Replication stress underlies genomic instability at CTCF/cohesin-binding sites in cancer

Elangoli Ebrahimkutty Faseela, Dimple Notani*, Radhakrishnan Sabarinathan*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, 560065, India

*Correspondence: dnotani@ncbs.res.in, sabari@ncbs.res.in

Abstract

CCCTC-binding factor (CTCF) and cohesin play a significant role in the formation of chromatin loops and topologically associating domains (TADs), which influence gene expression and DNA replication. CTCF/cohesin-binding sites (CBSs) present at the loop anchors and TAD boundaries are frequently mutated in cancer; however, the molecular mechanisms underlying this remain unclear. Here, we investigate whether the binding of CTCF/cohesin on DNA imposes constraints on DNA replication, leading to replication stress and genomic instability. Our results reveal that CTCF and cohesin remain bound to DNA during replication (S phase) in cancer cells (HeLa). Further, examination of replication stress through imaging revealed colocalization of DNA damage response proteins (MRE11 and STN1) with CTCF/cohesin in the S phase. In addition, ChIP-seq of the DNA damage response/repair proteins (MRE11, STN1, γH2AX and RAD51) showed high enrichment of these proteins at CBSs (as compared to their immediate flanking regions and control sites) and positively correlated with the binding strength of CTCF/cohesin at CBSs in the S phase. Moreover, analysis of somatic mutations from cancer genomes supports that the enrichment of mutations at CBSs is significantly higher in samples harbouring somatic copy number deletion in MRE11 and STN1 compared to wild-type samples. Together, these results demonstrate that the binding of CTCF/cohesin on the DNA during the S phase causes replication stress and DNA strand breaks, and this could lead to genome instability at CBSs observed in cancer.

Keywords: CTCF/cohesin-binding sites (CBSs); DNA replication; replication stress; DNA damage and repair; mutations; cancer

Introduction

Chromatin regulates the accessibility of DNA to various biological processes, including DNA replication, DNA damage and repair, and gene expression[1–3]. In eukaryotes, chromatin is organised into distinct structural units called topologically associating domains (TADs)[4]. TADs are composed of genes and their cis-regulatory elements that show frequent physical interactions between them through chromatin looping. This looping is mediated by two main architectural proteins: CTCF (CCCTC-binding factor) – a zinc finger protein that binds to its highly conserved motif sequences on the DNA; and cohesin – a ring-shaped protein complex that extrudes the chromatin loop until it encounters convergent CTCF bound motifs[5,6]. Thus, CTCF binding and its motif orientation on the DNA determine the positioning of cohesin to create loop anchors and TAD boundaries[7]. Moreover, the binding of CTCF at TAD boundaries can act as an insulator, limiting the chromatin interactions between genes and their regulatory regions predominantly within the TAD, thereby controlling expression of genes within the TAD[8,9].

Previous studies have shown that the CTCF/cohesin-binding sites (CBSs; i.e., DNA regions bound by both CTCF and cohesin), enriched at the boundaries of chromatin-loops and TADs, exhibit an increased somatic mutation rate in multiple tumour types[10–13]. However, the molecular mechanism underlying this observation is not well understood. Since the somatic mutations observed at CBSs are predominantly generated by the mutational processes active in the respective tumour type[12,13], we hypothesise that this enhanced activity of mutational processes at the CBSs could be influenced by the interplay between endogenous process (DNA replication) and binding of CTCF/cohesin to DNA. Recent studies have shown that the boundaries of chromatin-loops and TADs experience topological stress and undergo DNA double-strand breaks due to the activity of Topoisomerase II beta (TOP2B)[14,15]. Moreover, in yeast, cohesin has been shown to induce topological stress-dependent DNA damage during replication[16]. While these findings point to the topological stress imposed by the binding of CTCF and/or cohesin on

the DNA, their association with replication stress and genomic instability remains uncharacterized.

In this study, we found that CTCF and cohesin remain bound to the DNA during the replication (S) phase and are associated with replication stress in cancer cells (HeLa). Further, we show that the DNA damage response (DDR) proteins, MRE11 and STN1, associated with fork-stalling are enriched at CBSs during the S phase, with and without external replication stress. Following this, we observed an enrichment of γH2AX and RAD51 at the CBSs, indicating that these sites undergo DNA double-strand breaks and repair. Finally, by using somatic mutation data from whole-genome sequencing of patient tumours (PCAWG[17] cohort), we demonstrate an elevated mutation rate at CBSs, particularly in tumours harbouring somatic copy number deletions in DDR genes (such as MRE11 and STN1) as compared to tumours without deletions in those genes. Taken together, our results reveal that the binding of CTCF/cohesin proteins to the DNA during the replication phase imposes constraints on the progression of the DNA replication fork, leading to fork-stalling and DNA strand breaks. The subsequent activation of error-prone replication or repair could lead to increased somatic mutations at CBSs.

Results

CTCF/cohesin remain bound to DNA during the replication phase

To determine whether the CTCF and cohesin proteins are bound to the DNA during DNA replication or synthesis (S) phase, we used cervical cancer cell line HeLa (given the rich functional genomics data available for this cell line) to measure the expression level of the above proteins and their genome-wide binding. Initially, HeLa cells were synchronised to get cells at different stages of cell cycle (G1/S, Early S, Mid S, Late S, Mitotic) by using thymidine arrest/release or nocodazole block method (see Methods). The extent of synchronisation under each stage was verified by assessing the cell cycle profiles using flow cytometry with Propidium Iodide (PI) staining (Figure 1a, see Methods). In addition, the synchronisation of cells in the S phase was confirmed by immunostaining of a replication phase marker, Proliferating Cell Nuclear Antigen (PCNA)[18], that showed a punctate signal in 86% of the S phase synchronised cells, as compared to 21% in

asynchronous cells (Figures S1a and S1b). Subsequently, we performed chromatin fractionation followed by western blotting to check the endogenous level of chromatin-bound CTCF and RAD21 (a cohesin subunit) at different stages of cell cycle (Figure 1b). This showed that both CTCF and RAD21 were enriched in the chromatin-bound fraction (insoluble) at various time points of the replication phase (Early S, Mid S and Late S), similar to asynchronous cells (in which the majority of the cells are in the G1 phase; Figure 1a), suggesting that CTCF and cohesin are consistently bound to chromatin across cell cycle phases, including S phase. Further, to check if binding dynamics of CTCF in S phase is different from asynchronous cells[19], we performed Fluorescence Recovery After Photobleaching (FRAP) on GFP-tagged CTCF (CTCF-eGFP) (Figure 1c, see Methods). We observed that the fluorescence recovery rate of CTCF in the S phase (median half-time recovery $t_{1/2}$= 3.45 seconds, IQR = 1.8 seconds) was similar to that in the asynchronous cell population (median $t_{1/2}$ = 3.46 seconds, IQR = 1.7 seconds) (Figure 1d and Figure S1c), suggesting that CTCF dynamics does not change in S phase.

Next, to determine which genomic regions are bound by CTCF/cohesin during replication, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for CTCF and RAD21 in the Mid S phase HeLa cells. We identified 87,734 and 43,567 DNA binding sites for CTCF and RAD21, respectively, in the Mid S phase (Figures S1d-S1f, see Methods). Interestingly, the majority (80%) of RAD21 peaks overlapped with CTCF peaks in the S phase (Figures 1e and 1f). Moreover, around 83% of RAD21 and 62% of CTCF peaks in the S phase overlapped with RAD21 and CTCF peaks from the asynchronous (G1 enriched) condition[20] (Figures S1g and S1h), suggesting that the genome-wide binding of CTCF and cohesin is broadly consistent across interphase. Of the additional CTCF peaks seen in the S phase, around 45% were proximal (< 5 kb) to existing peaks in the asynchronous population of cells, whereas the remaining were distributed distally, likely due to the redistribution of CTCF binding during replication (Figure S1i).

Figure 1

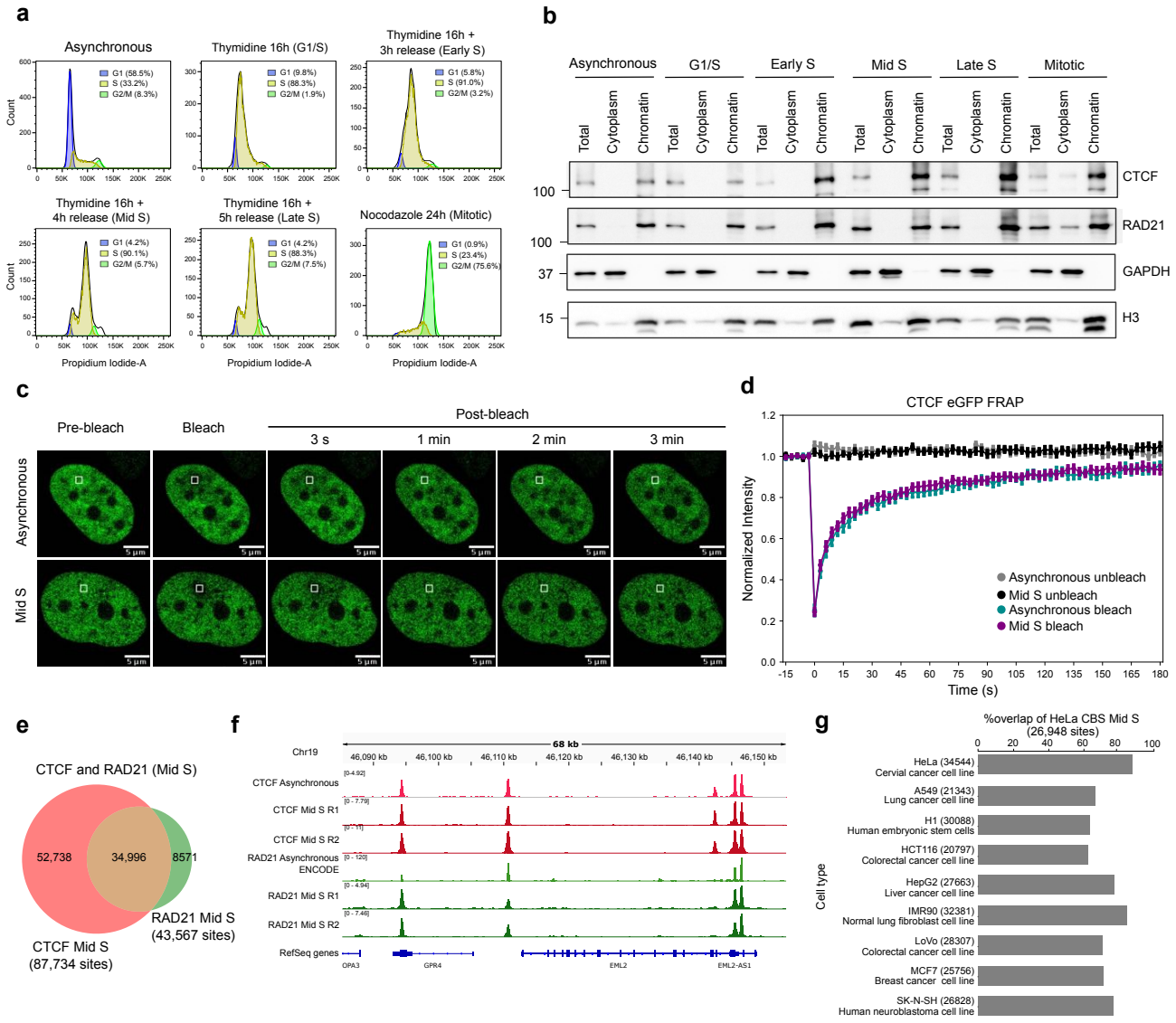

Figure 1: CTCF and cohesin remain bound to DNA in the replication phase.

a) HeLa cells were synchronised to various phases of the cell cycle by using thymidine arrest/ release or nocodazole block method (see Methods). The cell cycle profile in each stage was confirmed by staining with Propidium Iodide (PI) and analysed on a BD FACSVerse flow cytometer. In each histogram, cell count was plotted against DNA content measured from the PI fluorescence signal. The percentage of cells in each phase of the cell cycle is mentioned in brackets.

b) The chromatin-bound fraction of CTCF and RAD21 at various cell cycle phases of HeLa cells was checked by western blotting. GAPDH and Histone H3 were used as a loading control for cytoplasmic and chromatin-bound fractions, respectively.

c) Representative FRAP images of GFP-tagged CTCF in asynchronous (top panel) and Mid S (bottom panel) HeLa cells. The scale bar is 5µm.

d) Fluorescence recovery curve of GFP-tagged CTCF for 3 minutes post-bleaching in asynchronous and Mid S HeLa cells (37 and 41 cells in asynchronous and S phase, respectively) from 3 biological replicates. The normalised intensity value is plotted on the y-axis (each dot in the graph represents mean±1SD)
e) Venn diagram showing the overlapping peaks of CTCF and RAD21 in Mid S HeLa cells (see Methods).
f) An example genome browser snapshot of a region on chr19 showing the overlap between ChIP-seq peaks in asynchronous and Mid S cells. From top to bottom: ChIP-seq signal in CTCF asynchronous, CTCF Mid S (two replicates), RAD21 asynchronous (from ENCODE) and RAD21 (two replicates) in Mid S HeLa cells.
g) Overlap of CBSs in the S phase HeLa with CBSs defined from various cell lines in asynchronous condition (from ENCODE). The number of CBSs identified in each cell line is mentioned in brackets.

To obtain the specific DNA regions bound by both CTCF/cohesin (henceforth, CBSs) in the S phase, we focused on CTCF peaks which have a CTCF motif (see Methods and Figure S1j) and also have a RAD21 peak within 100 bp from the CTCF motif. This resulted in 26,948 CBSs in the S phase of HeLa cells. Around 88% of these CBSs in the S phase overlapped with the CBSs in asynchronous cells of HeLa (Figure 1g). Also, when we compared the CBSs from HeLa S phase with interphase CBSs of other cell lines (different tissue-of-origins, from ENCODE[20]), we observed 62% to 84% overlap, indicating that these sites are consistently bound by CTCF and RAD21 across multiple tissue types. Taken together, our results suggest that the CTCF and cohesin remain bound to the DNA throughout the cell cycle, including replication phase.

CTCF/cohesin binding to DNA cause replication stress

Next, we asked whether the binding of CTCF/cohesin on the DNA can affect the replication fork progression, leading to fork stalling and replication stress. To check this, we examined the enrichment of DNA replication stress-associated proteins (MRE11 and STN1) at the CBSs by using co-immunostaining and ChIP-seq approaches. MRE11 is a component of the MRN (MRE11-RAD50-NBS1) complex which acts as an initial DNA damage sensor and regulates DNA repair activity in response to replication fork stalling and DNA double-strand breaks (DSBs)[21]. STN1 (a part of the human CST: CTC1-STN1-TEN1 complex) is known to protect the stalled replication fork from excessive MRE11-mediated degradation[22].

Figure 2

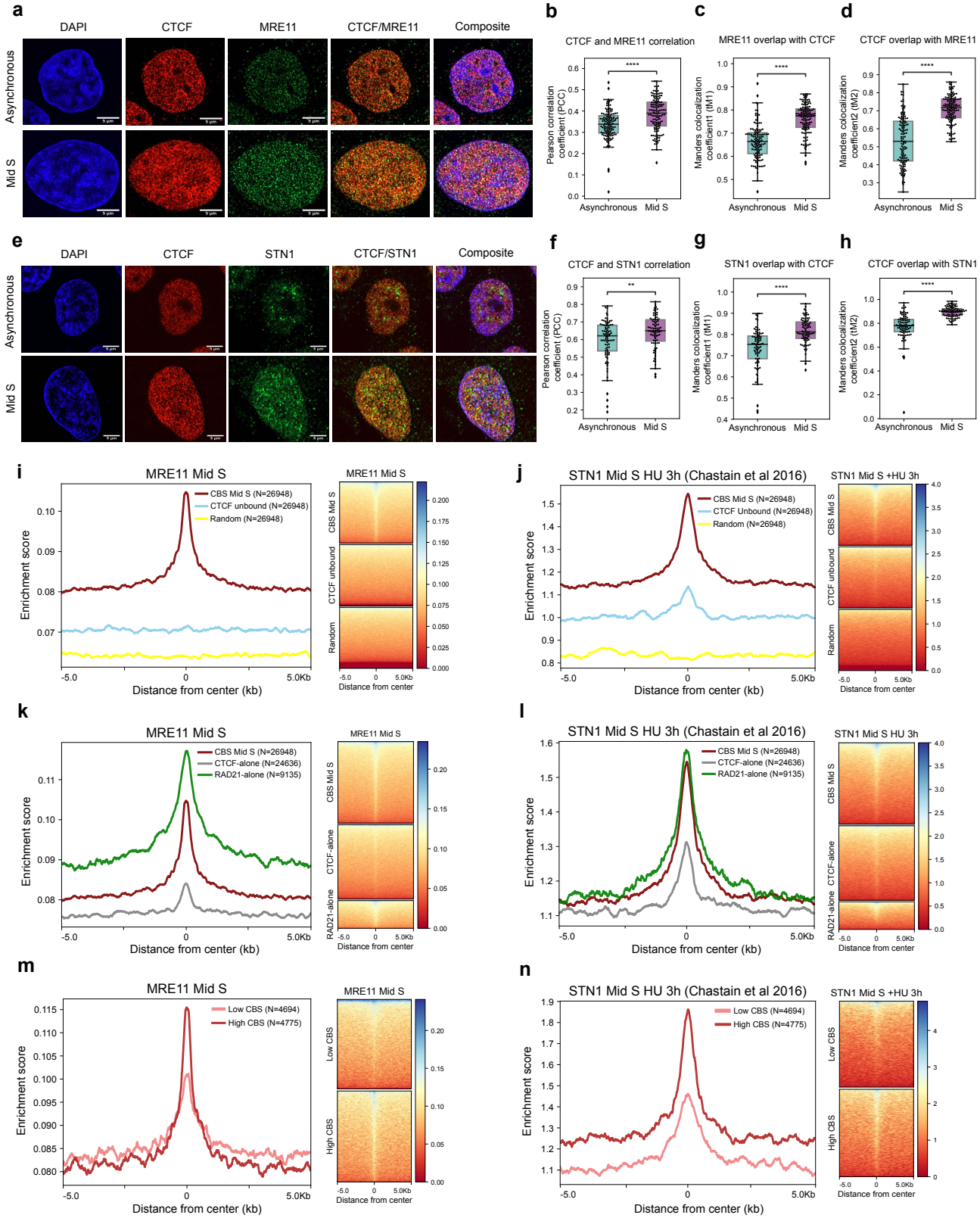
a
DAPI
CTCF
MRE11
CTCF/MRE11
Composite
Asynchronous
Mid S
b
CTCF and MRE11 correlation
Pearson correlation coefficient (PCC)
c
MRE11 overlap with CTCF
Manders colocalization coefficient1 (tM1)
d
CTCF overlap with MRE11
Manders colocalization coefficient2 (tM2)
e
DAPI
CTCF
STN1
CTCF/STN1
Composite
Asynchronous
Mid S
f
CTCF and STN1 correlation
Pearson correlation coefficient (PCC)
g
STN1 overlap with CTCF
Manders colocalization coefficient1 (tM1)
h
CTCF overlap with STN1
Manders colocalization coefficient2 (tM2)
Asynchronous
Mid S
i
MRE11 Mid S
CBS Mid S (N=26948)
CTCF unbound (N=26948)
Random (N=26948)
Enrichment score
Distance from center (kb)
CBS Mid S
CTCF unbound
Random
Distance from center
j
STN1 Mid S HU 3h (Chastain et al 2016)
STN1 Mid S +HU 3h
CBS Mid S (N=26948)
CTCF Unbound (N=26948)
Random (N=26948)
Enrichment score
Distance from center (kb)
k
MRE11 Mid S
CBS Mid S (N=26948)
CTCF-alone (N=24636)
RAD21-alone (N=9135)
Enrichment score
Distance from center (kb)
CTCF-alone
RAD21-alone
l
STN1 Mid S HU 3h (Chastain et al 2016)
STN1 Mid S HU 3h
m
MRE11 Mid S
Low CBS (N=4694)
High CBS (N=4775)
Enrichment score
Distance from center (kb)
Low CBS
High CBS
n
STN1 Mid S HU 3h (Chastain et al 2016)
STN1 Mid S +HU 3h

Figure 2: CTCF/cohesin binding to DNA cause replication stress.
a-d) Representative images of asynchronous and Mid S HeLa cells co-immunostained with CTCF (red, Alexa Fluor-568) and MRE11 (green, Alexa Fluor-488). DNA is counter-stained with DAPI (blue). Images were acquired with ZEISS LSM 980 with Airyscan 2 microscope in super-resolution (SR) mode. Scale bar, 5µm. No. of cells n=135, and No. of biological replicates N=3. Quantification of nuclear colocalization between CTCF and MRE11 by **b)** Pearson's Correlation Coefficient (PCC: $P=2.21x10^{-07}$) and Mander's overlap coefficients **c)** tM1 for MRE11 i.e., the fraction of MRE11 overlapping with CTCF ($P=1.05x10^{-22}$) and **d)** tM2 for CTCF i.e., the fraction of CTCF overlapping with MRE11 ($P=1.44x10^{-24}$).
e-h) CTCF co-stained with STN1 in asynchronous and S-phase synchronised cells. No. of cells n=88, and No. of biological replicates N=3. Colocalization quantified by **f)** PCC ($P= 6.99x10^{-03}$) and Mander's **g)** tM1 ($P=2.35x10^{-10}$) and **h)** tM2 ($P=4.31x10^{-19}$). Colocalization was quantified using the JACoP plugin in Fiji. DAPI staining was used to draw ROI to get only the nuclear signals. The threshold signal was determined using Costes' automatic threshold method and manual adjustment. Centerline: median; Box limits: 1st and 3rd quartile; whiskers: maximum and minimum values: 1.5xIQR. p-values were determined by a two-sided Mann-Whitney U test. p-value annotation legend: ns: 0.05 < P <= 1, *: 0.01 < P <= 0.05, **: 0.001 < P <= 0.01, ***: 0.0001 < P <= 0.001, ****: $P <= 1x10^{-04}$.
i) ChIP-seq profile of MRE11 in Mid S HeLa cells plotted across a 5kb window centred at CBSs from Mid S HeLa (maroon), CTCF unbound sites (sky blue) and random genomic regions (yellow). 26,948 sites were in each category.
j) Similarly, the ChIP-seq profile of STN1 in Mid S upon HU treatment (from Chastain et al.[23]) plotted surrounding CBSs and control sites (+/-5kb).
k-l): ChIP-seq signals of **k)** MRE11 and **l)** STN1 in Mid S phase plotted surrounding CBSs (maroon; 26,948 sites), CTCF-alone sites (grey; 24,636 sites) and RAD21-alone sites (green; 9,135 sites).
m) MRE11 profile at CBSs based on CTCF and RAD21 binding strength. Light red: CBSs with both CTCF and RAD21 low-binding (4,694 sites). Dark red: CBSs with both CTCF and RAD21 high-binding CBSs (4,775 sites).
n) STN1 occupancy at CTCF/RAD21 low- or high-binding CBSs spanning a 5kb window.

At first, we performed co-immunostaining of CTCF with MRE11 and STN1 separately in asynchronous and Mid S HeLa cells to check for their colocalization by using super-resolution microscopy imaging (LSM980 with Airyscan 2, see Methods) (Figures 2a-2h). The colocalization was measured using two different approaches: Pearson's Correlation Coefficient (PCC)[24] which measures the linear relationship between the signal intensities of two proteins, and Mander's Colocalization Coefficient (MCC; tM1 for channel 1 and tM2 for channel 2)[25] which measures the co-occurrence of signals in one channel with another[26]. Both approaches revealed that the CTCF and MRE11 colocalization was

significantly higher in the S phase as compared to asynchronous cells (Mann-Whitney U test, two-sided, P = 2.21×10^{-07}, 1.05×10^{-22}, 1.44×10^{-24} for PCC, tM1 and tM2, respectively) (Figures 2b-2d). Similarly, CTCF and STN1 colocalization was significantly higher in the S phase as compared to asynchronous cells (PCC: P=6.99×10^{-03}, tM1: P=2.35×10^{-10}, tM2: P=4.31×10^{-19}, Figures 2f-2h). Moreover, the colocalization of MRE11/STN1 with RAD21 was also higher in the S phase than in asynchronous cells (Figures S2a-S2h).

Further, to test the protein overlap seen by immunostaining to chromatin level (genome-wide enrichment of MRE11 at the CBSs), we performed a ChIP-seq of MRE11 in Mid S HeLa cells. We examined the MRE11 coverage signal across a 5kb window centred on the CBSs in the S phase and control sites (Figure 2i). To check if this signal is non-random, we used two different control sites for comparison: a) CTCF-unbound (that is, CTCF motif predictions without ChIP-seq signals) to match for GC content, and b) random sites, where genomic regions (without CTCF motif or ChIP-seq signals) were randomly sampled with size equal to the CBSs. Interestingly, MRE11 displayed a higher enrichment at positions near CBSs while unbound and random regions showed a lower and evenly distributed binding (Figures 2i and S2i). For STN1, due to the unavailability of ChIP-grade antibody, we used the publicly available ChIP-seq data of STN1 with Myc-tag in the HeLa cells during the S phase upon replication damage with Hydroxyurea (HU)[23]. Similar to MRE11, we observed high enrichment of STN1 at CBSs relative to flanking regions and control sites (Figures 2j and S2j).

We then inquired whether the individual binding of CTCF or RAD21, or their combined presence, could influence the replication stress. For this, we checked the MRE11 and STN1 ChIP-seq signals at CTCF-alone and RAD21-alone sites compared to CBSs (both CTCF and RAD21 present). Surprisingly, this showed a higher MRE11 and STN1 signal at RAD21-alone sites as compared to CBSs and CTCF-alone sites (Figures 2k and 2l), indicating that the stress at the CBSs could be contributed mainly by cohesin (RAD21). We further extended this analysis by checking the occupancy at CBSs based on the binding strength of CTCF and RAD21 (Figures S2k and S2l). For this, we ranked the CBSs by CTCF and RAD21 peak strength and observed that those regions with both CTCF and RAD21 high binding have higher MRE11 (and STN1) enrichment than CBSs

with CTCF/RAD21 low binding (Figures 2m and 2n), indicating the binding strength of CTCF and cohesin (RAD21) is positively associated with the level of replication stress.

Enrichment of DNA double-strand breaks at CTCF/cohesin-binding sites in the replication phase

Persistent replication stress leads to replication fork collapse and DSBs[27]. To check whether the CBSs undergo DSBs, we measured the enrichment of γH2AX (phosphorylation of Histone variant H2AX)[28,29], a known marker associated with DSBs, through ChIP-seq in the S phase HeLa cells. We also measured the levels of H2AX (through ChIP-seq) as a baseline control. An overall enrichment of γH2AX (normalised to H2AX signal) at positions around CBSs was observed as compared to CTCF unbound and random sites (Figures 3a and S3a). A similar signal of γH2AX was displayed at RAD21-alone sites (Figure 3b) suggesting DSBs at RAD21 binding sites. Further analysis demonstrated that the increased γH2AX enrichment observed at CBSs is mainly from the regions with high CTCF and RAD21 binding (Figure 3c).

Broken replication forks are repaired by Homologous recombination (HR) repair[30,31]. BRCA1 and RAD51 are the two main players in HR repair whose signals were found to be enriched at CBSs in asynchronous cells by re-analysing respective ChIP-seq datasets from ENCODE (Figures S3b-S3f). We asked if the HR repair proteins are recruited at CBSs in the S phase. For this, we performed a ChIP-seq of RAD51 in the Mid S phase HeLa cells. Strikingly, our ChIP-seq profile of RAD51 in the S phase displayed a higher signal at CBSs relative to its flanking regions as well as control sites (Figures 3d and S3g). In addition, the RAD21-alone sites also showed high enrichment of RAD51 (similar to MRE11 and γH2AX) (Figure 3e), suggesting a preferential repair at these sites. Further, the enrichment of RAD51 signals at CBSs positively correlates with CTCF/RAD21 binding strength at these sites (Figure 3f).

Further, to check whether the CBSs undergo replication re-initiation, we utilised Okazaki fragments sequencing (OK-seq) data (from HeLa cells[32]) that maps short fragments of nascent DNA from replication initiation sites. We plotted the OK-seq signal at CBSs (separated by the presence of CTCF motif on Watson or Crick strand) and control sites

Figure 3

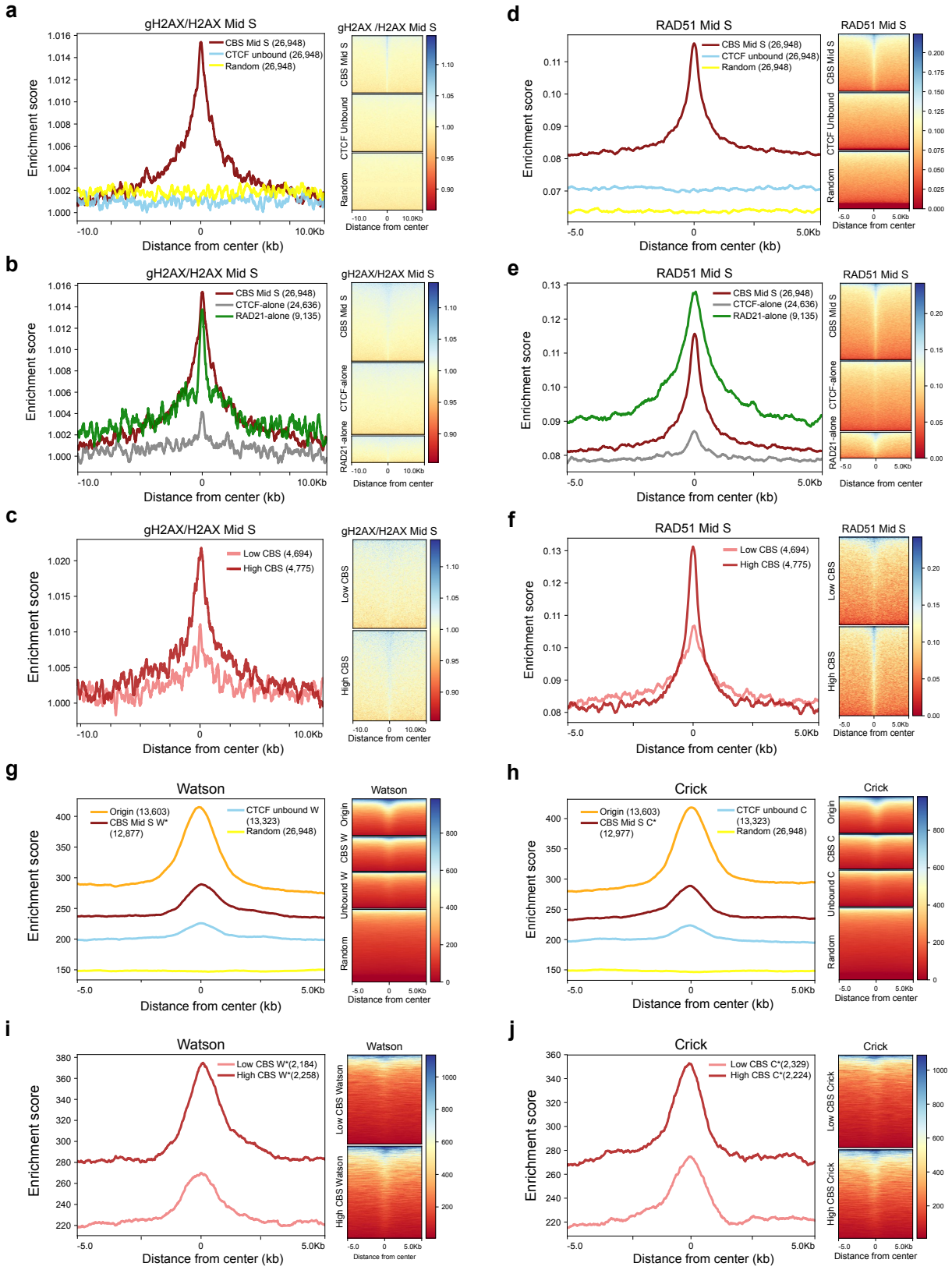
a
gH2AX/H2AX Mid S
CBS Mid S (26,948)
CTCF unbound (26,948)
Random (26,948)
Enrichment score
Distance from center (kb)
CBS Mid S
CTCF Unbound
Random
Distance from center
b
gH2AX/H2AX Mid S
CBS Mid S (26,948)
CTCF-alone (24,636)
RAD21-alone (9,135)
CTCF-alone
RAD21-alone
c
gH2AX/H2AX Mid S
Low CBS (4,694)
High CBS (4,775)
Low CBS
High CBS
d
RAD51 Mid S
CBS Mid S (26,948)
CTCF unbound (26,948)
Random (26,948)
e
RAD51 Mid S
CBS Mid S (26,948)
CTCF-alone (24,636)
RAD21-alone (9,135)
f
RAD51 Mid S
Low CBS (4,694)
High CBS (4,775)
g
Watson
Origin (13,603)
CBS Mid S W* (12,877)
CTCF unbound W (13,323)
Random (26,948)
Origin
CBS W
Unbound W
Random
h
Crick
Origin (13,603)
CBS Mid S C* (12,977)
CTCF unbound C (13,323)
Random (26,948)
Origin
CBS C
Unbound C
Random
i
Watson
Low CBS W*(2,184)
High CBS W*(2,258)
Low CBS Watson
High CBS Watson
j
Crick
Low CBS C*(2,329)
High CBS C*(2,224)
Low CBS Crick
High CBS Crick

Figure 3: Enrichment of DNA damage response and repair proteins at CBSs.
a) ChIP-seq signal profile and heatmap of γH2AX (normalised with H2AX) from Mid S HeLa cells plotted at CBSs (26,948 sites), CTCF unbound (26,948 sites) and random regions (26,948 sites).
b) ChIP-seq signal of γH2AX (normalised with H2AX) plotted at CBS Mid S (26,948 sites), CTCF-alone (24,636 sites) and RAD21-alone (9,135 sites).
c) ChIP-seq signal of γH2AX (normalised with H2AX) at low (4,694 sites) and high CBSs (4,775 sites). +/-10kb flanks were considered for the γH2AX signal.
d-f) Same as a-c, but for RAD51 ChIP-seq signal from Mid S HeLa cells plotted at +/-5kb regions.
g) OK-seq signal profile and heatmap on Watson strand (W) plotted at origins (13,603 sites), CBS (motif is present on Watson strand; 12,877 sites non-overlapping with origin), CTCF unbound sites on Watson strand (13,323 sites) and random sites (26,948 sites).
h) OK-seq signal on Crick strand (C) plotted at origins (13,603 sites), CBS (motifs on Crick strand; 12,997 sites without origin), CTCF unbound sites on Crick strand (13,625 sites) and random sites (26,948 sites). Similarly OK-seq signal enrichment at low and high CBS on Watson and Crick strands plotted in **i** and **j** respectively. Asterisk (*) indicates that sites overlapping with origins were excluded.

(Figure 3g and 3h). As a positive control, we checked for enrichment of OK-seq signals at replication origins with ORC1 ChIP-seq data in HeLa[33]. For this analysis, the sites overlapping with ORC1 peaks were removed from CBSs to eliminate the signals from the origin of replication. As expected, origins showed a higher signal on both strands. Interestingly, CBSs also displayed a higher signal at the centre compared to flanks as well as negative control sites, suggesting replication re-initiation near CBSs. Similar to DNA damage response/repair protein profile at CBSs, OK-seq signals were also correlated with the binding strength of both CTCF and RAD21 at CBSs (Figures 3i and 3j). Together, our results support that the binding of CTCF/cohesin imposes replication stress, and CBSs are vulnerable to DNA damage. As a result, the DNA damage response/repair proteins are recruited to CBSs likely to prevent genomic instability.

Exogenous replication stress increases DNA damage at CTCF/cohesin binding sites

Next, we asked whether the external replication stress could increase DNA damage or strand breaks at CTCF/cohesin sites during replication. For this, we performed ChIP-seq of DDR proteins (γH2AX, MRE11 and RAD51) in drug-induced replication stress

conditions in HeLa cells through HU (which causes fork stalling) and Etoposide (ETO, which inhibits the re-ligation activity of topoisomerase II) treatments, separately. Overall, the enrichment of DDR proteins upon HU treatment was similar to the above results which were without exogenous stress in the Mid S phase (Figures S4-S6). γH2AX signal at CBSs and other regions followed the same pattern after the induction of replication stress by 2 mM HU (Figure S4). γH2AX enrichment was more prominent in prolonged HU treatment (16 hours) compared to that in the Mid S phase for 3 hours (Figures S4a and S4b). MRE11 also displayed a higher signal at CBSs compared to control sites in HU-treated cells (Figures S5a and S5b). Similar to untreated cells, RAD21-alone sites showed highest occupancy of MRE11 (Figures S5d and S5e). At high CBSs, MRE11 enrichment was more localised around the centre, whereas at low CBSs, the signal was more spread to flanking sites (Figures S5g and S5h). However, we observed a decrease in RAD51 binding at CBSs on HU treatment (Figure S6a) likely due to the persistent DNA damage and slower or impaired repair. Upon ETO treatment, Mid S HeLa cells also showed an enrichment of DNA breaks at CBSs (as seen from the γH2AX signal; Figures S4c, S4f and S4i) similar to previous findings in asynchronous cells[14,15,34]. However, MRE11 and RAD51 did not show any prominent localisation at CBSs (Figures S5c and S6b), possibly due to the perturbed DNA damage sensing and repair.

Increased somatic mutations at CTCF/cohesin-binding sites in tumours with loss of MRE11/STN1 genes

The association of DNA damage response proteins (MRE11 and STN1) with CTCF/ cohesin (Figure 2) in the S phase indicates that these proteins might be protecting CBSs from replication stalling and genomic instability. We wondered whether the rate of somatic mutations at CBSs would increase upon perturbations in MRE11 and STN1 (gene names MRE11A and OBFC1, respectively). Therefore, we explored somatic mutations at CBSs in patient tumours harbouring somatic copy number deletion in MRE11 and/or STN1 from the PCAWG[17] cohort. We computed the somatic substitutions at +/-1kb region centred around CBSs, CTCF unbound sites, CTCF-alone sites and RAD21-alone sites in tumour types with at least four samples in the following conditions: a) both MRE11 and STN1 wild-type (no alterations), b) either MRE11 or STN1 deletion, and c) both MRE11 and

STN1 deletions. To account for the influence of DNA sequence composition on the mutation rate, we computed the expected mutation frequency based on the probabilities of each tri-nucleotide to be mutated and calculated mutation fold change (fc: ratio of observed mutations by expected mutations at the centre +/-20bp) (see Methods). This showed that the enrichment of somatic mutation at CBSs in wild-type samples appeared to be high compared to flanking regions and control sites (Figure 4a), as shown previously[10–12]. Strikingly, we noticed that the tumours with deletions in MRE11-STN1 displayed a greater incidence of mutations at CBSs than wild-type samples in certain tumours (stomach, esophageal, liver, pancreatic and breast cancers; Figures 4b and 4c).

For instance, in esophageal adenocarcinoma, samples with deletions in both MRE11 and STN1 genes exhibited a dramatic increase (fc=16.7, P=0) in somatic substitution rate compared to MRE11-STN1 wild-type samples (fc=6.3, P=0) (Figures 4a and 4b). Interestingly, neither the MRE11-STN1 wildt-ype or deletion samples showed any increase in mutations at CTCF-unbound sites (both wild-type: fc=0.93, P=0.25; both deletion: fc=1.01, P=0.895) or RAD21-alone sites (both wild-type: fc=1.42, P=0; both deletion: fc=1.52, P=0.068). CTCF-alone sites and other transcription factor binding sites (TFBS) displayed a slight increase in mutation rate in MRE11 and STN1 deletion tumours (Figure 4b, Figures S7a and S7b), suggesting that the mutations are accumulated specifically at CBSs upon MRE11 and STN1 gene deletion. Moreover, the increase in mutation rate at CBSs was relatively higher in samples with both MRE11 and STN1 deletion as compared to samples with either MRE11 or STN1 deletion (Figure 4c). In contrast, mutations at control sites in MRE11-STN1 deletions were similar to that of wild-type samples in most tumours (Figure S7c). For example, stomach adenocarcinoma samples with either MRE11 or STN1 deletion (fc=3.41, P=0) exhibited a 1.8-times increase in mutations, while samples having both MRE11 and STN1 loss (fc=9.37, P=0) showed a 5.8-times increase in mutations at CBSs compared to MRE11 and STN1 wild-type samples (fc=1.86 P=0) of the same tumour type (Figure 4c).

Figure 4

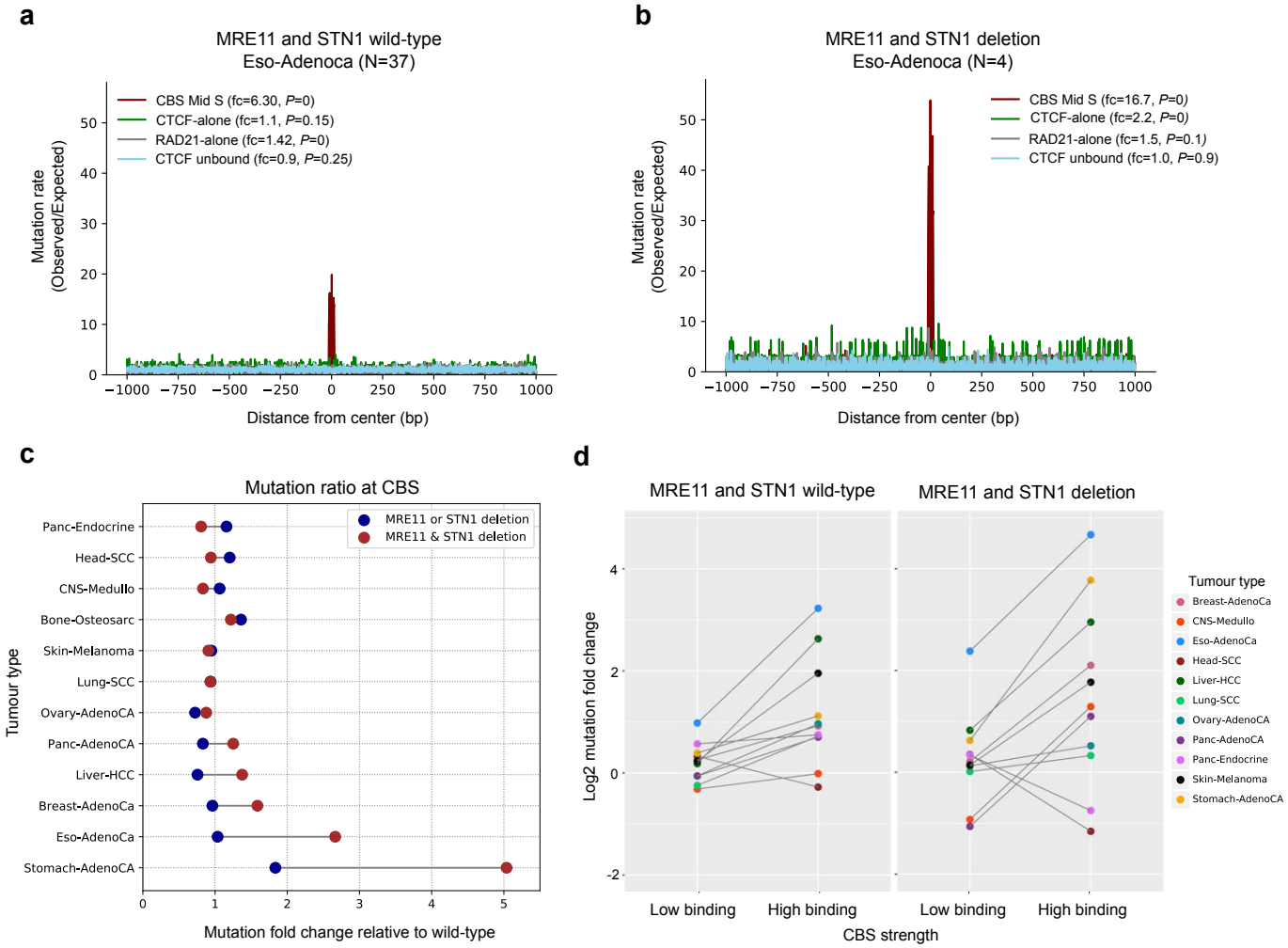

Figure 4: Enrichment of somatic mutations at CBSs in STN1/MRE11 deleted tumours.
a-b) Mutation rate plotted at +/-1kb region of CBSs Mid S and control sites (CTCF unbound, CTCF-alone and RAD21-alone sites) in Esophageal adenocarcinoma samples with **a)** both MRE11 and STN1 wild-type (No. of samples, N=37) and **b)** both MRE11 and STN1 deletion (No. of samples, N=4).
c) Mutation fold change at CBSs (+/-20bp) core in tumour samples with either MRE11 or STN1 deletion (or both deletion) relative to mutation fold change in MRE11 and STN1 wild-type samples across different tumour types.
d) Mutation fold change at CBSs (+/-20bp) core stratified based on the CTCF and RAD21 binding strength. On the x-axis, the high binding represents CBSs with both CTCF and RAD21 strong binding and low binding represents CBSs with both CTCF and RAD21 weak binding.

As enrichment of DDR proteins at CBSs was associated with CTCF/RAD21 binding strength (Figures 2 and 3), we wondered whether the mutation rate at CBSs also follows the same trend. As expected, the high CBSs displayed an increased mutation rate compared to low CBSs across tumours (Figure S7d). Analysing mutations at low and high CBSs in MRE11-STN1 wild=type and deletion samples revealed a drastic increase in mutation rate at high CBSs in MRE11-STN1 deletion samples in many tumours suggesting the functional significance of DNA damage response proteins enrichment at CBSs (Figure 4d). Together, this suggests that the sequence compositions or binding of CTCF or RAD21-alone do not determine the elevated mutation rate at CBSs. Instead, the presence of both CTCF and RAD21 contributes to a high mutation rate at CBSs and the mutation rate is strongly associated with their binding strength.

Mutation in the TP53 gene was previously associated with an increased somatic mutation rate at CBSs in colorectal cancers[12]. We wondered whether the increased mutation rate in esophageal cancer is influenced by TP53 mutation status. Towards this, we checked the mutations in samples having TP53 deletion ensuring that those samples do not have any alterations in MRE11 or STN1 genes (i.e. MRE11 and STN1 wild-type). Interestingly, the eso-adenocarcinoma samples with TP53 deletion did not show any increase in mutations at CBSs, suggesting that irrespective of the TP53 gene status, MRE11 and STN1 deletion leads to accumulation of mutations at CBSs in esophageal cancer (Figures S7e and S7f). However, we noticed biliary adenocarcinoma and stomach adenocarcinoma displaying high mutations at CBSs in samples with TP53 deletion (Figure S7g). We further explored these samples to see whether they harbour heterozygous (shallow deletion) or homozygous (deep deletion) in TP53 deletion. Interestingly, samples with homozygous TP53 deletion exhibited a higher mutation rate at CBSs compared to heterozygous condition (Figures S7h-S7j). Together, this suggests that the perturbation in DNA damage sensing/response genes could lead to increased somatic mutation rate at CBSs, because these sites experience high levels of DNA damage/DSBs.

Discussion

CTCF and cohesin are involved in diverse functions like genome folding, transcriptional regulation[7,35], DNA replication[36,37], DNA damage response[38] and repair[39,40,41]. Despite these roles in the maintenance of genomic integrity, the local DNA regions bound by both CTCF and cohesin (CBS) are found to be a hotspot for genomic instability[12,14,15,34] and somatic mutations in multiple cancers[10–12,42]. In particular, the CBSs at the TAD boundaries and late replicating regions showed high somatic mutations[13]. In some cases, mutations at TAD boundaries have been shown to disrupt TAD insulation, leading to the activation of proto-oncogenes[43]. However, the molecular mechanism underlying high somatic mutation rates at the CBSs remains unclear. Here, we show that the CBSs experience a higher amount of replication stress compared to the immediate neighbouring sites. This is associated with increased DNA double-strand breaks and enrichment of somatic mutations in cancer genomes.

Though the binding of CTCF and cohesin to chromatin are extensively studied in the G1 phase and mitosis, the presence of these proteins during the S phase and the hindrance posed by them to DNA replication are not yet fully understood. We find that the CTCF and cohesin proteins are present in chromatin-bound nuclear fraction throughout the cell cycle (Figure 1b). FRAP experiments revealed that the dynamics of CTCF present in the nucleus during S phase are similar to that of asynchronous cells (Figures 1c and 1d). This result is in line with the CTCF-FRAP experiment by Hansen et al[44] in mESCs. However, the cell cycle synchronisation method in our study ensures the selection of cells undergoing DNA replication, as compared to the aforementioned study where the S and G2 cells cannot be distinguished (due to the FUCCI system). Further, to map the genome-wide binding of CTCF and RAD21 in the S phase, we generated ChIP-seq data in synchronised HeLa cells. This revealed that most of the CTCF and RAD21 binding sites in G1, including the persistent or constitutively bound CTCF/cohesin sites in all cell types[45,46], are occupied in the S phase too, providing independent validation of CTCF/cohesin binding in the replication phase. However, further studies are required to determine the binding of CTCF on newly synthesised DNA and whether chromatin loops are re-established in the S phase.

Next, we tracked the genome-wide binding of replication stress and DNA repair-associated proteins in the S phase to monitor endogenous stress at CBSs. ChIP-seq of DDR and repair proteins is challenging as the protein binding is not sequence-specific and thus the binding is mostly transient. Previous studies that performed ChIP-seq of DDR proteins were mainly in the exogenous stress conditions or enzyme-induced break scenarios. For instance, Chastain et al[23] uncovered that the STN1 is recruited to GC-rich fragile genomic regions upon replication stress (induced by exogenous HU treatment) to promote fork progression. Given that CTCF binding sites have a high GC content, we utilised their STN1 ChIP-seq dataset and observed that not only STN1 signal is high at CBSs but also correlated with the binding strength of CTCF/cohesin, suggesting that the CTCF/cohesin binding cause replication fork stalling. Complementing this, our MRE11 ChIP-seq data also showed a high enrichment at CBSs in the replication phase with and without any exogenous damage, confirming replication stress/fork stalling at these sites (Figure 2). This high enrichment of MRE11 at CBSs could be partly explained by the physical interaction between CTCF and MRE11 at DNA damage sites as shown previously by Hwang et al[39]. Further, our γH2AX ChIP-seq data revealed persistent endogenous DSBs during replication at CTCF/cohesin binding sites. Although γH2AX signals enrich differently at various genomic sites depending on the kind of exogenous replication stress[47], we find a consistent occupancy of γH2AX at CBSs upon external replication stress induced by HU and ETO. In addition, we observed that the enrichment of DNA damage response/repair protein is higher in RAD21-alone sites, followed by CBSs (both CTCF and cohesin binding) and CTCF-alone sites, suggesting that the binding of RAD21 is strongly associated with replication fork stalling and DNA strand breaks. This complements the previous finding that cohesin (RAD21) binding is associated with TOP2B-mediated DNA double-strand breaks at chromatin-loop anchors in asynchronous cells[14,15,48].

Finally, we analysed the essentiality of replication stress-associated proteins at CBSs by assessing the somatic mutation rates at CBSs in patient tumours harbouring MRE11 and STN1 gene deletion or not (wild-type) from the PCAWG[17] cohort. We found that MRE11 and STN1 deletion is associated with increased mutation rates at CBSs in certain tumours (stomach, esophageal, liver, pancreatic and breast cancers), suggesting that this mechanism is specific to those cell types. Loss of DNA damage sensors like MRE11 and

STN1 leads to inefficient replication stress response and repair that increases the likelihood of mutation rate across the genome. However, the control sites used in this study (CTCF unbound, CTCF-alone and RAD21-alone sites) did not show any drastic increase in mutations relative to CBSs in MRE11-STN1 deletion samples even after correcting for background mutations, highlighting the requirement of CTCF and cohesin together to cause genomic instability. This further supports the previous findings which showed that the enrichment of somatic mutations at CBSs is relatively higher as compared to CTCF or RAD21-alone sites[12] and the recurrent mutation at CBSs is associated with chromosomal instability in the nearby regions[42].

During replication, DNA-binding proteins ahead of the fork are temporarily removed from the DNA to ensure the proper propagation of the replication fork[3,49–52]. Interestingly, CpGs flanking CTCF/cohesin sites are marked by hemi-methylation to facilitate the nucleosome re-assembly and positioning post-DNA replication[53]. A recent study in mouse embryonic stem cells found that CTCF is capable of rapidly binding to DNA immediately after replication to re-establish nucleosome positioning[54]. Previous studies have shown that the binding of transcription factors to DNA damage can impair DNA repair activity[10,55–57]. Linking the re-occupancy of CTCF/cohesin onto nascent DNA immediately after the replication with impaired repair activity provides us with a hint for the source of elevated somatic mutations at CBSs. We speculate that cohesin-induced topological stress is a major cause of DNA breaks at loop anchors as DDR proteins are recruited/enriched at RAD21 binding sites (including CBSs and RAD21-alone sites). However, these sites might undergo faithful repair after DNA replication, resulting in a low mutation rate. In contrast, at CBSs (i.e., RAD21+CTCF sites), the rapid re-binding of CTCF could impair the repair of errors introduced during replication, resulting in the accumulation of mutations at these sites, particularly in MRE11/STN1 deleted conditions (Figure 5).

However, the limitations of the study include: a) the DNA replication fork stalling and DNA strand breaks were measured indirectly through the DNA damage sensing/repair proteins, but this can be further improved with proximity ligation assay involving replication complex protein and cohesin, or by measuring the replication speed in the presence and absence of CTCF/cohesin binding; b) given that the genome organisation and chromatin loops are

Figure 5

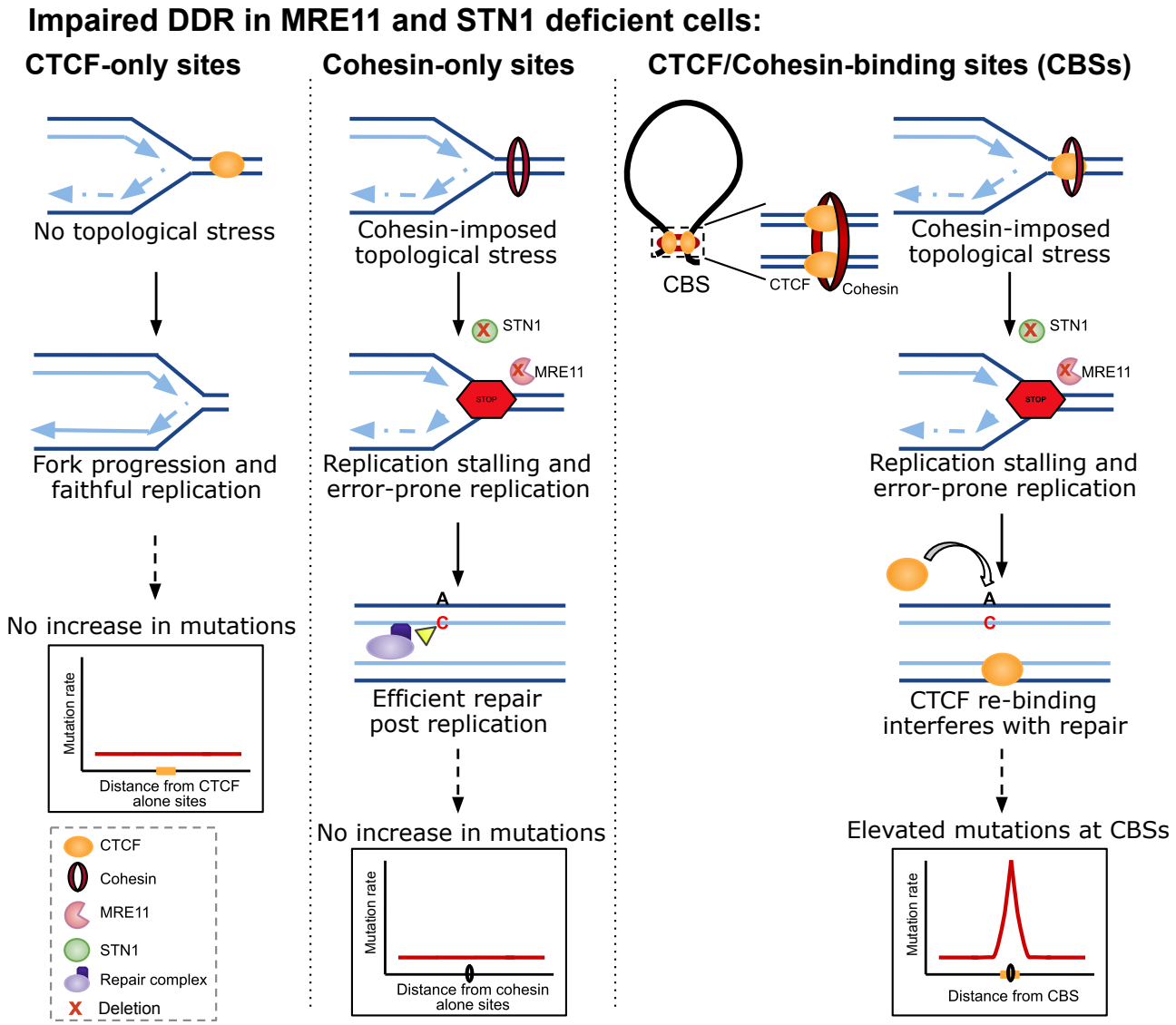

Figure 5: Schematic model for the mechanism behind elevated somatic mutations at CBSs in cancers upon MRE11 and STN1 deletion. At CTCF-only bound sites, the replication fork progresses without any constraints and undergoes faithful replication, thus preventing the enrichment of somatic mutations at these sites. Whereas cohesin-only bound sites can induce topological stress and cause replication stalling. This can lead to error-prone replication; however, faithful repair post-replication prevents the enrichment of somatic mutations at these sites. However, at CTCF/cohesin-binding sites (CBSs), cohesin-induced stress can lead to replication stalling and error-prone replication. Moreover, the immediate re-occupancy of CTCF post-replication at these sites could interfere with mismatch recognition and repair, leading to an elevated mutation rate at CBSs.

correlated with the replicating timing[36], the DNA strand breaks (and enrichment of histone marks associated with that) around CBSs could be partly associated with the replication initiation; c) whether the activity of error-prone replication (e.g., break-induced replication) or error-prone repair (e.g., translesion synthesis) that underlies the source of DNA mismatch remains to be characterised; d) the somatic mutations analysed here were predominantly from primary tumours, thus extending our analysis to normal tissues could help to understand the baseline rate of these events, and to metastatic tumours (which undergone topoisomerase inhibitor or other chemotherapy treatment) to study the impact of treatment induced stress at CBSs.

In summary, our study demonstrates that binding of the CTCF/cohesin on DNA imposes constraints on the DNA that affect the progression of replication fork, leading to fork stalling and DSBs. This could be a potential source of mutations at CBSs, and the likelihood of this is higher in cells lacking DNA damage sensing/repair proteins (such as MRE11 and STN1). This finding has important implications in understanding how chromatin architecture influences the somatic mutational processes across the genome and also for the development of better background mutation models to predict non-coding cancer drivers.

Materials and Methods

Cell lines and culture conditions: HeLa cells (obtained from ATCC) were grown in High Glucose Dulbecco's modified Eagle's medium (DMEM) (Gibco, ThermoFisher, #10569044), supplemented with 10% Fetal Bovine Serum (FBS, Gibco, Invitrogen #16000044) and 1% penicillin-streptomycin (Gibco, Invitrogen #15140163). Cells were cultured at 37°C in a humidified 5% CO_2 incubator and passaged every 3-4 days before reaching confluency. Cells were routinely tested for free of mycoplasma by using LookOut Mycoplasma PCR detection kit (Sigma, #MP0035).

Cell cycle synchronisation: Cell cycle synchronisation was achieved following widely used thymidine arrest/release or nocodazole block method[58]. Briefly, HeLa cells were cultured at 60-70% confluency and treated with 2 mM Thymidine (Sigma #T1895) for 16 hours. Cells were released from the Thymidine block by washing twice with PBS and incubating in fresh DMEM (as mentioned above) for either 3 (Early S), 4 (Mid S), or 5 (Late S) hours, before harvesting. G1/S cells were obtained by collecting cells after the Thymidine block without release. To obtain mitotic cells, 200µg/ml of Nocodazole (Sigma, #M1404) was added to the culturing media and incubated for 24 hours. Cell cycle synchronisation was confirmed by the flow cytometry profile. Replication stress was induced by treating Mid S synchronised cells with 2 mM HU (Sigma, #H8627) or 25µM etoposide (Sigma, #E1383) for 3 hours. For prolonged HU treatment, cells were incubated with 2 mM HU overnight.

Flow cytometry: Synchronised HeLa cells were trypsinized and collected in DMEM. The cell pellet was washed twice with ice-cold PBS, resuspended in DMEM and fixed with 80% ethanol with constant vortexing, followed by incubation on ice for 10-15 minutes. Fixed cells were washed with PBS, stained with 50µg/ml Propidium Iodide (Invitrogen, #P3566) containing RNase A (Qiagen, #19101), incubated at RT for 30 minutes, and measured the DNA content using a flow cytometer (BD FACS Verse) at CIFF Facility, NCBS. The FCS files obtained for DNA content profiling were analysed for the proportion of cells in each cell cycle stage using the Watson Pragmatic model[59] in FlowJo™ Software (BD Life Sciences).

Immunostaining: HeLa cells were seeded at a density of $0.3x10^5$ cells on coverslips and synchronised to the Mid S phase as described above. Cells were then fixed using pre-chilled methanol at room temperature (RT) for 7 minutes, followed by three PBS washes. Cells were then incubated with a permeabilization buffer: 0.5% Triton X-100 (Sigma, #T8787) in 1% BSA (Invitrogen, #AM2616) at RT for 15 minutes. Next, primary antibodies (CTCF: D31H2 CST 1:600, MRE11: ab214 Abcam 1:600, PCNA: ab29 Abcam 1:500, RAD21: ab992 Abcam 1:500, STN1: sc-376450 1:200) were added in a 1% BSA in PBS solution overnight at 4°C or RT for 3 hours. Following primary antibody incubation, cells were washed with PBS-T (PBS + 0.1% Tween 20) three times, each of 5 minutes. Next, cells were incubated with a secondary antibody (Anti-Rabbit Alexa Fluor-568: Invitrogen A10042; Anti-Mouse Alexa Fluor-488: Invitrogen #A11001) at 1:500 dilution in PBS-T for 1 hour at RT and then washed three times in PBS-T. Nuclei were stained with 0.5µg/ml DAPI (Sigma, #10236276001) for 5 minutes and coverslips were then mounted onto glass slides using 90% glycerol. Images were acquired using ZEISS LSM 980 confocal microscope at the CIFF Facility, NCBS. Images were acquired in 0.15µm intervals with a minimum of 5 stacks with Airyscan mode to obtain super-resolution images. Image analysis was performed using Fiji[60].

Colocalization analysis: The region of interest (ROI) was generated around the nuclear boundary on the DAPI channel to get the nuclear signal from every single cell. Colocalization was quantified using the JACoP (Just another Colocalization Plugin)[61] plugin on Fiji[60]. Colocalization was quantified with two different approaches: Pearson's correlation coefficient (PCC)[24] and Mander colocalization coefficient[25]. PCC measures the correlation between intensity variation in channel 1 compared to channel 2. Mander's coefficients (M1 and M2; where M1 is the fraction of channel 1 overlapping with channel 2 and M2 is the fraction of Channel 2 overlapping with channel 1) quantify the co-occurrence of signals[62]. Threshold values for Mander's coefficients (tM1 and tM2) Costes automatic threshold methods[63] and visually checked for appropriate adjustments.

A minimum of 30 cells (n) were analysed for colocalization per biological replicate (N). The results were verified by performing each experiment independently at least three times. A two-sided non-parametric Mann-Whitney U test was performed to check whether

there is a difference in colocalization in asynchronous and Mid S phase cells. The plots were generated using the Seaborn package in Python.

Fluorescence Recovery after photobleaching (FRAP): HeLa cells cultured on coverglass dishes were transfected with pKS070-pCAGGS-3XFLAG- (human)CTCF eGFP, a gift from Elphege Nora (Addgene plasmid #1564488)[6], using Lipofectamine 2000 (Invitrogen, 11668030) and OptiMEM (Gibco, #11058021). Cells were synchronised to the S phase 24 hours post-transfection by thymidine block and released into phenol red-free media for 4 hours (live imaging starts from 3.30 hours of release and ends at 4.30 hours). FRAP experiments were performed on Olympus FV3000 6 laser microscope at the CIFF facility, NCBS. A rectangular region of interest (ROI) of ~1µm area was bleached using a 488 nm laser with 100% power. Throughout the imaging, cells were maintained at 37°C and 5% CO2. The images were captured before bleaching (5 frames), during bleaching (5th frame) and post-bleaching (rest of the frames) every 3 seconds for 3 minutes. Fluorescence recovery in the ROI across the frames was calculated by compensating background intensity as well as the variations in intensities in the same nuclei using the FRAP Norm plugin in Fiji. Data were derived from three biological replicates. For plotting the curve, the intensity obtained from above was normalised with the whole-cell intensity. As a control, the intensity at an unbleached ROI of the same area was also calculated and plotted similarly. The half time of recovery ($t_{1/2}$) was calculated by fitting an exponential curve to the recovery data.

Fractionation: Synchronised HeLa cells were washed twice with PBS, and spun down at 2900 rpm at 4°C for 5 min. To the pellet, Buffer A (100 mM NaCl, 300 mM Sucrose, 3 mM MgCl2, 10 mM PIPES pH 6.8, 1 mM EGTA and 0.2% Triton X-100) and 1x Protease Inhibitor Cocktail (PIC) were added and incubated on ice for 30 min with a gentle tap on the tube every 5 minutes. This was followed by centrifugation at 2900 rpm for 5 minutes at 4°C to get the cytoplasmic protein fraction in the supernatant. To the supernatant, a 6x Laemlli sample buffer (375 mM Tris-HCl pH 6.8, 9% SDS, 50% Glycerol, 0.03% Bromophenol blue and 9% beta-mercaptoethanol) was added and heated at 95°C for 15 minutes. The pellet was washed twice with Buffer A followed by incubation with Buffer B (1 M HEPES pH 7.6, 50% Glycerol, 3 M MgCl2, 0.5 M EDTA, 10% NP-40 and 2 M

Potassium acetate) containing 1x PIC on ice for 30 minutes with a gentle tap on the tube every 5 minutes. The resulting solution was centrifuged at 12,000 rpm for 12 minutes and washed twice with Buffer B. To the pellet, which has insoluble chromatin-bound proteins, 2x Laemlli sample buffer was added and preheated at 95°C for 15 minutes.

Western blotting: Cells were washed with PBS, scraped off the plate and then lysed and collected in a 2x Laemlli sample buffer. Cell lysates were pre-heated at 95°C for 5 minutes and spun down at 13,400 rpm for 1 minute at RT. Samples were separated on SDS-PAGE (stacking 4% and resolving (8% top and 15% bottom)) and the protein-resolved gel was transferred onto a PVDF membrane (Millipore, #IPVH00010) in a 1X transfer buffer (25 mM Tris, 192 mM glycine, pH 8.3 and 20% methanol) at 20 V for 1 hour. The membrane was blocked in 5% skimmed milk (Santa Cruz, sc-2324) prepared in 1x TBS-T at RT for 1 hour on the rocker. Blots were then incubated with primary antibody (Mouse anti-GAPDH Santa Cruz sc-47724; 1:6000, Rabbit anti-CTCF Cell Signalling Technology D31H2; 1:6000, Rabbit anti-RAD21 abcam ab992 1:2000, Rabbit anti-Histone H3 Sigma H0164; 1:6000) overnight at 4°C on the rocker. After primary incubation, blots were washed thrice with 1x TBS-T for 10 minutes each and incubated with HRP conjugated secondary antibody (anti-Rabbit HRP Invitrogen, #32460; 1:10,000 or anti-Mouse HRP Invitrogen, #32430; 1:10,000) prepared in 1xTBS-T for 1 hour at RT on the rocker. Blots were further washed thrice with 1x TBS-T and developed using Clarity Western ECL (BioRad, #1705060). Images were captured using the ImageQuant LAS 4000 gel documentation system. Images were then processed using ImageJ software.

Chromatin immunoprecipitation (ChIP): Cells were cross-linked using 1% formaldehyde (Sigma, #F8775) for 10 minutes and then were quenched with 125 mM glycine (Sigma, #50046) for 5 minutes at 60 rpm at RT. Cells were washed thrice with ice-cold PBS, scraped and pelleted down at 2000 rpm at 4°C for 5 minutes. The cell pellet was either stored at -80°C or directly used in further steps. First, the cells were lysed with L1A buffer (10 mM HEPES/KOH pH 7.9, 95 mM KCl, 1 mM EDTA pH 8 and 1% NP-40) for 10 minutes on ice and centrifuge at 3500 rpm for 5 minutes. Nuclei were obtained by adding a nuclear lysis buffer (L2) (50 mM Tris-HCl pH 7.4, 1% SDS and 10 mM EDTA pH 8) supplemented with 1x PIC, to the pellet for 10 minutes on ice. The chromatin was

fragmented to an average size of 400-500 bp by 22 cycles of 30 seconds ON and 30 seconds OFF pulses using a Diagenode Bioruptor sonicator. Fragment size was visualised by running the chromatin on 1% Agarose gel. The DNA concentration was checked in NanoDrop. Chromatin lysate was cleared by centrifugation at 12000 rpm at 4°C for 12 minutes. One hundred µg of chromatin lysate was diluted in the Dilution buffer (20 mM Tris-HCl pH 7.4, 100 mM NaCl, 2 mM EDTA pH 8 and 0.5% Triton X-100) in a 1:1.5 ratio and 10% of chromatin lysate was taken as input. The antibody was added and incubated overnight at 4°C with a constant rotation. Antibodies used are: CTCF (D31H2: Cell Signaling Technology, 1µg), RAD21 (ab992: Abcam, 1µg), MRE11 (ab208020: Abcam, 2µg), γH2AX (ab81299: Abcam, 2µg) and RAD51 (ab176458: Abcam, 2µg).

The next day, Protein A/G Dynabeads (Invitrogen, #10001D and #10003D) which were blocked in 5% BSA (Invitrogen, #AM2616) for 1 hour, were added (15µl per tube) for 4 hours at 4°C at 12 rpm. This was followed by washes with wash buffer I (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% SDS, 2 mM EDTA pH 8 and 1% Triton X-100), wash buffer II (20 mM Tris-HCl pH 7.4, 500 mM NaCl, 2 mM EDTA pH 8 and 1% Triton X-100), wash buffer III (20 mM Tris-HCl pH 7.4, 250 mM LiCl, 1% NP40, 1 mM EDTA pH 8 and 10% Sodium deoxycholate) and 1xTE buffer (10 mM Tris pH 8, 1 mM EDTA pH 8) in sequence for 5 minutes at 4°C. Flow through was discarded after each wash step. Then the protein-DNA complex was eluted in 200µl of Elution Buffer (100 mM NaHCO3, 1% SDS) at 37°C. Decrosslinking was performed by adding 14 µl of 5M NaCl (Sigma, #746398) to the eluate at 65°C overnight. The DNA was purified using the Phenol:Chloroform:IAA (Invitrogen, #AM9732) method followed by ethanol precipitation. Finally, the pellet was dissolved in water and the extracted DNA was used in a qPCR reaction to verify the ChIP and proceeded for sequencing.

ChIP-sequencing and analysis: ChIP sequencing was carried out at the NGS Facility, NCBS. Library preparation was done using the NEBNext® Ultra™ II DNA Library preparation kit (NEB, #E7103L). Libraries were sequenced on single-end 50 bp (1x50 bp) reads on the HiSeq 2500 Illumina platform. ChIP-seq data were analysed using the nf-core/ChIPseq[64] pipelines. The command used was ‘nextflow run nf-core/ChIPseq -r 1.2.1 -profile docker --input design.csv --genome GRCh37 --single_end --narrow_peak’ for

narrow peaks. Briefly, the pipeline trims the adapters (Trim Galore!), aligns the reads to reference genome (BWA), marks duplicates (picard), filters (SAMtools and BEDTools), performs alignment-level QC and library complexity estimation (picard, Preseq), creates normalised bigWig files (BEDTools, bedGraphToBigWig) and calls peaks (MACS2). For CTCF and RAD21 ChIP, MACS2 peak calling was run by specifying 'narrowpeak' while all other ChIP peaks were called by 'broadpeak'. Overlapping ChIP-seq peaks between the biological replicates were computed with BEDTools[65] intersect by keeping the fraction of overlap as 0.5.

Somatic mutation analysis: Somatic mutation and somatic copy number alteration (SCNA, gene level) data of 37 tumour types were obtained from PCAWG[17]. Samples with mutations in the exonuclease domain of POLE were excluded from the analysis. Single base substitution mutations were extracted and intersected with each of the regions (i.e., CBSs, CTCF unbound sites, CTCF-alone sites, RAD21-alone sites and TFBS) spanning 1000 bp upstream and downstream (+/-1kb) from the motif/peak centre, using BEDTools intersect to compute the rate of substitutions at respective sites. The samples were grouped into three categories according to the SCNA values of MRE11 (MRE11A) and STN1 (OBFC1) genes (samples with SCNA value of zero for both MRE11A and OBFC1 was considered as wild-type or no alterations; and samples with SCNA value less than 0 was considered as deletion). Samples with amplifications (SCNA >0) in MRE11 or STN1 were not considered for the analysis. Also, tumour types with less than four samples in any of the groups were excluded. Similarly, samples with TP53 gene deletion were classified as above. Further, the samples with SCNA values -1 and -2 were categorised as likely heterozygous and homozygous deletion, respectively.

To assess whether the mutations observed at CBSs (or other TFBS) are due to the local DNA sequence contexts, we calculated the background mutations based on the probability of each trinucleotide to be mutated (as described in Frigola et al[66]). Briefly, for each sample group (wild-type or STN1-MRE11 deletion samples), we computed the mutation probability by using the frequency of mutations observed in the specific tri-nucleotide context normalised by the abundance of that tri-nucleotide context in the genome. Fold change (fc) or the enrichment of mutations over background was computed

as the ratio of observed mutations over expected mutations at +/-20bp from midpoint (core region). The significance value was calculated by using G-test goodness-of-fit (frequency of observed mutations in core and flanking regions were compared to the frequency of expected mutations in those regions).

Predicting CBS in the S phase: To identify the CTCF/cohesin binding sites in the S phase (CBSs Mid S), we first performed motif analysis in 87,734 CTCF peaks (from HeLa Mid S phase) to check the presence of CTCF motif using MOODS (Motif Occurrence Detection Suite)[67] in Python with a p-value cut-off of 0.0001. The position frequency matrix (PFM) file of CTCF motif in humans (Matrix ID: MA0139.1) was obtained from JASPAR[68,69]. In case of a peak region with multiple CTCF motif matches, the one with maximum motif match score was retained. Further, we defined CTCF/cohesin-binding sites (CBSs, n=26,948) as those CTCF peaks with CTCF motif match and have a RAD21 peak within 100 bp from the CTCF motif centre (calculated using BEDTools closest function). We separated the CBS Mid S into 3 quantiles; low, medium and high based on CTCF and RAD21 ChIP-seq signals. Those CBSs with both CTCF and RAD21 low binding strength were defined as Low CBS and both CTCF and RAD21 high binding strength were considered as High CBS. CBSs from other cell lines were also obtained by following the same method as above on the ChIP-seq data obtained from ENCODE (see “External datasets” section below).

Control sites were defined as follows: CTCF unbound sites were obtained by taking the CTCF prediction sites from JASPAR[68] and removing all overlapping CTCF ChIP-seq sites in the S phase. The origin of replication sites (from ORC1 ChIP-seq in HeLa[33]) were also removed from these regions and randomly sampled 26,948 sites to match the number of CBSs in the S phase. Random sites used in this study were the randomly sampled genomic regions of 26,948 sites with no CTCF/RAD21 ChIP enrichment or replication origin sites.

ChIP enrichment analysis: We used deepTools computeMatrix to check the occupancy of proteins centred at specific genomic regions. ChIP-seq bigwig files were provided as the score files (-S) and genomic sites of interest in bed format as region files (-R). The enrichment was plotted by plotHeatmap and plotProfiles functions.

External datasets: To define CBSs in other cell lines, CTCF and RAD21 ChIP-seq peaks from various human cell lines were downloaded from ENCODE[20] from the GEO: GSE51334, which includes HeLa (cervical cancer cell line), A549 (lung cancer cell line), H1 (human embryonic stem cells), HCT116 (colorectal cancer cell line), Hep G2 (liver cancer cell line), IMR 90 (normal lung fibroblast cell line), MCF7 (breast adenocarcinoma cell line) and SK-N-SH (human neuroblastoma cell line). CBSs in LoVo (colorectal cancer cell line) were from Katainen et al 2015[12]. STN1 ChIP-seq data in HeLa cells treated with HU[23] were obtained from the GEO: GSE82123. The origin of replication sites in HeLa was taken from the ORC1 ChIP-seq data[33] from the GEO: GSM922790. OK-seq data[32] was downloaded from the GitHub page (https://xfer.curie.fr/get/SbyS2ueTaTz/OKSeq.Data.zip).

Acknowledgements

We thank Aswin Sai Narain Seshasayee, Madhusudhan Srinivasan, and the members of RS and DN labs for feedback and suggestions on the manuscript. We acknowledge the Next Generation Genomics Facility (NGGF) and Central Imaging & Flow Cytometry Facility (CIFF) at NCBS for their support and help. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Funding

This work was supported by the DBT/Wellcome Trust India Alliance Fellowship [grant number IA/I/20/1/504928] awarded to RS. We also acknowledge support of the Department of Atomic Energy, Government of India, under Project Identification No. RTI 4006 and intramural funds from NCBS-TIFR.

Author contributions

EEF and RS conceived the study. EEF, DN and RS designed the experiments. EEF performed all the experiments and data analysis, interpreted the results, and prepared figures. DN and RS jointly supervised the study and interpreted the results. FEE and RS wrote the manuscript with inputs from DN. All authors read and approved the final manuscript.

References

1. Hauer, M.H., and Gasser, S.M. (2017). Chromatin and nucleosome dynamics in DNA damage and repair. Genes Dev. *31*, 2204–2221.

2. Adar, S., Hu, J., Lieb, J.D., and Sancar, A. (2016). Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. Proc. Natl. Acad. Sci. U. S. A. *113*, E2124–E2133.

3. MacAlpine, D.M., and Almouzni, G. (2013). Chromatin and DNA replication. Cold Spring Harb. Perspect. Biol. *5*, a010207.

4. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

5. Li, Y., Haarhuis, J.H.I., Sedeño Cacciatore, Á., Oldenkamp, R., van Ruiten, M.S., Willems, L., Teunissen, H., Muir, K.W., de Wit, E., Rowland, B.D., et al. (2020). The structural basis for cohesin-CTCF-anchored loops. Nature *578*, 472–476.

6. Nora, E.P., Caccianini, L., Fudenberg, G., Kameswaran, V., Nagle, A., Uebersohn, A., So, K., Hajj, B., Le Saux, A., Coulon, A., et al. Molecular basis of CTCF binding polarity in genome folding. 10.1101/2019.12.13.876177.

7. Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.e22.

8. Dekker, J., and Mirny, L. (2016). The 3D genome as moderator of chromosomal communication. Cell *164*, 1110–1121.

9. Krivega, I., and Dean, A. (2017). CTCF fences make good neighbours. Nat. Cell Biol. *19*, 883–885.

10. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature *532*, 264–267.

11. Poulos, R.C., Thoms, J.A.I., Guan, Y.F., Unnikrishnan, A., Pimanda, J.E., and Wong, J.W.H. (2016). Functional mutations form at CTCF-cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif. Cell Rep. *17*, 2865–2872.

12. Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., et al. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. *47*, 818–821.

13. Kaiser, V.B., Taylor, M.S., and Semple, C.A. (2016). Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. PLoS Genet. *12*, e1006207.

14. Uusküla-Reimand, L., Hou, H., Samavarchi-Tehrani, P., Rudan, M.V., Liang, M., Medina-Rivera, A., Mohammed, H., Schmidt, D., Schwalie, P., Young, E.J., et al. (2016). Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. Genome Biol. *17*, 182.

15. Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S.S.P., et al. (2017). Genome Organization Drives Chromosome Fragility. Cell *170*, 507–521.e18.

16. Minchell, N.E., Keszthelyi, A., and Baxter, J. (2020). Cohesin causes replicative DNA damage by trapping DNA topological stress. Mol. Cell *78*, 739–751.e8.

17. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

18. Schönenberger, F., Deutzmann, A., Ferrando-May, E., and Merhof, D. (2015). Discrimination of cell cycle phases in PCNA-immunolabeled cells. BMC Bioinformatics *16*, 180.

19. Oomen, M.E., Hansen, A.S., Liu, Y., Darzacq, X., and Dekker, J. (2019). CTCF sites display cell cycle-dependent dynamics in factor binding and nucleosome positioning. Genome Res. *29*, 236–249.

20. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

21. Syed, A., and Tainer, J.A. (2018). The MRE11-RAD50-NBS1 complex conducts the orchestration of damage signaling and outcomes to stress in DNA replication and repair. Annu. Rev. Biochem. *87*, 263–294.

22. Lyu, X., Lei, K.-H., Biak Sang, P., Shiva, O., Chastain, M., Chi, P., and Chai, W. (2021). Human CST complex protects stalled replication forks by directly blocking MRE11 degradation of nascent-strand DNA. EMBO J. *40*, e103654.

23. Chastain, M., Zhou, Q., Shiva, O., Fadri-Moskwik, M., Whitmore, L., Jia, P., Dai, X., Huang, C., Ye, P., and Chai, W. (2016). Human CST facilitates genome-wide RAD51 recruitment to GC-rich repetitive sequences in response to replication stress. Cell Rep. *16*, 1300–1314.

24. Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. Philos. trans. R. Soc. Lond. *187*, 253–318.

25. Manders, E.M.M., Verbeek, F.J., and Aten, J.A. (1993). Measurement of co-localization of objects in dual-colour confocal images. J. Microsc. *169*, 375–382.

26. Dunn, K.W., Kamocka, M.M., and McDonald, J.H. (2011). A practical guide to evaluating colocalization in biological microscopy. Am. J. Physiol. Cell Physiol. *300*, C723–C742.

27. Khurana, S., and Oberdoerffer, P. (2015). Replication Stress: A Lifetime of Epigenetic Change. Genes *6*, 858–877.

28. Podhorecka, M., Skladanowski, A., and Bozko, P. (2010). H2AX Phosphorylation: Its Role in DNA Damage Response and Cancer Therapy. J. Nucleic Acids *2010*. 10.4061/2010/920161.

29. Valdiglesias, V., Giunta, S., Fenech, M., Neri, M., and Bonassi, S. (2013). γH2AX as a marker of DNA double strand breaks and genomic instability in human population studies. Mutat. Res. *753*, 24–40.

30. Prado, F. (2018). Homologous Recombination: To Fork and Beyond. Genes *9*. 10.3390/genes9120603.

31. Costes, A., and Lambert, S.A.E. (2012). Homologous recombination as a replication fork escort: fork-protection and recovery. Biomolecules *3*, 39–71.

32. Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.-L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat. Commun. *7*, 10208.

33. Dellino, G.I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., and Pelicci, P.G. (2013). 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.

34. Gothe, H.J., Bouwman, B.A.M., Gusmao, E.G., Piccinno, R., Petrosino, G., Sayols, S., Drechsel, O., Minneker, V., Josipovic, N., Mizi, A., et al. (2019). Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. Mol. Cell *75*, 267–283.e12.

35. Davidson, I.F., and Peters, J.-M. (2021). Genome folding through loop extrusion by SMC complexes. Nat. Rev. Mol. Cell Biol. *22*, 445–464.

36. Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

37. Sakata, R., Niwa, K., Ugarte La Torre, D., Gu, C., Tahara, E., Takada, S., and Nishiyama, T. (2021). Opening of cohesin's SMC ring is essential for timely DNA replication and DNA loop formation. Cell Rep. *35*, 108999.

38. Han, D., Chen, Q., Shi, J., Zhang, F., and Yu, X. (2017). CTCF participates in DNA damage response via poly(ADP-ribosyl)ation. Sci. Rep. *7*, 43530.

39. Hwang, S.Y., Kang, M.A., Baik, C.J., Lee, Y., Hang, N.T., Kim, B.-G., Han, J.S., Jeong, J.-H., Park, D., Myung, K., et al. (2019). CTCF cooperates with CtIP to drive homologous recombination repair of double-strand breaks. Nucleic Acids Res. *47*, 9160–9179.

40. Lang, F., Li, X., Zheng, W., Li, Z., Lu, D., Chen, G., Gong, D., Yang, L., Fu, J., Shi, P., et al. (2017). CTCF prevents genomic instability by promoting homologous recombination-directed DNA double-strand break repair. Proc. Natl. Acad. Sci. U. S. A. *114*, 10912–10917.

41. Hilmi, K., Jangal, M., Marques, M., Zhao, T., Saad, A., Zhang, C., Luo, V.M., Syme, A., Rejon, C., Yu, Z., et al. (2017). CTCF facilitates DNA double-strand break repair by enhancing homologous recombination repair. Sci Adv *3*, e1601898.

42. Guo, Y.A., Chang, M.M., Huang, W., Ooi, W.F., Xing, M., Tan, P., and Skanderup, A.J. (2018). Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat. Commun. *9*, 1520.

43. Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

44. Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. Elife *6*. 10.7554/elife.25776.

45. Li, Y., Huang, W., Niu, L., Umbach, D.M., Covo, S., and Li, L. (2013). Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. BMC Genomics *14*, 553.

46. Khoury, A., Achinger-Kawecka, J., Bert, S.A., Smith, G.C., French, H.J., Luu, P.-L., Peters, T.J., Du, Q., Parry, A.J., Valdes-Mora, F., et al. (2020). Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat. Commun. *11*, 54.

47. Lyu, X., Chastain, M., and Chai, W. (2019). Genome-wide mapping and profiling of γH2AX binding hotspots in response to different replication stress inducers. BMC Genomics *20*, 579.

48. Canela, A., Maman, Y., Huang, S.-Y.N., Wutz, G., Tang, W., Zagnoli-Vieira, G., Callen, E., Wong, N., Day, A., Peters, J.-M., et al. (2019). Topoisomerase II-Induced Chromosome Breakage and Translocation Is Determined by Chromosome Architecture and Transcriptional Activity. Mol. Cell *75*, 252–266.e8.

49. Groth, A., Rocha, W., Verreault, A., and Almouzni, G. (2007). Chromatin challenges during DNA replication and repair. Cell *128*, 721–733.

50. Alabert, C., and Groth, A. (2012). Chromatin replication and epigenome maintenance. Nat. Rev. Mol. Cell Biol. *13*, 153–167.

51. Ramachandran, S., and Henikoff, S. (2015). Replicating Nucleosomes. Sci Adv *1*. 10.1126/sciadv.1500587.

52. Stewart-Morgan, K.R., Reverón-Gómez, N., and Groth, A. (2019). Transcription Restart Establishes Chromatin Accessibility after DNA Replication. Mol. Cell *75*, 284–297.e6.

53. Xu, C., and Corces, V.G. (2018). Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. Science *359*, 1166–1170.

54. Owens, N., Papadopoulou, T., Festuccia, N., Tachtsidi, A., Gonzalez, I., Dubois, A., Vandormael-Pournin, S., Nora, E.P., Bruneau, B.G., Cohen-Tannoudji, M., et al. (2019). CTCF confers local nucleosome resiliency after DNA replication and during mitosis. Elife *8*. 10.7554/eLife.47898.

55. Mielko, Z., Zhang, Y., Sahay, H., Liu, Y., Schaich, M.A., Schnable, B., Morrison, A.M., Burdinski, D., Adar, S., Pufall, M., et al. (2023). UV irradiation remodels the specificity landscape of transcription factors. Proc. Natl. Acad. Sci. U. S. A. *120*, e2217422120.

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

57. Reijns, M.A.M., Kemp, H., Ding, J., de Procé, S.M., Jackson, A.P., and Taylor, M.S. (2015). Lagging-strand replication shapes the mutational landscape of the genome. Nature *518*, 502–506.

58. Ma, H.T., and Poon, R.Y.C. (2011). Synchronization of HeLa cells. Methods Mol. Biol. *761*, 151–161.

59. Watson, J.V., Chambers, S.H., and Smith, P.J. (1987). A pragmatic approach to the analysis of DNA histograms with a definable G1 peak. Cytometry *8*, 1–8.

60. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

61. Bolte, S., and Cordelières, F.P. (2006). A guided tour into subcellular colocalization analysis in light microscopy. J. Microsc. *224*, 213–232.

62. Adler, J., and Parmryd, I. (2010). Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. Cytometry A *77*, 733–742.

63. Costes, S.V., Daelemans, D., Cho, E.H., Dobbin, Z., Pavlakis, G., and Lockett, S. (2004). Automatic and quantitative measurement of protein-protein colocalization in live cells. Biophys. J. *86*, 3993–4003.

64. Ewels, P., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Ulysse Garcia, M., Di Tommaso, P., and Nahnsen, S. (2022). The nf-core framework for community-curated bioinformatics pipelines (Zenodo) 10.5281/ZENODO.3240506.

65. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

66. Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. Nat. Genet. *49*, 1684–1692.

67. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. (2009). MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics *25*, 3181–3182.

68. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173.

69. Pizzi, C., Rastas, P., and Ukkonen, E. (2011). Finding significant matches of position weight matrices in linear time. IEEE/ACM Trans. Comput. Biol. Bioinform. *8*, 69–79.

Supplementary Figures:

Figure S1

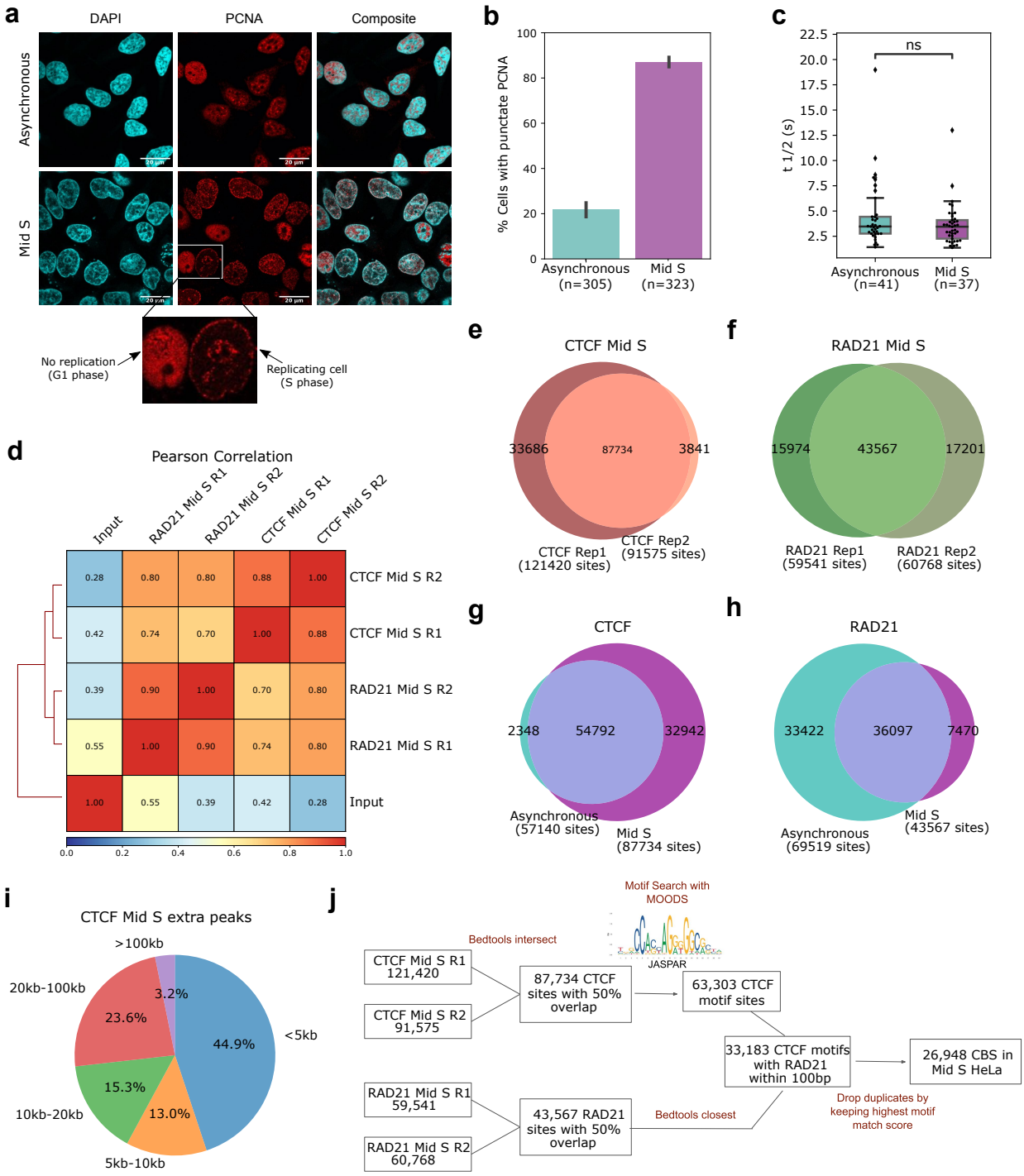

Figure S1: Profiling of CTCF/cohesin binding in replication phase. **a)** Representative images of PCNA stained asynchronous and Mid S synchronised HeLa cells. Nuclei were counterstained with DAPI. The scale bar is 20μM. **b)** Quantification of punctate PCNA in asynchronous and Mid S synchronised cells from five biologically independent experiments. The bar represents mean value and the error bar represents SD. **c)** Half time recovery ($t_{1/2}$) of CTCF FRAP in asynchronous and Mid S HeLa cells. Centerline: median; Box limits: 1st and 3rd quartile; whiskers: maximum and minimum values: 1.5xIQR. The p-value was calculated by two-sided Mann-Whitney U test: ns: $5.00x10^{-02} < p <= 1$ **d)** Hierarchically clustered correlation heatmap of CTCF and RAD21 ChIP-seq replicates in Mid S HeLa cells. DeepTools' multiBamSummary followed by plotCorrelation functions were used to find the Pearson correlation between the samples. **e), f)** Venn diagram showing the number of shared (with 50% overlap between the peaks) and unique peaks of **e)** CTCF and **f)** RAD21 biological replicates in the Mid S HeLa cells. **g)** Intersection of CTCF ChIP-seq peaks in asynchronous and Mid S HeLa cells (see Methods). **h)** Intersection of RAD21 ChIP-seq peak in asynchronous (from ENCODE) and Mid S HeLa cells. **i)** Pie chart depicting the distance between extra CTCF ChIP-seq peaks observed in Mid S with the common peaks from Mid S and asynchronous cells. **j)** Workflow for defining CTCF/cohesin binding sites (CBSs; i.e., regions bound by both CTCF and RAD21) in Mid S HeLa cells.

Figure S2

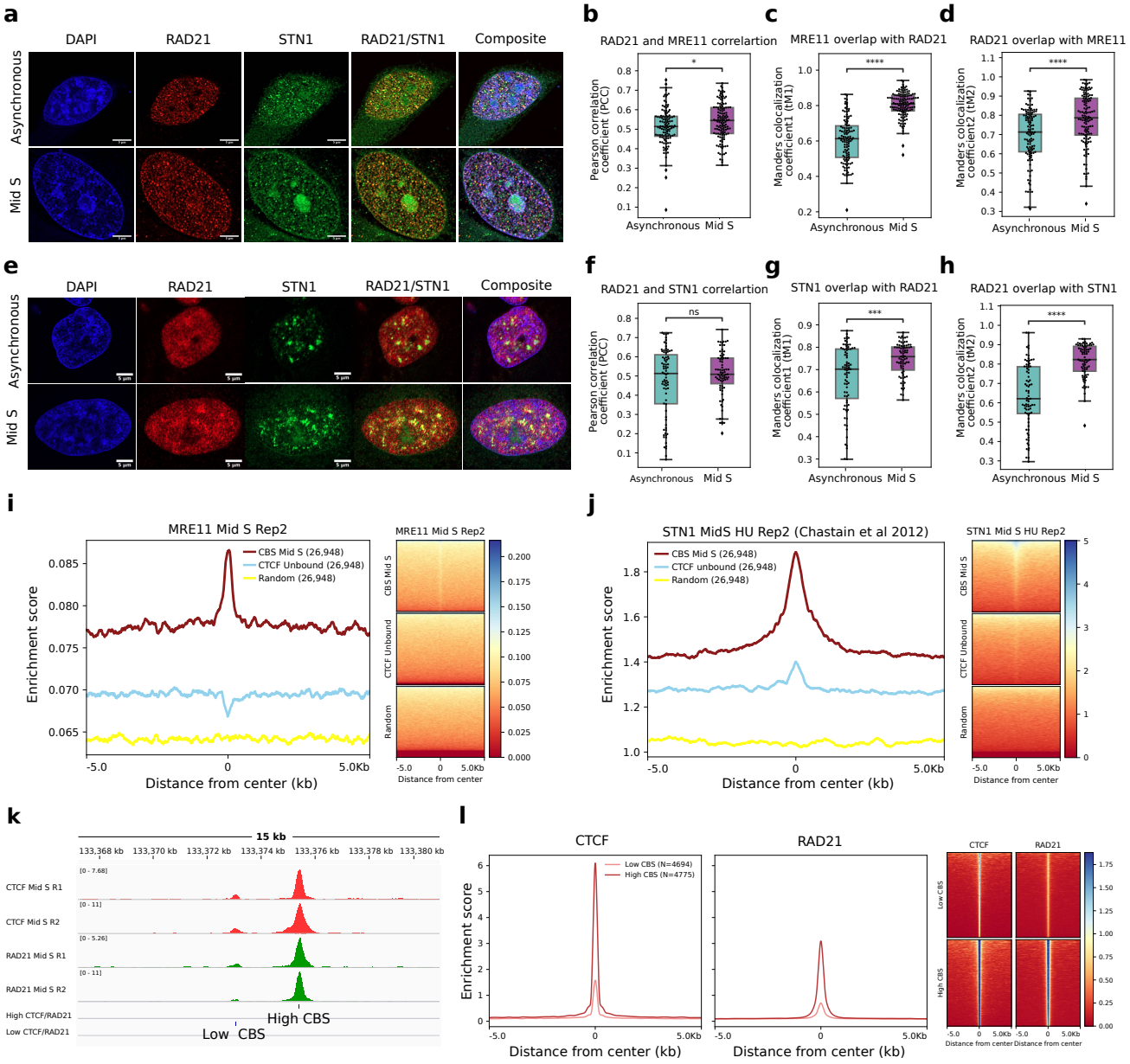

Figure S2: Enrichment of replication stress associated proteins with CTCF/cohesin binding to DNA. a) Representative images of asynchronous and Mid S HeLa cells co-immunostained with RAD21 and MRE11 from three biologically independent experiments. DNA was counter-stained with DAPI. Images were acquired with LSM 980 Airyscan microscope in super-resolution (SR) mode. The scale bar is 5µm. No. of asynchronous cells=118, No. of Mid S cells=129. Quantification of nuclear colocalization between RAD21 and MRE11 by **b)** Pearson's Correlation Coefficient (PCC: $P=3.24x10^{-02}$) and Mander's overlap coefficients **c)** tM1 for RAD21 ($P=1.28x10^{-29}$) and **d)** tM2 for MRE11 ($P=1.12x10^{-5}$). **e)** Representative images of RAD21 co-

stained with STN1 in asynchronous and S-phase synchronised cells. No. of cells n=79, No. of biological replicates N=2. Colocalization quantified by **f)** PCC ($P=4.1x10^{-01}$) and Mander's **g)** tM1 ($P=6.02x10^{-4}$) and **h)** tM2 ($P=1.44x10^{-12}$). Colocalization was quantified using the JACoP plugin in Fiji. DAPI staining was used to draw ROI to get only the nuclear signals. The threshold signal was determined with Costes' automatic threshold method and manual adjustment. Centerline: median; Box limits: 1st and 3rd quartile; whiskers: 1.5xIQR. p-values were determined by a two-sided non-parametric Mann-Whitney test. p-value annotation legend: ns: $5.00x10^{-02} < p <= 1$, *: $1.00x10^{-02} < p <= 5.00x10^{-02}$, **: $1.00x10^{-03} < p <= 1.00x10^{-02}$, ***: $1.00x10^{-04} < p <= 1.00x10^{-03}$, ****: $p <= 1x10^{-04}$. **i)** Second biological replicate of MRE11 ChIP-seq signal in Mid S HeLa cells plotted across a 5kb window centred at Mid S CBSs (maroon), CTCF unbound sites (sky blue) and random genomic regions (yellow). 26,948 sites were in each category. **j)** Similarly, the ChIP-seq profile of STN1 (second biological replicate) in Mid S upon HU treatment from Chastain et al. [23] plotted around CBSs and control sites (+/-5kb). **k)** Genome browser snapshot of two CBSs showing low- or high-binding of CTCF and RAD21. **l)** ChIP-seq profile and heatmap of CTCF and RAD21 at low- and high-binding sites.

Figure S3

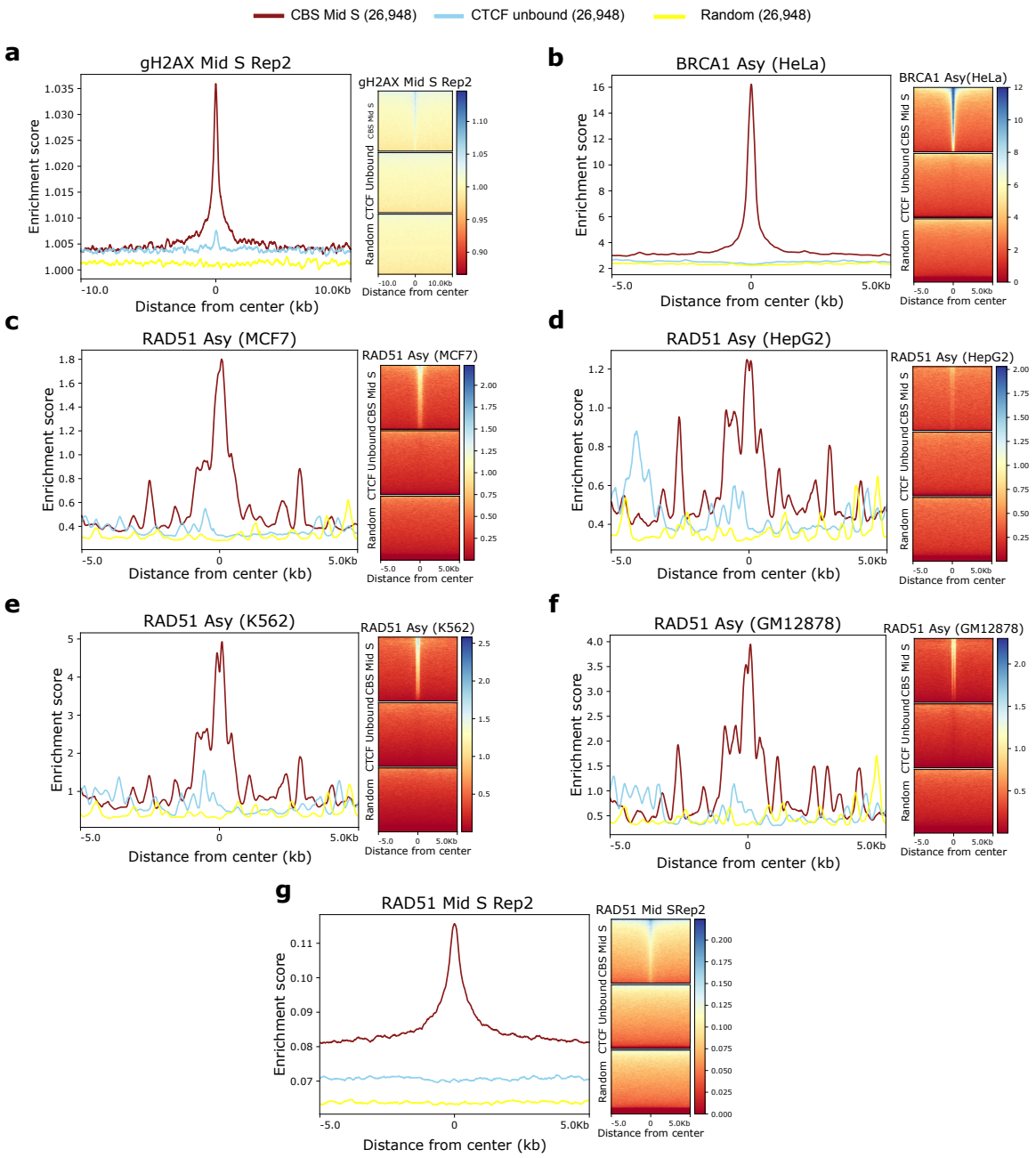

Figure S3: Enrichment of γH2AX, BRCA1 and RAD21 ChIP-seq signals at CBSs. a) γH2AX ChIP-seq signal in Mid S HeLa cells from second biological replicate plotted at CBS Mid S, CTCF unbound sites and random sites. **b)** BRCA1 ChIP-seq signal from the asynchronous population of HeLa cells downloaded from ENCODE plotted at CBS and control sites. RAD51 ChIP-seq from ENCODE **c)** MCF7 **d)** HepG2 **e)** K562 and **f)** GM12878 cells plotted at CBSs and control sites in asynchronous. **g**) Second replicate of RAD51 ChIP-seq signal from Mid S synchronised HeLa cells at CBS Mid S and control sites.

Figure S4

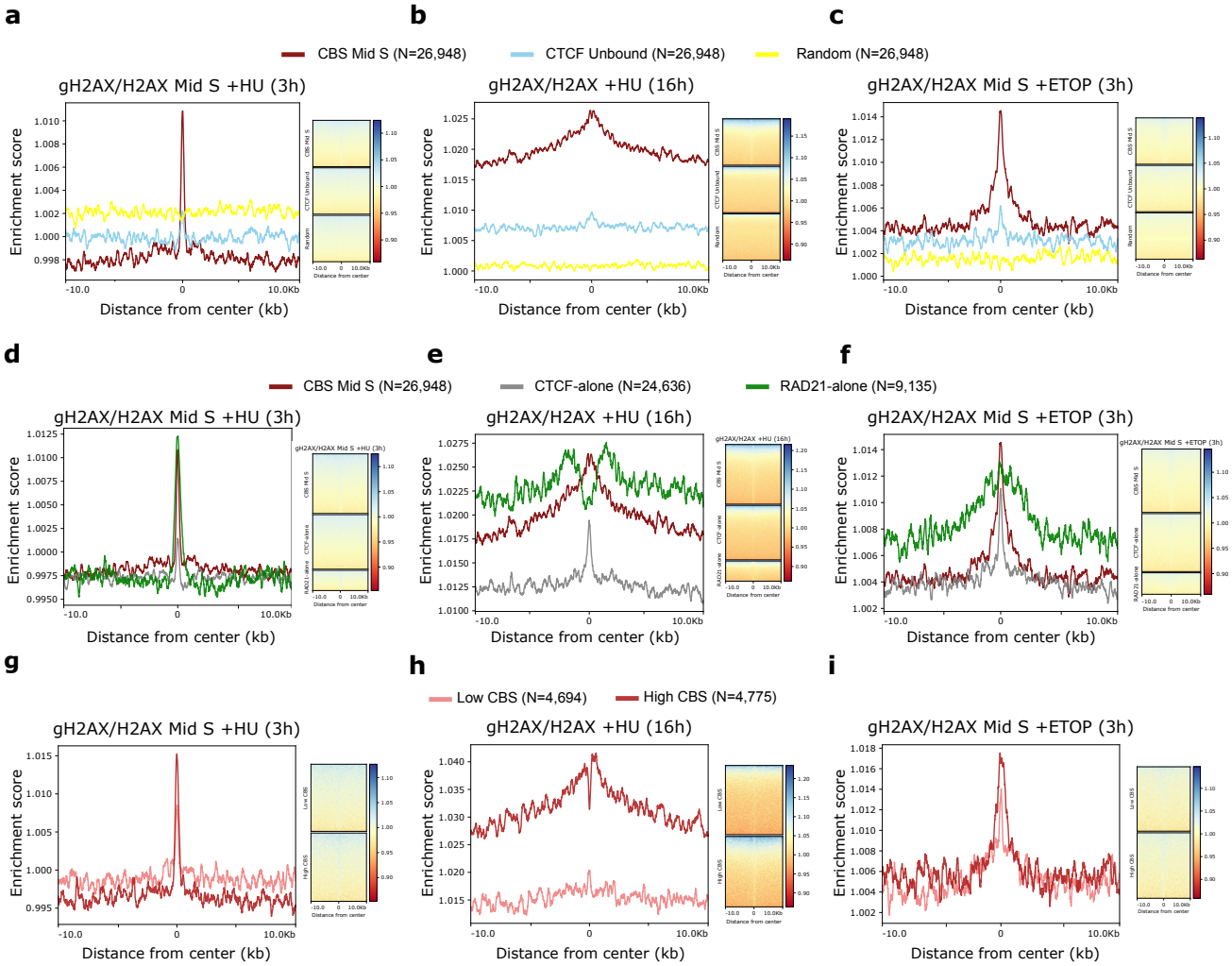

Figure S4: ChIP-seq signal of γH2AX in exogenous replication stress conditions. ChIP-seq profile and heatmap of γH2AX (normalised with H2AX) from **a)** Mid S HeLa cells treated with 2 mM HU for three hours **b)** HeLa cells treated with 2 mM HU for 16 hours and **c)** Mid S HeLa cells treated with 25μM Etoposide for 3 hours plotted at CBSs (26,948 sites), CTCF unbound (26,948 sites) and random regions (26,948 sites). ChIP-seq profile and heatmap of γH2AX (normalised with H2AX) from **d)** Mid S HeLa cells treated with 2 mM HU for three hours **e)** HeLa cells treated with 2 mM HU for 16 hours and **f)** Mid S HeLa cell treated with 25μM Etoposide for 3 hours plotted at CBS Mid S (26,948 sites), CTCF-alone (24,636 sites) and RAD21-alone sites (9,135 sites). ChIP-seq signal of γH2AX (normalised with H2AX) at low (4,694 sites) and high CBSs (4,775 sites) in +/-10kb window in **g)** Mid S HeLa cells treated with 2 mM HU for three hours **h)** HeLa cells treated with 2 mM HU for 16 hours and **i)** Mid S HeLa cell treated with 25μM Etoposide for 3 hours.

Figure S5

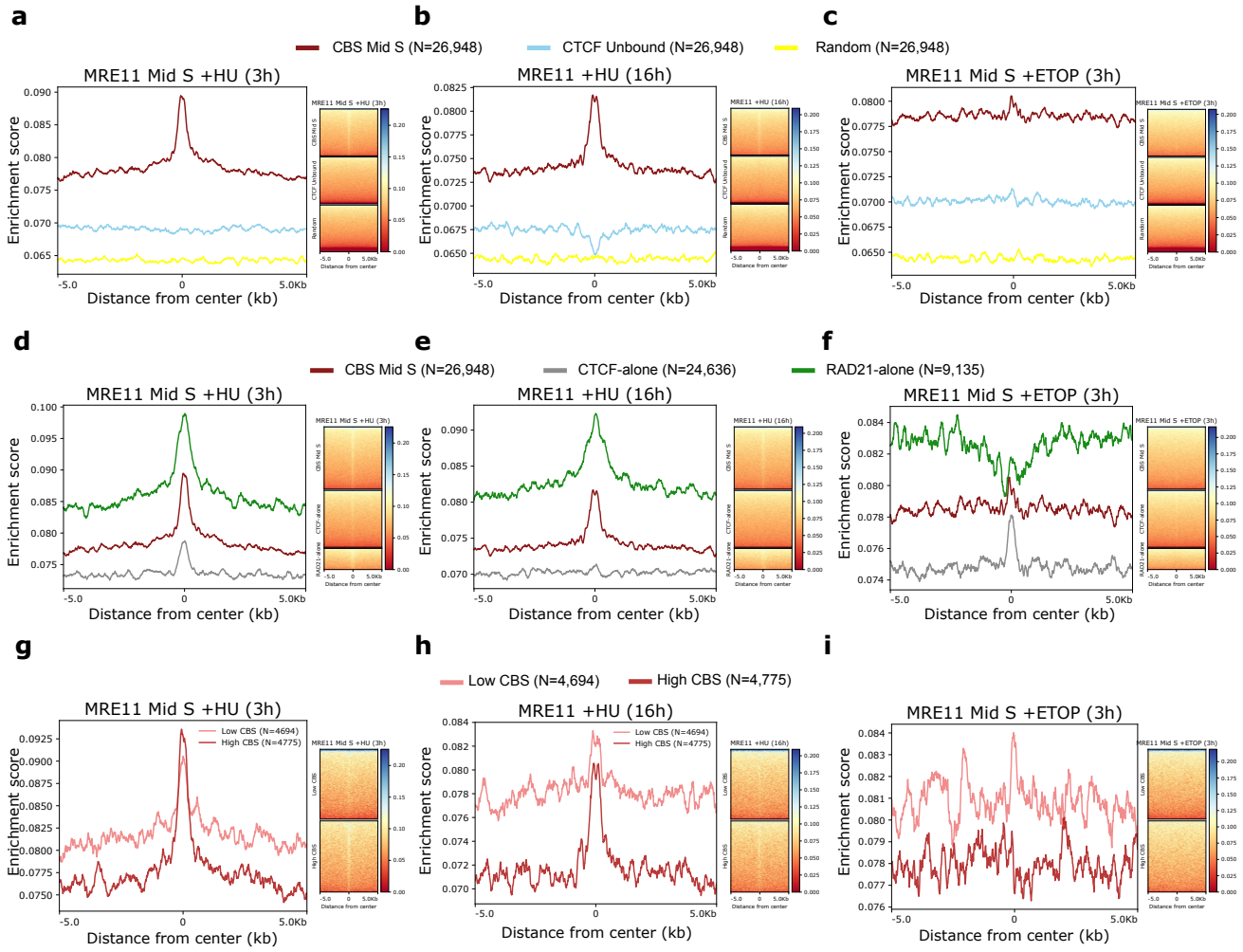

Figure S5: ChIP-seq signal of MRE11 in exogenous replication stress conditions. ChIP-seq profile and heatmap of MRE11 from **a)** Mid S HeLa cells treated with 2 mM HU for three hours **b)** HeLa cells treated with 2 mM HU for 16 hours and **c)** Mid S HeLa cells treated with 25 μM Etoposide for 3 hours plotted at CBSs (26,948 sites), CTCF unbound (26,948 sites) and random regions (26,948 sites). ChIP-seq profile and heatmap of MRE11 from **d)** Mid S HeLa cells treated with 2mM HU for three hours **e)** HeLa cells treated with 2 mM HU for 16 hours and **f)** Mid S HeLa cells treated with 25 μM Etoposide for 3 hours plotted at CTCF-alone sites (24,636 sites), RAD21-alone sites (9,135 sites) and CBS Mid S (26,948 sites). ChIP-seq signal of MRE11 at low (4,694 sites) and high CBSs (4,775 sites) in +/-10kb window in **g)** Mid S HeLa cells treated with 2 mM HU for three hours **h)** HeLa cells treated with 2 mM HU for 16 hours and **i)** Mid S HeLa cell treated with 25 μM Etoposide for 3 hours.

Figure S6

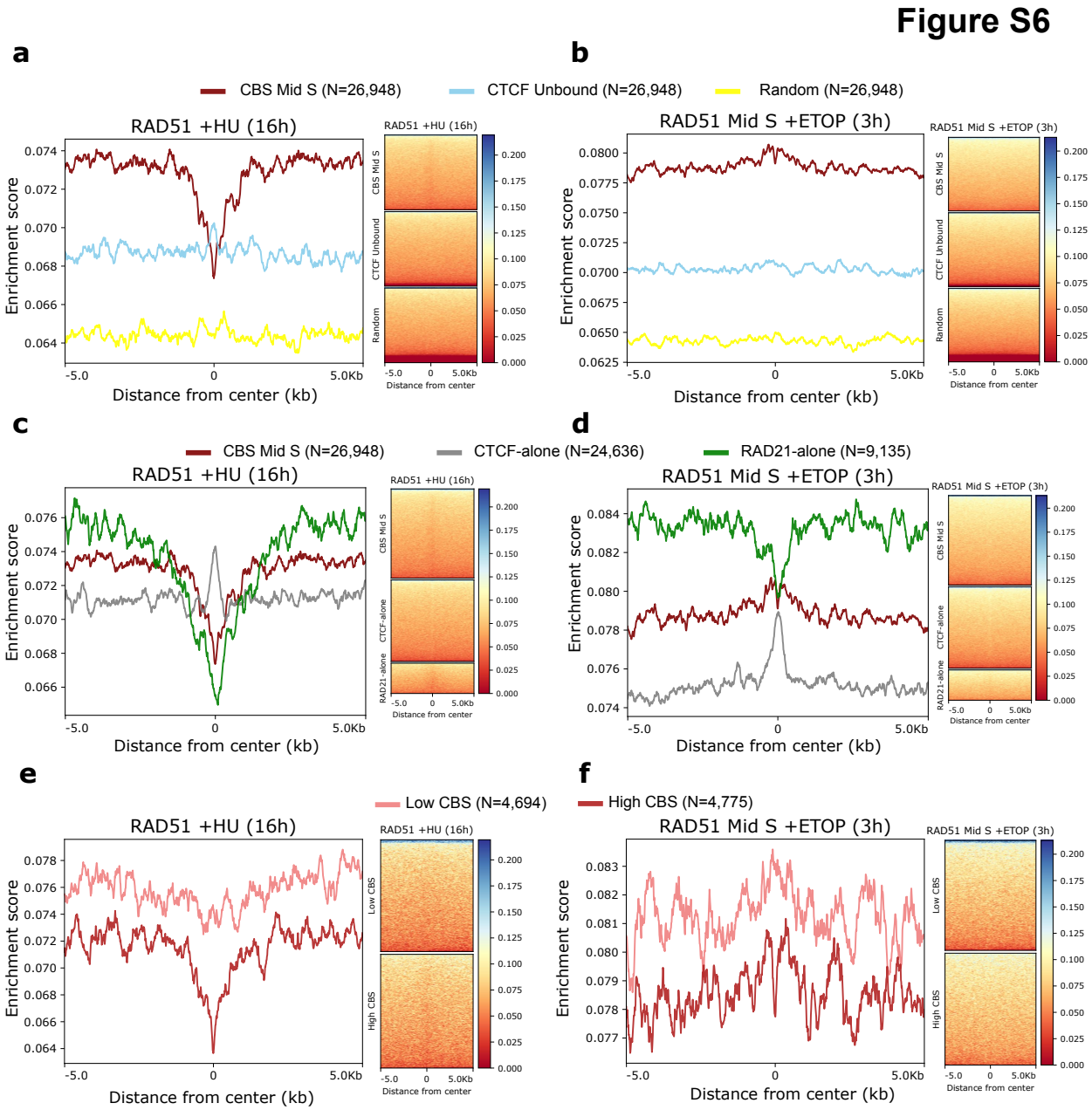

Figure S6: ChIP-seq signal of RAD51 in exogenous replication stress conditions. ChIP-seq profile and heatmap of RAD51 from **a)** HeLa cells treated with 2 mM HU for 16 hours and **b)** Mid S HeLa cell treated with 25 μM Etoposide for 3 hours at CBSs (26,948 sites), CTCF unbound (26,948 sites) and random regions (26,948 sites). ChIP-seq profile and heatmap of RAD51 from **c)** HeLa cells treated with 2 mM HU for 16 hours and **d)** Mid S HeLa cell treated with 25 μM Etoposide for 3 hours plotted at CTCF-alone sites (24,636 sites), RAD21-alone sites (9,135 sites) and CBS Mid S (26,948 sites). ChIP-seq signal of RAD51 at low (4,694 sites) and high CBSs (4,775 sites) in +/-10kb window in **e)** Mid S HeLa cells treated with 2 mM HU for three hours and **f)** Mid S HeLa cell treated with 25 μM Etoposide for 3 hours.

Figure S7

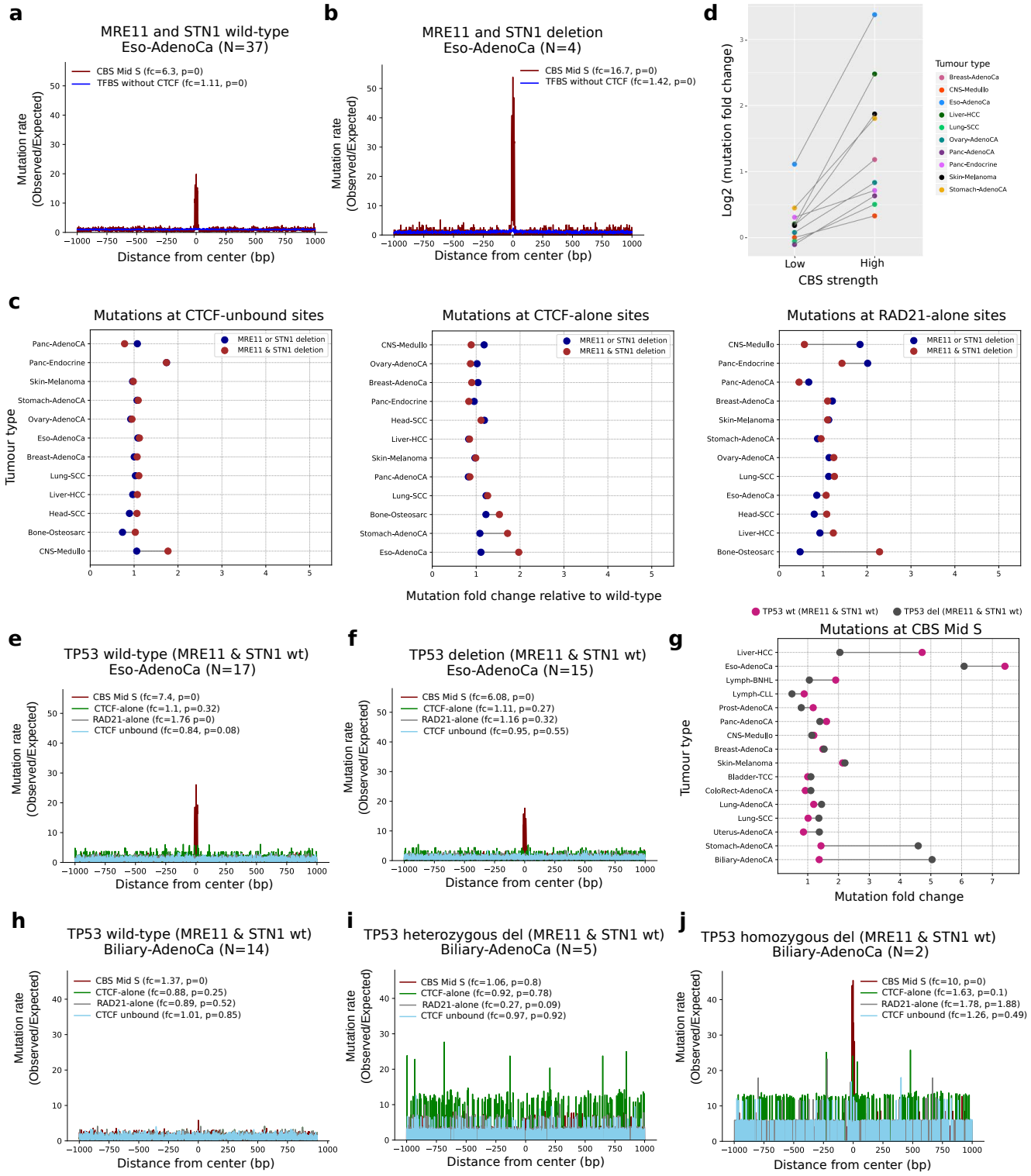
a
MRE11 and STN1 wild-type
Eso-AdenoCa (N=37)
CBS Mid S (fc=6.3, p=0)
TFBS without CTCF (fc=1.11, p=0)
Mutation rate (Observed/Expected)
Distance from center (bp)
b
MRE11 and STN1 deletion
Eso-AdenoCa (N=4)
CBS Mid S (fc=16.7, p=0)
TFBS without CTCF (fc=1.42, p=0)
Mutation rate (Observed/Expected)
Distance from center (bp)
d
Log2 (mutation fold change)
CBS strength
Low
High
Tumour type
Breast-AdenoCa
CNS-Medullo
Eso-AdenoCa
Liver-HCC
Lung-SCC
Ovary-AdenoCA
Panc-AdenoCA
Panc-Endocrine
Skin-Melanoma
Stomach-AdenoCA
c
Mutations at CTCF-unbound sites
Mutations at CTCF-alone sites
Mutations at RAD21-alone sites
MRE11 or STN1 deletion
MRE11 & STN1 deletion
Tumour type
Mutation fold change relative to wild-type
e
TP53 wild-type (MRE11 & STN1 wt)
Eso-AdenoCa (N=17)
CBS Mid S (fc=7.4, p=0)
CTCF-alone (fc=1.1, p=0.32)
RAD21-alone (fc=1.76 p=0)
CTCF unbound (fc=0.84, p=0.08)
Mutation rate (Observed/Expected)
Distance from center (bp)
f
TP53 deletion (MRE11 & STN1 wt)
Eso-AdenoCa (N=15)
CBS Mid S (fc=6.08, p=0)
CTCF-alone (fc=1.11, p=0.27)
RAD21-alone (fc=1.16 p=0.32)
CTCF unbound (fc=0.95, p=0.55)
Mutation rate (Observed/Expected)
Distance from center (bp)
g
TP53 wt (MRE11 & STN1 wt)
TP53 del (MRE11 & STN1 wt)
Mutations at CBS Mid S
Tumour type
Mutation fold change
h
TP53 wild-type (MRE11 & STN1 wt)
Biliary-AdenoCa (N=14)
CBS Mid S (fc=1.37, p=0)
CTCF-alone (fc=0.88, p=0.25)
RAD21-alone (fc=0.89, p=0.52)
CTCF unbound (fc=1.01, p=0.85)
Mutation rate (Observed/Expected)
Distance from center (bp)
i
TP53 heterozygous del (MRE11 & STN1 wt)
Biliary-AdenoCa (N=5)
CBS Mid S (fc=1.06, p=0.8)
CTCF-alone (fc=0.92, p=0.78)
RAD21-alone (fc=0.27, p=0.09)
CTCF unbound (fc=0.97, p=0.92)
Mutation rate (Observed/Expected)
Distance from center (bp)
j
TP53 homozygous del (MRE11 & STN1 wt)
Biliary-AdenoCa (N=2)
CBS Mid S (fc=10, p=0)
CTCF-alone (fc=1.63, p=0.1)
RAD21-alone (fc=1.78, p=1.88)
CTCF unbound (fc=1.26, p=0.49)
Mutation rate (Observed/Expected)
Distance from center (bp)

Figure S7: Mutation rate plotted at +/-1 kb region of CBS Mid S and Transcription factor binding sites (TFBS excluding CTCF) in Esophageal adenocarcinoma samples with **a)** both MRE11 and STN1 wild-type (No. of samples N=37) **b)** both MRE11 and STN1 deletion sample (N=4). **c)** Mutation ratio between MRE11 or/and STN1 deletion to MRE11 and STN1 wild-type samples at CTCF unbound sites, CTCF-alone sites and RAD21-alone sites across tumour types. **d)** Mutation rate at around low and high binding CTCF/cohesin sites (+/-20 bp) computed in various tumour types based on MRE11-STN1 deletion status. **e, f:** Mutation rate at CBSs and control sites in TP53 **e)** wild-type (N=17) and **f)** TP53 deletion samples (N=15) from Esophageal adenocarcinoma with no alterations in MRE11 and STN1 (i.e., MRE11 and STN1 wild-type). **g)** Mutations fold change at CBSs in TP53 wild-type and deletion samples from different tumour types **h, i, j:** Mutation rate plotted at CBS Mid S and control sites in Biliary adenocarcinoma samples (with background of MRE11 and STN1 wild-type) with **h)** TP53 wild-type (N=14) **i)** TP53 deletion in heterozygous condition (N=5) **j)** TP53 deletion in homozygous condition (N=2).