Nucleic Acids Research, 2024, **52**, 8086–8099
https://doi.org/10.1093/nar/gkae530
Advance access publication date: 2 July 2024
Computational Biology

Machine learning enables pan-cancer identification of mutational hotspots at persistent CTCF binding sites

Wenhan Chen[1], Yi C. Zeng[2,3], Joanna Achinger-Kawecka[1,3], Elyssa Campbell[1], Alicia K. Jones[1], Alastair G. Stewart[2,3], Amanda Khoury[1,3,*] and Susan J. Clark[1,3,*]

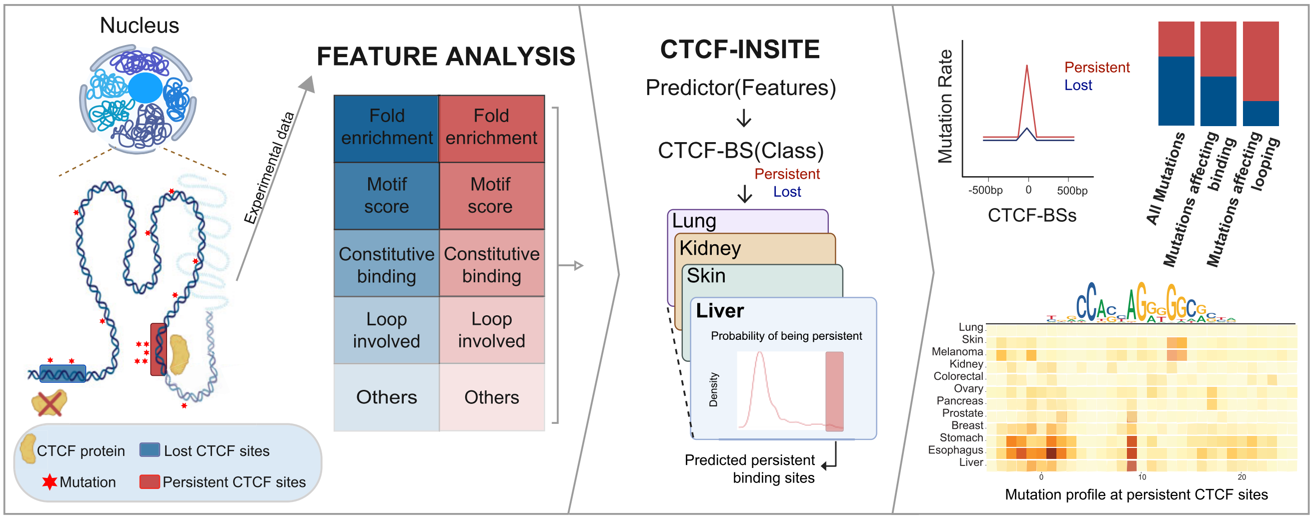
[1] Epigenetics Laboratory, Garvan Institute of Medical Research, Sydney 2010 New South Wales, Australia
[2] Structural Biology Laboratory, Victor Chang Cardiac Research Institute, Sydney 2010 New South Wales, Australia
[3] St Vincent's Clinical School, UNSW, Sydney 2010 New South Wales, Australia

[*] To whom correspondence should be addressed. Email: a.khoury@garvan.org.au
Correspondence may also be addressed to Susan J. Clark. Email: s.clark@garvan.org.au
Present address: Wenhan Chen, Computational Biology Group, Children's Cancer Institute, Sydney 2031 New South Wales, Australia.

Abstract

CCCTC-binding factor (CTCF) is an insulator protein that binds to a highly conserved DNA motif and facilitates regulation of three-dimensional (3D) nuclear architecture and transcription. CTCF binding sites (CTCF-BSs) reside in non-coding DNA and are frequently mutated in cancer. Our previous study identified a small subclass of CTCF-BSs that are resistant to CTCF knock down, termed persistent CTCF binding sites (P-CTCF-BSs). P-CTCF-BSs show high binding conservation and potentially regulate cell-type constitutive 3D chromatin architecture. Here, using ICGC sequencing data we made the striking observation that P-CTCF-BSs display a highly elevated mutation rate in breast and prostate cancer when compared to all CTCF-BSs. To address whether P-CTCF-BS mutations are also enriched in other cell-types, we developed CTCF-INSITE—a tool utilising machine learning to predict persistence based on genetic and epigenetic features of experimentally-determined P-CTCF-BSs. Notably, predicted P-CTCF-BSs also show a significantly elevated mutational burden in all 12 cancer-types tested. Enrichment was even stronger for P-CTCF-BS mutations with predicted functional impact to CTCF binding and chromatin looping. Using *in vitro* binding assays we validated that P-CTCF-BS cancer mutations, predicted to be disruptive, indeed reduced CTCF binding. Together this study reveals a new subclass of cancer specific CTCF-BS DNA mutations and provides insights into their importance in genome organization in a pan-cancer setting.

Graphical abstract



Introduction

The CCCTC-binding factor (CTCF) is an 11-zinc finger DNA-binding protein ubiquitously expressed in eukaryotes. CTCF recognizes a consensus DNA sequence motif (1,2) and binds to tens of thousands of open chromatin sites across the human genome in a cell-type specific manner (3). There is

Received: September 29, 2023. **Revised:** May 15, 2024. **Editorial Decision:** May 19, 2024. **Accepted:** June 13, 2024

substantial heterogeneity amongst the vast number of CTCF binding sites (CTCF-BSs), including variable binding affinity, cell-type specificity, conservation between mammalian species, presence of an upstream binding motif and involvement in three-dimensional (3D) chromatin structures (1,3–5). By interacting with the cohesin complex (6,7), CTCF plays a crucial role in regulating 3D genomic organization through which it can influence gene expression via various mechanisms, including regulating enhancer-promoter looping (5,8–11), demarcation of active and repressive domains (12–14), and mediating the boundary formation between consecutive topologically associating domains (TADs) (12,15). Loss of CTCF binding has been shown to result in structural re-organization of the 3D genome (16–20) and, in some cases, changes to gene expression (16–18,20).

Previously, we and others have shown potent CTCF knockdown results in wide-spread loss of CTCF binding at sites, here termed lost CTCF binding sites (L-CTCF-BSs); in contrast a small subset of CTCF binding sites remain persistently bound, here termed persistent CTCF binding sites (P-CTCF-BSs) (18,21,22). Interestingly P-CTCF-BSs displayed distinct genetic and epigenetic properties compared to L-CTCF-BSs, including stronger binding intensity, cell-type constitutive binding and high enrichment at chromatin loop anchors and TAD boundaries (18). Moreover, we found that CRISPR mediated loss of P-CTCF-BSs at the Kallikrein locus boundary in prostate cancer cells resulted in reduction of 3D chromatin looping and coordinate upregulation of genes (18).

CTCF-BSs are among the non-coding DNA sequences that are frequently mutated in cancer (23,24). CTCF-BS mutations can also reduce the ability of CTCF to bind (25) which can destabilize the 3D genome, including loss of looping and loss of insulation of domain boundaries which can result in oncogenic gene activation (26,27). CTCF-BS mutations have been linked to development of gastrointestinal cancers and melanoma (28,29), however, the vast majority of CTCF-BS mutations are unexplored.

In this study, we investigated whether the P-CTCF-BS subclass of CTCF-BSs are more frequently mutated and whether this is a pan-cancer phenomenon since their binding is highly conserved across varied normal and cancer cell-types. We employed machine learning approaches to predict CTCF binding persistence to knockdown. Using experimentally defined P-CTCF-BSs, we evaluated genetic and epigenetic features that explain CTCF binding persistence. We evaluated 15 features in determining binding persistence and created a computational tool, called CTCF-INSITE (IN-Silico Investigation of persisTEnt binding), which predicts P-CTCF-BSs *in silico*. Using CTCF-INSITE, we generated persistence metrics for ENCODE CTCF ChIP-seq data from various normal tissue types, which enabled analysis of the mutational burden at P-CTCF-BSs from International Cancer Genome Consortium (ICGC) sequencing data from matched tumour types. We verified that candidate P-CTCF-BS cancer mutations predicted to be disruptive also reduced CTCF binding in *in vitro* assays. We discovered that mutations at P-CTCF-BSs are highly enriched across 12 different cancer types relative to L-CTCF-BSs. We further found that P-CTCF-BS mutations are significantly associated with loop disruption suggesting these mutations may play a role in dysregulation of the 3D genome in cancer.

Materials and methods

Datasets

Experimentally-determined P-CTCF-BSs

CTCF ChIP-seq data from 3 cell lines, LNCaP, IMR-90 and MCF7, were used to define P-CTCF-BSs. LNCaP and IMR-90 data were previously generated in-house (18). MCF7 data was obtained from an independent study (7). All cell lines had been subjected to two transfection conditions: 1. CTCF RNAi 2. non-targeted RNAi, as control. Transfections were conducted for a period of 144 hours for LNCaP and IMR-90, and 48 hours for MCF7. CTCF knockdown was confirmed by western blot (7,18). After transfection, samples were processed through immunoprecipitation and sequenced using the Illumina Genome Analyzer II (40bp; paired-end) or HiSeq 2500 (50–75bp, single-end). P-CTCF-BSs were identified by comparing the CTCF ChIP-seq peaks remaining in the knockdown versus those lost from control samples from the same cell line, respectively (see Khoury *et al.* for method details (18)). Different knockdown efficiencies were achieved: 11% of 25602 CTCF-BSs were identified as P-CTCF-BSs for LNCaP, 23% of 20179 for IMR-90 and 33% of 37472 for MCF7, possibly due to differences in knockdown stringencies.

Public ChIP-seq data

The 90 public tissues and cell line CTCF ChIP-seq data were downloaded from ENCODE (30) and NCBI. Data sources are summarized in Supplementary Table S1. We used the unperturbed CTCF ChIP-seq data as input for prediction of persistent binding.

Public ChIA-PET data

CTCF ChIA-PET data for GM12878 (31) as well as RAD21 ChIA-PET data for LNCaP and MCF7 were downloaded from ENCODE with accession codes provided in Supplementary Table S1. The latter two cell lines contain a much smaller number of inter Paired-End Tags (PETs) compared to GM12878 ChIA-PET.

Public WGS data

The high-coverage WGS of GM12878 is available from platinum genome project at https://hgdownload.soe.ucsc.edu/gbdb/hg19/platinumGenomes/. The high-coverage WGS for LNCaP and MCF7 (32) were generated in house. Variants were identified using GATK HaplotypeCaller with the following parameter, in_base_quality_score of 10.

Cancer mutations

Summary data of simple mutations (i.e. SNV and indels) identified from WGS were downloaded from ICGC by setting 'Donor Analysis Type' to WGS from the ICGC data portal at https://dcc.icgc.org/search. Quality controls were performed at the cohort and individual levels. More specifically, cohorts with a substantially smaller number of mutations per individual were filtered. These are likely to be Whole Exome Sequencing data, but mis-labelled, as the mutations map to exome annotations. Individuals with a very high number of mutations were also removed because they are likely to have distinct cancer aetiologies but the sample size is too small to address this. The Interquartile Range (IQR; IQR = 75th–25th quartile) approach was used to identify and filter out outliers, which are <25th quantile—1.5IQR or >75th quantile + 1.5IQR. This two-step QC resulted in 24 cohorts and 3218 patients to use in our study. Mutations from cohorts of the same cancer type were then merged together into 12 solid cancer types. The sample size per cancer type varied quite substantially from 30 for colon to 686 for breast. The summary of datasets can be found in Supplementary Table S2, including the assignment of cohorts to cancer groups and sample size after QC.

Pairing ICGC cancer types to ENCODE ChIP-seq tissue types

Healthy tissue CTCF ChIP-seq data were used from ENCODE for all cancers, except for melanoma and kidney. CTCF ChIP-seq data from melanoma cell line, COLO829, and differentiated nephron progenitor cells were used for melanoma and kidney analyses, respectively (Supplementary Table S2).

Prediction of P-CTCF-BSs

Features

We investigated 15 distinct features (Figure 1) for the cell lines LNCaP, IMR-90, and MCF7. Six were genomic features, measuring the relative positions to different genic domains, including promoter, 5′ UTR, 3′ UTR, exon or intron, and proximity to transcriptional start site (TSS). The genomic domain annotation was downloaded from UCSC (33). Two features related to chromatin interactions were considered: (i) proximity to TAD boundaries (18) and (ii) frequency of overlap with chromatin loop anchors. Two features related to binding affinity were included, namely (a) fold enrichment from MACS2 (34) and (b) motif score from DeepBind (35). Additional features were replication timing quantified using Repli-Seq from a previous study (36), conservation score measured as Genomic Evolutionary Rate Profiling (GERP) (37) from UCSC (33), number of CpG sites found at the CTCF core motif, and constitutive binding defined as the frequency of a CTCF peak found in a reference panel of 40 tissue and cell line ChIP-seq data in (3). Replication timing data was standardized to a value between 0 and 100 at a 1kb resolution. The replication timing score for a CTCF peak was calculated as the mean value of the overlapping intervals. Similarly, the conservation score for a peak was calculated as the mean value of the GERP scores.

Prediction model

We used the above features to create two types of machine learning models, namely a logistic regression model (R package stat v3.6.3) and a random forest model (R package randomForest v4.6) (38). The experimentally-determined L-/P-CTCF-BSs from LNCaP, were split into training and testing datasets in a 9:1 ratio. The testing dataset was then used for an in-sample validation of the performance measured in Area Under the Curve (AUC) and the Area Under the Precision-Recall Curve (AUC-PR). Splitting was iterated 100 times to generate standard errors for the AUC and AUC-PR metrics. Five-fold cross-validation was used to fine-tune the ‘mtry’ parameter for the RF model, which represents the number of predictors sampled for splitting at each node (38). This analysis identified the optimal value as 3. In addition to the in-sample test (i.e. splitting a dataset into two for training and testing), models were tested on out-of-sample by assessing the models on a separate dataset, P-/L- CTCF-BSs from IMR-90. Note that chromatin looping and replication timing are unavailable for IMR-90 (and are generally unavailable for other public data sets); therefore, we used 13 features instead of 15 to create the models for this out-of-sample test. We also created the CTCF-INSITE tool with these 13 features.

Prediction

Both models output the probability for a CTCF-BS to maintain persistent binding, described as ‘persistence’ *in silico*. We applied the RF model to predict the persistence of the binding sites from ENCODE CTCF ChIP-seq data. A CTCF-BS was defined as persistent if that site demonstrated greater persistence than a pre-defined threshold. Herein, varying thresholds were used to derive P-CTCF-BSs at different stringencies.

Calling ChIP-seq peaks

The raw sequencing data of the immunoprecipitated sample and the matched control sample (the input sample without immunoprecipitation) were processed through standard pipelines: adaptor trimming using Trim Galore v0.6.6; alignment to hg19 reference using Bowtie v1.3.0 (39) and peak calling using MACS2 v2.2.7 (34) with paired immunoprecipitated and control samples, a fixed shift size of 200 bp and a *q*-value cutoff of 0.05 (i.e. macs2 –nomodel –extsize 200 –qvalue 0.05). Peaks identified from MACS2 (34) were annotated with genomic location, fold enrichment and adjusted *p*-value. Any peaks overlapping the ENCODE Blacklist from UCSC (40) were excluded due to poor mappablity. The size of the peaks ranged from 200 bp to 2000 bp.

Allele-specific binding analysis based on ChIP-seq data

Heterozygous loci (either germ-line or somatic) were identified from WGS data for LNCaP and GM12878. Allelic read depths from ChIP-seq data were then used to quantify the degree of allele-specific binding, as described in Tang *et al.* (41). The data processing pipeline for ChIP-seq data is the same as described above, except for mapping to a masked hg19 reference genome. Specifically, common SNPs from dbSNP v151 (42) were masked as ‘N’ to avoid mapping biases towards the alternative allele. GATK CollectAllelicCounts (v4.2.3) (43) was used to extract allelic read counts while filtering those with read depth <5. Low-mappability and imprinted control regions were excluded to avoid confounding effects. Indels were excluded as required by GATK CollectAllelicCounts. A binomial test was performed to assess the significance of allelic imbalance, i.e. read depth of A1 $\sim$ Binomial (total read depth, 0.5).

Calling chromatin loops based on ChIA-PET data

ChIA-PET data was processed through the CHIA-PIPE pipeline (44) which automated analyses including the adaptor trimming, linker identification, alignment using BWA mem and aln (45), loop calling by clustering PETs, calling for TF binding sites using (SPP or MACS2 v2.2.7 (34)), loop calling refinement based on the binding sites and identification of allele-specific loops. Briefly, a loop was defined as a locus with ≥ 3 supporting intra-ligated PETs (with an insertion size > 8 kb) which were clustered within <1000 bp from each other. Loop calling was then refined by confirming that one or both loop anchors contained a CTCF-BS. Loop anchors were identified at a <1 kb resolution (Supplementary Figure S1).

To enable analyses of allelic bias in looping, we made two modifications to CHIA-PIPE, (i) aligning reads to a masked hg19 reference genome for common SNPs and (ii) removing self-ligated PETs with insertions <8 kb when counting reads mapped to either allele at a heterozygous site. The latter modification is introduced because the self-ligated PETs can confound allele-specific binding with allele-specific looping. Similar to the allelic binding analysis, we filtered low-mappability and imprinted control regions, used GATK CollectAllelicCounts (v4.2.3) (43) to extract read counts and performed significance testing using a binomial test, i.e. read depth of A1 ∼ Binomial (total read depth, 0.5). An imbalance in allelic read depths demonstrates differential looping between alleles. We examined loops mediated by CTCF through analyses of allelic biases using ChIP-seq and CHIA-PET data for GM12878. The LNCaP and MCF7 ChIP-PET data were not deep enough for such allele-specific analysis.

Binding motif analyses

Positioning of the CTCF core motif in a ChIP-seq peak

TFBSTools (R package v1.24.0) (46) and the 19 bp CTCF core motif from JASPAR (ID: MA0139.1) (47) were used to identify the binding location within in a ChIP-seq peak ranging from 200 to 2000 bp.

DeepBind

DeepBind (v0.11) (35) software scores binding affinity for a given DNA/RNA sequence. A motif score was calculated for each ChIP-seq peak region. Notably, DeepBind uses multiple pre-calculated motifs, rather than one motif and provides a combined score for the binding affinity *in silico*. It does not provide the position of a particular motif found in a sequence. Scores from DeepBind can be negative values, which reflects how unlikely the binding is. DeepBind has a limitation of sequence length <1000 bp, so peaks >1000 bp were trimmed at both ends to fit within this range.

Binding score

The sequences for peak regions (denoted by its chromosome, start and end positions) were obtained from the hg19 reference genome. DeepBind was then used to determine the score for the sequence which is referred to as the reference score. When there was a mutation at the peak, the reference sequence was altered in accordance with the mutation, i.e. changing the original base to a new base. A mutation score was then calculated based on the mutated sequence. If multiple mutations were found within a single peak, the mutation score that led to the biggest change was retained i.e. only the impact of a single variant was considered, and potential compounding effects were disregarded.

Fluorescence polarization DNA binding (FPDB) assay

CTCF plasmids

Plasmid 6xHis-SUMO ZF1-11 CTCF kindly provided by Dr Peter Jones, was cloned from the following plasmids: CTCF ZF1–11 (pXC1441), a gift from Drs Xiaodong Cheng and John Horton (MD Anderson) and pDONR223 CTCF WT a gift from Drs Jesse Boehm, William Hahn, and David Root (Addgene plasmid # 81789). The CTCF ZF1-11 fragment was cloned into pSUMO vectors for expressing 6xHis-SUMO tagged CTCF proteins, as described in Thomas *et al.* (48).

Protein expression

Tagged human CTCF ZF1-11 protein was expressed in the *Escherichia coli* strain BL21-CodonPlus (DE3)-RIPL (230280, Agilent) and colonies were grown on agar plates containing ampicillin (100 μg/ml) and chloramphenicol (25 μg/ml) overnight. Colonies were inoculated into 8 L of LB medium containing ampicillin (100 μg /ml), chloramphenicol (25 μg/ml) and 25 μM ZnCl2 and cultured for 4–5 h at 28°C to OD 600 of ∼0.8 and cooled to 16°C. Protein expression was induced with 0.1 mM isopropyl-D-1-thiogalactopyranoside (IPTG) overnight at 16°C. Cells were harvested by centrifugation at 4000 × g for 20 min at 4°C.

Protein purification

Cell pellets were resuspended in 150 ml of lysis buffer (20 mM Tris–HCl (pH 8.0), 25 mM imidazole, 1 M NaCl, 5% (v/v) glycerol, 0.5 mM Tris (2-carboxyethyl) phosphine hydrochloride (TCEP), 25 μM $ZnCl_2$) with 1 mM phenylmethanesulfonylfluoride (PMSF). Resuspended cells were lysed in a cell disrupter (Constant Systems) and clarified by centrifugation at 12 100 × g at 4°C for 40 min. The clarified supernatant was DNase treated before being loaded onto 6 ml of Ni-NTA Agarose (QIAGEN), pre-equilibrated with lysis buffer and washed with 5× column volumes of washing buffer (20 mM Tris–HCl (pH 8.0), 50 mM imidazole, 1 M NaCl, 5% (v/v) glycerol, 0.5 mM TCEP, 25 μM $ZnCl_2$). Bound proteins were eluted with elution buffer (20 mM Tris–HCl (pH 8.0), 250 mM imidazole, 1 M NaCl, 5% (v/v) glycerol, 0.5 mM TCEP, 25 μM $ZnCl_2$). Pooled protein was subsequently concentrated and loaded onto a Superdex 200 16/600 column (GE28-9893-35, Cytiva) pre-equilibrated with size-exclusion buffer (20 mM Tris–HCl (pH 8.0), 1 M NaCl, 5% (v/v) glycerol, 0.5 mM TCEP, 25 μM $ZnCl_2$). The protein eluted at a volume of 78 ml. Peak fractions were pooled, concentrated, and flash-frozen in 25 μl aliquots.

FPDB assay

Double-stranded DNA oligos were purchased from IDT (sequences are shown in Supplementary Table S3). The strand containing the CTCF motif was labeled with 6-carboxyfluorescein (FAM). Double-stranded oligos were diluted in DNA binding buffer (20 mM Tris–HCl (pH 7.5), 300 mM NaCl, 5% (v/v) glycerol, and 0.5 mM TCEP) and each oligo (5 nM) was incubated for 15 min at 25°C with a serial dilution of CTCF protein in the range of 0.0005–1 μM. Triplicate protein serial dilutions were set up for each oligo in a 384-well black assay plate (#3575, Corning). The plates were read on a PHERAstar FS (BMG LABTECH). Polarisation and anisotropy values were determined by the instrument (excitation at 485 nm and emission/polarisation at 520 nm). Delta anistropy was calculated by subtracting the anisotropy value for the lowest protein concentration from all concentrations.

Enrichment analysis

Permutation test

A permutation test was used to compare the mutational burden between P- and L-CTCF-BSs. The null hypothesis posits that there is no significant difference in mutational burden between the two. Under this assumption, a background distribution of mutation counts was generated by counting the number of mutations found at each randomly-sampled subset of all CTCF-BSs. The sampling was repeated 100 000 times with each subset selected at a size equal to the number of P-CTCF-BSs. A normal distribution was fitted to the mutation counts. Given the observed number of mutations (k) at all P-CTCF-BSs, a P-value was calculated as the proportion of subsets having a mutation count greater or equal to k (i.e. $P(K>=k)$).

Positional enrichment

CTCF peak regions were divided into 41 consecutive 40bp-tiles centered at the CTCF core motif. The number of mutations from all peaks were summed up for each tile. The significance of mutational enrichment at the core motif was calculated from a χ^2 test comparing the observed number of mutations at the core and flanking to the expected numbers derived from the null hypothesis of no enrichment at the core motif (i.e. uniform distribution). Under this assumption, the expected numbers are 1/41 and 40/41 of the total number of mutations for core and flanking respectively.

Gene set enrichment analysis

Gene set enrichment analysis was conducted using g:Profiler (49) software with the hallmark gene set panel from the Gene Set Enrichment Analysis (GSEA) database (50,51).

Creating gene lists

Genes adjacent to CTCF-BSs were defined as genes located within 1kb upstream or downstream of any CTCF-BSs of interest. Using this approach, we generated two distinct sets of neighbouring genes: those neighbouring P-CTCF-BSs and those neighbouring all CTCF-BSs.

Enrichment analysis

Gene sets containing a significant overlap with genes neighbouring P-CTCF-BSs were compared against the overlap with genes neighbouring all CTCF-BSs (background gene set) A gene set was significantly enriched if it met a p-value threshold of 0.05, adjusted for multiple testing using the g:SCS method provided by g:Profiler (49).

Simulating mutations based on the trinucleotide context

Using the SigProfilerSimulator software (52), we calculated the trinucleotide mutational context for each patient. We simulated background mutation rate for each patient and then aggregated these to create a background mutation rate at CTCF-BSs for each cancer . The simulation was repeated 1000 times to generate the sampling mean and variance.

Results

P-CTCF-BSs can be predicted from distinct genomic and epigenomic features

We first evaluated the genomic and epigenomic features of experimentally defined P-CTCF-BSs, to predict CTCF binding persistence to knockdown using machine learning approaches. We curated comprehensive molecular datasets for LNCaP, MCF7 and IMR90 cell lines, either in-house or from the public domain. These included Whole Genome Sequencing (WGS), CTCF ChIP-seq, RNAi-mediated CTCF knockdown ChIP-seq (RNAi CTCF-ChIP-seq), RAD21 or CTCF ChIA-PET and DNA replication timing data. P-CTCF-BSs are defined as the location of CTCF ChIP-seq peaks that remain largely unchanged following CTCF-knockdown, compared to non-targeted CTCF ChIP-seq experiments; whereas L-CTCF-BSs are defined as the locations where CTCF ChIP-seq peaks are lost (18). P-CTCF-BSs account for 11% out of 25 602 CTCF-BSs for LNCaP, 23% out of 20 179 sites for IMR-90 and 33% out of 37 472 sites for MCF7, potentially due to relative differences in knockdown conditions (7,18).

We next quantified the power of previously explored (18,21) and additional features in delineating P-/L-CTCF-BSs based on experimental data from LNCaP and MCF7 cell lines (Features are summarized in Materials and methods). The additional features include conservation score, co-location with chromatin loop anchors and replication timing, which have all been studied in relation to CTCF-BSs (23,53). The discriminating power was measured in the variance explained, in particular, McFadden's Pseudo r^2 from a univariant logistic regression between each feature and the classification of CTCF-BS. Whilst most features showed significant correlation with CTCF-BS classification as persistent or lost, the discriminating power varied substantially (Figure 1A). The top three features: (i) fold enrichment of reads at each ChIP-seq peak (fold enrichment); (ii) motif score and (iii) constitutive binding, emerged as the strongest predictors for both cell lines with Pseudo $r^2 > 8.5\%$ compared to $r^2 < 3.5\%$ for remaining features. Interestingly, whilst all three top features are measures of binding affinity, they were only moderately correlated with each other (Pearson's $r^2 < 0.1$; Supplementary Figure S2). This suggests that binding affinity is a key determinant of persistence, but is only partially captured by the above features. Note that the fold enrichment from ChIP-seq data is not a perfect measure of binding affinity due to technical noise such as the GC-bias (Supplementary Figure S2). We also observed that significantly more P-CTCF-BSs are located at chromatin loop anchors, TAD boundaries, regions of late replication timing and have higher conservation scores, compared to the L-CTCF-BSs. Of these features, co-location of CTCF-BSs with chromatin loops was the strongest predictor of persistence (pseudo $r^2 = \sim 3–6\%$) (Figure 1A).

We next sought to develop a computational prediction of P-CTCF-BSs. We used LNCaP data as a training set to develop logistic regression and Random Forest (RF) models to predict P-CTCF-BSs from genomic and epigenomic data. For the model development, chromatin looping and replication timing were excluded as these datasets are not readily available for most cell types. The models output a probability for a CTCF-BS to maintain persistent binding, described as 'persistence' *in silico,* which is then converted to a binary outcome, persistent or not persistent, for a chosen cut-off. We assessed performance for in-sample (in which a subset of CTCF-BSs was left out as a validation set from the LNCaP training set) and out-sample tests (in which the model trained is used to predict persistence in an independent dataset- IMR-90), measuring the Area Under Curve (AUC) and the Area Under Precision Recall Curve (AUC-PR) (Materials and methods). AUC-PR is used since L-CTCF-BSs and P-CTCF-BSs are proportionally misbalanced. The RF model combining all features achieved the best performance in the out-sample test with an AUC of ∼0.8 compared to random guess of 0.5 and AUC-PR of ∼0.6 compared to random guess of ∼0.2 (Figure 1B). The AUC-PR was only slightly smaller than that from the in-sample test, showing that the RF model is robust. Performance using only the top three features was inferior than using all features for both models, showing that the remaining features provide additional classification power. Indeed, feature

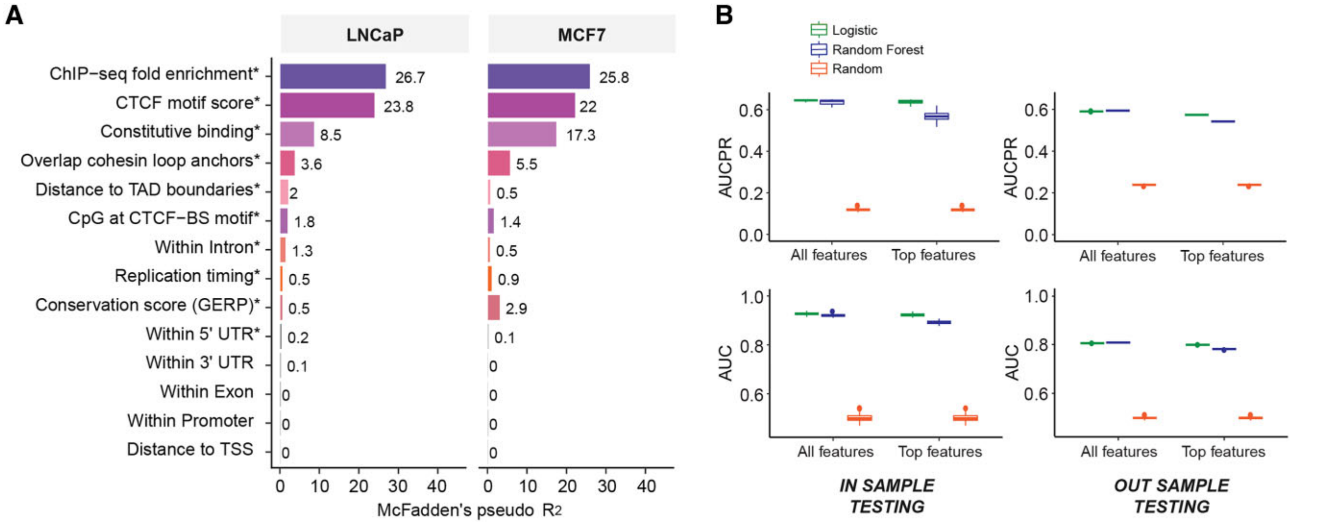


Figure 1. Development of CTCF-INSITE to predict CTCF binding persistence. (**A**) Each bar shows the individual predictive power of a feature in distinguishing P-/L-CTCF-BSs. These were quantified using McFadden's Pseudo R^2 (scaled up by 100) derived from univariate logistic regression between each feature and the bi-partition of CTCF-BSs, experimentally defined from LNCaP and MCF7 cell lines. Features showing significance (χ^2 test P-value < 0.01) in both cell lines are marked with an asterisk (*) in their labels. (**B**) Predictive models were built using logistic regression (green bars) and Random Forest (blue bars) methods for all features or top three features (as identified in A) and compared with random guess (orange bars). The bar graphs illustrate the predictive performance as measured in AUC-PR (upper panels) and AUC (lower panels) for the in-sample test (left panels) and out-of-sample test (right panels).

importance analyses from the RF model confirm that these contributions are not redundant (Supplementary Figure S3). We implemented both the logistic regression and RF models into a tool, called CTCF-INSITE (IN-Silico Investigation of persisTEnt binding), and a web server version can be accessed at: https://when.shinyapps.io/ctcf-insight/. In the later analyses we employ the CTCF-INSITE RF model as it displayed better performance for predicting persistent CTCF binding *in silico*.

P-CTCF-BSs show a highly elevated mutation rate

Since CTCF-BSs are among the non-coding DNA sequences that are frequently mutated in cancers (23), we addressed if P-CTCF-BSs display an elevated mutation rate compared to L-CTCF-BSs. We examined the relationship between persistence and mutation rates by intersecting CTCF-ChIP-seq peak regions (usually 300–400 bp in length) for LNCaP and MCF7 with all mutations from WGS data obtained from the ICGC (sample size $n = 536$ and $n = 686$ for prostate and breast cancers, respectively). This revealed that a large fraction of CTCF-ChIP-seq peak regions contain ≥ 1 mutations (17% prostate cancer and 30% breast cancer), but rarely contain $\geq$3 mutations (<2%) (Supplementary Table S4). Compared to peaks containing L-CTCF-BSs, P-CTCF-BSs exhibited a higher mutation rate of 1.34 and 1.25 times per CTCF-BS for LNCaP and MCF7 respectively (Supplementary Table S4), and the elevated rate is highly significant (P-value $< 1 \times 10^{-8}$) as determined by a permutation test (Materials and methods) (Figure 2A, B).

We performed additional tests to determine the extent that enrichment is confounded by (i) passenger mutations, that is mutations that occur randomly and do not contribute to cancer or (ii) by location-specific variation of mutation rates, for example open chromatin regions (54,55) and late-replicating regions have a significantly higher mutational rate (56). First, we repeated the permutation test for mutations occurring $\geq$2 or $\geq$3 in the ICGC cancer cohorts, as passenger mutations are less likely to reoccur. While the extent of enrichment at P-CTCF-BSs became weaker, it remained significant (Supplementary Figure S4). Second, to control for positional effects, we performed enrichment analyses comparing the mutation rate at the CTCF core motif (i.e. a 40 bp region centered on JASPAR motif MA0139.1) within the ChIP-seq peak with that of the flanking regions (Figure 2C, D; Materials and methods). The elevated mutation rate was only present at the core motif in both cell lines. Once again P-CTCF-BSs displayed a highly elevated mutation rate, ~2-fold higher at the core compared to flanking regions (P-value $< 1 \times 10^{-12}$; Supplementary Table S5); compared to L-CTCF-BSs (P-values of 0.09 for LNCaP and 0.0006 for MCF7). We confirmed that this increase is not due to background trinucleotide mutational rate using SigProfilerSimulator (Materials and methods, Supplementary Figure S5). Next, we examined the CTCF core motif using sequence logo plots with stacked bar plots for LNCaP and MCF7, which indicate the type and frequency of mutations for P-CTCF-BSs compared to L-CTCF-BSs (Figure 2E, F). P-CTCF-BSs display elevated mutations at the upstream, downstream and 9th base of the core motif, however such a signal was absent from L-CTCF-BSs. A clear mutational profile only emerges within the P-CTCF-BSs implying that these sites are selected for mutations. Together these results suggest that the mutational enrichment is not driven by location-specific effects and is only partly influenced by passenger mutations.

We next sought to evaluate the robustness of CTCF-INSITE. To do this we used CTCF-INSITE to predict P-CTCF-BSs in LNCaP and MCF7 cells. First, we confirmed the equivalence in mutational rates between the experimentally-defined and predicted P-CTCF-BSs for prostate and breast cancer (Figure 2G,H). Second, since the prediction model allows persistence to be assessed at varying stringencies, we examined the mutation rates at different stringencies. We found that mutation rate increased cubically as the stringency of per-

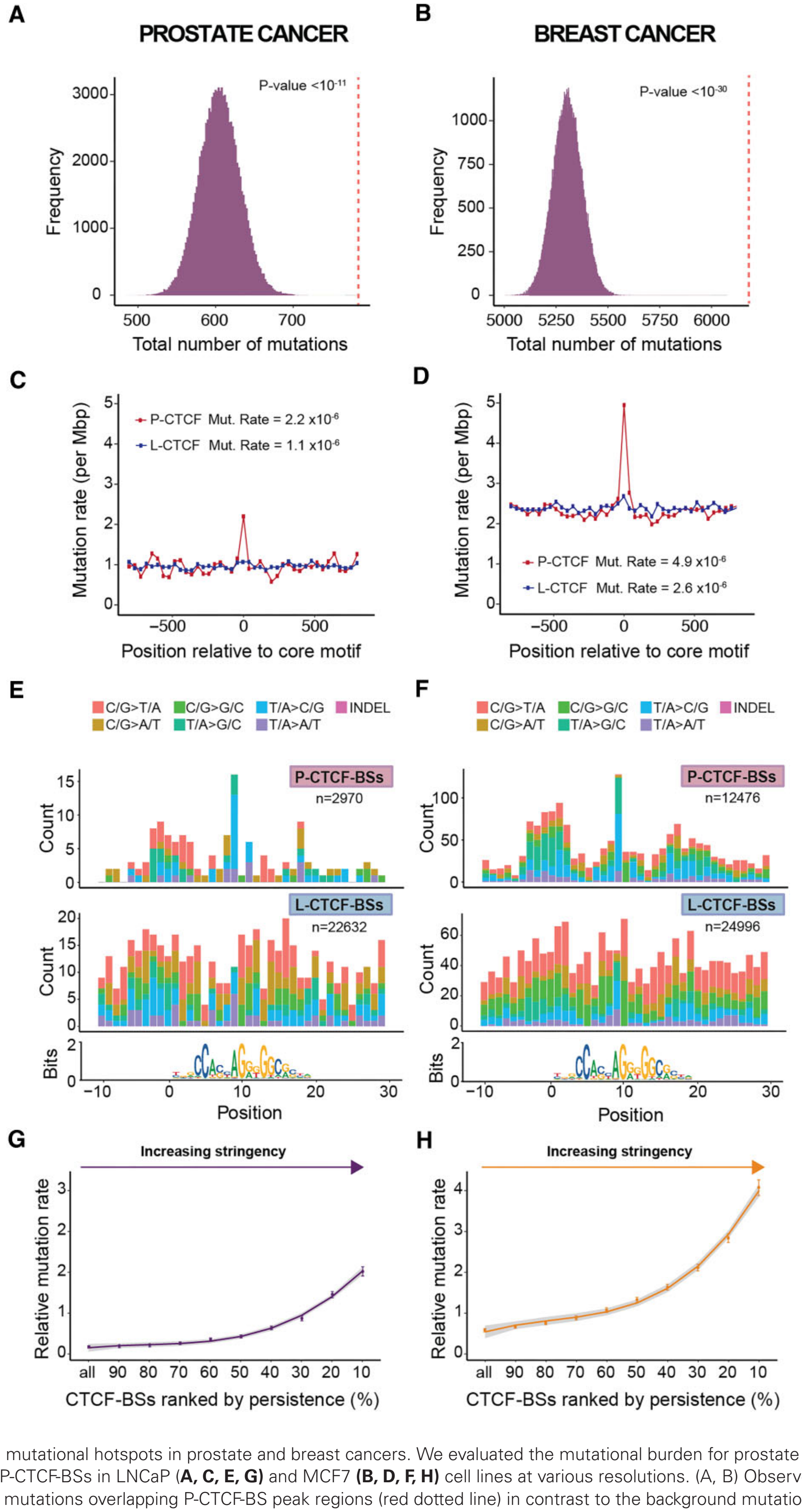


Figure 2. P-CTCF-BSs are mutational hotspots in prostate and breast cancers. We evaluated the mutational burden for prostate and breast cancer at experimentally-identified P-CTCF-BSs in LNCaP (**A, C, E, G**) and MCF7 **(B, D, F, H)** cell lines at various resolutions. (A, B) Observed number of ICGC prostate or breast cancer mutations overlapping P-CTCF-BS peak regions (red dotted line) in contrast to the background mutation count distribution. This distribution was generated from 100 000 peak regions randomly sampled from all CTCF ChIP-seq peaks, operating under the null hypothesis that there is no difference in mutational burden between P- and L-CTCF-BSs. *P*-values were calculated using a permutation test (Materials and methods). (C, D) CTCF-BS peak regions were divided into 41 consecutive 40bp intervals centered at the CTCF core motif. The dots show mutation rates calculated for each interval for P- and L-CTCF-BSs (red and blue, respectively). (E, F) A sequence logo plot of the CTCF core motif. Stacked bar plots show the number of mutations (y-axis) for each base (x-axis) at the core motif for P-CTCF-BSs (top) and L-CTCF-BSs (bottom). Coloured segments of the stacked bar plot correspond to different mutations, categorized by the original base and its respective mutant. (G, H) Line graphs show relative mutation rates (core vs. flanking regions) at P-CTCF-BSs predicted at varying stringencies using CTCF-INSITE.

sistence increases from all CTCF-BSs to the top 10% predicted P-CTCF-BSs. Third, we confirmed equivalent mutational patterns can also be robustly detected using P-CTCF-BSs predicted from other prostate and breast cell lines (Supplementary Figure S6).

Finally, we examined other features that have previously been associated with elevated mutation rates, including colocation with chromatin loop anchors (23,24,29,53), and high CTCF ChIP-seq binding strength (29). Although both features were associated with high mutation rates, we found that the enrichment was weaker than that between CTCF-BS persistence and CTCF-BS mutation rates (Supplementary Figure S7; Supplementary Table S5). CTCF-BSs overlapping loop anchors (18–24% of all CTCF-BSs for both cell lines) exhibited a relative mutation rate of 1.4x for LNCaP and 1.98x for MCF7, which is lower than the respective 1.7× and 3.8× mutation rates at P-CTCF-BSs predicted by CTCF-INSITE for the top 20% most persistent CTCF-BSs (chosen as a similar proportion to all CTCF-BSs that overlap loop anchors) (Supplementary Table S5). Similar enrichment was found for CTCF ChIP-seq fold enrichment in the top 20% (1.4× and 1.8× for LNCaP and MCF7 respectively). Together our analyses demonstrate that these hotspots in breast and prostate cancer may be driven primarily by P-CTCF-BSs.

P-CTCF-BS mutations are enriched at sites of potential 3D genome dysregulation

To next understand the functional importance of P-CTCF-BSs in cancer, we examined the enrichment of functional mutations at P-CTCF-BSs versus L-CTCF-BSs. We defined mutations that were potentially functional as those that (i) led to alteration in CTCF binding and (ii) were co-located at chromatin loop anchors and therefore could alter chromatin conformation. Changes in CTCF binding can be assessed from the allele-specific imbalance using ChIP-seq data or the difference in allele-specific motif scores (Δscore), illustrated schematically in Figure 3A. Similarly, the impact on looping can be assessed from allelic analysis of ChIA-PET data (Materials and methods). Importantly, the ChIP-seq and ChIA-PET data needs to be of sufficient depth to ensure coverage at individual alleles when performing allelic analysis. As public CTCF ChIA-PET data of sufficient depth in breast or prostate cell lines was not available we performed the following analysis using high depth ChIP-seq and ChIA-PET data for GM12878 cells. We first compared read depths for reference and non-reference alleles at all heterozygous sites (germline or somatic variants) and found a strong correlation between changes in motif scores and biases in allelic read frequency (Spearman's $r_s{}^2 = 0.41$) (Figure 3B top panel; Supplementary Figure S8). Specifically, a motif score decrease of ≥2 for the non-reference allele led to a substantial decrease in allelic read depth for this allele. Similarly, a motif score increase ≥2 led to a substantial increase in allelic read depth. Second, we found that Δscore was also correlated with allelic differences in looping (Spearman's $r_s{}^2 = 0.35$) (Figure 3B bottom panel). Moreover, we identified 91 sites that were common to both ChIP-seq and ChIA-PET datasets that showed significant allelic imbalance in binding; 78% of these displayed significant bias in chromatin looping between alleles (Figure 3C). As the motif score correlates well with disrupted binding, as measured by allelic imbalance, it can be used without the need for ChIP-seq or ChIA-PET experimental data. Taken together, this analysis reveals that a mutation causing |Δscore| ≥2 is an appropriate cutoff to define disruptive mutations. We classified ICGC mutations as 'disruptive' if they led to a |Δscore| ≥2 and identified 40.4% disruptive mutations within the P-CTCF-BS subset relative to only 26.5% disruptive mutations within the L-CTCF-BSs (Figure 3D; Supplementary Figure S9). Moreover, most of these mutations led to a decrease in binding rather than an increase in binding.

To provide functional evidence that motif score accurately predicts a disruption to CTCF binding affinity, we performed Fluorescence Polarisation DNA Binding (FPDB) *in vitro* assays, using the recombinant truncated CTCF protein encoding the DNA binding domain (11 zinc-finger domain; see Materials and methods). We selected four candidate ICGC mutations that Deepbind predicted to disrupt DNA binding by CTCF. The selected oligos harboured different mutations across the CTCF motif (Supplementary Table S6) and all showed lower binding affinity in the FPDB assay when compared to the reference oligos (Supplementary Figure S10). For example, the breast cancer mutation corresponding to the deletion of the entire CTCF motif greatly reduced the binding affinity, with little change in anisotropy detected even at the highest CTCF concentration of 1 μM (Supplementary Figure S10). The point mutations, regardless of their position in the CTCF motif, also led to substantial reduction in affinity (Supplementary Figure S10).

Next, we found P-CTCF-BSs, compared to L-CTCF-BSs, were enriched for both disruptive mutations and localization at chromatin loop anchors in both prostate and breast cancer (Figure 3E, F; Supplementary Figure S9). Specifically, in the prostate cancer model, 1% of CTCF-BSs (294/25602) contain disruptive mutations, with a ratio of 0.33 (74/220) P- to L- CTCF-BSs (Figure 3F). This represents a significant enrichment of P-CTCF-BS disruption (*P*-value < 0.01) compared to the background ratio of 0.25 (183/830) for all mutated CTCF-BSs (Figure 3F). We found 27% of the disrupted CTCF-BSs were potentially functional as they were also located at loop anchors. Remarkably, we also found for these functional mutations there was an even greater enrichment of P- to L- CTCF-BSs with a ratio of 1.2 (43:37). Similar enrichment of potentially functional mutations was observed in breast cancer at P-CTCF-BSs, with a P-/L-CTCF-BS ratio of 3.2 (223:70), compared to the background ratio of 0.75 (1585:2113) for all mutated CTCF-BSs (Supplementary Figure S9). In contrast, at CTCF-BSs containing disruptive mutations, but located outside of loop anchors, the P-/L-CTCF-BS ratios, 0.17 (31:183) for prostate and 0.85 (393: 460) for breast cancer, were not significantly different from the background ratios (Figure 3F; Supplementary Figure S9). Overall, the enrichment of potentially functional mutations within P-CTCF-BSs suggests their important role in 3D genome dysregulation in cancer.

Finally, we sought to establish a mechanistic relationship between mutated P-CTCF-BSs and oncogenesis. We compiled lists of genes that were within +/-1kb of mutated P-CTCF-BSs in breast and prostate cancer and performed gene set enrichment analysis (Materials and methods) to assess whether there was enrichment of hallmark genes from the Gene Set Enrichment Analysis (GSEA) database (50,51). In breast cancer we found enrichment of genes downregulated in response to UV and for prostate cancer the enrichment was for the gene set that defines epithelial to mesenchymal transition.

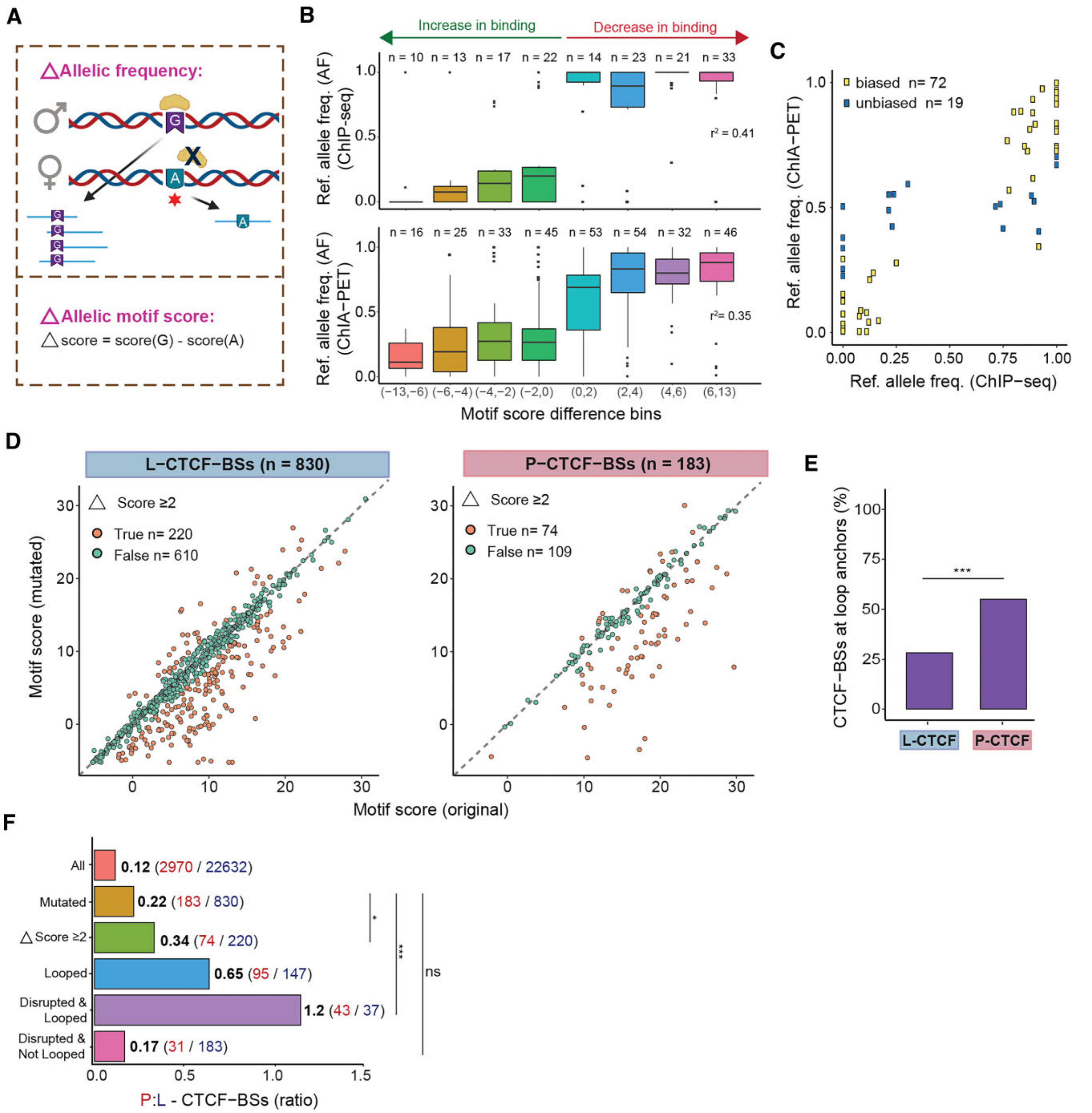


Figure 3. Allelic analyses identify that functional mutations are enriched at P-CTCF-BSs. (**A**) Schematic representation of method used in (B) to determine differences in allelic frequencies and motif scores for reference and non-reference alleles. Created with Biorender.com. (**B**) Boxplots show the allelic difference in DeepBind motif scores (reference minus non-reference)(x-axis) vs. reference allelic frequency (AF) (y-axis). AFs at heterozygotes determined from WGS data were calculated from allelic depths obtained for GM12878 ChIP-seq (top panel) and ChIA-PET (bottom panel) data. The correlation is indicated for eight bins that were divided by motif score differences. Arrows indicate where the non-reference allele leads to an increase or decrease in CTCF binding. (**C**) Correlation of reference AF from ChIP-seq (x-axis) and ChIA-PET (y-axis) data for CTCF motifs at CTCF-BSs with motif score differences ≥2. A binomial test was used to determine the CTCF-BSs that display a significant bias in looping (yellow) and those that do not (blue) (Materials and methods). (**D**) Reference vs. non-reference motif scores for L-CTCF-BSs (left) and P-CTCF-BSs (right) defined in prostate cancer. Orange and blue dots signify CTCF-BSs with |Δmotif scores|≥2, or |Δmotif scores|<2, respectively. (**E**) Proportion of L-CTCF and P-CTCF that are located at loop anchors determined from LNCaP ChIA-PET data. Significance of enrichment of loop anchors to be at P-CTCF-BSs vs. L-CTCF-BSs was calculated from a χ^2 test based on a 2 × 2 contingency table. '***' represents *P*-value < 0.0001. (**F**) Ratio of P- and L- CTCF-BSs under varying selection criteria (y-axis). Pairwise comparisons between ratios generated for different criteria (marked by vertical lines) were performed, with *P*-values calculated using χ^2 tests. Significance levels are indicated as: 'ns' for not significant, '*' for *P*-value < 0.01 and '***' for *P*-value < 0.0001.

(Supplementary Figure S11). Altogether these results suggest that P-CTCF-BS mutations potentially cause loss of CTCF binding that leads to dysregulation of cancer related genes.

P-CTCF-BS mutations are enriched across multiple cancer types

We next assessed whether our observed mutation enrichment at P-CTCF-BSs is a pan-cancer phenomenon. We compiled simple mutations from WGS of the different cancer cohorts from ICGC, performed quality control to exclude the microsatellite instable cancer data (Materials and methods), and grouped the mutations from the cohorts into 12 cancer types according to tissue of origin (Supplementary Table S2). We curated ENCODE CTCF ChIP-seq data for each of the cancer types, and used CTCF-INSITE to predict the top 10% most persistent CTCF-BSs for the pan-cancer enrichment analyses. As expected, the different cancer types varied substantially in mutation rate within CTCF-BSs (Figure 4A). However, all cancers had a significantly increased mutation rate at P-CTCF-BSs relative to L-CTCF-BSs (Figure 4B). Interestingly, the prevalence of mutations at P-CTCF-BSs was not dependent on the overall mutational burden. For example, the elevated mutation rate at P- vs L-CTCF-BSs is three times higher in oesophageal and stomach cancers, despite a lower mutational burden (Figure 4A, B). To exclude that the increased mutational rate might be driven by a small number of patients, we stratified the data into three subgroups of patients with higher mutational loads i.e. top 10%, 10–20% and 20–30% and evaluated the enrichment of mutations in the P-CTCF-BS core motif versus flanking regions. We observed a consistent enrichment of mutations in the core motif for all three groups, indicating a robustness of enrichment of P-CTCF-BS mutations throughout the cohorts (Supplementary Figure S12). Furthermore, P-CTCF-BSs also displayed a significant enrichment of disruptive mutations in most cancer types (Figure 4C and Supplementary Table S7), similar to our observations in breast and prostate cancer (Figure 3F).

We next investigated the per-base mutation rate at the CTCF core motif with a ±5 bp flanking region. To normalize for differences in mutation rates (Figure 4A), we calculated relative mutation rate by dividing the observed value by an expected value that was simulated using trinucleotide mutation rates for each cancer (Materials and methods). In 8/12 cancers, we identified an elevated mutation rate at the 9th (predominantly A/T → G/C or A/T → C/G mutations), 8th, −2nd to 2nd and 17th to 18th bases (Figure 4D). This mutational pattern has been reported at CTCF-BSs in gastrointestinal cancers (23,25,26,28); however, this is the first report in prostate and breast cancers, likely because the pattern is weak when examining all CTCF-BSs (Supplementary Figure S13). Mutational enrichment at the 7th, 13th, 14th (G bases) in skin cancer and melanoma has also been reported and is linked to UV impairment of Nucleotide Excision and Mismatch Repair (NEMR) (29). No enrichment was observed at the binding motifs in SCLC (Figure 4D), despite a high background mutation rate (Figure 4A). This is potentially due to the small cohort size or that this cancer is caused by distinct carcinogens compared to other cancer type (57). Indeed, the different mutational signatures observed across the cancers might highlight different or, in some cases, shared aetiologies among them. Taken together our results confirm that P-CTCF-BSs are mutational hotspots in a pan-cancer context.

Discussion

There is substantial heterogeneity amongst CTCF binding sites (CTCF-BSs) across the genome, including variable binding affinity, cell-type specificity (3), conservation, and involvement in 3D chromatin structures. However, a subset of CTCF-BSs has recently been identified by us and others that display persistent and strong CTCF binding in contrast to the majority of CTCF-BSs, that are sensitive to robust CTCF experimental depletion using different methodologies (18,21,22). Characterisation of P-CTCF-BSs revealed high conservation and enrichment at chromatin loop anchors and TAD boundaries, suggesting a possible fundamental role in constitutive chromatin architecture (18). In this study, using computational modelling to predict P-CTCF-BSs we show for the first time that this subclass of CTCF-BSs is highly enriched for mutations across multiple different cancer types when compared to all CTCF-BSs. Furthermore, these mutations are suggested to be functional by potential disruption of associated chromatin loops and reduced binding in *in vitro* binding assays.

To predict CTCF binding persistence we developed a software CTCF-INSITE which implemented two machine learning models by training on the genetic and epigenetic features of experimentally-defined P-CTCF-BSs. First, we found that while binding affinity, as measured by ChIP-seq fold enrichment, motif score and constitutive binding, is the strongest predictor of persistence, all features combined outperformed the models using the top three features alone. Second, we validated that predicted P-CTCF-BSs from CTCF-INSITE produce concordant results to those from experimentally-derived P-CTCF-BSs, including the enrichment of mutations and mutational profile at the core motif. In addition, the prediction model also allows persistence to be assessed at varying stringencies without the need for further experimental studies. Third, using predicted P-CTCF-BSs from CTCF-INSITE we made the notable discovery that mutations are highly enriched at P-CTCF-BSs across all the cancer types relative to L-CTCF-BSs. Even though CTCF-BS mutations in cancer have been described previously (23–26,28,29,55), our analyses demonstrate that these hotspots across multiple cancer types may be driven primarily by P-CTCF-BSs which is strongly supported by our observation that P-CTCF-BSs have a highly elevated rate of mutation when compared to L-CTCF-BSs.

The enriched mutational signal at P-CTCF-BSs can be useful to study mutational profiles. For example, focusing on per-base mutation rate, we observed a canonical mutational profile in 8 cancer types with frequent mutations at A/T bases in the CTCF core motif. This profile has been previously described at all CTCF-BSs for gastrointestinal cancers (25). However, we found the same mutational profile in other cancer types, such as prostate and breast cancer when focusing only on P-CTCF-BSs. The notable exceptions to the canonical mutational profile are in skin, melanoma and lung cancer. The first two displayed elevated mutation rates at 7th, 13th, 14th (G bases), a pattern previously linked to UV excision misrepair (26); whereas lung showed no specific pattern, which may reflect the aetiology of this cancer which is commonly caused by environmental pollutants, e.g. smoking (57).

Lastly, we demonstrate that P-CTCF-BSs are significantly enriched with disruptive mutations which likely prevent CTCF binding, and are also enriched at loop anchors compared to L-CTCF-BSs. Importantly, we also showed that disrupted CTCF binding was correlated with loss of looping. The fact that disruptive mutations were only enriched at loop

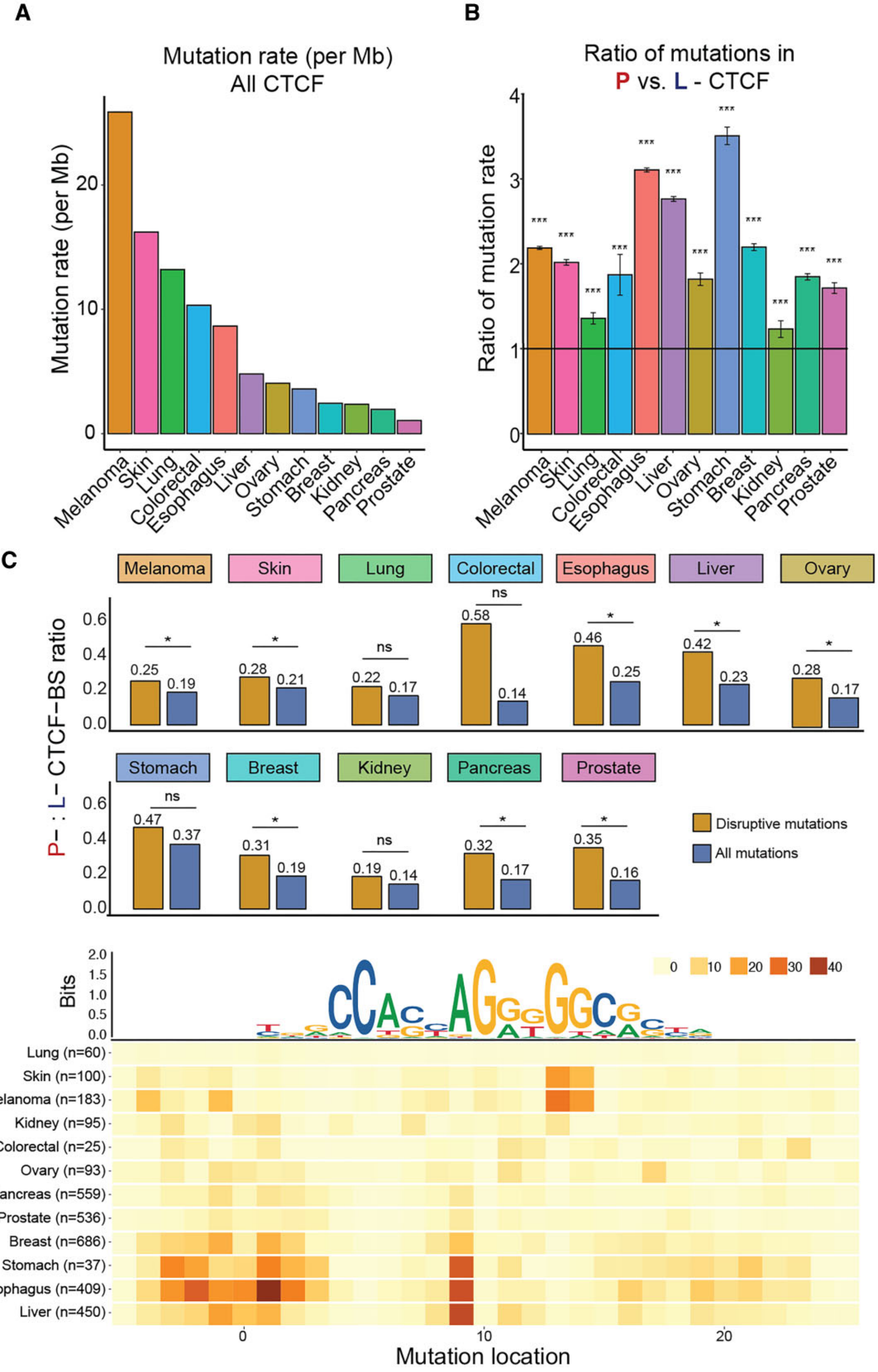


Figure 4. P-CTCF-BSs are pan-cancer mutational hotspots. CTCF-INSITE was used to predict P-CTCF-BSs from all CTCF-BSs from 12 ChIP-seq datasets for various tissues downloaded from ENCODE. These tissues match the 12 solid cancers from ICGC. (**A**) Mutation rate (per Mb) at CTCF binding sites for various cancers from ICGC. (**B**) Relative mutation rate at predicted P-CTCF-BSs compared to all CTCF-BSs calculated at a 40-bp interval centered at the core motif. *P*-values were calculated using χ^2 tests, '***' indicates *P*-value < 0.0001. (**C**) Ratio of P- :L- CTCF-BSs for functional mutation (orange) and all mutation (blue) categories. *P*-values were calculated using χ^2 tests from pairwise comparison between ratios. Significance levels are indicated as: 'ns' for not significant and '*' for *P*-value < 0.01. (**D**) Adjusted mutation rate per base (columns) for each cancer type (rows) calculated by dividing the observed mutations counts by the expected mutations counts derived from simulation (Materials and methods). A sequence logo plot of the CTCF core motif is placed to indicate the location relative to the core.

anchors indicates that loop disruption may provide a selective advantage. Taken together this suggests that P-CTCF-BS mutations may also promote oncogenic programs by altering looping of cancer-related genes; initially proposed as a function for exemplary CTCF-BSs (26,58). Given that P-CTCF-BSs mediate constitutive chromatin loops and domains (18), it is interesting to speculate that cell-type constitutive structures are the target of cancer mutations, however this will require further studies.

In summary, our study suggests that among the tens of thousands of CTCF-BSs across the genome, P-CTCF-BSs predicted by CTCF-INSITE provide worthy candidates to prioritize for experimental manipulation in the pursuit of new biological insights into cancer aetiology.

Data availability

For results generated from public sequencing data, data information has been provided in Supplementary Table S1. WGS for LNCAP is deposited in Sequence Read Archive (SRA): SRR26235471.

The code used in this manuscript is available at https://github.com/Yves-CHEN/CTCF-INSITE and https://doi.org/10.5281/zenodo.11275819.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank members of the Clark Lab for helpful discussions regarding the manuscript and the Garvan Institute of Medical Research and National Computer Infrastructure (NCI) for research computing. Stacey Thomas and Tinghai Xu from Peter Jones Lab at Van Andel Institute, USA, for advice relating to CTCF protein purification. Graphical Abstract & Figure 3A were created with Biorender.com.

Author contributions: Conception: S.J.C. and A.K.; Design: W.C., A.K. and S.J.C.; Experimental: A.K., Y.C.Z., A.G.S. & A.K.J.; Data analyses: W.C., E.C. & J.A-K.; Writing the manuscript: W.C., A.K. & S.J.C.

Funding

National Health and Medical Research Council (NHMRC) Fellowship [1156408 to S.J.C.]; NHMRC Ideas Grant [2012072, 2 020 334 to S.J.C.]. The contents of the published material are solely the responsibility of the Administering Institution, a Participating Institution or individual authors and do not reflect the views of the NHMRC. Funding for open access charge: NHMRC.

Conflict of interest statement

None declared.

References

1. Nakahashi,H., Kieffer Kwon,K.R., Resch,W., Vian,L., Dose,M., Stavreva,D., Hakim,O., Pruett,N., Nelson,S., Yamane,A., *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.*, **3**, 1678–1689
2. Soochit,W., Sleutels,F., Stik,G., Bartkuhn,M., Basu,S., Hernandez,S.C., Merzouk,S., Vidal,E., Boers,R., Boers,J., *et al.* (2021) CTCF chromatin residence time controls three-dimensional genome organization, gene expression and DNA methylation in pluripotent cells. *Nat. Cell Biol.*, **23**, 881–893.
3. Chen,H., Tian,Y., Shu,W., Bo,X. and Wang,S. (2012) Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One*, **7**, e41374.
4. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Goncalves,A., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages (vol 148, pg 335, 2012). *Cell*, **148**, 832–832.
5. Ong,C.T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234–246.
6. Merkenschlager,M. and Nora,E.P. (2016) CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genomics Hum. Genet.*, **17**, 17–43.
7. Schmidt,D., Schwalie,P.C., Ross-Innes,C.S., Hurtado,A., Brown,G.D., Carroll,J.S., Flicek,P. and Odom,D.T. (2010) A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.*, **20**, 578–588.
8. Kurukuti,S., Tiwari,V.K., Tavoosidana,G., Pugacheva,E., Murrell,A., Zhao,Z.H., Lobanenkov,V., Reik,W. and Ohlsson,R. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10684–10689.
9. Bell,A.C. and Felsenfeld,G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405**, 482–485.
10. Hou,C., Zhao,H., Tanimoto,K. and Dean,A. (2008) CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 20398–20403.
11. Schuijers,J., Manteiga,J.C., Weintraub,A.S., Day,D.S., Zamudio,A.V., Hnisz,D., Lee,T.I. and Young,R.A. (2018) Transcriptional dysregulation of MYC reveals common enhancer-docking mechanism. *Cell Rep.*, **23**, 349–360.
12. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
13. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F., *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
14. Cuddapah,S., Jothi,R., Schones,D.E., Roh,T.Y., Cui,K. and Zhao,K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.
15. Zuin,J., Dixon,J.R., van der Reijden,M.I., Ye,Z., Kolovos,P., Brouwer,R.W., van de Corput,M.P., van de Werken,H.J., Knoch,T.A., van,I.W.F., *et al.* (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 996–1001.
16. Kubo,N., Ishii,H., Xiong,X., Bianco,S., Meitinger,F., Hu,R., Hocker,J.D., Conte,M., Gorkin,D., Yu,M., *et al.* (2021) Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. *Nat. Struct. Mol. Biol.*, **28**, 152–161.
17. Franke,M., De la Calle-Mustienes,E., Neto,A., Almuedo-Castillo,M., Irastorza-Azcarate,I., Acemel,R.D., Tena,J.J., Santos-Pereira,J.M. and Gomez-Skarmeta,J.L. (2021) CTCF knockout in zebrafish induces alterations in regulatory landscapes and developmental gene expression. *Nat. Commun.*, **12**, 5415.

18. 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.
19. 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.
20. Lupianez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R., *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
21. Luan,J., Xiang,G., Gomez-Garcia,P.A., Tome,J.M., Zhang,Z., Vermunt,M.W., Zhang,H., Huang,A., Keller,C.A., Giardine,B.M., *et al.* (2021) Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Rep.*, **34**, 108783.
22. Marina-Zarate,E., Rodriguez-Ronchel,A., Gomez,M.J., Sanchez-Cabo,F. and Ramiro,A.R. (2023) Low-affinity CTCF binding drives transcriptional regulation whereas high-affinity binding encompasses architectural functions. *iScience*, **26**, 106106.
23. Katainen,R., Dave,K., Pitkanen,E., Palin,K., Kivioja,T., Valimaki,N., Gylfe,A.E., Ristolainen,H., Hanninen,U.A., Cajuso,T., *et al.* (2015) CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.*, **47**, 818–821.
24. 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.
25. Umer,H.M., Cavalli,M., Dabrowski,M.J., Diamanti,K., Kruczyk,M., Pan,G., Komorowski,J. and Wadelius,C. (2016) A significant regulatory mutation burden at a high-affinity position of the CTCF motif in gastrointestinal cancers. *Hum. Mutat.*, **37**, 904–913.
26. 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.
27. Ji,X., Dadon,D.B., Powell,B.E., Fan,Z.P., Borges-Rivera,D., Shachar,S., Weintraub,A.S., Hnisz,D., Pegoraro,G., Lee,T.I., *et al.* (2016) 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell*, **18**, 262–275.
28. Guo,Y.A., Chang,M.M., Huang,W.T., Ooi,W.F., Xing,M.J., Tan,P. and Skanderup,A.J. (2018) Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.*, **9**, 1520.
29. 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.
30. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
31. Dekker,J., Belmont,A.S., Guttman,M., Leshyk,V.O., Lis,J.T., Lomvardas,S., Mirny,L.A., O'Shea,C.C., Park,P.J., Ren,B., *et al.* (2017) The 4D nucleome project. *Nature*, **549**, 219–226.
32. Achinger-Kawecka,J., Valdes-Mora,F., Luu,P.L., Giles,K.A., Caldon,C.E., Qu,W., Nair,S., Soto,S., Locke,W.J., Yeo-Teh,N.S., *et al.* (2020) Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nat. Commun.*, **11**, 320.
33. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
34. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
35. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
36. Du,Q., Bert,S.A., Armstrong,N.J., Caldon,C.E., Song,J.Z., Nair,S.S., Gould,C.M., Luu,P.L., Peters,T., Khoury,A., *et al.* (2019) Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun.*, **10**, 416
37. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP plus. *PLoS Comput. Biol.*, **6**, e1001025.
38. Breiman,L., (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
39. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
40. Amemiya,H.M., Kundaje,A. and Boyle,A.P. (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 9354.
41. Tang,Z., Luo,O.J., Li,X., Zheng,M., Zhu,J.J., Szalaj,P., Trzaskoma,P., Magalska,A., Wlodarczyk,J., Ruszczycki,B., *et al.* (2015) CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, **163**, 1611–1627.
42. Sherry,S.T., Ward,M. and Sirotkin,K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.*, **9**, 677–679.
43. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
44. Lee,B., Wang,J., Cai,L., Kim,M., Namburi,S., Tjong,H., Feng,Y., Wang,P., Tang,Z., Abbas,A., *et al.* (2020) ChIA-PIPE: A fully automated pipeline for comprehensive ChIA-PET data analysis and visualization. *Sci. Adv.*, **6**, eaay2078.
45. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
46. Tan,G. and Lenhard,B. (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, **32**, 1555–1556.
47. Tan,G. (2014) JASPAR2014: Data package for JASPAR.
48. Thomas,S.L., Xu,T.H., Carpenter,B.L., Pierce,S.E., Dickson,B.M., Liu,M., Liang,G. and Jones,P.A. (2023) DNA strand asymmetry generated by CpG hemimethylation has opposing effects on CTCF binding. *Nucleic Acids Res.*, **51**, 5997–6005.
49. Kolberg,L., Raudvere,U., Kuzmin,I., Adler,P., Vilo,J. and Peterson,H. (2023) g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.*, **51**, W207–W212.
50. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E., *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
51. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
52. Bergstrom,E.N., Barnes,M., Martincorena,I. and Alexandrov,L.B. (2020) Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinf.*, **21**, 438.
53. Liu,E.M.W., Martinez-Fundichely,A., Diaz,B.J., Aronson,B., Cuykendall,T., MacKay,M., Dhingra,P., Wong,E.W., Chi,P.,

Apostolou,E., *et al.* (2019) Identification of cancer drivers at CTCF insulators in 1,962 whole genomes. *Cell Syst.*, **8**, 446–455.

54. Kikutake,C. and Suyama,M. (2022) Pan-cancer analysis of mutations in open chromatin regions and their possible association with cancer pathogenesis. *Cancer Med.*, **11**, 3902–3916.
55. Lee,C.A., Abd-Rabbo,D. and Reimand,J. (2021) Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes. *Genome Biol.*, **22**, 133.
56. Nesta,A.V., Tafur,D. and Beck,C.R. (2021) Hotspots of human mutation. *Trends Genet.*, **37**, 717–729.
57. Pesch,B., Kendzia,B., Gustavsson,P., Jöckel,K.H., Johnen,G., Pohlabeln,H., Olsson,A., Ahrens,W., Gross,I.M. and Brüske,I. (2012) Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case–control studies. *Int. J. Cancer*, **131**, 1210–1219.
58. Oh,S., Shao,J., Mitra,J., Xiong,F., D'Antonio,M., Wang,R., Garcia-Bassets,I., Ma,Q., Zhu,X., Lee,J.H., *et al.* (2021) Enhancer release and retargeting activates disease-susceptibility genes. *Nature*, **595**, 735–740.

Received: September 29, 2023. **Revised:** May 15, 2024. **Editorial Decision:** May 19, 2024. **Accepted:** June 13, 2024