Supplementary Materials

Table S1. Sequencing and alignment statistics for paired-end Illumina sequencing libraries prepared from MCF10A cells.

Table S2. Sequencing and alignment statistics for paired-end Illumina sequencing libraries prepared from Jurkat cells.

Table S3. Summary of publicly available data used in this study.

Table S4. Summary of CTCF ChIP-seq binding sites (all sites or CTCF motif-containing sites) for each cell line used in this study.

Figure S1. Reproducibility of genome-wide DSB mapping/sequencing in MCF10A and Jurkat cells

Figure S2. DSBs are significantly enriched at strong CTCF binding sites in five cell types, and strong alternative DNA secondary structures are also significantly present at these sites.

Figure S3. Strong CTCF binding sites are enriched for DSBs and shared between cell lines.

Figure S4. Distribution of strong (top 10%) and weak (bottom 10%) CTCF binding sites among genic and intergenic sites do not explain DSB enrichment.

Figure S5. No differences in gene expression were observed among genic CTCF binding sites binned by CTCF binding strength, and gene expression does not cause increased DSBs in strong CTCF binding sites.

Figure S6. DSBs are enriched at strong, but not weak, CTCF binding sites in a dose-dependent manner after exposure to etoposide.

Figure S7. Etoposide treatment decreases CTCF expression in MCF10A cells while increasing damage markers.

Figure S8. Validation of lost, gained, and unchanged CTCF binding sites in shCTCF and shLuc MCF10A cells by CTCF ChIP-qPCR.

Figure S9. TAD boundary-associated CTCF binding sites are enriched for DSBs, G-quadruplexes, and TOP2 binding compared to loop-associated CTCF binding sites.

Figure S10. TAD boundary-associated CTCF binding sites are enriched among strong CTCF binding sites and are enriched for DSBs compared to loop-associated CTCF binding sites.

Table S1. Sequencing and alignment statistic for DSB paired-end Illumina sequencing libraries prepared from MCF10A cells.

Treatment	Replicates		# Sequenced read pairs[a]	% Alignment rate[b]	# Proper read pairs[c]	% Duplication[d]	# Mapped DSBs[e]	% Mapped DSBs[f]
	Biological	Technical						
UT	N1	1	6647933	91.58	5797063	9.95	4791320	72.07
	N2	1	20619652	72.44	9750389	29.98	6476801	31.41
	N3	1	11973646	91.07	10105182	35.68	5916836	49.42
	N4	1	11162025	95.25	10041620	12.83	7783827	69.73
		2	11068591	95.63	10027803	12.52	7805911	70.52
0.15µM ETO	N1	1	15731579	86.53	12840923	43.32	6214941	39.51
	N2	1	28738165	82.23	22148286	67.43	6146314	21.39
	N3	1	11998194	91.20	10073548	36.32	5886143	49.06
	N4	1	11174549	96.04	10214225	27.50	6452374	57.54
		2	10686941	95.94	9748291	26.42	6278009	58.74
15µM ETO	N1	1	15873188	87.97	13208864	29.14	8221660	51.80
	N2	1	20199989	91.29	17170220	62.57	5610764	27.78
	N3	1	15946006	71.46	10280992	50.84	4446327	27.88
	N4	1	9773149	94.56	8567131	30.94	5080006	51.98
		2	9274183	94.92	8189631	29.71	4966564	53.55
shLuc	N1	1	16426275	92.86	14626110	34.90	8568137	52.16
	N2	2	17277728	93.88	15560475	24.38	10656435	61.68
shCTCF	N1	1	13178237	93.18	11707603	36.78	6892369	52.30
	N2	2	15076213	92.16	13277674	27.57	8969817	59.50

[a] Number of raw paired read1 and read2 following Illumina paired-end sequencing and quality filtering.

[b] Percentage of reads that had at least one alignment to the hg38 genome assembly as processed and reported by bowtie2 alignment tool.

[c] Following the quality control removal of all unmapped, non-primary, supplementary, and low-quality reads, the remaining number of paired read1 and read2 are indicated.

[d] PCR duplicates are marked and removed meaningfully using read1 and read2 alignment, and "% Duplication" is based on original "# Sequenced read pairs".

[e] For each non-duplicated pair, only read1 is kept, and the 5' most nucleotide of read1 that defines the DNA break position.

[f] "% Mapped DSBs" was calculated by dividing "# Mapped DSBs" by "# Sequenced read pairs".

Table S2. Sequencing and alignment statistic for DSB paired-end Illumina sequencing libraries prepared from Jurkat cells.

Cell type	Biological replicates	# Sequenced read pairs[a]	% Alignment rate[b]	# Proper read pairs[c]	% Duplication[d]	# Mapped DSBs[e]	% Mapped DSBs[f]
Jurkat	N1	14449233	90.43	12111851	57.68	4555910	31.53
	N2	21158037	94.39	18842415	43.01	9793050	46.29
	N3	24864775	93.87	21898447	47.91	10328841	41.54
	N4	30315964	41.56	11968089	10.63	9964927	32.87
	N5	16859985	94.70	14987915	44.69	7572104	44.91

[a] Number of raw paired read1 and read2 following Illumina paired-end sequencing and quality filtering.
[b] Percentage of reads that had at least one alignment to the hg38 genome assembly as processed and reported by bowtie2 alignment tool.
[c] Following the quality control removal of all unmapped, non-primary, supplementary, and low-quality reads, the remaining number of paired read1 and read2 are indicated.
[d] PCR duplicates are marked and removed meaningfully using read1 and read2 alignment, and "% Duplication" is based on original "# Sequenced read pairs".
[e] For each non-duplicated pair, only read1 is kept, and the 5' most nucleotide of read1 that defines the DNA break position.
[f] "% Mapped DSBs" was calculated by dividing "# Mapped DSBs" by "# Sequenced read pairs".

Table S3. Summary of publicly available data used in this study.

Species / Cell Line	Data Type	Accession	Reference
H. sapiens / GM12878	CTCF ChIP-seq	ENCSR000AKB	ENCODE
	RNA-seq	SRR1153470	Tilgner *et al*.
H. sapiens / GM13069	DSB Mapping	PRJNA497476	Szlachta *et al*.
H. sapiens / HaCaT	BG4 ChIP-seq	GSE99205	Hansel-Hertsch *et al*. (2018)
H. sapiens / HeLa	CTCF ChIP-seq	ENCSR000A0A	ENCODE
	RNA-seq	GSE95452	Tchasovnikarova *et al*.
	DSB Mapping	PRJNA579071	Singh *et al.*
H. sapiens / Jurkat	CTCF ChIP-seq	GSE68976	Hnisz *et al*.
H. sapiens / K562	BG4 ChIP-seq	GSE107690	Mao *et al.*
H. sapiens / MCF10A	CTCF ChIP-seq	GSE98551	Fritz *et al.*
	CTCF ChIP-seq	GSE183381	Lebeau *et al.*
	RNA-seq	GSE45258	Kang *et al*.
	TOP2B ChIP-seq	SRR5136803	Dellino *et al*.
H.sapiens / NHEK	BG4 ChIP-seq	GSE76688	Hansel-Hertsch *et al*. (2016)
H. sapiens / NPC	CTCF ChIP-seq	GSM3498323	ENCODE
	RNA-seq	PRJNA591220	Michel *et al*.
	DSB Mapping	PRJNA542485	Szlachta *et al.*
H. sapiens / RPE-1	CTCF ChIP-seq	SRR299281/604593	ENCODE
	CC-seq	SRP187576	Gittens *et al*.
H.sapiens / U2OS	BG4 CUT&Tag	GSE181373	Hui *et al*.

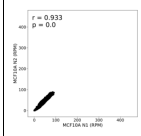
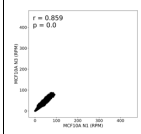
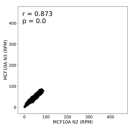
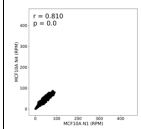
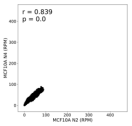
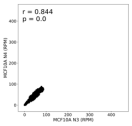
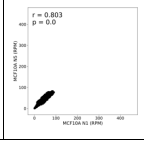
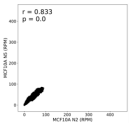
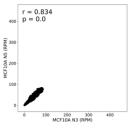
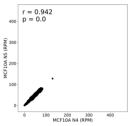
Table S4. Summary of CTCF ChIP-seq binding sites (all sites or CTCF motif-containing sites) for each cell line used in this study.

Cell line	GM12878	HeLa	MCF10A	NPC	Jurkat
CTCF ChIP-seq binding peaks*	40,191	69,117	60,115	98,416	68,203
CTCF ChIP-seq binding peaks with CTCF motifs**	35,288	53,864	48,781	69,117	55,931
% ChIP-seq peaks with CTCF motifs	87.8	77.9	81.1	70.2	82.0

* The publicly available data (Table S3) for CTCF ChIP-seq from GM12878, HeLa, MCF10A, NPC, and Jurkat, and each associated input data, were downloaded and aligned to the GRCh38/hg38 genome using bowtie2 (v 2.3.4.1). Binding peaks were called by the macs2 tool (v 2.2.9.1) using the default setting with each dataset controlled for the matching input data (-c).

**BEDtools (v 2.27.1) intersect between called CTCF ChIP-seq peaks and a list of determined genome-wide CTCF motifs (n = 887,981) (Fang *et al.* 2020) was performed, and CTCF binding peaks were refined by excluding the binding sites that lack CTCF motifs

A

MCF10A, r	N1	N2	N3	N4-T1	N4-T2
N1		0.933	0.859	0.810	0.803
N2			0.873	0.839	0.833
N3				0.844	0.834
N4-T1					0.942
N4-T2					

B

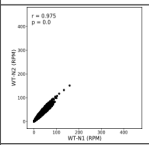
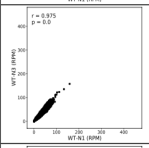
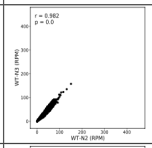
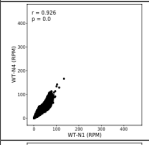
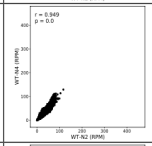
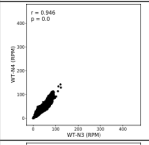
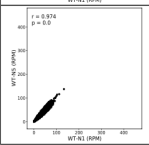
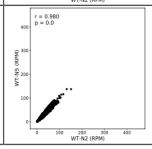
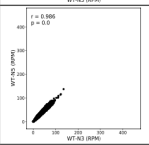
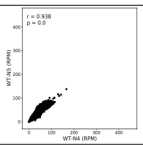
Jurkat, r	N1	N2	N3	N4	N5
N1		0.975	0.975	0.926	0.974
N2			0.982	0.949	0.980
N3				0.949	0.989
N4					0.938
N5					

Figure S1. Reproducibility of genome-wide DSB mapping/sequencing in MCF10A (A) and Jurkat (B) cells. Scatter plot of genome-wide DSB mapping/sequencing reads from biological replicates (untreated N1-N4 for MCF10A cells, Table S1; and N1-N5 for Jurkat cells, Table S2) show a strong correlation (Pearson's correlation r = 0.803-0.942 for MCF10A, r = 0.926-0.989 for Jurkat, $p \sim 0$). Read-normalized coverage for each preparation was calculated for 100kb genome-wide, non-overlapping windows (n = 30,895), and Pearson's correlation was calculated.

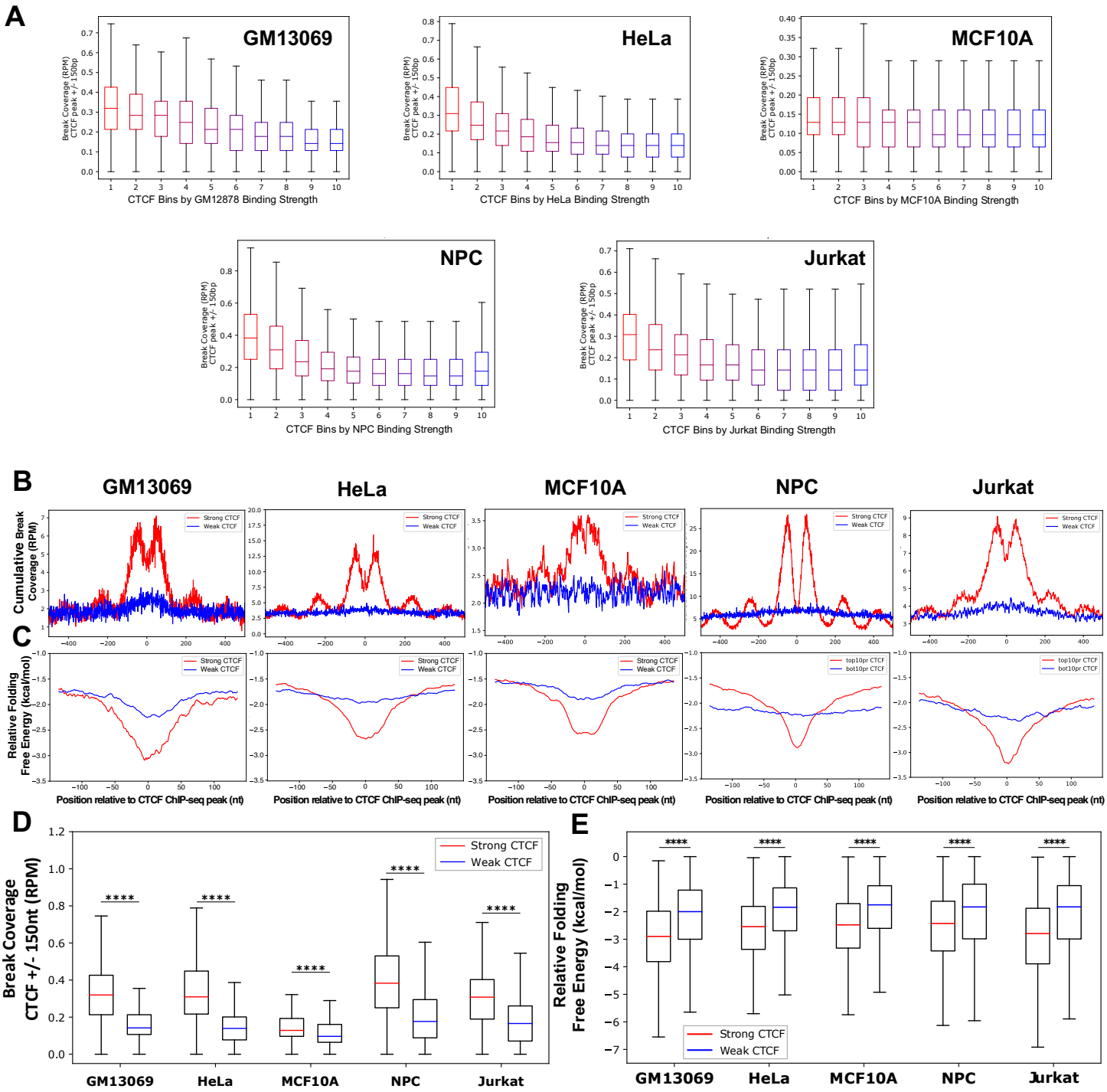


Figure S2. DSBs are significantly enriched at strong CTCF binding sites in five cell types, and alternative DNA secondary structures are also significantly present at these sites. (A) Mapped DSBs in five untreated cell lines are enriched in stronger CTCF binding sites and medians decrease with CTCF binding strength (± 150 bp, RPM). **(B)** DSBs are enriched at the top 10% strongest CTCF binding sites (strong, red), but not at the 10% weakest CTCF binding sites (weak, blue) in untreated GM13069 (n = 4019), HeLa (n = 6911), MCF10A (n = 6011), NPC (n = 9841), and Jurkat (n = 6820) cells, as demonstrated by cumulative DSB coverage (RPM, Reads Per Million) at these sites. **(C)** DNA sequences around strong CTCF binding sites (red, ± 150 nt) form more energetically favorable structures (ΔG, kcal/mol) than sequences around weak CTCF binding sites (blue, ± 150 nt), as determined by folding predictions of single-stranded DNA using ViennaRNA with DNA thermodynamic parameters and a 30 nt sliding window with a 1 nt step; a low ΔG (kcal/mol) indicates sequences are more favorable to form alternative DNA secondary structure. **(D)** Read-normalized DSB coverage (RPM, reads per million) was significantly greater at the strong (red) versus the weak (blue) CTCF binding sites (± 150 nt) in each cell line. **(E)** Relative folding free energy was significantly more favorable at the strong (red) versus the weak (blue) CTCF binding peaks (± 150 nt) for each cell line. Boxes denote 25th and 75th-percentiles, middle lines show medians, and whiskers span from 5% to 95%; **** indicates $p \sim 0$; two-sample, Kolmogorov-Smirnov test.

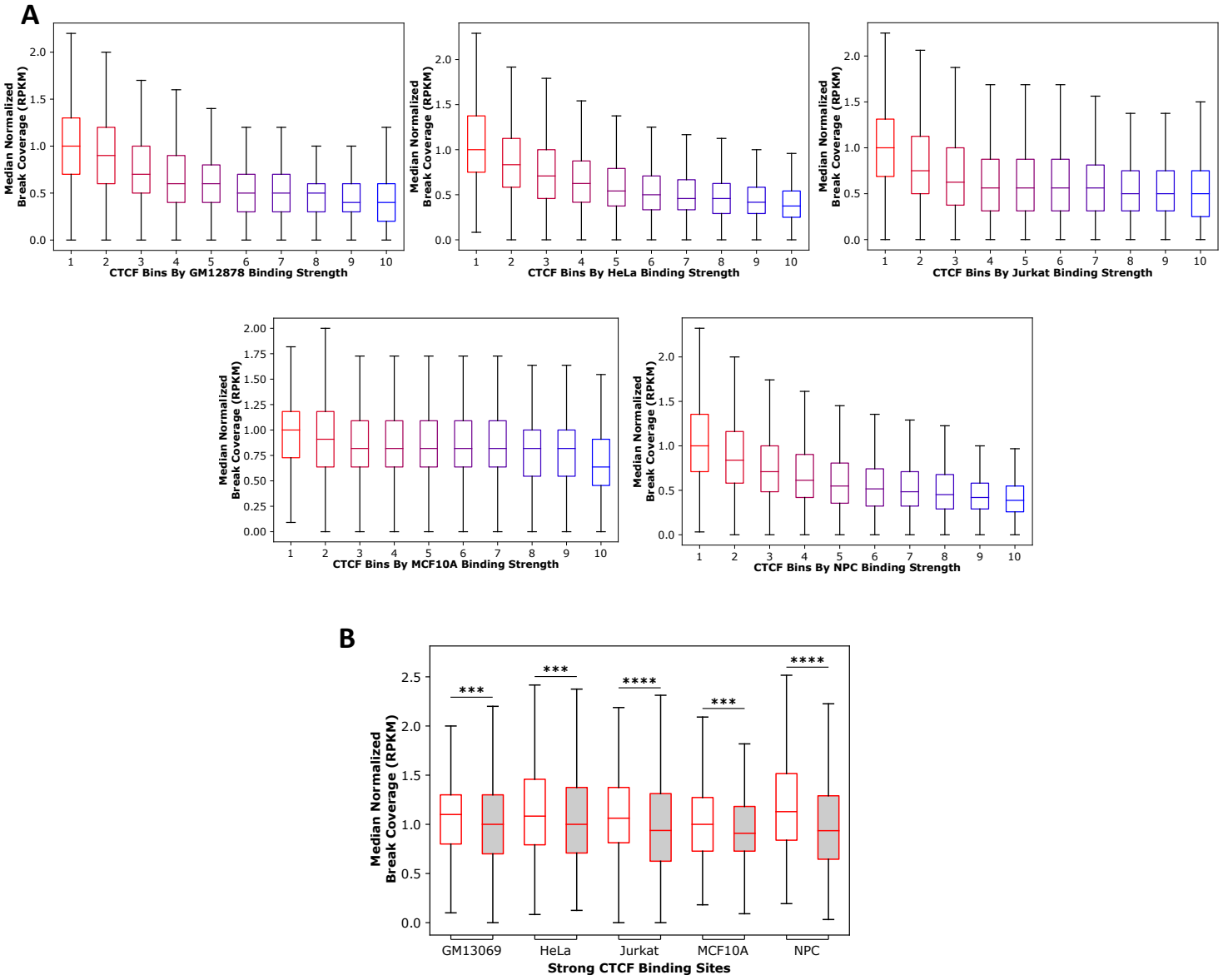


Figure S3. Strong CTCF binding sites are enriched for DSBs and shared between cell lines. (A) Mapped DSBs in untreated cells, normalized to median value of the top 10% CTCF (Bin 1) for each cell line, are enriched in strong CTCF binding sites and medians decrease with CTCF binding site strength assessed across the union set of CTCF sites from the five cell lines (RPKM, Reads Per Kilobase per Million) (n = 6884 per bin). We created the union set of sites between five cell lines resulting in 68,841 CTCF binding sites, and then these sites were divided into deciles based on the binding strength determined by the DiffBind 3.0 R package in each cell line. **(B)** Common strong CTCF binding sites (white, n = 2100) present in strongest 10% binding in five cell lines as assessed from the union CTCF binding site set, show greater DSB enrichment than the non-common strong CTCF binding sites (grey, n = 4784) for each cell line (RPKM, Reads Per Kilobase per Million). Boxes denote 25th and 75th-percentiles, middle lines show medians, and whiskers span from 5% to 95%; *** indicates $p < 0.001$, **** indicates $p \sim 0$, Kruskal-Wallis with Dunn post-hoc test.

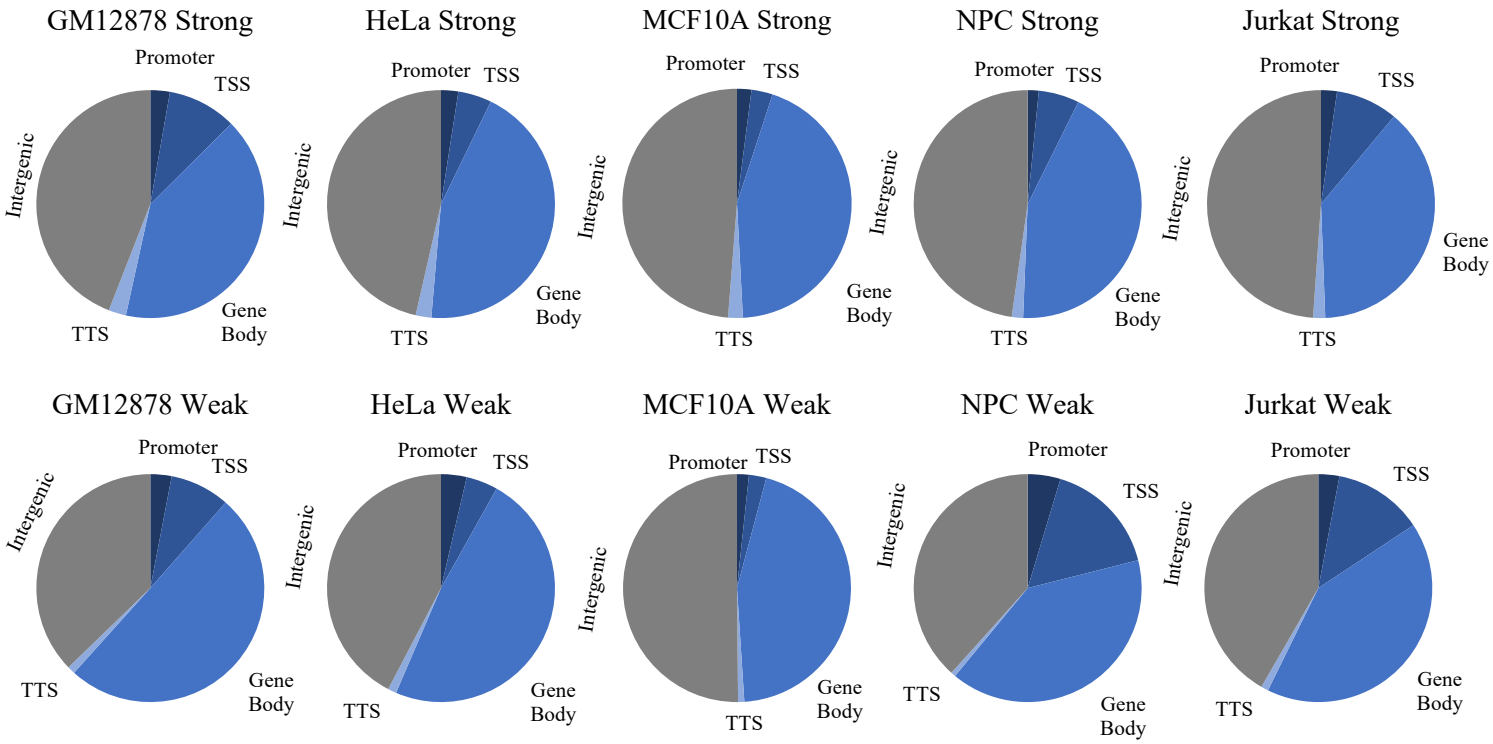


Figure S4. Distribution of strong (top 10%) and weak (bottom 10%) CTCF binding sites among genic and intergenic sites do not explain DSB enrichment. Strong (top) and weak (bottom) CTCF binding sites for GM12878 (n = 4019), HeLa (n = 6911), MCF10A (n = 6011), NPC (n = 9833), and Jurkat (n = 6820) were annotated for genomic features [promoter, transcription start site (TSS), gene body, transcription termination site (TTS), and intergenic regions]. The definitions used for each genomic feature was described in the Materials and Methods section under "Genomic region definitions".

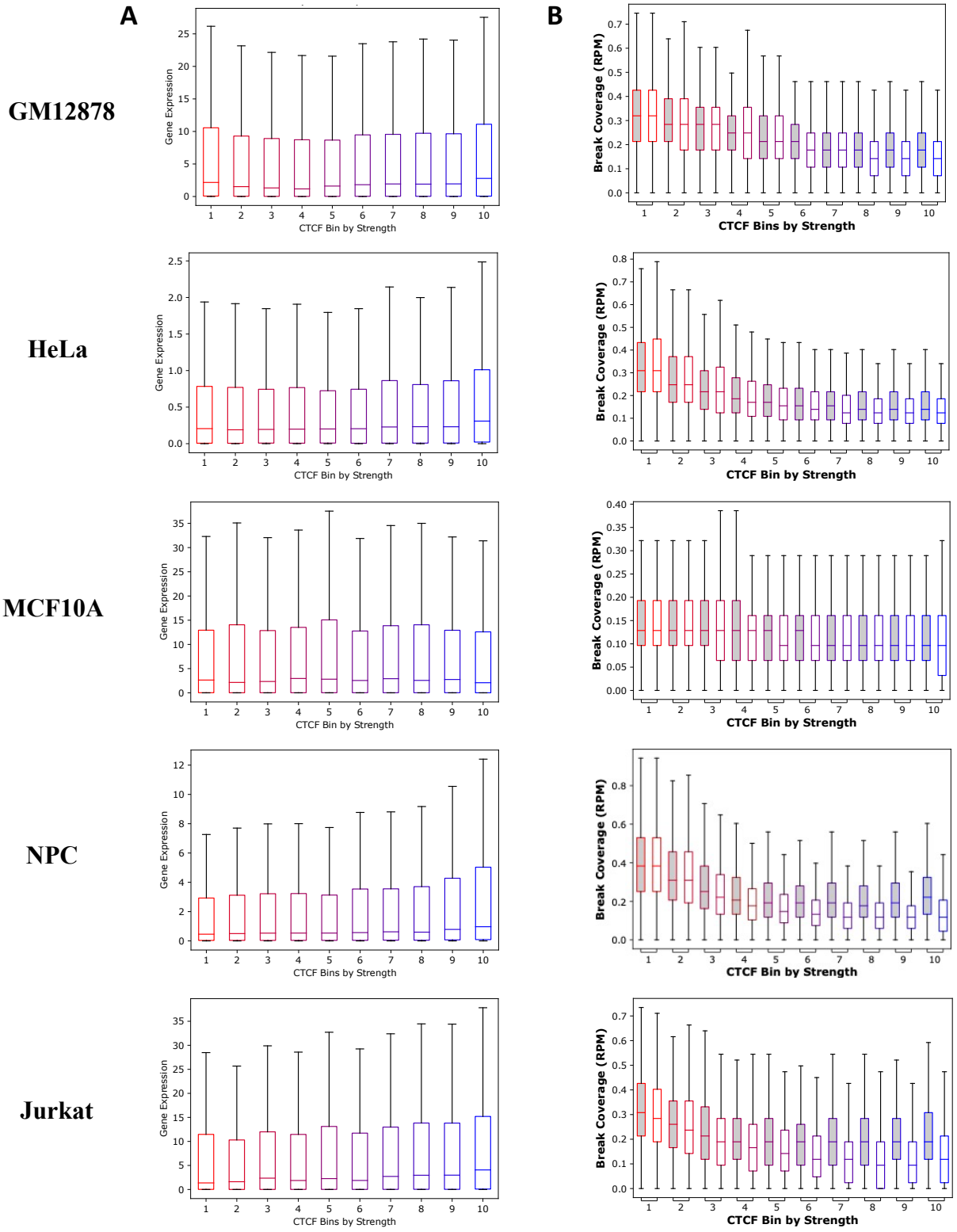


Figure S5. No differences in gene expression were observed among genic CTCF binding sites binned by CTCF binding strength, and gene expression does not cause increased DSBs in strong CTCF binding sites. (A) Gene expression medians between bins are not higher in strong CTCF binding sites compared to weak sites in all five cell liens. **(B)** Strong CTCF binding sites are not differently enriched for DSBs between genic (grey) and intergenic (white) distributions. Weak CTCF binding sites do show enriched DSBs in genic sites compared to intergenic sites in five cell lines (RPM, Reads Per Million). Boxes denote 25th and 75th-percentiles, middle lines show medians, and whiskers span from 5% to 95%.

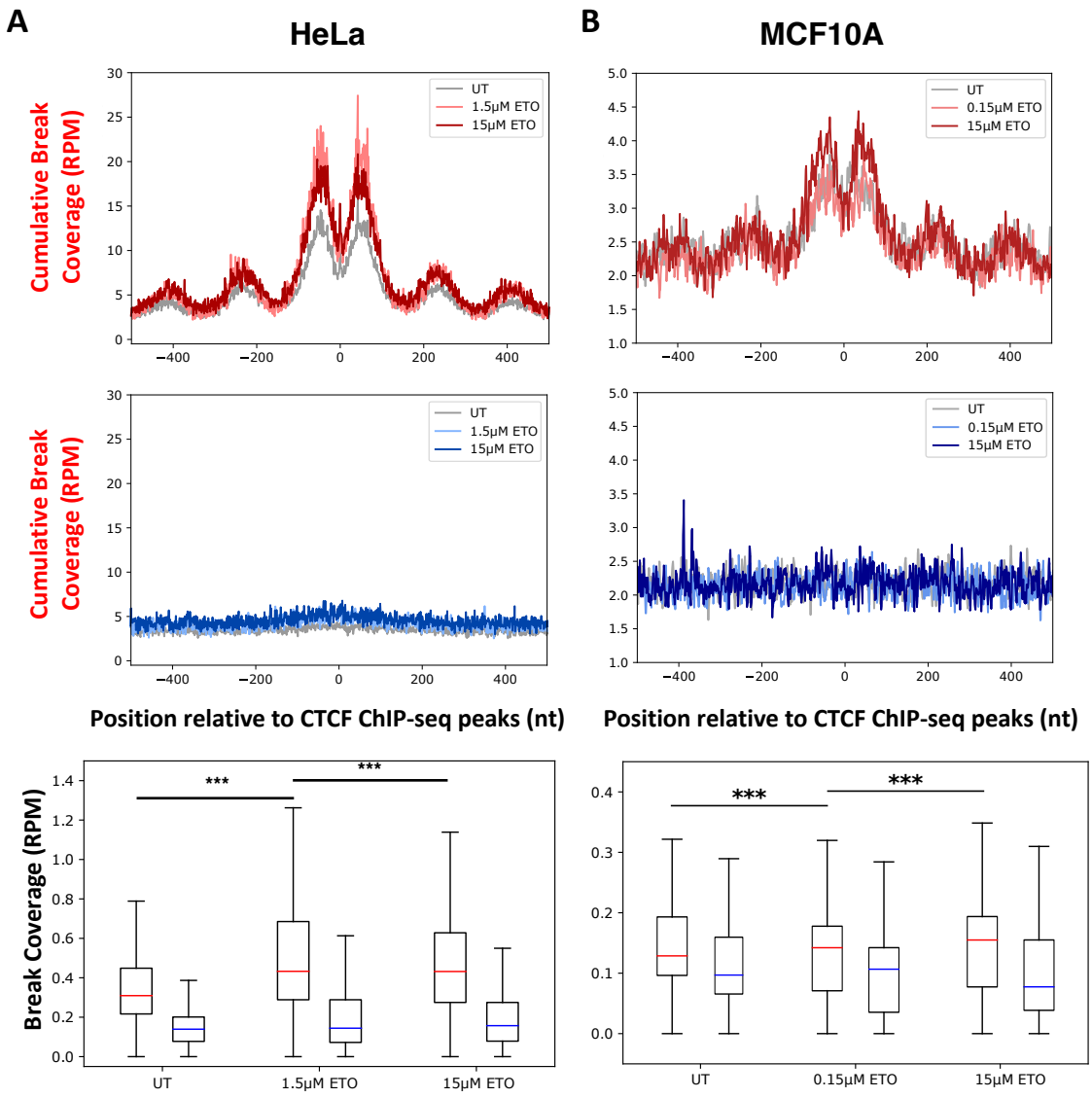


Figure S6. DSBs are enriched at strong, but not weak, CTCF binding sites in a dose-dependent manner after exposure to etoposide. Read-normalized DSBs (RPM) were mapped in untreated and etoposide-treated (0.15 μM, 1.5 μM, or 15 μM) **A)** HeLa and **B)** MCF10A cells at the strong (top 10%, top panels) and the weak (bottom 10%, middle panels) CTCF binding sites (n = 6911 and 6011 binding sites for HeLa and MCF10A, respectively). Quantification of break coverage (RPM; bottom panels) at the strong (red) and the weak (blue) CTCF binding sites (± 150 nt) for both MCF10 and HeLa demonstrated that etoposide treatment resulted in a significant increase of DSBs in a dose-dependent manner at strong sites. Boxes denote 25th and 75th-percentiles, middle lines show medians, and whiskers span from 5% to 95%; *** indicates $p < 0.001$; two-sample, Kolmogorov-Smirnov test.

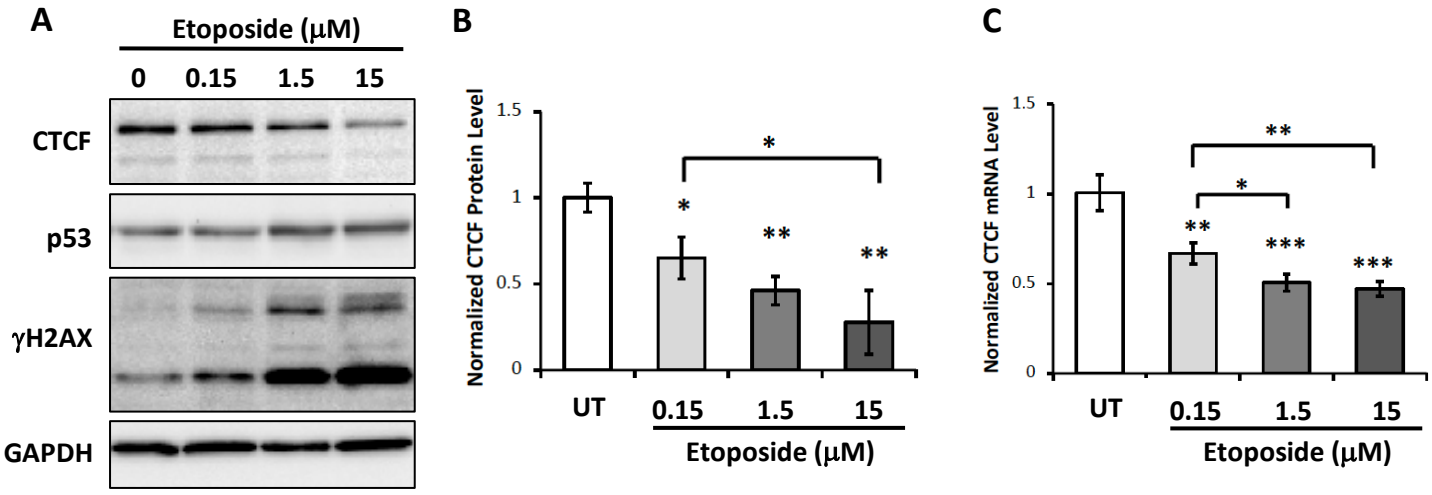


Figure S7. Etoposide treatment decreases CTCF expression in MCF10A cells while increasing damage markers. (A) Representative western blot showing CTCF (top), p53 (top-middle), γH2AX (bottom-middle), and GAPDH (bottom) in MCF10A cells treated with etoposide at indicated doses for 24 hours. **(B)** Quantification of CTCF protein level normalized to GAPDH loading control and untreated samples (n = 3). **(C)** Quantification of CTCF mRNA level from RT-qPCR normalized by the ΔΔCq method (n = 3). Bar plots show means, and error bars indicate ± standard deviation; All treatments compared to UT and each other, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, Student's t-test.

A

Region	Location	Primer F (5' → 3')	Primer R (5' → 3')
FHAD1	chr1:15329390-15329459	F: GGTCGACCTTCTTCAGCAC	R: GACGTCCAGCACCACCTT
DNAH11	chr7:21551244-21551317	F: TGGGATCAGGCTTGTCTTCT	R: GGAAACAAGCTTGCAGATGG
SOS1 #1	chr2:39120343-39120412	F: AGCAGCTGCCCTACGAGTT	R: AGCGCAGGCACCAGTAGT
SOS1 #2	chr2:39124616-39124691	F: ACGCCAGTGTGAGTTCTTGA	R: GCAGCCACAGTGATCCTTCT
ANKRD22	chr10:88835697-88835768	F: TCAGCGTTAGTGCGACTCTC	R: GCTGTTCATTGCTGATCGTG
LIPA	chr10:89217128-89217207	F: TTTGACAAAGAATGTCTGAGCA	R: GCCAAATGAATTTGAAATGGT

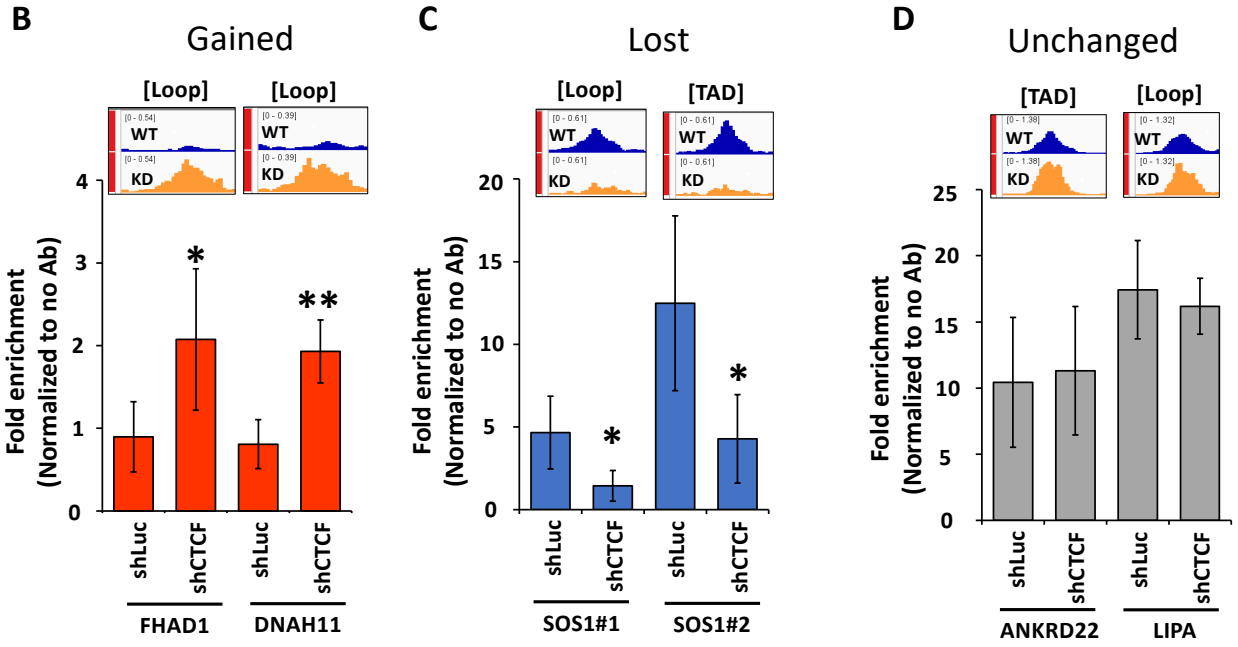


Figure S8. Validation of lost, gained, and unchanged CTCF binding sites in shCTCF and shLuc MCF10A cells by CTCF ChIP-qPCR. (A) The genomic coordinates of qPCR primers of the six CTCF binding sites. The primer-amplified regions contain CTCF motifs (FHAD1, DNAH11, LIPA) or are within 70 bp of CTCF motifs (SOS1#1, SOS1#2, ANKRD22). Fold enrichment of CTCF for gained **(B)**, lost **(C)**, and unchanged **(D)** CTCF binding sites. Top panels illustrate the CTCF-ChIP-seq data for individual locus [CTCF-WT (WT, blue) and CTCF-KD (KD, orange) in MCF10A cells], marked with the location at either TAD boundaries or loops. Bottom panels are CTCF fold enrichment normalized to no antibody (no Ab) control. At least three biological repeats were performed for each locus. Bar plots show means, and error bars indicate ± standard deviation; All shCTCF samples were compared to respective shLuc samples, * indicates $p < 0.05$, ** indicates $p < 0.01$, Student's t-test.

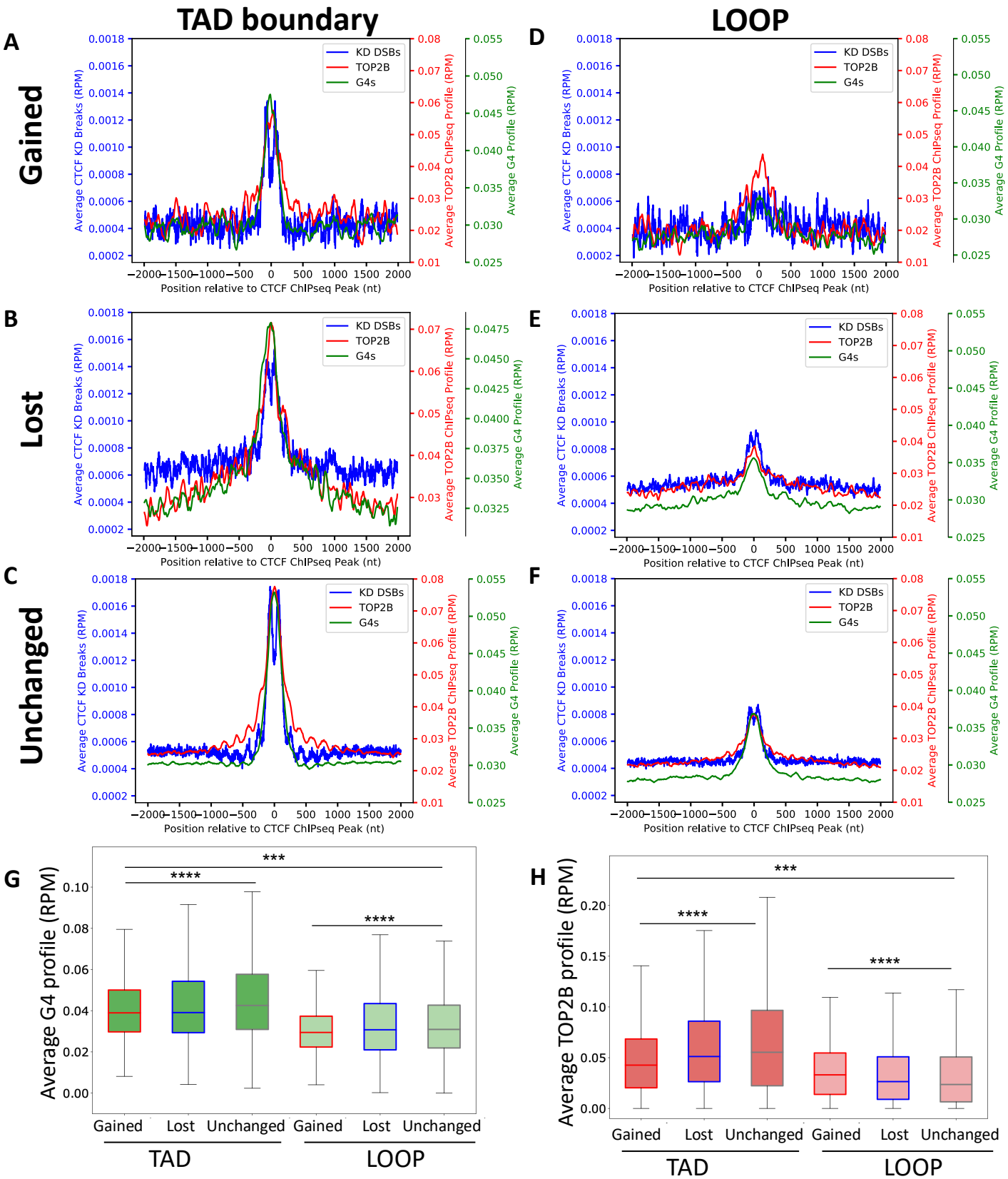


Figure S9. TAD boundary-associated CTCF binding sites are enriched for DSBs, G-quadruplexes, and TOP2 binding, compared to loop-associated CTCF binding sites. Average DSB profile from CTCF knockdown MCF10A (blue, RPM), average TOP2B binding (red, RPM), and average G-quadruplex profile (green, G4) are plotted over TAD boundary-associated **(A)** gained, **(B)** lost, and **(C)** unchanged CTCF binding sites and loop-associated **(D)** gained, **(E)** lost, and **(F)** unchanged CTCF binding sites. Quantification of G4 coverage (**G**) and TOP2B coverage (**H**) at TAD boundary- and loop-associated CTCF binding sites. Boxes denote 25th and 75th-percentiles, middle lines show medians, and whiskers span from 5% to 95%; *** indicates $p < 0.001$, **** indicates $p \sim 0$, Kruskal-Wallis followed by Dunn post-hoc test.

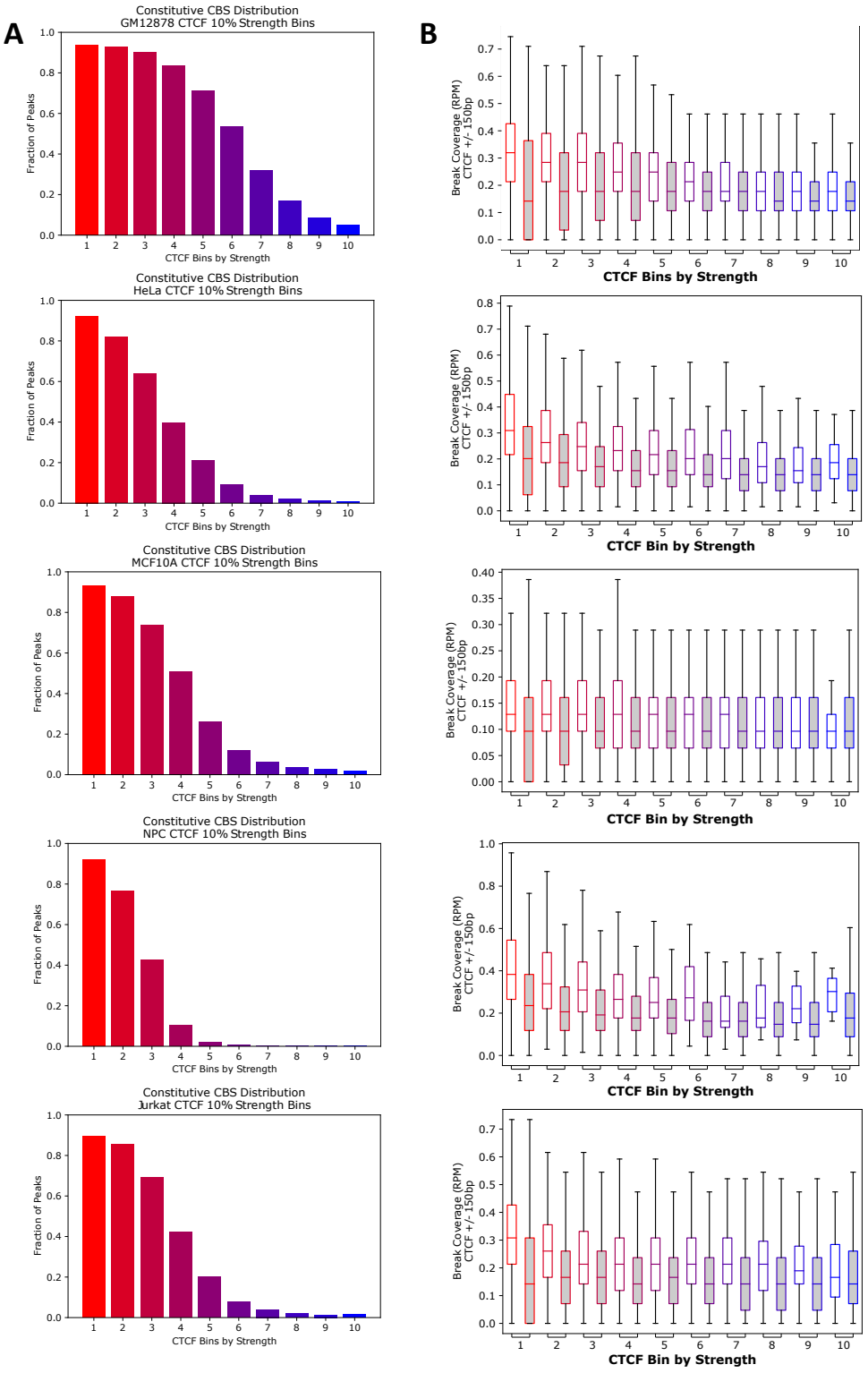


Figure S10. TAD boundary-associated CTCF binding sites are enriched among strong CTCF binding site and are enriched for DSBs compared to loop-associated CTCF binding sites. (A) Distribution of TAD boundary-associated CTCF binding sites (n = 22,097) in CTCF binding site bins based on CTCF binding strength in each cell line. **(B)** DSBs are enriched in TAD boundary-associated CTCF binding sites (white) compared to loop-associated CTCF binding sites (grey) across all decile strengths in five cell lines, while still showing binding strength differences (RPM, ± 150 bp). Boxes denote 25th and 75th-percentiles, middle lines show medians, and whiskers span from 5% to 95%.