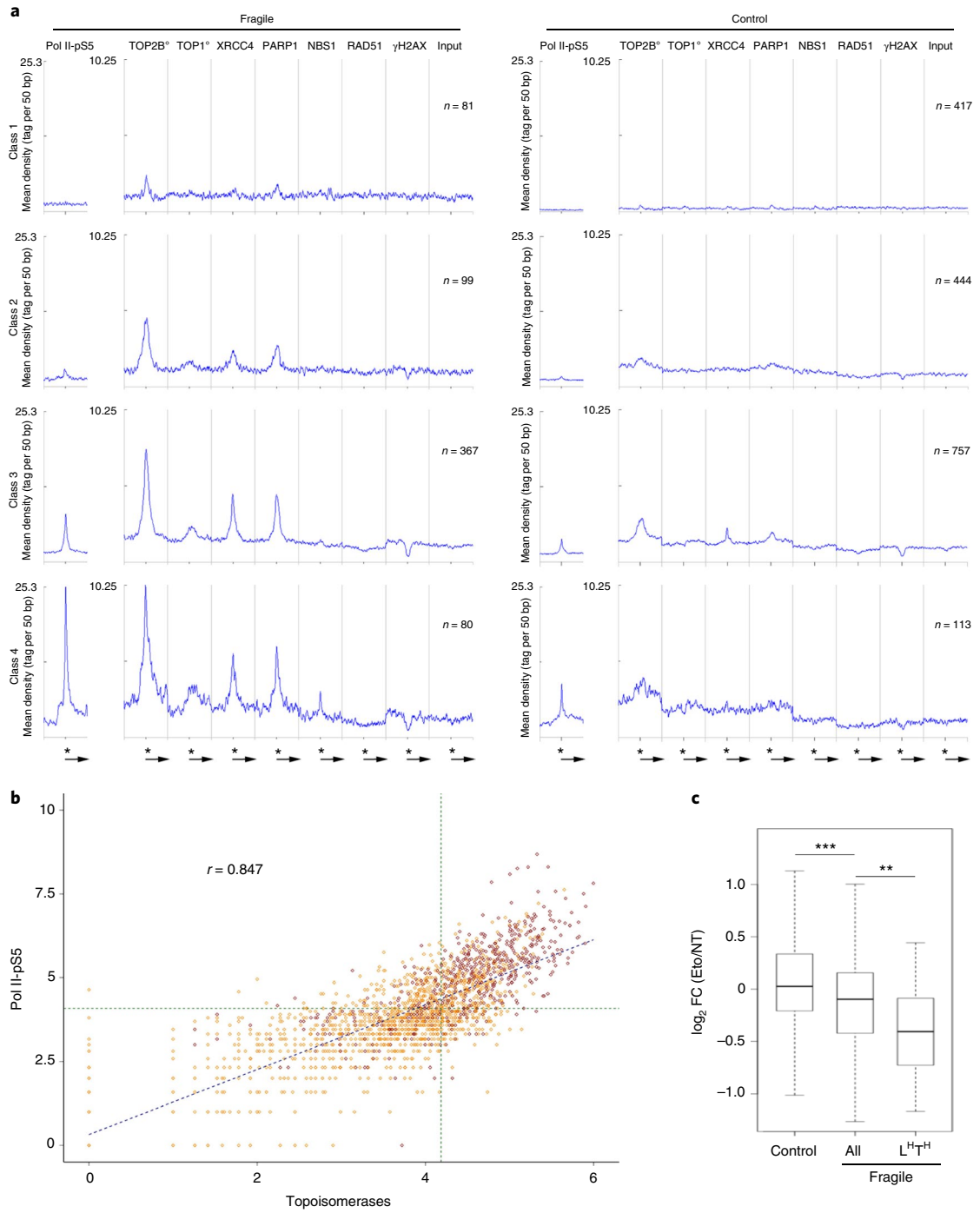

Fig. 4 | Pol II-pS5, topoisomerases and DNA-repair factors at fragile and control promoters from different transcription classes, and effect of etoposide treatment on gene expression. a, Mean-density profiles (±5 kb from the TSS) of Pol II-pS5, TOP2B, TOP1, XRCC4, PARP1, NBS1, RAD51, γH2AX normalized ChIP-seq signals, and input DNA, for each transcription class (1–4, based on GRO-seq data), as indicated. TSS (star), direction of transcription (arrow), pre-treatment of cells with etoposide (TOP2°) or camptothecin (TOP1°), as indicated. **b**, Scatterplot showing levels of Pol II-pS5 and topoisomerases (square root of the product of TOP1 and TOP2B ChIP-seq values) at control ($n = 1{,}731$) and fragile ($n = 627$) promoters (yellow and red circles, respectively). Pearson correlation coefficient (r) as indicated. **c**, Box plot showing $\log_2$FC (FC, fold change) of expression levels measured in control (NT) and etoposide-treated (Eto) cells at the first exon (RPKM > 1) of genes with control ($n = 718$) or fragile promoters: all genes (All; $n = 327$), or the longest among the most transcribed genes (L^HT^H; > median gene length and upper quartile of GRO-seq transcription levels; $n = 52$), as indicated. $^{***}P = 6.85 \times 10^{-9}$; $^{**}P = 3.54 \times 10^{-5}$ (one-sided Wilcoxon rank-sum test). In the box plots, the center line, box edges and whiskers indicate the median, upper and lower quartiles, and 1.5× interquartile range, respectively.

that intron length favors DSB accumulation at 5′ splice site: among the 95 distal-promoter D^+ introns, 54 were first, 32 were second and 9 were third introns (median of 46, 41 and 6 kb, respectively), much longer than the corresponding upstream introns (<1 kb; $P = 1.44 \times 10^{-14}$, 9.94×10^{-12} and 9.11×10^{-5}). The remaining promoter D^+ introns (proximal-promoter introns; $n = 459$ out of 554)

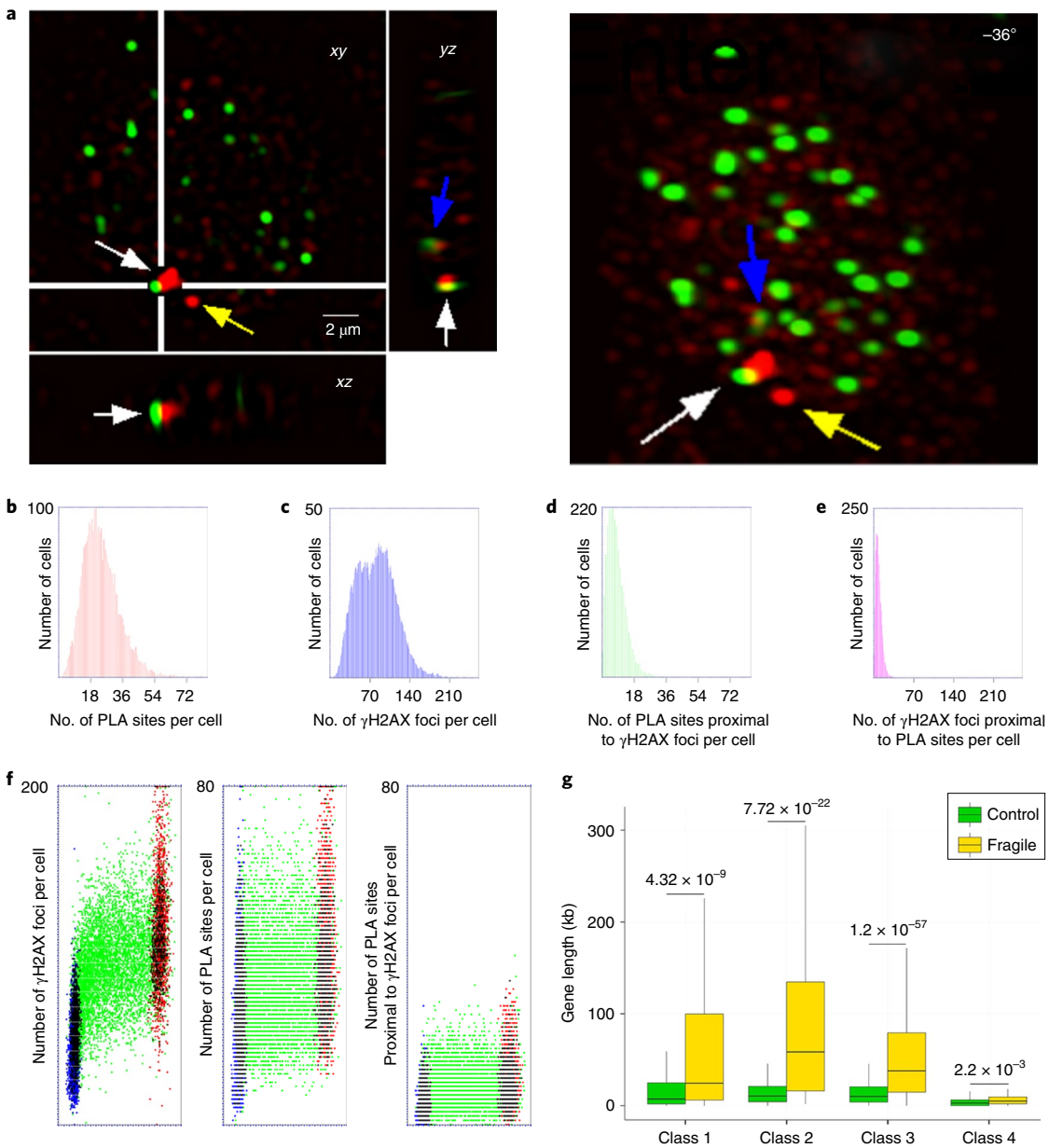

Fig. 5 | In situ localization of TOP2B-XRCC4 interaction sites, and length of genes with fragile or control promoters. a, 3D orthogonal views and projection (−36° around *y* axis) of computationally identified PLA sites (green) and γH2AX foci (red spots). γH2AX focus overlapping with a PLA site cut by the orthogonal planes (white arrow), γH2AX spot proximal to a PLA site (blue arrow) and γH2AX focus non-overlapping with a PLA site (yellow arrow), as indicated. Scale bar, 2 μm. This experiment was repeated three times with similar results. **b**–**e**, Distribution of the number of cells containing the number of PLA sites (**b**), γH2AX foci (**c**), PLA sites proximal (<400 nm) to γH2AX foci (**d**) and γH2AX foci proximal (<400 nm) to PLA sites (**e**) per cell, indicated on the *x* axis. **f**, Distribution of the number of γH2AX foci, XRCC4–TOP2B PLA sites and XRCC4–TOP2B PLA sites proximal (<400 nm) to a γH2AX focus per cell, as indicated. G1 (blue), S (green) and G2 (red) cells, based on DNA content and EdU incorporation, as indicated. **g**, Box plot showing length distribution (in kb) of genes with class 1–4 control ($n = 417, 444, 757, 113$, respectively) or fragile ($n = 81, 99, 367, 80$, respectively) promoters, as indicated. *P* values (one-sided Wilcoxon rank-sum test), as indicated. In the box plots, the center line, box edges and whiskers indicate the median, upper and lower quartiles, and 1.5× interquartile range, respectively.

were almost exclusively first introns and particularly long (median of 33 kb; Supplementary Fig. 17). Promoter D^+ introns were longer than most of the control DSB-free genes (33–46 versus 9 kb; Fig. 5g and Supplementary Fig. 17), suggesting that promoter D^+ introns contribute significantly to the main predictor of promoter fragility, namely gene length.

Finally, we analyzed gene body D^+ introns (that is, with 5′ splice site > 2.5 kb from the TSS; $n = 1{,}589$ out of 2,143; Supplementary Table 16). Approximately 10% of them contained at least one fragile enhancer ($n = 142$ out of 1,589; $P < 1 \times 10^{-6}$; Supplementary Note), which showed Pol II-pS5/TOP2B/XRCC4 accumulation, with typical response to DRB administration and removal (Supplementary Fig. 18). The remaining gene body D^+ introns, instead, showed neither accumulation of Pol II-pS5 at their 5′ splice site (not shown) nor the asymmetrical DSB distribution observed at promoter D^+ introns (Supplementary Fig. 16a), suggesting other mechanisms of DSB formation. Thus, accumulation of DSBs at the 5′ splice site of promoter-associated long introns, TSSs or active enhancers is

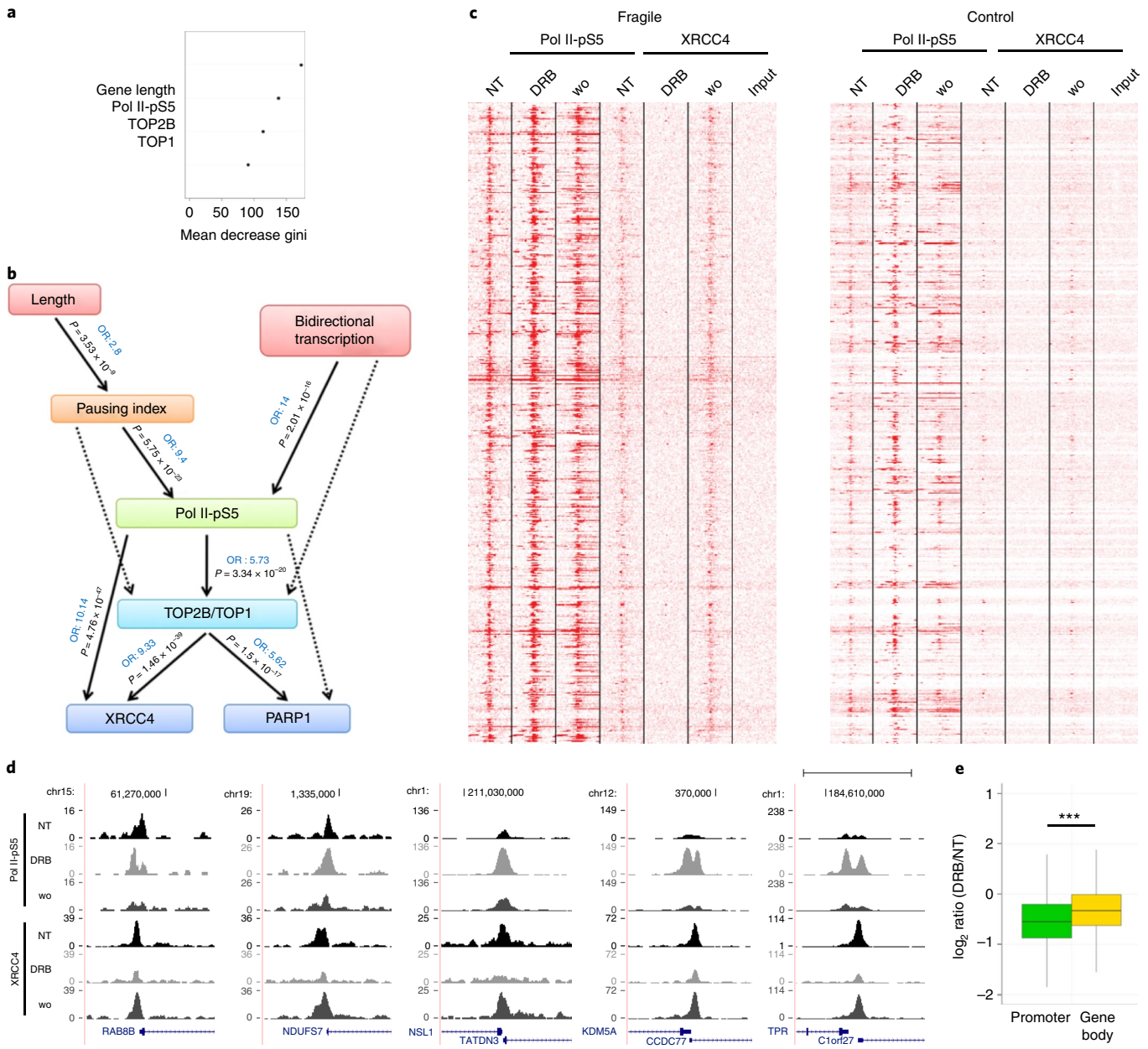

Fig. 6 | Mechanisms of endogenous DSB formation/processing, and effect of DRB administration and washout on Pol II-pS5 and XRCC4 ChIP-seq signals at fragile and control promoters. a, Average of MDG over four cross-validation cycles, which measures relative contribution of gene length, Pol II-pS5, TOP2B and TOP1 to the prediction accuracy of the random-forest classifier (Supplementary Tables 6–8 and Supplementary Note). **b**, Causal model of promoter fragility. When the upstream factor of each arrow shows levels higher than its own risk threshold both in control ($n = 870$) and fragile ($n = 447$) promoters (class 3 and 4), levels of the downstream factor are significantly associated with promoter DSB occurrence (measured by ChIP-seq levels of XRCC4 and PARP1). Solid and dotted arrows indicate causal associations either validated (direct effects) or not confirmed (indirect effects) by structural equation model fitting, respectively. Odds ratios (OR) and *P* values, as indicated (two-sided Fisher's exact test; Supplementary Tables 11, 13 and 14 and Supplementary Note). **c**, Heat map of genomic distribution around the TSS (±5 kb) of Pol II-pS5 and XRCC4 normalized ChIP-seq signals, prior (NT), after DRB administration (DRB) or following DRB removal (wo), and input DNA, at fragile ($n = 627$) and control ($n = 1{,}731$) promoters in MCF10A-AsiSIER cells. **d**, Screenshots from the UCSC Genome Browser showing Pol II-pS5 and XRCC4 normalized ChIP-seq signals prior (NT) or after DRB administration (DRB) and following DRB removal (wo), at representative fragile promoters with increasing (left to right) Pol II-pS5 levels in NT cells. Number of overlapping reads (on *y* axis) and RefSeq genes, as indicated. Scale bar, 5 kb. **e**, Box plot showing the $\log_2$(fold change) of BLISS signal following DRB administration (20 μM, 30 min) compared to untreated cells, at the fragile promoters (promoter) and gene bodies (gene body) ($n = 627$); $^{***}P = 8.43 \times 10^{-12}$ (one-sided Wilcoxon rank-sum test). In the box plots, the center line, box edges and whiskers indicate the median, upper and lower quartiles, and 1.5× interquartile range, respectively.

associated with enrichment of Pol II-pS5, TOP2B and XRCC4, and, in all cases, release of Pol II pausing is the main determinant of DSB formation.

Translocation breakpoints and topologically associating domain boundary-flanking regions. To test whether damaged introns in MCF10A cells are associated with translocations in breast cancers,

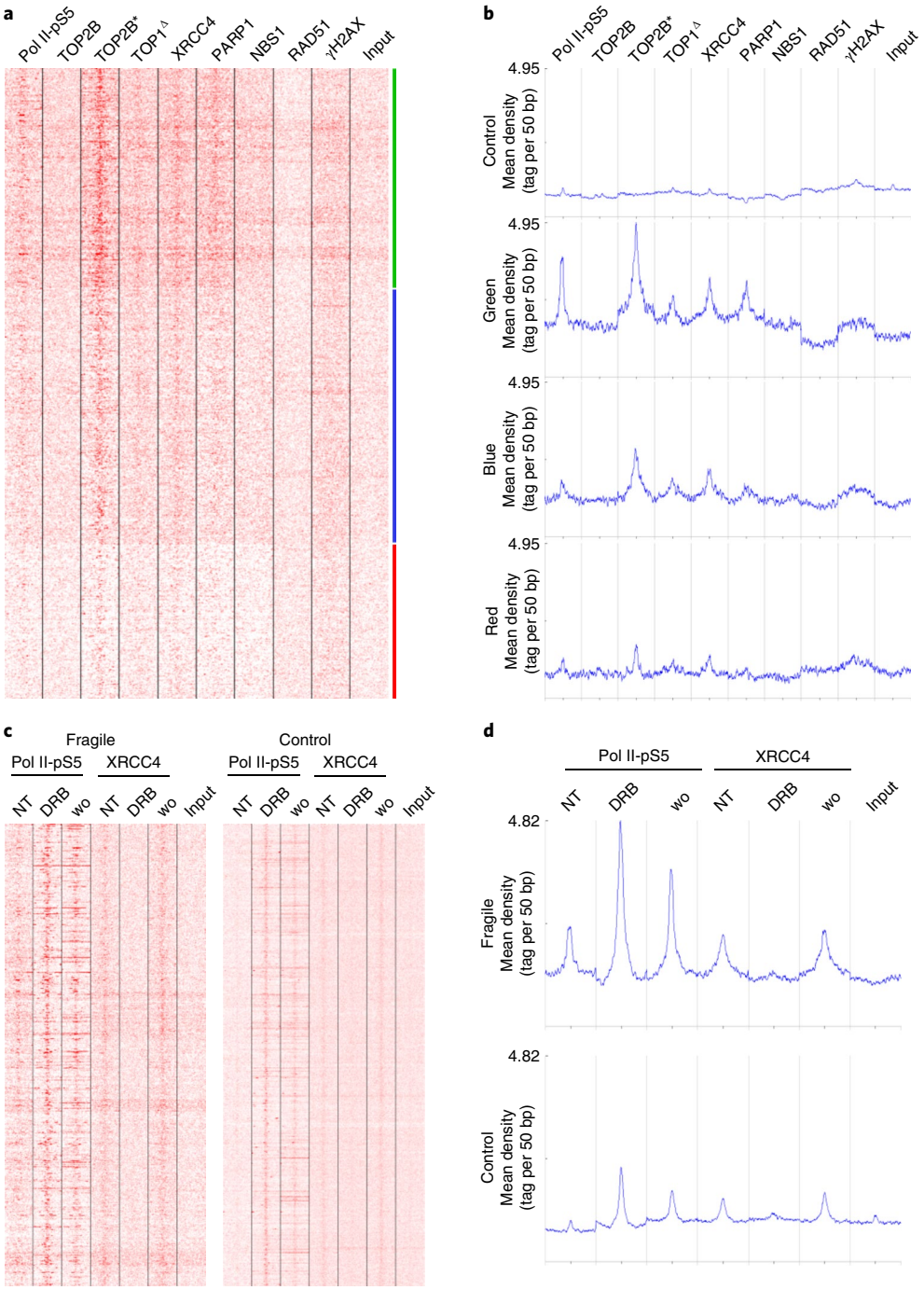

Fig. 7 | Characterization of fragile and control enhancers. a, Heat map of genomic distribution at the three clusters (green, blue, red) of fragile enhancers ($n = 799$; ±5 kb from the H3K27ac peak summit) of Pol II-pS5, TOP2B, TOP1, XRCC4, PARP1, NBS1, RAD51 and γH2AX normalized ChIP-seq signals, and input DNA, as indicated. Asterisks, triangles: ChIP-seq assays performed with etoposide- or camptothecin-treated cells, respectively. **b**, Mean-density profiles (tag per 50 bp; ±5 kb from the H3K27ac peak summit) of Pol II-pS5, TOP2B, TOP1, XRCC4, PARP1, NBS1, RAD51 and γH2AX normalized ChIP-seq signals, and input DNA, at control and fragile (green, blue and red cluster as in panel **a**) enhancers, as indicated. Asterisks, triangles as in panel **a**. **c**,**d**, Normalized ChIP-seq signals of Pol II-pS5 and XRCC4, and input DNA (**c**) and mean-density profiles (**d**) ±5 kb from the H3K27ac peak summit of control and fragile enhancers, before (NT) and after DRB administration (DRB), or following DRB removal (wo).

we interrogated a dataset of 2,822 fusion transcripts identified in patients with breast cancer[41]. Both promoter- and enhancer-associated gene body D^+ introns were highly enriched in translocation breakpoints ($n = 112$ out of 554; $P < 1 \times 10^{-6}$, and $n = 39$ out of 142; $P = 8.23 \times 10^{-4}$, respectively), while the remaining gene body D^+ introns were not ($n = 1{,}447$; $P = 0.2$; Supplementary Tables 16 and 17). Since only promoter- and enhancer-associated D^+ introns showed increased levels of Pol II-pS5/TOP2B/XRCC4 (D^+Pol II^+ introns; $n = 696$), translocations are directly linked to DSBs generated upon release of paused Pol II.

Only 20% of the D^+Pol II^+ introns also contain breakpoints (B) (151 D^+B^+Pol II^+ versus 545 D^+B^-Pol II^+; Supplementary Table 16).

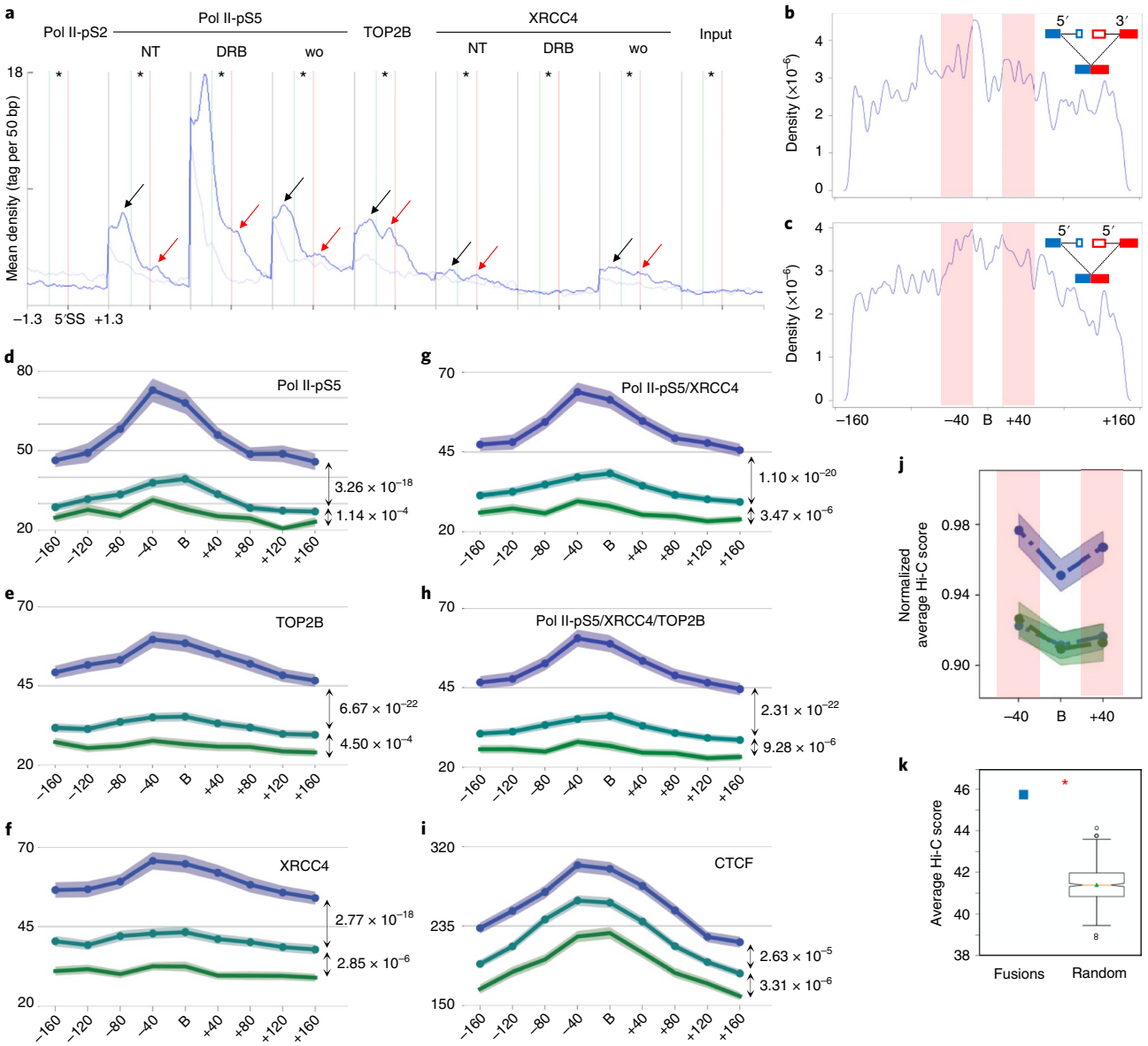

Fig. 8 | Characterization of the damaged introns associated with translocation breakpoints. a, Distribution of Pol II-pS5 and XRCC4 ChIP-seq signals before (NT) and after DRB administration (DRB), or following drug removal (wo), around the 5′ splice sites (5′ SSs) (±1.3 kb) of D^+ (dark blue) or D^- (light blue) distal-promoter introns with unambiguous 5′ splice sites, as indicated. ChIP-seq signals of TOP2B following etoposide administration, of Pol II-pS2 in untreated cells, and input DNA are shown. Red (5′ splice site of distal-promoter introns) and green (−600 bp from 5′ splice site) vertical lines set the limits of the TSS-free region (star); black and red arrows indicate upstream (TSS-associated) and downstream (5′ splice site-associated) peaks, respectively, of the analyzed factors. **b**,**c**, Distribution of translocation breakpoints within the 864 D^{pos}/B^{pos} BFRs. Breakpoints ($n = 1{,}796$) were mapped using the 5′ splice site or 3′ splice site of the B^+ introns (**b**) or the 5′ splice site of both B^+ introns involved in each translocation event (**c**), as indicated in the inset. Density is shown in number of breakpoints per bp/total number of breakpoints. Pink-shaded areas indicate the most proximal 40-kb regions to the boundary-containing 40-kb region (B). **d**–**i**, ChIP-seq signal distribution (measured within 40-kb bins) of the indicated factors, measured individually (**d**–**f**,**i**) or in combinations (**g**,**h**) within the MCF10A D^{pos}/B^{pos} (blue; $n = 864$), D^{pos}/B^{neg} (dark green; $n = 1{,}179$) and D^{neg}/B^{neg} (light green; $n = 827$) TAD boundaries (B; 40 kb) and the corresponding BFRs (±160 kb), as indicated. *P* value (one-sided Mann–Whitney *U*-test) is the highest value (that is, the least significant) among the points on the *x* axis. The shaded area and the central line show the standard error of the distribution within each bin and the mean value, respectively. **j**, Normalized average Hi-C scores measured at TAD boundaries (B; 40 kb) and their most proximal regions (±40 kb) at D^{pos}/B^{pos} (blue), D^{pos}/B^{neg} (dark green) and D^{neg}/B^{neg} (light green) boundaries, as in panels **d**–**i**. Pink-shaded areas and central line as in panels **d**–**i**. $P = 0.0002$ (one-sided Mann–Whitney *U*-test; Supplementary Note). **k**, Average Hi-C scores measured at the 5′ splice sites of the intron pairs showing the highest (that is, upper quartile) Pol II-pS5 levels involved in 162 translocations (Fusions) or at randomly positioned control intron pairs (one-sided randomization test; $n = 500$ iterations; $P < 2 \times 10^{-3}$; Supplementary Note). In the box plot,the center line, box edges, whiskers and circles indicate the median, upper and lower quartiles, 1.5× interquartile range and outliers, respectively.

However, comparable levels of Pol II-pS5 were found at D^+B^+Pol II^+ and D^+B^-Pol II^+ introns (Supplementary Fig. 19). To identify distinguishing features, we tested physical proximity by analyzing the distribution of DSBs and breakpoints at the boundaries of 'topologically associating domains' (TADs) and their flanking regions. TADs represent physically and functionally isolated units of genome

organization[42] and accumulate etoposide-induced DSBs[15]. Analyses of 'chromosome conformation capture' (Hi-C) contact maps of MCF10A cells (at 40-kb resolution)[43] revealed enrichment of both DSBs and breakpoints within TAD boundaries and the two flanking 160-kb intervals ('boundary-flanking regions' (BFRs); $P<0.0001$; Supplementary Fig. 20), suggesting higher probability of translocations at the DSBs located within BFRs. Intron 5′ splice sites and 3′ splice sites involved in translocations showed random distribution within the 864 DSB-positive/breakpoint-positive (D^{pos}/B^{pos}) BFRs identified in MCF10A cells (Fig. 8b). The 5′ splice sites of the same introns, instead, showed increased density within the most proximal 40-kb intervals flanking the TAD boundaries (Fig. 8c), suggesting that the identified BFRs are regions of high Pol II-pS5 density. Levels of Pol II-pS5, TOP2B and XRCC4, individually or in combination, were markedly increased at the most proximal BFRs, compared to the D^{pos}/B^{neg} BFRs ($n=1{,}179$; $P=3.26\times10^{-18}$ to $P=2.31\times10^{-22}$; Fig. 8d–h) and showed overlapping profiles with breakpoints. Increased Pol II-pS5 levels at the D^{pos}/B^{pos} BFRs were consistent with their enrichment in $D^{+}B^{+}$ introns, compared to D^{pos}/B^{neg} boundaries ($P=1.36\times10^{-8}$, OR = 5.339; two-sided Fisher's exact test), and the equal distribution of $D^{+}B^{-}$ introns within D^{pos}/B^{pos} and D^{pos}/B^{neg} boundaries ($P=0.143$, OR = 1.234; two-sided Fisher's exact test). In conclusion, we identified 864 $D^{pos}B^{pos}$Pol II^{pos} BFRs showing enrichments of DSBs, $D^{+}B^{+}$Pol II^{+} introns and translocation breakpoints. Notably, the corresponding TAD boundaries showed features of strong insulators, as shown by high levels of CTCF (Fig. 8i), a boundary-binding protein that stabilizes chromatin interactions within TADs[42].

Finally, we investigated the frequency of interactions with distal genomic regions of the two 40-kb intervals flanking the D^{pos}/B^{pos} boundaries and showing the highest density of intron 5′ splice sites involved in translocations. Their interaction frequencies were higher than those shown by their D^{pos}/B^{neg} counterparts, as shown by Hi-C score analyses ($P=0.0002$; Fig. 8j). Of note, the 5′ splice sites of the two introns involved in each translocation event and showing the highest Pol II-pS5 levels ($n=162$ translocations) also interacted with each other more frequently than control breakpoint pairs ($P=2\times10^{-3}$; Supplementary Note and Fig. 8k). These data suggest that paused Pol II and TOP2B cooperate in DSB formation, and that erroneous end-joining of intron DSBs at chromatin regions showing high interaction frequencies with distal regions increases the probability of translocation events.

Discussion

We identified ~8,000 DSBs within diploid mammary epithelial cells grown under unperturbed conditions. The DSBs were not randomly distributed across the genome, were consistently found at the same positions in independent cultures and were significantly enriched at promoters, intron 5′ splice sites and active enhancers, suggesting that DSB accumulation is an intrinsic property of the genomic regions involved. Endogenous DSBs have been previously associated with transcription[26,44]. However, >50% of the DSB-free genes in mammary cells showed moderate to high transcription levels, suggesting that transcription has no causal effect on DSB formation, as was also demonstrated by our logistic models and causal-association tests. We found, instead, that release of Pol II pausing is the main causal factor of endogenous DSBs, regardless of transcription levels. Consistently, fragile promoters of silenced or poorly transcribed genes also showed high levels of Pol II-pS5.

Paused Pol II is a unique feature of fragile promoters, enhancers and 5′ splice sites. In all cases its release induces DSBs, regardless of the underlying mechanisms of Pol II pausing. Thus, release of paused Pol II by intracellular or extracellular signals might impose cell-type-specific patterns of DSB distribution across the genome. Levels of paused Pol II at damaged sites strongly correlated with topoisomerases (mainly TOP2B), consistent with their reported physical interaction[45,46], suggesting that topoisomerases contribute to endogenous DSB formation. Notably, TOP2B is required for the early events of transcription elongation at fragile promoters under steady-state conditions, suggesting that formation of DSBs is necessary for the transition into productive elongation of physiologically paused Pol II at specific promoters and, possibly, at enhancers and 5′ splice sites.

We identified two main intrinsic causal factors of Pol II pausing at fragile promoters: gene length and bidirectional transcription. Block or attenuation of transcription elongation at long genes or converging/diverging transcription might favor Pol II pausing at TSSs and require topoisomerases for its resolution, with formation of endogenous DSBs and transition into productive elongation[35–37]. Fragile introns are significantly longer than most DSB-free genes, suggesting that they contribute significantly to the effect of gene length on Pol II pausing/release, possibly conferring topological tension, which might also require topoisomerases for its resolution. However, our causal-association models suggest the existence of additional mechanisms of DSB formation (for example, other endonucleases[47], the intrinsic fragility of persistent single strand DNA at promoters with paused Pol II, etc.), consistent with the fact that BLISS does not distinguish between topoisomerase-dependent and -independent DSBs.

While exogenously induced DSBs at AsiSI sites lead to canonical DDR activation, endogenous persistent DSBs do not activate DNA-damage checkpoints and accumulate in proliferating MCF10A cells, suggesting that they trigger unique signaling pathways. Exogenously induced DSBs can be repaired through NHEJ or homologous recombination, as shown by the correlation of RAD51 and XRCC4 signal intensities at AsiSI-induced DSBs (Supplementary Fig. 21). Endogenous DSBs, instead, are mainly processed by c-NHEJ, as shown by accumulation of XRCC4, but not RAD51. Notably, mammalian NHEJ proteins form a multiprotein complex with Pol II and are intrinsically associated with the transcription machinery[32], suggesting that endogenous DSBs are continuously formed after release of paused Pol II and repaired by NHEJ. Consistently, in intact cells we observed XRCC4–TOP2B interactions at sites of transcription, with extremely poor overlap with γH2AX foci. DSBs are weakly associated with PARP1 (Supplementary Table 12), suggesting a minor contribution of alt-EJ, although PARP1 has functions beyond DNA repair, including transcription elongation[48].

During the catalytic cycle of type II topoisomerases, a short-lived cleavage complex is formed, with the enzyme covalently linked to the newly generated DNA 5′ termini[49]. Topoisomerase inhibitors or agents promoting an excess of DNA lesions favor formation of protein-linked DNA breaks (PDBs)[50], where topoisomerases are trapped on DNA termini. Similarly, high frequency of nearby DNA lesions within genomic regions showing intrinsic frailty (that is, promoters, enhancers, etc.) might induce stabilization of the cleavage complex into PDBs, where DSB ends are masked and not sensed by intracellular checkpoints. Interestingly, PDBs formed upon exposure to TOP2 poisons are efficiently repaired by tyrosyl DNA phosphodiesterase 2 (TDP2), which generates ligatable DNA termini processed by NHEJ[51]. Similarly, endogenous PDBs might be processed by TDP2 and then directly repaired by the Pol II-interacting NHEJ proteins[32]. Importantly, only TOP2B-linked DNA breaks with ligatable DNA ends, such as those processed by TDP2, can be identified by BLISS.

All intragenic DSBs in MCF10A cells, including those initially identified as promoter associated, were indeed located within introns. We identified two groups of damaged introns, on the basis of the presence of paused Pol II-pS5/TOP2B/XRCC4. Only Pol II-pS5/TOP2B/XRCC4-positive introns were enriched in translocation breakpoints. DSBs at these introns were mechanistically linked to the release of Pol II pausing at either 5′ splice sites or active enhancers,

suggesting that the association with paused Pol II-pS5/TOP2B/XRCC4 identifies a specific mechanism of DSB formation, and is a prerequisite for the generation of chromosomal translocations. Within the 891 genes damaged in MCF10A cells and involved in translocations in breast cancer, we identified 410 damaged introns containing breakpoints. The co-occurrence of Pol II-pS5/TOP2B/XRCC4 and DSBs was found in ~60% of these introns.

However, only one-fifth of introns with DSBs and Pol II-pS5/TOP2B/XRCC4 contain translocation breakpoints. Do these introns possess distinguishing features or do recombination events occur by chance? DSBs, Pol II-pS5, TOP2B, XRCC4 and breakpoints are enriched at the most proximal regions flanking TAD boundaries, suggesting that recombination events occur specifically within topologically defined chromatin domains. Hi-C score analyses showed that these regions possess high interaction frequencies with distal loci, including their translocation partners. Thus, translocation events may involve Pol II-positive DSBs occurring within either the same TAD or different TADs. In the latter case, DSB formation might contribute to disruption of the TAD structures involved, thus leading to proximity of the damaged sites, a necessary condition for erroneous ligation by NHEJ.

In conclusion, our findings are consistent with a model whereby release of paused Pol II at specific genomic loci (5′ splice sites of promoter-associated long introns or active enhancers within gene-body introns) and at specific chromatin domains (TAD boundary-flanking regions) increases the probability of abnormal DNA recombinations, leading to cancer-associated chromosomal translocations.

References

1. Aguilera, A. & Garcia-Muse, T. Causes of genome instability. *Annu. Rev. Genet.* **47**, 1–32 (2013).
2. Kim, N. & Jinks-Robertson, S. Transcription as a source of genome instability. *Nat. Rev. Genet.* **13**, 204–214 (2012).
3. Di Leonardo, A., Linke, S. P., Clarkin, K. & Wahl, G. M. DNA damage triggers a prolonged p53-dependent G1 arrest and long-term induction of Cip1 in normal human fibroblasts. *Genes Dev.* **8**, 2540–2551 (1994).
4. Ishizaka, Y., Chernov, M. V., Burns, C. M. & Stark, G. R. p53-dependent growth arrest of REF52 cells containing newly amplified DNA. *Proc. Natl Acad. Sci. USA* **92**, 3224–3228 (1995).
5. Huang, L. C., Clarkin, K. C. & Wahl, G. M. Sensitivity and selectivity of the DNA damage sensor responsible for activating p53-dependent G1 arrest. *Proc. Natl Acad. Sci. USA* **93**, 4827–4832 (1996).
6. Khanna, K. K. & Jackson, S. P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.* **27**, 247–254 (2001).
7. Aparicio, T., Baer, R. & Gautier, J. DNA double-strand break repair pathway choice and cancer. *DNA Repair (Amst.)* **19**, 169–175 (2014).
8. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell Biol.* **14**, 8096–8106 (1994).
9. Iacovoni, J. S. et al. High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. *EMBO J.* **29**, 1446–1457 (2010).
10. Shanbhag, N. M., Rafalska-Metcalf, I. U., Balane-Bolivar, C., Janicki, S. M. & Greenberg, R. A. ATM-dependent chromatin changes silence transcription in cis to DNA double-strand breaks. *Cell* **141**, 970–981 (2010).
11. Berkovich, E., Monnat, R. J. Jr. & Kastan, M. B. Roles of ATM and NBS1 in chromatin structure modulation and DNA double-strand break repair. *Nat. Cell Biol.* **9**, 683–690 (2007).
12. van Sluis, M. & McStay, B. A localized nucleolar DNA damage response facilitates recruitment of the homology-directed repair machinery independent of cell cycle stage. *Genes Dev.* **29**, 1151–1163 (2015).
13. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase chromatin domains involved in DNA double-strand breaks in vivo. *J. Cell Biol.* **146**, 905–916 (1999).
14. Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol.* **26**, 52–64 (2016).
15. Canela, A. et al. Genome organization drives chromosome fragility. *Cell* **170**, 507–521.e18 (2017).
16. Wu, H. Y., Shyy, S. H., Wang, J. C. & Liu, L. F. Transcription generates positively and negatively supercoiled domains in the template. *Cell* **53**, 433–440 (1988).
17. Ju, B. G. et al. A topoisomerase IIβ-mediated dsDNA break required for regulated transcription. *Science* **312**, 1798–1802 (2006).
18. Haffner, M. C. et al. Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat. Genet.* **42**, 668–675 (2010).
19. Bunch, H. et al. Transcriptional elongation requires DNA break-induced signalling. *Nat. Commun.* **6**, 10191 (2015).
20. Madabhushi, R. et al. Activity-induced DNA breaks govern the expression of neuronal early-response genes. *Cell* **161**, 1592–1605 (2015).
21. Puc, J. et al. Ligand-dependent enhancer activation regulated by topoisomerase-I activity. *Cell* **160**, 367–380 (2015).
22. Bastus, N. C. et al. Androgen-induced TMPRSS2:ERG fusion in nonmalignant prostate epithelial cells. *Cancer Res.* **70**, 9544–9548 (2010).
23. Chiarle, R. et al. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–119 (2011).
24. Crosetto, N. et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365 (2013).
25. Canela, A. et al. DNA breaks and end resection measured genome-wide by end sequencing. *Mol. Cell* **63**, 898–911 (2016).
26. Lensing, S. V. et al. DSBCapture: in situ capture and sequencing of DNA breaks. *Nat. Methods* **13**, 855–857 (2016).
27. Schwer, B. et al. Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc. Natl Acad. Sci. USA* **113**, 2258–2263 (2016).
28. Ambrosio, S. et al. Cell cycle-dependent resolution of DNA double-strand breaks. *Oncotarget* **7**, 4949–4960 (2016).
29. Yan, W. X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 15058 (2017).
30. Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).
31. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).
32. Chakraborty, A. et al. Classical non-homologous end-joining pathway utilizes nascent RNA for error-free double-strand break repair of transcribed genes. *Nat. Commun.* **7**, 13049 (2016).
33. Leuchowius, K. J., Weibrecht, I. & Soderberg, O. In situ proximity ligation assay for microscopy and flow cytometry. *Curr. Protoc. Cytom.* **56**, 9.36.1–9.36.15 (2011).
34. Furia, L., Pelicci, P. G. & Faretta, M. A computational platform for robotized fluorescence microscopy (II): DNA damage, replication, checkpoint activation, and cell cycle progression by high-content high-resolution multiparameter image-cytometry. *Cytometry A* **83**, 344–355 (2013).
35. Joshi, R. S., Pina, B. & Roca, J. Topoisomerase II is required for the production of long Pol II gene transcripts in yeast. *Nucleic Acids Res.* **40**, 7907–7915 (2012).
36. King, I. F. et al. Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501**, 58–62 (2013).
37. Pannunzio, N. R. & Lieber, M. R. RNA polymerase collision versus DNA structural distortion: twists and turns can cause break failure. *Mol. Cell* **62**, 327–334 (2016).
38. Zhu, Y. et al. Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev.* **11**, 2622–2632 (1997).
39. Fraser, N. W., Sehgal, P. B. & Darnell, J. E. DRB-induced premature termination of late adenovirus transcription. *Nature* **272**, 590–593 (1978).
40. Henriques, T. et al. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev.* **32**, 26–41 (2018).
41. Yoshihara, K. et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).
42. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).
43. Barutcu, A. R. et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.* **16**, 214 (2015).
44. Gaillard, H. & Aguilera, A. Transcription as a threat to genome integrity. *Annu. Rev. Biochem.* **85**, 291–317 (2016).
45. Mondal, N. & Parvin, J. D. DNA topoisomerase IIα is required for RNA polymerase II transcription on chromatin templates. *Nature* **413**, 435–438 (2001).

46. Baranello, L. et al. RNA polymerase II regulates topoisomerase 1 activity to favor efficient transcription. *Cell* **165**, 357–371 (2016).
47. Lin, C. et al. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139**, 1069–1083 (2009).
48. Gibson, B. A. et al. Chemical genetic discovery of PARP targets reveals a role for PARP-1 in transcription elongation. *Science* **353**, 45–50 (2016).
49. Deweese, J. E. & Osheroff, N. The DNA cleavage reaction of topoisomerase II: wolf in sheep's clothing. *Nucleic Acids Res.* **37**, 738–748 (2009).
50. Ashour, M. E., Atteya, R. & El-Khamisy, S. F. Topoisomerase-mediated chromosomal break repair: an emerging player in many games. *Nat. Rev. Cancer* **15**, 137–151 (2015).
51. Gomez-Herreros, F. et al. TDP2-dependent non-homologous end-joining protects against topoisomerase II-induced DNA breaks and genome instability in cells and in vivo. *PLoS Genet.* **9**, e1003226 (2013).

Acknowledgements

We thank R. Mirzazadeh for initial training on the BLISS method; I. Pallavicini and T. Kallas for technical assistance with cell culture; L. Rotta and T. Capra of the Sequencing Facility at the IEO Genomic Unit; E. Colombo for helpful discussions; and P. Dalton and S. Averaimo for critical review of the manuscript. F.P. was supported by a fellowship from Fondazione Umberto Veronesi (grant no. FUV 2018). N.C. acknowledges support from the Karolinska Institutet, the Swedish Research Council (grant no. 521-2014-2866), the Swedish Cancer Research Foundation (grant no. CAN 2015/585) and the Ragnar Söderberg Foundation. M.F. acknowledges support from Italian Ministry of Health grant no. RF-2011-02347946. This study was supported by European Research Council advanced grant no. 341131 (to P.G.P.).

Author contributions

R.P. and B.A.M.B. performed the BLISS assays under the supervision of N.C. R.P., G.I.D. and G.D.C. performed the ChIP-seq and RNA-seq assays. F.P. performed statistical analyses and machine learning-based approaches. G.I.D., F.P., L.L., G.M. and D.C. analyzed the sequencing data. A.M.C. and S.B. performed the Hi-C analyses under the supervision of M.N. D.G. contributed to the statistical analyses. L.G. aligned the sequencing data. L.F. performed the immunofluorescence. M.F. performed the imaging analyses. P.G.P. and G.I.D. wrote the manuscript. G.I.D. and P.G.P. contributed to study design and oversaw the study.

Competing interests

The authors declare no competing interests.

Methods

Cell culture and treatment with 4-hydroxytamoxifen or topoisomerase inhibitors. MCF10A-AsiSI[28] were grown in DMEM/Ham's F12 medium (1:1) supplemented with 5% horse serum, 50 ng ml^{-1} penicillin/streptomycin (Lonza), 10 μg ml^{-1} insulin (Roche), 0.5 μg ml^{-1} hydrocortisone (Sigma-Aldrich), 100 ng ml^{-1} cholera toxin (Sigma-Aldrich) and 20 ng ml^{-1} epidermal growth factor (PeproTech) at 37 °C in humidified atmosphere with 5% CO_2. AsiSI-dependent DSBs were induced by adding 4-hydroxytamoxifen 300 nM (final concentration) directly into the culture medium of MCF10A-AsiSI cells for 2 h. Treatments with topoisomerase inhibitors were performed by adding either DRB 20 μM (final concentration) or flavopiridol 200 nM (final concentration) directly into the culture medium for 30 min. DRB-treated cells were then washed with PBS and incubated in DRB-free fresh medium for 1 h (washout).

Immunofluorescence and automated image cytometry analysis of the effects of AsiSI digestion. Cells were grown on glass coverslips coated with 0.5% gelatin (wt/vol.) in PBS, and fixed for 10 min in 4% paraformaldehyde (wt/vol.). Fixed coverslips were washed twice in PBS and permeabilized for 10 min in 0.1% Triton X-100 in PBS. After blocking (5% BSA in PBS), cells were incubated for 1 h at room temperature with primary antibodies: anti-γH2AX (613406, Biolegend); anti-p53 (fl-393, Santa Cruz Biotechnologies); anti-HA (Roche Applied Science, 12CA5). After washing (3×) in PBS, cells were incubated for 1 h at room temperature with secondary antibodies: anti-mouse Alexa 488- or Alexa 647-conjugated IgGs (Life Technologies) or anti-rabbit Cy3-conjugated IgGs (Jackson ImmunoResearch). Finally, after washings, DNA was counterstained overnight with DAPI. Coverslips were then mounted in Mowiol-containing mounting media.

Images were collected by a BX61 fully motorized Olympus fluorescence microscope controlled by Scan^R software (v.2.2.09, Olympus). An oil-immersion ×60, 1.3 numerical aperture (NA) objective was employed for acquisition. Acquisition parameters were set to optimize subsequent analysis as described elsewhere[34,52,53]. The data collected were analyzed using the A.M.I.CO software package[34,52,53].

Immunofluorescence and microscopy. Cells were grown on gelatinized glass coverslips (see previous section). To detect active DNA replication and ongoing transcription, 5-ethynyldeoxyuridine (EdU) (Life Technologies) and 5-bromodeoxyuridine (BrdU) (Sigma-Aldrich) were added to the culture medium (final concentration 10 μM and 10 mM, respectively) and cells were incubated for 30 min before fixation in 4% paraformaldehyde. Simultaneous immunostaining and PLA was performed as previously described[34,52,53]. Briefly, coverslips were permeabilized in PBS containing 0.1% Triton X-100 (Sigma-Aldrich) and processed using the Click-iTTM EdU Imaging kit (Life Technologies) plus Pacific-Blue azide according to the manufacturer's instructions. After reaction with EdU, samples were processed for in situ PLA according to the manufacturer's instructions (Sigma-Aldrich) using the DuoLink in situ Orange detection reagent. Primary antibodies employed for PLA were rabbit anti-XRCC4 (Sigma-Aldrich, HAP006801) and mouse anti-TOP2B (Santa Cruz, sc-25330). After PLA assay, cells were incubated with rat anti-BrdU (Serotec, OBT0030G) and chicken anti-γH2AX (Byorbit, ORB195374-100) and, after washes, with Alexa 488 donkey anti-rat (Jackson ImmunoResearch) and Alexa 647 donkey anti-chicken (Jackson ImmunoResearch, 703606155) secondary antibodies. For cell-cycle distribution, cell nuclei were stained with DAPI and Chromomycin A3 (10 μM in PBS, 70 mM $MgCl_2$). Images were collected with a BX61 fully motorized fluorescence microscope controlled by Scan^R software (v.2.2.09, Olympus). An oil-immersion ×60, 1.3 NA objective was employed for acquisition. Confocal microscopy data were collected with oil-immersion ×63, 1.4 NA objectives by a SP5 laser scanning spectral confocal microscope equipped with a resonance scanning unit. Image acquisition parameters were set to minimize fluorescence cross-talk (Leica Microsystems) and to optimize subsequent image deconvolution.

Image cytometry analysis. The image cytometry experiments for simultaneous detection of cell-cycle distribution, DNA replication, transcription, PLA detection and γH2AX content and spatial localization were performed as previously described[34,52,53]. Automated wide-field fluorescence microscopy was employed to obtain high content and statistical sampling analysis (more than 10,000 cells were analyzed in each experiment). Images were analyzed by dedicated macros developed in the ImageJ software[34,52,53]. To validate the results obtained with improved three-dimensional spatial resolution, stacks were then collected by confocal microscopy on a selected cell population (about 100 cells per analysis). To evaluate cell-cycle distribution, cells were classified according to the DNA and EdU content (G1: 2 N EdU-negative; G2: 4 N EdU-negative; S phase: EdU-positive). Confocal stacks were deconvolved to optimize signal-to-noise ratio (Huygens, SVI) before analysis. γH2AX and PLA (TOP2B-XRCC4 or TOP2B–Pol II-pS5) spots were detected by applying a two-dimensional (wide-field) or three-dimensional (confocal) Laplace of Gaussian filter on background-subtracted images. Co-localization of targeted spots was evaluated by calculating mutual distances of the fluorescence barycenter ranging from 300 to 500 nm (smaller than the sum of the radius of the spots). A cut-off distance of 400 nm was chosen for this work. All three-dimensional image-processing steps (for example, background subtraction, smoothing, projections, etc.) were performed by ImageJ analysis software.

BLISS. A detailed BLISS protocol has been published[29]. Briefly, MCF10A-AsiSIER cells were grown directly on 22 × 22 mm^2 coverglasses and fixed for 10 min in 4% formaldehyde at room temperature. After permeabilization, cells were incubated in a blunting reaction mix (NEB, E1201L) for 1 h at room temperature, followed by in situ DSB ligation in a T4 DNA ligase reaction mix (NEB, M0202M) for 16–18 h at 16 °C. The next day, genomic DNA was fragmented in situ by incubating the samples with HaeIII (NEB, R0108L) for 3 h at 37 °C. Afterwards, the cells were scraped off the coverglass and genomic DNA purified using proteinase K (NEB, P8107S). Purified genomic DNA was linearly amplified using the T7 RNA polymerase (ThermoFisher, AM1334), followed by library preparation using a modified Illumina TruSeq Small RNA Library Prep Kit (RS-200-0012).

ChIP assays. ChIP assays were performed as previously reported[54]. Briefly, MCF10A-AsiSI cells were cross-linked by adding formaldehyde to the culture medium to a final concentration of 1% (8 min at room temperature). Only for anti-TOP2B and anti-TOP1 ChIP assays, cells were pre-treated (1 h), or not, with etoposide (5 μM) or camptothecin (10 μM), respectively. Cross-linking was stopped by addition of glycine to a final concentration of 125 mM. Cells were washed twice with PBS and lysed in SDS buffer: 100 mM NaCl, 50 mM Tris–HCl (pH 8.1), 5 mM EDTA (pH 8), 0.5% SDS and protease inhibitors. Chromatin lysates were then pelleted and resuspended in immunoprecipitation buffer (100 mM NaCl, 100 mM Tris–HCl at pH 8.1, 5 mM EDTA at pH 8, 0.3% SDS, 1.7% Triton X-100). Cells were sonicated directly in immunoprecipitation buffer before overnight incubation with the following antibodies: anti-XRCC4 (Sigma, HPA006801), anti-NBS1 (Abcam, ab32074), anti-RAD51 (Santa Cruz, sc-8349), anti-PARP1 (Active Motif, 39559), anti-γH2AX (Biolegend, 613401), anti-Pol II-pS5 (Abcam, ab5131), anti-Pol II-pS2 (Bethyl Laboratories, A300-654A), anti-TOP2B (Abcam, ab58442), anti-H3K4me3 (Active Motif, 39159), anti-H3K4me1 (Abcam, ab8895) and anti-H3K27ac (Abcam, ab4729).

ChIP-seq analyses and peak calling. DNA libraries of NBS1, XRCC4, RAD51, γH2AX, H3K4me3, H3K4me1, H3K27ac, TOP2B, Pol II-pS5P and Pol II-pS2 were prepared for HiSeq 2000 sequencing as previously described[55]. A total of 51 bp single-end reads were pre-processed checking their quality with FastQC 0.11.5 and were filtered based on the Illumina filter. Then they were aligned to the reference genome (hg18) using BWA (v.0.6.2-r126)[56] with default parameters. Reads with mapping quality <20 were removed (SAMtools)[57] and those aligning to the same position were counted only once to avoid potential PCR bias. NBS1 and XRCC4 highly enriched genomic regions (peaks) were called using MACS v.1.4.1 with default settings. Peaks were annotated with their nearest gene using the R package ChIPseeker;[58] all differential peak sets, signal tracks and genome-wide coverage tracks to be displayed in the UCSC Genome Browser were generated using a combination of bedtools functions[59]. For each ChIP-seq dataset of γH2AX broad peaks were identified using MACS2 (ref. [60]) (parameters: -g hs --extsize 150 --nomodel --slocal 0 --llocal 5000000). Resulting peaks were filtered for a number of supporting reads >5. Filtered peaks were intersected with 50-kb genomic windows. Windows having more than five intersecting peaks were retained and clustered using bx-python find_clusters function (mincols = 10,000). Cluster boundaries were then expanded by 50 kb and the resulting overlapping regions were merged. Domains from multiple ChIP-seq dataset were merged to create a universal superset of regions. Read count on the superset was evaluated using bedtools multicov[61]. The resulting matrix of counts was analyzed with edgeR[62] to identify regions with statistically significant enrichment of ChIP-seq signal in 4-OHT-treated cells.

RNA isolation and quantitative PCR with reverse transcription (RT–qPCR). For RNA isolation, total RNA was extracted using Quick-RNA Miniprep kit (Zymo Research, R1055) with addition of the DNase treatment. The RNA integrity number was determined using the TapeStation System (Agilent Technologies). RNA quantification was obtained with a Nanodrop spectrophotometer (Life Technologies). Reverse transcription was carried out using ImProm-II reverse transcriptase (Promega) as per the manufacturer's instructions using random hexamer primers (Promega) and 1 μg RNA per 20 μl reaction. Complementary DNA (1 μl per reaction) was used for RT–qPCR with the Fast SYBR Green Master mix (ThermoFisher and Applied Biosystem 7500 Fast). Real-time primers spanning exon–intron junctions (Supplementary Table 18) were designed using the IDT primer-designing software PrimerQuest on the Integrated DNA Technologies (IDT) website (http://www.idtdna.com). All the primers were tested for their specificity, both in silico (http://genome.ucsc.edu) and by standard PCR. A total of 40 PCR cycles were performed in a two-step cycling procedure with an initial denaturation step at 94 °C for 3 min and subsequent steps of 94 °C for 15 s and 60 °C for 30 s. Final values for each probe (using primers spanning exon1–intron1 junctions of GAPDH, TPR or RAB8B genes) in DRB-treated cells were plotted relative to the value in control cells, which was set to 1.0, normalized to the total levels of GAPDH transcript measured by Taqman probe (Hs02786624_g1) in DRB-treated and control cells.

Preparation of RNA sequencing libraries. For the preparation of RNA-seq libraries, total RNA (600 ng) was processed using Illumina TruSeq RNA Sample

Prep Kit (RS-122-2002). Briefly, the poly(A) containing mRNA molecules were purified using poly(T) oligonucleotide-attached magnetic beads. Following purification, the mRNA was fragmented into small pieces using divalent cations at high temperature. The cleaved RNA fragments were copied into first-strand cDNA using reverse transcriptase and random primers. This was followed by second-strand cDNA synthesis. These cDNA fragments then went through an end-repair process, the addition of a single 'A' base and then ligation of the adapters. The products were then purified and enriched with PCR to create the final cDNA library. The quality of each library was analyzed by Bioanalyzer using a High Sensitivity DNA chip (Agilent).

RNA-seq analyses. RNA-seq libraries were sequenced with the Illumina Hiseq 2000 system (51 nucleotide paired-end). After quality control performed with FASTQC (ref. [63]), the reads were aligned to the human reference genome (NCBI36/hg18) using TopHat264 (ref. [64]) and the annotation table 'refGene.gtf', downloaded at the UCSC site (http://hgdownload.soe.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz).

Further quality checks based on the alignment information, such as GeneBodyCoverage, were obtained using the RSeQC package[65]. Raw gene expression values were then obtained with HTseq[66] and used to measure the differential gene expression between etoposide-treated (5 μM, 1 h) and untreated cells, with the edgeR R package, applying the TMM normalization[62].

For the expression analysis of the first exon of the human genes, we used the same analysis pipeline as described above and a custom version of the original human 'refGene.gtf' annotation file that considers the genomic coordinates of the first exon from all genes. To analyze the effect of the etoposide treatment on transcription, we used the expression values (reads per kilobase per million fragments mapped (RPKM)) measured for both whole-gene and first-exon RNA-seq analyses of different classes of genes. Only expressed (RPKM > 0) genes, or first exons, in treated and/or untreated cells were considered: (1) 1,263 of 1,731 and 581 of 627 for the 'whole-gene' analysis of genes associated with control and fragile promoters, respectively, and (2) 718 of 1,723 and 327 of 625 for the 'first-exon' analysis of genes associated with control and fragile promoters. Among the 327 genes, those showing the highest transcription levels (transcripts per kilobase million (TPM) ≥ 5.76, with 5.76 corresponding to the third quartile of the distribution of all GRO-seq transcription levels measured in MCF10A cells) were selected ($n = 104$) and of these, the longest (>82 kb, corresponding to the median length of the 104 genes) were also analyzed ($n = 52$). *P* values of the comparison between different $\log_2$(fold change) distributions were obtained with a Wilcoxon rank-sum statistics in R using the wilcox.test function with alternative = 'l' parameter setting.

BLISS analyses and DSB calling. Alignment of BLISS reads was performed as for ChIP-seq reads (see above), with one modification: duplicates (reads with same start–end position) deriving from the same DSB end independently generated in different cells were distinguished from potential PCR artifacts due to the presence, within the BLISS adapters, of random 8-nucleotide sequences serving as unique molecular identifiers[29]. Only duplicated reads with different unique molecular identifiers were considered[29]. Raw BLISS signal was analyzed using the findPeaks algorithm from the HOMER analysis suite[67], enclosed inside a customizable Python-based pipeline[68,69] used for data pre- and post-processing. Calling of BLISS enrichments was independently performed on each strand, to consider the correct orientation of BLISS tags (that is, reverse tag enrichment upstream and forward tag enrichment downstream of the damaged site). Calling of raw tag enrichments was performed using both local settings, using 10 kb sliding windows (fold enrichment ≥ 2, $P \leq 1 \times 10^{-4}$), to account for local signal variability, and genome-wide settings (fold enrichment ≥ 1.5, $P \leq 1 \times 10^{-5}$). Significant BLISS enrichments of 4-OHT-treated samples were called using BLISS signals from untreated samples.

After raw calling, reverse and forward enrichments with proper orientation and closer than 5 kb from each other were paired. Paired tag enrichments were defined as DSB clusters (or DSBs), since they might include multiple damage events.

If reverse or forward enrichments missed their counterparts (orphan calls), due to the presence of genomic regions containing high density of recognition sites for the HaeIII restriction endonuclease (used for genomic DNA fragmentation during the BLISS procedure) or highly repetitive DNA sequence elements, they were paired with the closest HaeIII restriction site, or repeat, found within 5 kb. BLISS signal clearance, defined as the proportion of bases covered by individual HaeIII sites (or flanked by two HaeIII sites closer than 50 bp, corresponding to the length of each sequenced read), and/or highly repetitive DNA sequence, within ±100 bp and ±2 kb from the center of the DSB, respectively, was calculated to identify HaeIII- or repeat-rich AsiSI sites, which were excluded from further analyses (31 out of 134).

DSBs were ranked on the basis of their strength score (w_j), defined for each *j*th DSB as the product:

$$w_j = P_j \log_2(\hat{n} + \hat{n}_j^- \quad _j^+)$$

where $P_j = \max[-\log_{10}(\text{reverseStrand\_p-value}), -\log_{10}(\text{forwardStrand\_p-value})]$, multiplied by $\log_2$ of the sum of the significant BLISS signal (that is, tag count) supporting the *j*th damage site, from both reverse and forward strands. We established a minimum strength threshold ($w = 75$), defined as the *w* value that maximizes the true positives (that is, DSBs detected at digested AsiSI sites) and minimizes the number of false discoveries (that is, DSBs detected at non-digested AsiSI sites).

Calling of endogenous DSBs and definition of endogenous DSB tiers. Endogenous DSBs were called as described above, with the following modification: fold-enrichments of BLISS signals were calculated over random background. From an initial pool of 190,160 raw endogenous DSBs, 88,845 had a paired (P_j) $P \leq 1 \times 10^{-4}$ (Tier1 + Tier2 + Tier1), 55,436 also had $w_j \geq 75$ (Tier2 + Tier1), while 8,132 (Tier1) also showed also fold enrichment ≥2, when comparing ChIP-seq signals of at least one of three repair factors (NBS1, XRCC4, RAD51) within ±1 kb from the center of the BLISS signal, to the input DNA. To assess the reproducibility of the BLISS, the peak calling was repeated using an independent biological replicate.

Detection of active enhancers. H3K27ac peaks were called using SICER v.1.1 (ref. [70]) running the SICER-rb.sh algorithm with *E* value 150 (that is, the expected number of islands detected by random background fluctuations), 200-bp window size, 50-bp fragment size and an effective genome fraction of 0.769, as reported in the SICER documentation. All the other settings were set to default values. We then calculated H3K4me3 tag density distribution of H3K27ac peaks mapping within or outside gene promoters (±2.5 kb from known TSSs), and determined the threshold of H3K4me3 enrichment (160 tags) as the equality point of the two density distributions. The H3K27ac peaks mapping outside gene promoters and showing H3K4me3 signal above the threshold were flagged as non-annotated TSSs and removed from the following analyses. H3K27ac peaks mapping outside gene promoters and showing H3K4me3 signal below the threshold, instead, were intersected with H3K4me1 peaks (detected using the same SICER settings as for H3K27ac peaks). Finally, H3K4me1-positive H3K27ac peaks were considered as active enhancers ($N_E = 28{,}531$).

Reference, control and fragile gene/promoter datasets. We considered a whole reference set of 20,396 genes annotated to the human genome assembly NCBI36/hg18. For better comparisons between GRO-seq and poly(A$^+$) RNA-seq signals, micro RNA genes were excluded from the analyses and only the canonical (that is, the longest) transcript of all the remaining genes was considered. If a DSB occurred at an alternative TSS we modified genomic coordinates of the damaged transcript accordingly.

Damaged genes contain at least one Tier1 DSB within their promoter and/or gene body, while fragile promoters contain at least one Tier1 DSB within ±2.5 kb from the TSS, regardless of the presence of DSBs within the body of the associated gene. Controls are defined as DSB-free genes, both within their gene body and promoter; three control datasets were defined: control I (Tier1 + Tier2 + Tier3 DSB-free genes; $n = 2{,}032$), control II (Tier1 + Tier2 DSB-free genes; $n = 10{,}224$) and control III (Tier1 DSB-free genes; $n = 17{,}335$). High-confidence control I genes ($n = 1{,}731$ out of 2,032) were also negative to TOP2B, XRCC4 and PARP1 ChIP-seq signals at their promoter, while the remaining 301 control I genes, with TOP2B-, XRCC4- and PARP1-positive promoters, were excluded from further analyses. Of note, they contained HaeIII- and/or repeat-rich promoters.

Enrichment of DSBs at known genomic regions. Enrichment of Tier1 DSBs within genes, promoters, gene bodies or enhancers was measured by means of Fisher's exact test (Supplementary Table 2). $K = \{\text{gene, promoter, gene\_body, enhancer}\}$ was the feature set. For each *j*th feature in *K*, we first counted the occurrence of Tier1-positive $n(K_j+)$ and of Tier1-free elements $n(K_j-)$, and then calculated their expected values $E(K_j+)$ and $E(K_j-)$, respectively. We defined the expected value $E(K_j+)$ as the number of damaged features detected by randomly choosing a genomic position, in N_D independent trials, where $N_D = 8{,}132$, the number of Tier1 DSBs. The prior probability for each *j*th feature to occur at a genomic position was calculated as the genomic coverage for that feature in our reference feature set, using the size of sequenced genome (3.1 gigabases (Gb)) as reported in the UCSC Genome Browser database (accessed on 16 December 2016). This yielded the following priors (Mb, megabases): $P_G = 1.16$ Gb/3.1 Gb = 0.374 for genes, $P_T = 92.85$ Mb/3.1 Gb = 0.03 for promoters, $P_B = 1.11$ Gb/3.1 Gb = 0.358 for gene bodies and $P_E = 62.84$ Mb/3.1 Gb = 0.02 for enhancers. We then calculated $E(K_j+) = P_{K_j} \cdot N_D$, and $E(K_j-) = N_{K_j} - E(K_j+)$ and a Fisher's exact test was applied, as shown in Supplementary Table 2.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed data are available under accession number GSE93040. Previously published data used in this work are: GRO-seq: E-MTAB-742; γH2AX ChIP-seq data of 4-OHT-treated cells ($t = 2$ h, replicate no. 1 in Supplementary Fig. 2a) are available under accession number GSE71447.

References

52. Furia, L., Pelicci, P. G. & Faretta, M. A computational platform for robotized fluorescence microscopy (I): high-content image-based cell-cycle analysis. *Cytometry A* **83**, 333–343 (2013).
53. Marchesini, M. et al. PML is required for telomere stability in non-neoplastic human cells. *Oncogene* **35**, 1811–1821 (2016).
54. Dellino, G. I. et al. Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res.* **23**, 1–11 (2013).
55. Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
57. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
59. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
60. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
61. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–11.12.34 (2014).
62. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
63. Andrews, S. FastQC: a quality control tool for high throughput sequence data v.0.11.7 (Babraham Bioinformatics); http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
64. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
65. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
66. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
67. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
68. Python v.2.7.14 (Python Software Foundation); https://www.python.org/psf/
69. van Rossum, G. *The Python Language Reference Manual* (Network Theory Ltd, 2011).
70. Xu, S., Grullon, S., Ge, K. & Peng, W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol. Biol.* **1150**, 97–111 (2014).