CANCER

Activation of proto-oncogenes by disruption of chromosome neighborhoods

Denes Hnisz,1,2 Abraham S. Weintraub,1,2a Daniel S. Day,1 Anne-Laure Valton,3 Rasmus O. Bak,4 Charles H. Li,1,2 Johanna Goldmann,1 Bryan R. Lajoie,1 Zh Peng Fan,1,5 Alla A. Sigova,1 Jessica Reddy,1,2 Diego Borges-Rivera,1,2 Tong Ihn Lee,1 Rudolf Jaenisch,1,2 Matthew H. Porteus,4 Job Dekker,5,6 Richard A. Young1,2

Oncogenes are activated through well-known chromosomal alterations such as gene fusion, translocation, and focal amplification. In light of recent evidence that the control of key genes depends on chromosome structures called insulated neighborhoods, we investigated whether proto-oncogenes occur within these structures and whether oncogene activation can occur via disruption of insulated neighborhood boundaries in cancer cells. We mapped insulated neighborhoods in T-cell acute lymphoblastic leukemia (T-ALL) and found that tumor cell genomes contain recurrent microdeletions that eliminate the boundary sites of insulated neighborhoods containing prominent T-ALL proto-oncogenes. Perturbation of such boundaries in nonmalignant cells was sufficient to activate proto-oncogenes. Mutations affecting chromosome neighborhood boundaries were found in many types of cancer. Thus, oncogene activation can occur via genetic alterations that disrupt insulated neighborhoods in malignant cells.

Typically constrained within larger CCCTC-binding factor (CTCF) cohesin-mediated loops called insulated neighborhoods (3–10), which in turn can form clusters that contribute to topologically associating domains (TADs) (11, 12) (fig. S1A). This recent understanding of chromosome structure led us to hypothesize that silencing proto-oncogenes located within insulated neighborhoods might be achieved in cancer cells via loss of an insulated neighborhood boundary, with consequent aberrant activation by enhancers that are normally located outside the neighborhood (Fig. 1A, lowest panel).

To test this hypothesis, we used chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) to map neighborhoods and other cis-regulatory interactions in a cancer cell genome (Fig. 1B and table S1). A T-cell acute lymphoblastic leukemia (T-ALL) Jurkat cell line was selected for these studies because key T-ALL oncogenes and genetic alterations are well known (13, 14). The ChIA-PET technique generates a high-resolution (~5 kb) chromatin interaction map of sites in the genome bound by a specific protein factor (8, 15, 16). Cohesin was selected as the target protein because it is involved in both CTCF-CTCF interactions and enhancer-promoter interactions (5–7) and has proven useful for identifying insulated neighborhoods (8, 10) (fig. S1, A and B). The cohesin ChIA-PET data were processed using multiple analytical approaches (figs. S1 to S4 and table S1), and their analysis identified 9757 high-confidence interactions, including 9038 CTCF-CTCF interactions and 379 enhancer-promoter interactions (fig. S4C). The CTCF-CTCF loops had a median length of 270 kb, contained on average two or three genes, and covered ~52% of the genome (table S2). Such CTCF-CTCF loops have been called insulated neighborhoods because disruption of either CTCF boundary causes dysregulation of local genes due to inappropriate enhancer-promoter interactions (8, 10).

Consistent with this, the Jurkat chromosome structure data showed that the majority of cohesin-associated enhancer-promoter interactions had end points that occurred within the CTCF-CTCF loops (Fig. 1C and fig. S2H). These results provide an initial map of the three-dimensional (3D) regulatory landscape of a tumor cell genome.

We next investigated the relationship between genes that have been implicated in T-ALL pathogenesis and the insulated neighborhoods. The majority of genes (40 of 55) implicated in T-ALL pathogenesis, as curated from the Cancer Gene Census and individual studies (table S3), were located within the insulated neighborhoods identified in the Jurkat cell genome (Fig. 1D and fig. S5); 27 of these genes were transcriptionally active and 13 were silent, as determined by RNA sequencing (RNA-seq) (Fig. 2A and table S4). Active oncogenes are often associated with super-enhancers (17, 18), and we found that 13 of the 27 active T-ALL pathogenesis genes were associated with super-enhancers (Fig. 2, A and B, and fig. S5A). Silent genes have also been shown to be protected by insulated neighborhoods from active enhancers located outside the neighborhood, and we found multiple instances of silent proto-oncogenes located within CTCF-CTCF loop structures in the Jurkat genome (Fig. 2, A and C, and fig. S2B). Thus, both active oncogenes and

1Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. 2Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. 3Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA. 4Department of Pediatrics, Stanford University, Stanford, CA, USA. 5Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. 6Howard Hughes Medical Institute. *Corresponding author. E-mail: young@wi.mit.edu


ACKNOWLEDGMENTS

We thank M. Hume, Y.-H. Hsu, Y. Shen, and D. Balcha for technical assistance and A. Gimelbrant for helpful discussions. We are grateful to the Exome Aggregation Consortium for making its data publicly available prior to publication. This work was supported by the National Institutes of Health (grants NHGRI RO1 HG003985 to M.L.B. and T.H. and PG0 HG004233 to M.V. and D.E.H.), an A*STAR National Science Scholarship to K.H.K., and National Science Foundation Graduate Research Fellowships to A.L.B. and J.V.K. TF PBM data have been deposited into UniPROBE (publication data set accession BARISA). GST negative control PBM 8-mer data are provided in table S7. M.L.B. is a coinventor on the A*STAR National Science Scholarship to K.H.K., and the National Institutes of Health (grants NHGRI R01 HG003985 to M.L.B. and T.H. and PG0 HG004233 to M.V. and D.E.H.).

Corresponding author. E-mail: young@wi.mit.edu

SUPPLEMENTARY MATERIALS

silent proto-oncogenes are located within insulated neighborhoods in these T-ALL cells.

If some insulated neighborhoods function to prevent proto-oncogene activation, some T-ALL tumor cells may have genetic alterations that perturb the CTCF boundaries of neighborhoods containing T-ALL oncogenes. To investigate this possibility, we identified recurrent deletions in T-ALL genomes that span insulated neighborhood boundaries, using data from multiple studies (table S5A) and filtered for relatively short deletions (<500 kb) so as to minimize collection of deletions that affect multiple genes (fig. S6A). Among the 438 recurrent deletions identified with this approach, 113 overlapped at least one boundary of insulated neighborhoods identified in T-ALL, and 6 of these affected neighborhoods containing T-ALL pathogenesis genes (fig. S6B and table S5B). Examples of two such genes, TAL1 and LMO2, are shown in Fig. 3, A and G.

If deletions overlapping neighborhood boundaries can cause activation of proto-oncogenes within the loops, then site-specific deletion of a loop boundary CTCF site at the TAL1 locus should be sufficient to activate these proto-oncogenes in nonmalignant cells. TAL1 encodes a transcription factor that is overexpressed in ~50% of T-ALL cases and is a key oncogenic driver of this cancer (19, 20). TAL1 can be activated by deletions that fuse a promoterless TAL1 gene to the promoter of STIL (19), and this was observed in many patient deletions (Fig. 3A). Several patient deletions, however, retained the TAL1 promoter.

Fig. 1. 3D regulatory landscape of the T-ALL genome. (A) Mechanisms activating proto-oncogenes. (B) Hi-C interaction map, TADs defined in human embryonic stem cells (H1), cohesin ChIA-PET interactions (intensity of blue arc represents interaction significance), CTCF and H3K27Ac chromatin immunoprecipitation sequencing (ChIP-seq) profiles and peaks, and RNA-seq in Jurkat cells at the CD3D locus. ChIP-seq peaks are denoted as bars above ChIP-seq profiles. (C) ChIA-PET interactions at the RUNX1 locus displayed above the ChIP-seq profiles of CTCF, cohesin (SMC1), and H3K27Ac. FDR, false discovery rate.

Fig. 2. Active oncogenes and silent proto-oncogenes occur in insulated neighborhoods. (A) T-ALL pathogenesis genes. Colored boxes indicate whether a gene is located within a neighborhood, expressed, and associated with a superenhancer. (B) Insulated neighborhood at the active TAL1 locus. The cohesin ChIA-PET interactions are displayed above the ChIP-seq profiles of CTCF, cohesin (SMC1) H3K27Ac, and RNA-seq profile. A model of the insulated neighborhood is shown on the right. (C) Insulated neighborhood at the silent LMO2 locus.
Fig. 3. Disruption of insulated neighborhood boundaries is linked to proto-oncogene activation. (A) Cohesin ChIA-PET interactions and CTCF and cohesin (SMC1) binding profiles at the TAL1 locus in Jurkat cells. Patient deletions described in (22) are shown as bars below the gene models. The deletion on the bottom indicates the minimally deleted region identified in (26). (B) ChIP-seq profiles of CTCF, H3K27Ac, p300, CBP, and RNA-seq at the TAL1 locus in HEK-293T cells. The region deleted using a CRISPR/Cas9-based approach is highlighted in a gray box. (C) Quantitative reverse transcription polymerase chain reaction (qRT-PCR) analysis of TAL1 expression in wild-type HEK-293T cells (wt) and in cells where the neighborhood boundary highlighted in (B) was deleted. (D) Model of the neighborhood and perturbation at the TAL1 locus. (E) 5C contact matrices in wild-type HEK-293T cells and TAL1 neighborhood boundary-deleted cells. An arrow indicates the position of the region removed in the mutant cells. (F) Distance-adjusted z-score difference (5C) maps at the TAL1 locus (ΔCTCF = wild-type HEK-293T). Note the increase in the 5C signal adjacent to the deleted region. CTCF and H3K27Ac binding profiles in wild-type cells are displayed for orientation. (G) Cohesin ChIA-PET interactions and CTCF and cohesin (SMC1) binding profiles at the LMO2 locus. Patient deletions described in (22) are shown as bars below the gene models. (H) ChIP-Seq binding profile of CTCF, H3K27Ac, p300, CBP, and RNA-seq at the LMO2 locus in HEK-293T cells. The region deleted by a CRISPR/Cas9-based approach is highlighted in a gray box. (I) qRT-PCR analysis of LMO2 expression in wild-type HEK-293T cells and in cells where the neighborhood boundary highlighted in (H) was deleted. (J) Model of the neighborhood and perturbation at the LMO2 locus. (K) 5C contact matrices in wild-type HEK-293T cells and LMO2 neighborhood boundary-deleted cells. An arrow indicates the position of the region removed in the mutant cells. (L) Distance-adjusted z-score difference (5C) maps at the LMO2 locus (ΔCTCF = wild-type HEK-293T). Note the increase in the 5C signal adjacent to the deleted region. CTCF and H3K27Ac binding profiles in wild-type cells are displayed for orientation. In (C) and (I), data from n = 3 independent biological replicates are displayed as means ± SD; P < 0.01 between wild-type and boundary-deleted cells (two-tailed t test).
Fig. 4. Somatic mutations of neighborhood boundaries occur in many cancers. (A) "Constitutive neighborhood" at the NOTCH1 locus. CTCF ChIP-seq and cohesin ChIA-PET interactions in Jurkat (T-ALL), GM12878 (lymphoblastoid), and K562 (CML) cells are displayed. (B) Frequency of somatic mutations in the ICGC database at CTCF sites that form constitutive neighborhood boundaries (left) and CTCF sites that do not form neighborhood boundaries (right). (C) Somatic mutations in esophageal adenocarcinoma (ESAD-UK) at constitutive neighborhood boundary CTCF sites. (D) Somatic mutations in hepatocellular carcinoma (LIRI-JP) at constitutive neighborhood boundary CTCF sites. (E and F) Genes in constitutive neighborhood boundaries whose boundary is recurrently mutated in esophageal adenocarcinoma (E) and in hepatocellular carcinoma (F). The bars depict the number of mutations in the neighborhood boundary site. Proto-oncogenes annotated in the Cancer Gene Census are highlighted in red. (G and H) Mutations in the boundary sites of the neighborhood containing the LMO2 proto-oncogene in esophageal adenocarcinoma (G) and the FGFR1 proto-oncogene in hepatocellular carcinoma (H). The enrichment of mutations at the constitutive neighborhood boundary sites (±5 bp of the motif) shown in (B) to (D) relative to regions flanking the binding sites has a $P$ value of $<10^{-4}$ (permutation test).
common, as only a subset of CTCF sites form insu-
lated neighborhoods (8, 10, 24). CTCF cohesin- 
loop boundaries are largely preserved across cell 
types (8, 9, 24), and a set of ~10,000 constitutive 
CTCF-CTCF loops shaped by GM12878 lympho-
blastoid, Jurkat, and K562 (CML) cells (24) were 
identified for comparison (Fig. 4A, fig. S11, and 
table S8). We used the International Cancer Gen-
ome Consortium (ICGC) database—which contains 
data for ~50 cancer types, ~2300 whole-genome 
(WGS) samples, and ~13 million unique 
somatic mutations—to examine the boundaries 
of these neighborhoods for somatic point muta-
tions found in cancer genomes (table S9). We 
found a striking enrichment of mutations at the 
CTCF boundaries of constitutive neighborhoods 
(Fig. 4B, fig. S12A, and table S10) relative to 
regions flanking the boundary CTCF sites (~1 kb 
of the CTCF binding motif, P < 10−4, permutation 
test) (fig. S12B), and in many instances these 
developed a change in the consensus CTCF binding 
motif (fig. S12C). Nonboundary CTCF sites did 
not show such enrichment (Fig. 4B and figs. S12D 
and S14). The genomes of esophageal and liver 
carcinoma samples were particularly enriched 
for boundary CTCF site mutations (Fig. 4, C and 
D, fig. S12, D and E, fig. S13, and table S10), and 
there was no similar enrichment of mutations at 
the binding sites of other transcription factors 
(fig. S15). In these cancers, a considerable 
fraction of the mutated neighborhood boundary 
CTCF sites were affected by multiple mutations 
(≥3 mutations per site) [280/1826 (15%) in esoph-
geal carcinoma, 54/1030 (5%) in liver carcinoma 
and multitarget optimization, we developed a germline-targeting immunogen (eOD-GT8) 
that bind to relatively conserved epitopes 
on the HIV Env glycoprotein trimer and protect against challenge in animal models have
Activation of proto-oncogenes by disruption of chromosome neighborhoods
Science 351 (6280), 1454-1458. [doi: 10.1126/science.aad9024]
originally published online March 3, 2016

Editor's Summary

The spread of bad neighborhoods
Our genomes have complex three-dimensional (3D) arrangements that partition and regulate gene expression. Cancer cells frequently have their genomes grossly rearranged, disturbing this intricate 3D organization. Hnisz et al. show that the disruption of these 3D neighborhoods can bring oncogenes under the control of regulatory elements normally kept separate from them (see the Perspective by Wala and Beroukim). These novel juxtapositions can result in the inappropriate activation of oncogenes. Science, this issue p. 1454; see also p. 1398
Supplementary Materials for

Activation of proto-oncogenes by disruption of chromosome neighborhoods


*Corresponding author. E-mail: young@wi.mit.edu

Published 3 March 2016 on Science First Release
DOI: 10.1126/science.aad9024

This PDF file includes:
Materials and Methods
Figs. S1 to S16
Captions for tables S1 to S13
References

Other supplementary material for this manuscript includes the following:
Tables S1 to S13 (Excel format)
Materials and methods

Cell culture

Jurkat T-ALL cells were cultured in RPMI GlutaMAX (Invitrogen, 61870-127), supplemented with 10% fetal bovine serum, 100 U/ml penicillin and 100 µg/ml streptomycin (Invitrogen, 15140-122).

HEK-293T cells were cultured in DMEM (high glucose, pyruvate; Invitrogen, 11995-073) supplemented with 10% fetal bovine serum, 100 U/ml penicillin and 100 µg/ml streptomycin (Invitrogen, 15140-122).

CD3+ T-cells were isolated from buffy coats obtained from the Stanford School of Medicine Blood Center using a human Pan T-cell Isolation Kit (Miltenyi Biotec, San Diego, CA, USA) according to manufacturer’s instructions. CD3+ cells were cultured in X-VIVO 15 (Lonza, Walkersville, MD, USA) supplemented with 5% human serum (Sigma-Aldrich, St. Louis, MO, USA), 100 IU/ml human rIL-2 (Peprotech, Rocky Hill, NJ, USA), and 10 ng/ml human rIL-7 (BD Biosciences, San Jose, CA, USA). Directly after isolation, T-cells were activated for 3 days with immobilized anti-CD3 antibody (clone: OKT3, Tonbo Biosciences, San Diego, CA, USA) and soluble anti-CD28 antibody (clone: CD28.2, Tonbo Biosciences). T-cells were cultured at 37°C, 5% CO₂, and ambient oxygen levels.

ChIP-Seq

ChIP was performed as described in (27) with a few adaptations. Jurkat cells (~100 million cells, grown to a density of ~1 million cells/ml) or HEK-293T cells (~100 million cells at ~80% confluency) were crosslinked for 10 min at room temperature by the addition of one-tenth of the volume of 11% formaldehyde solution (11% formaldehyde, 50 mM HEPES pH 7.3, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0) to the growth media followed by 5 min quenching with 125 mM glycine. Cells were washed twice with PBS, then the supernatant was aspirated and the cell pellet was flash frozen in liquid nitrogen. Frozen crosslinked cells were stored at −80°C. 100 µl of Protein G Dynabeads (Life Technologies) were blocked with 0.5% BSA (w/v) in PBS. Magnetic beads were bound with 10 µg of anti-H3K27Ac antibody (Abcam ab4729), anti-CTCF antibody (Millipore 07-729), anti-RUNX1 antibody (Abcam ab23980) or anti-GATA3 (Santa Cruz sc-22206X) antibody. Nuclei were isolated as previously described (27), and sonicated in lysis buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, and 1% Triton X-100) on a Misonix 3000 sonicator for 10 cycles at 30s each on ice (18-21 W) with 60 s on ice between cycles. Sonicated lysates were cleared once by centrifugation and incubated overnight at 4°C with magnetic beads bound with antibody to enrich for DNA fragments bound by the indicated factor. Beads were washed with wash buffer A (50 mM HEPES-KOH pH7.9, 140 mM NaCl, 1 mM EDTA pH 8.0, 0.1% Na-Deoxycholate, 1% Triton X-100, 0.1% SDS), B (50 mM HEPES-KOH pH7.9, 500 mM NaCl, 1 mM EDTA pH 8.0, 0.1% Na-Deoxycholate, 1% Triton X-100, 0.1% SDS), C (20 mM Tris-HCl pH8.0, 250 mM LiCl, 1 mM EDTA pH 8.0, 0.5% Na-Deoxycholate, 0.5% IGEPAL C-630 0.1% SDS) and D (TE with 50 mM NaCl) sequentially. DNA was eluted in elution buffer (50 mM Tris-HCL pH 8.0, 10 mM EDTA, 1% SDS). Cross-links were reversed overnight. RNA and protein were digested using RNase A and Proteinase K, respectively and DNA was purified with phenol chloroform
extraction and ethanol precipitation. Purified ChIP DNA was used to prepare Illumina multiplexed sequencing libraries. Libraries for Illumina sequencing were prepared following the Illumina TruSeq DNA Sample Preparation v2 kit. Amplified libraries were size-selected using a 2% gel cassette in the Pippin Prep system from Sage Science set to capture fragments between 200 and 400 bp. Libraries were quantified by qPCR using the KAPA Biosystems Illumina Library Quantification kit according to kit protocols. Libraries were sequenced on the Illumina HiSeq 2500 for 40 bases in single read mode.

ChIP-seq data analysis

ChIP-Seq datasets were aligned using Bowtie (version 0.12.2) (28) to the human genome (build hg19, GRCh37) with parameter -k 1 -m 1 -n 2. We used the MACS version 1.4.2 (model-based analysis of ChIP-seq) (29) peak finding algorithm to identify regions of ChIP-seq enrichment over input DNA control with the parameter “--no-model --keep-dup=1”. A P-value threshold of 1e-09 was used for both H3K27Ac and CTCF. UCSC Genome Browser tracks were generated using MACS wiggle outputs with parameters “-w –S – space=50”. The browser snapshots of the ChIP-Seq binding profiles displayed throughout the study use the number of reads per kilobase per million mapped reads dimension (rpm/bp) on the y-axis. The SMC1 (cohesin) binding profiles show the number of reads in the merged ChIA-PET dataset and were generated with MACS2 (version 2.1.0.20150420) with the parameters “-q 0.05 –B” as part of the SMC1 ChIA-PET data processing using Mango (see below).

ChIP-seq enrichment heatmap

ChIP-seq read density (rpm/bp) for SMC1, MYB, RUNX1, GATA3, TAL1, RNAPII, H3K27Ac and CTCF at the SMC1-bound regions are displayed on Fig. S1B. The input-subtracted average ChIP-seq read density in 50 bp bins was calculated +/- 5 kb around the center of the SMC1-enriched regions exactly as previously described (8). In Fig. S1B, the summits of SMC1 peaks (79,976) identified as part of the SMC1 ChIA-PET data processing using Mango were used. Note that Mango merges the SMC1 peaks into 67,596 PET peaks (Fig. S1C) before identifying interactions.

ChIA-PET

ChIA-PET was performed using a modified version (16) of a previously described protocol (8). Jurkat cells (up to 500-800 million cells, grown to a density of ~1 million cells /ml) were crosslinked with 1% formaldehyde at room temperature for 10 min and then neutralized with 125mM glycine. Crosslinked cells were washed three times with ice-cold PBS, snap-frozen in liquid nitrogen, and stored at -80ºC before further processing. Nuclei were isolated as previously described (27), and chromatin was fragmented using a Misonix 3000 sonicator. The anti-SMC1 antibody (Bethyl, A300-055A) was used to enrich SMC1-bound chromatin fragments exactly as described at the ChIP-Seq section. A portion of ChIP DNA was eluted from antibody-coated beads for concentration quantification and for enrichment analysis using quantitative PCR. For ChIA-PET library construction ChIP DNA fragments were end-repaired using T4 DNA polymerase (NEB) followed by A-tailing with Klenow (NEB). A biotinylated bridge linker (F: /5Phos CGCGATATC/iBiodT/TATCTGACT; R: /5Phos GTCAGATAAGATATCGCGT) with T-overhangs
was added and the proximity ligation was performed overnight at 16°C in 1.5 mL volume. Unligated DNA was then digested with exonuclease and lambda nuclease (NEB). DNA was eluted off the beads in elution buffer (50 mM Tris-HCL pH 8.0, 10 mM EDTA, 1% SDS) followed by overnight crosslink reversal, RNase A treatment, and proteinase K digestion. A phenol:chloroform:isoamyl alcohol extraction was performed followed by an ethanol precipitation. Precipitated DNA was resuspended in Nextera DNA resuspension buffer (Illumina). The DNA was then tagmented with the Nextera Tagmentation kit (Illumina). The tagmented library was purified with a Zymo column and was bound to Streptavidin beads to enrich for ligation junctions (containing the biotinylated bridge linker). 12 cycles of the polymerase chain reaction were performed to amplify the library. The amplified library was size-selected (350-500 bp) with a Pippin prep machine and sequenced with either 100x100 (Replicate 1) or 125x125 (Replicate 2) paired-end sequencing on an Illumina Hi-Seq 2500 platform.

ChIA-PET data analysis using Mango

ChIA-PET analysis was carried out using a combination of in-house scripts and the Mango pipeline (Version: 1.0.4, (30)). Image analysis and base calling was done using the Solexa pipeline. Read pairs were examined for the presence of at least 10 base pairs of linker sequence (see above). Read pairs that did not contain linker in either mate were not processed further. Reads containing linker were trimmed using cutadapt (cutadapt -m 17 -a forward=ACGCGATATCTTATCTGACT –a reverse=AGTCAGATAAGATATCGCGT -- overlap 10) (http://code.google.com/p/cutadapt/). The trimmed mate pairs were then input into the Mango pipeline with the parameters (Rscript mango/mango.R --shortreads FALSE --peakslop 1500 --stages 2:5 --reportallpairs TRUE). The value of the peakslop parameter was adapted from a previous report (24). All other parameters were left as default. The interactions identified by Mango are found in Table S2A. In order to determine an appropriate FDR cutoff for the definition of high confidence interactions, we first identified interactions that were previously detected in the Jurkat T-ALL cells. A previous study has already found a looping interaction between two CTCF sites around the TAL1 gene in Jurkat T-ALL cells using chromosome conformation capture (3C) assay (31), and we found that this positive control interaction has an estimated FDR of 0.14 in our ChIA-PET dataset (Fig. 2B, Table S2A). We therefore defined high confidence interactions as having an FDR ≤ 0.2 for downstream analyses. We note the 0.2 FDR cutoff is likely an over-estimate of the true FDR and that interactions at even higher FDR values likely contain bona fide interactions. For example, the vast majority (>85%) of interactions with an 0.2 ≤ FDR ≤ 0.3 overlap genomic features such as enhancers, promoters or CTCF sites on each of their ends (Fig. S2G).

Mango determines a lower bound distance cutoff to exclude PETs that potentially arose from self-circularization of the same fragment during the proximity ligation step (30). Since self-circularized PETs are formed from a single DNA fragment, the mate pairs always map to opposite DNA strands. Mango uses the fraction of PETs that have reads on opposite strands to estimate the fraction of self-circularized PETs at various distances to estimate the distance at there is negligible bias to which DNA strand the mate pairs of the PETs map (30) (Fig. S2C).

The RAD21 (cohesin) ChIA-PET datasets in GM12878 and K562 were described in a previous study (24), and were processed as described for the Jurkat ChIA-PET data above.
Throughout the entire study, the interactions generated by the Mango pipeline are used for genome wide interaction analyses and visualization of interactions at individual loci unless otherwise noted (on Fig. S3, S4, Table S2B).

**ChIA-PET data analysis using the Dowen et al. pipeline**

All ChIA-PET datasets were also processed with an in house method (“Dowen et al. pipeline”) adapted from previous computational pipeline (8, 32). Image analysis and base calling was done using the Solexa pipeline. Reads were examined for the presence of at least 10 base pairs of linker sequence. Reads that did not contain linker were not processed further. Reads containing linker were trimmed using cutadapt (cutadapt -m 17 -a forward=ACGCGATATCTTATCTGACT -a reverse=AGTCAGATAAGATATCGCGT --overlap 10) (http://code.google.com/p/cutadapt/). Trimmed mate pairs were mapped independently to hg19 using Bowtie version 1.1.1 (bowtie -e 70 -k 1 -m 1 -v 2 -p 4 --best --strata -S) (28). Aligned reads were paired with mates with an in-house script using read identifiers. To remove PCR bias artifacts, reads were filtered for redundancy: PETs with identical genomic coordinates and strand information at both ends were collapsed into a single PET. The PETs were further categorized into intrachromosomal PETs or interchromosomal PETs. Regions of local enrichment (PET peaks) were called using MACS 1.4.2 (29) with the parameters “-p 1e-09 -no-lambda -no-model”. To identify long-range chromatin interactions, we first removed intrachromosomal PETs of length < 5 kb because these PETs may originate from self-ligation of DNA ends from a single chromatin fragment in the ChIA-PET procedure (8). We next identified PETs that overlapped with PET peaks at both ends by at least 1bp. These PETs were defined as putative interactions. A statistical model based upon the hypergeometric distribution was applied to identify high-confidence interactions, representing high-confidence physical contacts between the PET peaks. Specifically, the numbers of PET sequences that overlapped with PET peaks at both ends as well as the number of PETs within PET peaks at each end were counted. The PET count between two PET peaks represented the frequency of the interaction between the two genomic locations. A hypergeometric distribution was used to determine the probability of seeing at least the observed number of PETs linking the two PET peaks. A background distribution of interaction frequencies was then obtained through the random shuffling of the links between two ends of PETs, and a cutoff threshold for calling significant interactions was set to the corresponding p-value of the most significant proportion of shuffled interactions (at an FDR of 0.01). This method yielded similar number of interactions as the correction of p-values by the Benjamini-Hochberg procedure (33) to control for multiple hypothesis testing. Operationally, the pairs of interacting sites with three independent PETs were defined as high-confidence interactions in the SMC1 ChIA-PET merged dataset and with two independent PETs in the individual SMC1 ChIA-PET replicates (8). The steps of the data processing are displayed on Fig. S3A at the RUNXI locus. The high confidence interactions identified by the Dowen et al. pipeline are found in Table S2B.

**Comparison of ChIA-PET analytical approaches**

An important component of the analytical methods to identify interactions in chromatin contact data is the estimation of a background contact frequency, and the comparison of the observed contact frequencies to the estimated background using statistical models. There are
currently multiple approaches to estimate the background contact frequency. One major approach is based on previous studies using chromosome conformation capture methodologies that suggested that the genome has polymer-like properties such that the interaction frequency between two genomic loci decreases as a function of their linear genomic distance (34, 35). Several computational pipelines have thus been developed to estimate background contact frequencies based on this “genomic proximity bias,” both for chromosome conformation capture data (35, 36) and ChIA-PET data (30, 37). At the same time, other ChIA-PET analytical methods estimate the background contact frequency based on the ligation frequency of the immunoprecipitated chromatin fragments using a hypergeometric test (8, 15, 32, 38, 39). Studies using ChIA-PET analytical methods that estimate background contact frequencies based on the “genomic proximity bias,” have mostly revealed long-range CTCF-CTCF associated interactions that are consistent with Hi-C data (24, 30), whereas studies using ChIA-PET analytical methods that estimate the background contact frequency based on the ligation frequency of the immunoprecipitated chromatin fragments using a hypergeometric test have also identified short-range enhancer-promoter interactions that do not involve CTCF (8, 15, 38, 39).

We have analyzed our ChIA-PET data using two analytical approaches: one that estimates the background contact frequency based on the genomic proximity bias (Mango pipeline) adapted from previous reports (24, 30, 40), and an analytical approach that estimates the background contact frequency based on the ligation frequency of the immunoprecipitated chromatin fragments using a hypergeometric test (Dowen et al. pipeline). The Dowen et al. pipeline was adapted from a previous report (8) and was recently developed based on multiple previous studies (15, 32, 39). We found that the two analytical methods yielded largely similar high confidence interactions (Fig. S1C, S3A, S4A, S4B, Table S2A-B). As expected, the median length of the high confidence interactions identified using the Mango pipeline (~263 kb) were greater than the median length of the high confidence interactions identified using the using the Dowen et al. pipeline (~144 kb), and the Dowen et al. pipeline detected ~3x more enhancer-promoter interactions than the Mango pipeline (Fig. S4C). Furthermore, the short range high-confidence interactions detected by the Dowen et al. pipeline but not by the Mango pipeline included interactions that were previously detected by locus-specific 3C and perturbation approaches, e.g. at the CD3 and GATA3 loci (41, 42) (Table S2A-B). Since the key model investigated in this study involves CTCF-CTCF interactions, we have used the interactions identified by the Mango pipeline for the main analyses (Table S2A), and also deposit the interactions identified by the Dowen et al. pipeline, because these include ~3x more enhancer-promoter interactions (Table S2B) that may be a valuable resource for the scientific community.

**ChIA-PET replicate comparison**

For the comparison of ChIA-PET replicates displayed on Fig. S2A, we first binned the genome into 50kb non-overlapping bins. All unique PETs from each replicate ChIA-PET dataset were placed into the bins, and the number of reads in each bin was counted. The scatter plot on Fig. S2A shows the values in each bin in each of the two replicate dataset. The values were then used to calculate a Pearson correlation coefficient ($r = 0.99$).

To compare the interactions identified in the replicates we also performed the following analysis. We ranked the interactions identified in one of the replicates according to their FDR. We then calculated the cumulative percentage of interactions in one replicate ChIA-PET dataset
that overlapped with an interaction in the other replicate ChIA-PET dataset (regardless of the FDR value of the interaction in the latter). This analysis revealed that virtually all (>99%) of interactions characterized by a low FDR value (e.g. FDR ≤ 0.2 “high confidence interactions”) overlapped with an interaction in the other replicate, suggesting high similarity between the ChIA-PET replicate datasets (Fig. S2B).

CTCF motif orientation analysis

The DNA binding site of CTCF is asymmetric, and a previous study has suggested that interactions between CTCF sites occur primarily between sites with motifs in the convergent orientation (36). Therefore, we investigated the motif orientation of the CTCF sites connected by ChIA-PET interactions in our dataset, and indeed found that the majority (~80%) of interacting CTCF-CTCF sites are in a convergent motif orientation (Fig. S2F, S3E) suggesting high quality of the ChIA-PET data.

For this analysis, FIMO was first used to identify the location and orientation of the CTCF motifs at CTCF ChIP-seq peaks at a default P-value threshold of $10^{-4}$ (43). In the analysis, the canonical CTCF motif from the JASPAR CORE vertebrate motif database (ID. MA0139.1) (44) was used. The information of CTCF motif orientation at CTCF ChIP-seq peaks was next overlaid with PET peaks at two ends of CTCF-CTCF/cohesin ChIA-PET interactions. For simplicity, we only used the CTCF ChIP-seq peaks having a single CTCF motif for the analysis, and only the CTCF-CTCF/cohesin ChIA-PET interactions were used whose ends overlapped with only a single CTCF ChIP-seq peak by at least 1 base-pair at each end. The pairs of CTCF motifs at the two ends of CTCF-CTCF/cohesin ChIA-PET interactions were then classified into one of the four possible classes of motif orientation: a convergent orientation (forward-reverse), a divergent orientation (reverse-forward), the same direction on the forward strand (forward-forward) or the same direction on the reverse strand (reverse-reverse) (Fig. S2F, S3E).

Visualization of ChIA-PET interactions on the WashU Genome Browser

ChIA-PET interactions were visualized on the WashU Genome Browser (45). The depth of the color of the interactions on the Browser snapshots reflects the following value: (1-FDR)+0.1. The (1-FDR) operation was performed so that the interactions characterized by the lowest significance score are displayed as the darkest arcs, and to improve visualization, a pseudo-count of 0.1 was added. In Fig. 1B, 4A, 4G, 4H only the high confidence interactions are displayed (FDR≤0.2).

Comparison of the Jurkat T-ALL ChIA-PET interactions to interactions described in other cell types

Interactions detected using Hi-C data in GM12878, HeLa, HMEC, HUVEC, IMR90, K562, KBM7 and NHEK were downloaded from a previous study (36). For the overlap analysis displayed In Fig. S2D, S3C, an adaptation of the approach described in (36) was used: interactions were scored as overlapping across two cell types if they had a reciprocal overlap of at least 80% of the length of the interaction. The high confidence interactions detected in Jurkat T-ALL cells, but not in the other 8 cell types are listed in Table S2C.
Hi-C data analysis

Previously published Hi-C datasets in H1 human ESCs (11) were downloaded from GEO (GSM1267196). The raw reads from these datasets were mapped to the human genome build hg19 and filtered as previously described (46). Corrected contact probability matrices at 40-kb resolution were obtained using the hiclib library (https://bitbucket.org/mirnylab/hiclib). The corrected contact probability matrices displayed on the heatmap in Fig. 1B were generated by the image function in R. The topologically associating domain (TAD) coordinates in H1 hESC used in Fig. 1B, Fig. S2E, S3D were downloaded from a previous study (11) and were converted to hg19 co-ordinates using the liftover tool on the UCSC Genome Browser.

Topologically Associating Domain (TADs) boundary overlap analysis

To test whether the cohesin ChIA-PET interactions were contained within TADs, a permutation test was conducted. The percentage of cohesin ChIA-PET interactions contained within TADs was calculated by overlapping the interactions with the H1 TAD boundaries using bedtools intersect. An interaction was considered contained within a TAD if it intersected only with one TAD. After this percentage was calculated, the positions of ChIA-PET interactions were randomly shuffled across a chromosome 10,000 times using bedtools shuffle with the –chrom option (to ensure that chromatin loops were only shuffled chromosome-wide to control for the rate of chromatin loops observed per chromosome). For each randomly shuffled permutation, the percentage of ChIA-PET interactions contained within only one TAD was re-estimated to generate a random background. The p-value was estimated by calculating the frequency that a randomly generated permutation had a higher fraction of ChIA-PET interactions contained within TADs than the fraction observed in the actual data. (Fig. S2E, S3D).

Identification of enhancers and super-enhancers

Enhancers and super-enhancers in Jurkat cells were identified using H3K27Ac ChIP-Seq data as previously described (17). Briefly, enhancers were defined as H3K27Ac ChIP-Seq peaks identified using MACS. To identify super-enhancers, the H3K27Ac ChIP-Seq peaks (i.e. enhancers) were stitched together if they were within 12.5 kb, and the stitched enhancers were ranked by their ChIP-Seq read signal for H3K27Ac, using the ROSE algorithm (https://bitbucket.org/young_computation/rose) (18). ROSE separates super-enhancers from typical enhancers by identifying an inflection point of H3K27ac signal vs. enhancer rank (17, 18).

Assignment of genes to enhancers

For the assignment of genes to enhancers, the typical enhancers and super-enhancers generated by ROSE were used (see above). Typical enhancers and super-enhancers were assigned to promoters in two ways. When available, the enhancer-promoter ChIA-PET interactions were used to assign enhancers to their target genes. In the absence of a ChIA-PET interaction, typical enhancers and super-enhancers were assigned to the nearest active gene.
Active genes were defined as having an at least 0.5 mean rpm/bp H3K27ac ChIP-Seq density in a window 500 bases up- and downstream of the TSS, as previously described (8, 17).

**Assignment of interactions to regulatory elements**

To identify the association of long-range chromatin interactions with different regulatory elements, we assigned the PET peaks of interactions to different regulatory elements, including enhancers (H3K27Ac ChIP-Seq peaks), promoters (+/- 2 kb of the Refseq TSS), and CTCF binding sites as previously described (8). For the analysis displayed in Fig. S1D, S3B, if an anchor site overlapped with multiple regulatory elements priority was assigned as: (1) promoters, (2) enhancers, (3) CTCF sites. A minimum of 1 base-pair overlap was required. These anchor classifications represent the nodes in Fig. S1D, S3B. Next the edges were calculated by counting the number of interactions between the classified PET anchors. Note that this analysis does not include CTCF sites that overlap either enhancers or promoters in the “CTCF” node of the plot. The total number of CTCF-CTCF interactions displayed in Fig. S4C includes interaction between any two CTCF-bound sites, regardless whether they overlap enhancers or promoters or not.

**Insulated neighborhoods**

Candidate insulated neighborhoods were defined as two CTCF ChIP-Seq binding sites that have an at least 1bp overlap each with two PET peaks connected by a cohesin ChIA-PET interaction (8). A gene was considered to be inside an insulated neighborhood, if its transcription start site (TSS) is located within the neighborhood boundaries. When multiple TSSs were annotated for the same gene, the TSS of the longest transcript was used for further analysis.

**Heatmap representation of ChIA-PET interactions in insulated neighborhoods**

Heatmap representation of ChIA-PET interactions in Fig. S2H was created by mapping high-confidence ChIA-PET interactions across insulated neighborhoods using a previously described method (8). We created three types of regions: upstream, the insulated neighborhood, and downstream. Upstream and downstream regions are 20% of the insulated neighborhood’s length each. The upstream and downstream regions were divided into 10 equally sized bins each, and insulated neighborhoods were length normalized by dividing them into 50 equally sized bins. To calculate interactions in each bin the interactions were filtered in two ways: (1) we required interactions to have at least one end in the interrogated region. This removed interactions that are anchored outside of our region of interest. (2) We removed interactions that represent nested interactions (i.e. where one CTCF anchor site of two interactions are identical). The density of the whole span of ChIA-PET interactions in each bin was next calculated in the units of number of interactions per bin. The density of ChIA-PET interactions was row-normalized to the row maximum for each domain.

**RNA isolation and RNA-Seq**
Jurkat RNA was isolated and sequenced as previously described (47). RNA-Seq reads were aligned to the hg19 (GRCh37) reference genome using Tophat2 (48) version 2.0.11, using Bowtie (28) version 2.2.1.0 and Samtools version 0.1.19.0. RPKMs for each Refseq transcript were calculated from aligned reads using RPKM_count.py from RSeQC (49). For a gene to be considered expressed the cutoff of >1 RPKM was used (50). For the analysis displayed in Fig. 2A, if multiple TSSs were annotated for the same gene, the RPKM value of the longest transcript was considered (this method of collapsing TSSs produced qualitatively identical results compared to using the RPKM value of the highest-expressed transcript).

**CRISPR/Cas9 mediated genome editing**

Genome editing was performed using CRISPR/Cas9 essentially as described (51). Briefly, target-specific oligonucleotides were cloned into a plasmid carrying a codon-optimized version of Cas9 and either an mCherry or GFP expression cassette. SgRNA sequences were cloned into the BbsI recognition sites as described (http://www.genome-engineering.org/crispr/). The genomic sequences complementary to guide RNAs are listed below. Around 500,000 HEK-293T cells were transfected with two plasmids expressing Cas9 and sgRNA targeting regions around 200 basepairs up- and down-stream of the center of the targeted CTCF site at the TAL1 locus, and 200 basepairs up- and down-stream of the first and third CTCF binding sites at the LMO2 locus, respectively. One of the two guide RNAs were cloned into the Cas9 expression vector containing the mCherry, and the other into the Cas9 expression vector containing the GFP expression cassette. Transfection was carried out with the Lipofectamine 2000 reagent (Invitrogen) according to the manufacturer’s instructions. For the LMO2 locus 1 µl of a 10 µM repair template (160 bp ultramer with the desired deletion junction) was included in the transfection. Two days after transfections, cells positive for mCherry and GFP were FACs sorted, and replated at clonal density. Individual colonies were picked, expanded, and genotyped by PCR, and the edited alleles were verified by Sanger sequencing. The cell lines used for the expression analysis in Fig. 3 that carry a deletion allele at the TAL1 locus are homozygous, and the cell lines that carry a deletion allele at the LMO2 locus are heterozygous for the modification.

Sg1_TAL1: ACATTTCATTATATGTTAA
Sg2_TAL1: ATACTAGTTAAGCTTTTCTT
Sg1_LMO2: AAACAGCATTGCCACCTGG
Sg2_LMO2: CCAGGTGGCAATGCTGGTTT

LMO2 Repair Template: AGC CCC ATA GTT GGT GCT CAA TAA ATG CTA GTA ATA TTT ACT TGT GGC TTA CTG GTT CCT CAA GAT TCC TTA AAA TCT GAT GGC ATC AGA AGA GAC TAT CTC ACT GTT ATC ATG ACA TGG ACA TCC CGT GCA TGC CTG TAT TTG AAC ACT TGT CTC ATT G

**CRISPR/Cas9-mediated genome editing in CD3+ T-cells**

After three days of activation, T-cells were nucleofected using the Lonza Nucleofector 2b (program U-014) and the Human T-cell Nucleofector Kit (VPA-1002, Lonza). Nucleofection conditions were the following: 100 µl nucleofection solution, 10^6 cells, 2x5µg synthetic sgRNA and 15 µg Cas9 mRNA. Cas9 mRNA was purchased from TriLink BioTechnologies (San Diego,
CA, USA). All synthetic sgRNAs contained three terminal nucleotides at both the 5′ and 3′ ends with 2′ O-Methyl 3′ phosphorothioate modifications. Paired synthetic sgRNAs targeting CCR5 (named ‘D’ and ‘Q’, referred to as control deletion on Fig. S8D) were previously reported (52). TAL1-targeting sgRNAs (#3 and #4) were synthesized by TriLink BioTechnologies (San Diego, CA, USA). Paired synthetic sgRNAs were designed to flank the CTCF motif of the TAL1 insulated neighborhood boundary (Fig. S8C).

sgRNA sequences:

sgRNA #3:

5′ 2′OMe(A(ps)C(ps)A(ps)) UUU CAA UUA UAU GUU AAG UUU UAG AGC UAG AAA UAG CAA GUU AAA AUA AGG CUA GUC CGU UAU CAA CUU GAA AAA GUG GCA CCG AGU CGG UGC 2′OMe(U(ps)U(ps)U(ps)) U 3′

sgRNA #4:

5′ 2′OMe(A(ps)U(ps)A(ps)) CUA GUU AAG CUU UUC CUG UUU UAG AGC UAG AAA UAG CAA GUU AAA AUA AGG CUA GUC CGU UAU CAA CUU GAA AAA GUG GCA CCG AGU CGG UGC 2′OMe(U(ps)U(ps)U(ps)) U 3′

Abbreviations:

2′OMe: 2′ O-Methyl
ps: phosphorothioate

Target-specific complementary nucleotides are underlined. Editing was confirmed by PCR amplification of the targeted region (Fig. S8E).

RNA isolation and quantitative RT-PCR

Gene expression experiments displayed in Fig. 3C and Fig. 3I were performed on wild type and CTCF-site deleted clonal lines of HEK-293T cells.

For the expression analysis in primary human cells (Fig. S8D), T-cells were isolated from 9 independent donors, and transfected in two independent reactions per donor per time point. T-cells isolated from three donors were assayed for gene expression 4 days and 7 days post-transfection, T-cells isolated from three donors were assayed for gene expression 7 days post-transfection, and T-cells isolated from three donors were assayed for gene expression 10 days post-transfection, resulting in a total of 24 gene expression experiments. The PCR-based genotyping for three representative transfections is displayed on Fig. S8E.

RNA was isolated using the RNeasy Plus purification kit (Qiagen), and reverse transcribed using oligo-dT primers and SuperScript III reverse transcriptase (Invitrogen) according to the respective manufacturer’s instructions. Quantitative real-time PCR was performed on a 7000 AB Detection System using the following Taqman probes, according to the manufacturer’s instructions (Applied Biosystems):

\[
\text{GAPDH: hs02758991_g1}
\]
\[
\text{TAL1: hs01097987_m1}
\]
\[
\text{LMO2: hs00277106_m1}
\]
For the expression analysis in primary human T-cells, the differences in the Ct-values of *TAL1* and *GAPDH* are plotted on Fig. S8D. Samples in which no *TAL1* expression was detectable were excluded from the analysis, and only the samples where *TAL1* was above the detection limit are included on Fig. S8D, (9/24 mock samples, 14/24 control deletion samples, and 23/24 TAL1 neighborhood boundary -deletion samples).

Single molecule mRNA fluorescence in situ hybridization (FISH)

RNA FISH was performed as previously described (53). CRISPR/Cas9-edited primary human T-cells or Jurkat T-ALL cells were washed with PBS and subsequently fixed with paraformaldehyde (PFA) at a final concentration of 4%. The cells were incubated in 4% PFA for 10 minutes while rotating to avoid clumping of cells. After 10 minutes the cells were spun down for 2 min at 1000rpm. To permeabilize the cells, the cells were placed in 70% ethanol overnight. Cells of three independent transfections were pooled for further analysis. Cells from three independent transfections (described above) were pooled for the subsequent steps. The fixed cells were attached to chambered cover slides (Nunc Lab-Tek) coated with 0.1% poly-l-lysine (Sigma) prior to imaging. 20nt probes for *RUNX1* and *TAL1* were manually designed (sequences available upon request) and ordered through Biosearch Technologies, coupled with either A594-fluorophore (Invitrogen) or TMR-fluorophore (Invitrogen), respectively, and hybridized with standard FISH hybridization buffer containing 40% formamide. For hybridization conditions 75ng probes per µl of hybridization buffer were used. The probes were hybridized for 16 h at 30°C followed by two wash steps with wash buffer containing 40% formamide and 2x SSC. The cells were counterstained with Hoechst 33342. During imaging the cells were kept in a solution containing PBS, Glucose, Catalase and Trolox to avoid bleaching of fluorophores. All images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a 100X oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). Cells that seemed fragmented or had an excessive amount of background were excluded from quantification. Only transcripts that colocalized with the Hoescht signal were counted. Cells with more than 1 transcript for either *RUNX1* or *TAL1* were counted as positive for the respective transcript.

Chromosome conformation capture carbon copy (5C)

For each 5C library, 50 million cells were resuspended in 40 mL DMEM (high glucose, pyruvate; Invitrogen, 11995-073) followed by the addition of 4 mL 11% formaldehyde solution (11% formaldehyde, 50 mM HEPES pH 7.3, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0). Cells were cross-linked for 10 minutes with gentle agitation every 2 minutes. Cross-linking was quenched by the addition of 2 ml 2.5M glycine followed by two washes with ice cold PBS. Crosslinked cells were snap-frozen in liquid nitrogen and stored at -80°C until further processing.

5C experimental design

5C was carried out as previously described (54, 55). We investigated two 2Mb regions centered on the *TAL1* and *LMO2* loci. The *TAL1* 2Mb region is located on Chromosome 1 (hg19 chr1: 46740122-48740121) and the *LMO2* 2Mb region on Chromosome
11 (hg19 chr11: 33003550-35003549). Libraries were generated for three HEK-293T cell lines: wild type (wt), TAL1-deletion (TAL-ΔCTCF), and LMO2-deletion (LMO2-ΔCTCF) with two biological replicates for each line. Replicate data were pooled for processing.

**5C primer design**

5C primers were designed at HindIII restriction sites (AAGCTT) using 5C primer design tools previously developed and made publicly available online at the My5C website (http://my5C.umassmed.edu) (56). Primers were designed according to a new “double alternating scheme” (Fig. S10A). In this design, each restriction fragment has two primers, one primer designed on the 5’ end of the restriction fragment, and one primer designed on the 3’ end of the restriction fragment. If one fragment has a right (5’) forward (FOR) and a left (3’) reverse (LREV), the adjacent restriction fragment will have a left (3’) forward (LFOR) and a right (5’) reverse (REV) and so on (Figure S1). This design allows interrogation of all fragment-fragment interactions throughout the regions (with up to two independent interactions per pair of fragments), whereas the previously developed “alternating design” (54) allowed interrogation of interactions between even-numbered and odd-numbered fragments only. Primers settings were: U- BLAST, 3; S- BLAST, 50; 15-MER, 800; MIN_FSIZE, 100; MAX_FSIZE, 50,000; OPT_TM, 65; OPT_PSIZE, 40. The 5C primer tails uses were: (FOR/LFOR) T7 sequence: 5’-TAATACGACTCACTATAGCC-3’; (REV/LREV) T3 sequence 5’- TCCCTTTAG TGAGGGTTAATA-3’. The length of the forward primers was 60 bp and the length of the reverse primers was 61 bp. For the TAL1 locus, we designed 270 forward 5C primers and 285 reverse 5C primers for a possible of 76,950 interactions. For the LMO2 locus, we designed 367 forward 5C primers and 347 reverse 5C primers for a possible of 127,349 interactions. All primer information can be found in Table S6.

**Generation of 5C libraries**

3C was performed with HindIII restriction enzyme as previously described with some minor modifications (34, 57). Briefly, after digestion the restriction enzyme was inactivated 15 min at 65°C and immediately placed on ice. Ligation was performed in a final volume of 1,200 µL with 10 U of T4 ligase (Invitrogen) and 1% of Triton X-100. For 50 million cells, 10 ligation reactions were performed, and pooled after DNA purification. The 3C libraries were then interrogated by 5C (54, 58). 5C was performed as described (55) with the following changes. The multiplex annealing reaction was performed overnight at 50°C. Pairs of annealed 5C primers were ligated at the same temperature using Taq DNA ligase for 1 h. 7 independent ligation reactions were performed for each 5C library, each containing an amount of 3C template that represents 600,000 genome equivalents and 0.3 fmol of each primer. Ligated 5C primer pairs, which represent a specific ligation junction in the 3C library and thus a long-range interaction between the two corresponding loci, were then amplified using 20 cycles of PCR with T7 and T3R universal tail primers that recognize the common tails of the 5C forward and reverse primers. Four separate amplification reactions were carried out for each of 7 annealing reactions described above and all the PCR products were pooled together. This pool constitutes the 5C library. The libraries were concentrated using Amicon Ultra Centrifugal filters - 0.5ml 30K (Millipore) and purified with Qiaquick PCR purification kit. Index adaptors (Illumina) were ligated to the 5C library using the Illumina protocol (TruSeq Nano
DNA sample Prep kit). The linkered 5C libraries were then amplified by PCR (6 cycles), using the Illumina PCR mix. The 5C libraries were gel purified and sequenced on either the Illumina Miseq or the Illumina HiSeq 2000 platform, generating 50-bp paired-end reads. The indexed adapter was AD005 for WT, AD006 for TAL1 deletion and AD015 for LMO2 deletion.

5C read mapping

Sequencing data were obtained from both an Illumina MiSeq and an Illumina HiSeq 2000 machine and was processed by a custom pipeline to map and assemble 5C interactions, as previously described (55, 56). We used an updated version of the Novoalign mapping algorithm (V3.02.00). Data from the two biological replicates were pooled, producing a single interaction map for the wild type (wt), TAL1-deletion (TAL-ΔCTCF), and LMO2-deletion (LMO2-ΔCTCF) samples. The summary statistics and the read depth of each 5C libraries can be found in Table S7.

5C filtering and analysis

5C matrices were filtered using previously described methods (55, 56). To summarize, first we removed the diagonal from all matrices, which represent self-circularization of a restriction fragment. Second, we detected and flagged all outlier (singleton) pixel/interactions that are defined by a Z-score greater than 21 in each dataset. We then took the union of all outlier (singleton) pixel/interactions across the 3 5C matrices, and removed these pixels/interactions from all 3 datasets. Third, we detected and flagged all outlier (anchor) row/cols that are defined as having an aggregate (row/col) signal greater than or less than 1.5 * IQR (of the distribution of all row/col signals). We then took the union of all flagged (anchor) row/col outliers across the 3 5C matrices, and removed these (anchor) row/cols from all 3 datasets, excluding the regions that overlapped our LMO2 and TAL1 deletions in our LMO2 and TAL1 deletion matrices respectively. Fourth, the matrices were balanced according to the ICE method developed for Hi-C (46). Fifth, since each restriction fragment has two primers (one on the 3’ end and one on the 5’ end), an interactive between any two fragments can be represented by 2 possible primer ligations. We chose to calculate the mean of these two possible interactions thus collapsing the interaction map down to a fragment x fragment interaction map (as compared to a primer x primer interaction map). Sixth, to properly scale the matrix/heatmaps relative to genomic coordinates, the matrices were binned using a binsize=20 kb, binstep=2.5kb, binmode=median.

5C distance correction (deletions)

To rule out the possibility that the increased 5C interaction signal observed in the deletion samples was an artifact due to the fact that the regions are now closer in linear distance, we performed a distance-adjustment prior to the z-score / z-score difference calculation(s). For the LMO2-deletion 5C map and the TAL1-deletion 5C map, we first removed the primers contained in the deleted regions. We then adjusted for the 345bp and 26,628bp differences in the TAL1 and LMO2 regions respectively by shifting the coordinates of each restriction fragment downstream of the deleted region. We then transformed the interaction counts into a distance normalized z-score, using LOWESS to estimated signal per distance as previously described.
Next we merged primer interactions from the same fragments as described above in step 5, and binned the data using a binsize=20 kb, binstep=2.5kb, binmode=median. Finally, we calculated the difference of the z-score matrices to quantify the amount of the differential interactions observed between the WT, control and deletion datasets.

The plots without distance correction (“uncorrected”), and the plots generated with distance-adjustment (“distance adjusted”) are displayed on Fig. S10C and Fig. S10E, respectively. Fig. 3F and Fig 3L show the distance adjusted plots.

**Luciferase reporter assay**

Luciferase reporter assays were performed as previously described (51) with modifications. The candidate enhancer regions (~600bp) around the STIL and CMPK1 promoters were cloned into a pGL3 (Promega) reporter vector (BamHI-Sall sites) that contains a Firefly luciferase gene driven by a minimal c-MYC promoter (18). The candidate enhancer region around STIL was PCR-amplified using the following primer sequences (5’-3’): ATGTTACCCACCAACCTTCCC and AACTGTTTCTTCCGTTGCCG. The candidate enhancer region around CMPK1 was PCR-amplified using the following primer sequences (5’-3’): GATTCTCCTCTGCTCTCCACG and AAGACACGTTCGCTGACAGTG. HEK-293T genomic DNA was used as template DNA in the PCR reactions.

1*10⁵ HEK-293T cells were transfected with 490ng of the reporters using Lipofectamine 2000 (Invitrogen). 10ng of a Renilla luciferase control plasmid (pRL-SV40; Promega) was co-transfected as a normalization control. After 24 hours of incubation luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega). All luciferase reporter assays were performed in triplicates. Luciferase activity was normalized to the activity measured in cells transfected with a construct containing only the promoter (empty vector) (Fig. S8B).

2*10⁵ Jurkat T-ALL cells were transfected with 475ng of the reporters using MOLT4 Avalanche transfection reagent (EZ Biosystems). 25ng of a Renilla luciferase control plasmid (pRL-SV40; Promega) was co-transfected as a normalization control. After 40 hours of incubation luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega). All luciferase reporter assays were performed in triplicates. Luciferase activity was normalized to the activity measured in cells transfected with a construct containing only the promoter (empty vector) (Fig. S8B).

**T-ALL deletion catalog and overlap analysis**

Deletions in T-ALL genomes were compiled from multiple studies (22, 26, 59, 60). We filtered for relatively short deletions (<500 kb, around half the size of an average TAD (11)) in order to minimize deletions that affect multiple genes. The overlap with insulated neighborhood boundaries was analyzed as follows. A neighborhood boundary CTCF site was scored as overlapping a deletion, if the boundary site (i.e. the PET peak) overlapped at least one deletion by 1bp. A deletion was scored as overlapping a neighborhood (i.e. the PET peak) boundary if it overlapped a boundary site by at least 1bp. The deletion co-ordinates (hg19/GRCh37) and the source study are listed in Table S5.
To estimate significance of the overlap we determined whether the observed frequency of overlap is higher than expected (Fig. S6B). For this analysis, we did random shuffling of the 438 T-ALL deletions 10,000 times across the genome and calculated the overlap of deletions with the boundaries of 1) all insulated neighborhood and 2) insulated neighborhoods containing T-ALL Pathogenesis Genes. The P-value of each permutation test was estimated by calculating the fraction of 10,000 permutations that had a higher number of overlap with neighborhood boundaries than in the actual data.

**T-ALL Pathogenesis Genes**

To identify a set of genes whose mutations have been causally linked to T-ALL, we manually curated a list of genes using the Cancer Gene Census and individual studies. First, we downloaded the Cancer Gene Census on 2015.04.01 from www.cancer.sanger.ac.uk/cosmic. The complete Gene Census was filtered for genes that had “T-ALL” annotated in the “Tumor type” columns of the Gene Census. We added to the list the genes that were described as recurrently altered in T-ALL in (14). We then converted gene symbols into Refseq IDs for further analysis using the table described in the RNA-seq section. This resulted in a manually curated list of 55 genes (Table S3).

**T-ALL clinical expression data**

T-ALL clinical expression data (Fig. S7B, S9B) were downloaded from (22). The array probe IDs used are: 206283 PM.s_at (TAL1) and 204249 PM.s_at (LMO2). To estimate the significance of association between the chromosomal deletions that overlap the boundaries of the TAL1 or LMO2 insulated neighborhoods and elevated expression of those genes a Fisher’s exact test was performed (Fig. S7B, S9B). This test revealed a significant association at TAL1 (P-value=0.003276) but not at LMO2. We note that this analysis is likely limited by the data in the patient cohort, and that both TAL1 and LMO2 can be activated by complex chromosomal rearrangements not only deletions (13, 14). Previous studies have established an association with chromosomal deletions upstream of LMO2 and LMO2 activation in T-ALL (13, 14).

**Constitutive interactions across three cell types**

First, CTCF binding sites, and cohesin binding sites were identified in Jurkat, GM12878 and K562 cells (datasets listed in Table S13). Cohesin ChIA-PET in the three cell types were processed with the Mango pipeline as described above, and two CTCF bound sites that are connected by a cohesin ChIA-PET interaction were annotated as CTCF-CTCF/cohesin interactions in each cell type (i.e. candidate insulated neighborhoods). For the overlap analysis displayed in Fig. S11A-B, binding peaks in the respective datasets were considered shared if they overlapped by at least 1bp. In Fig. S11C, the CTCF-CTCF/cohesin interactions were scored as constitutive across two cell types if they had a reciprocal overlap of at least 80% of the length of the interaction. The ChIA-PET datasets are likely not saturated, suggesting that not every interaction found within a cell will be potentially represented in the dataset. Therefore, we defined candidate constitutive CTCF-CTCF/cohesin interactions as the set of CTCF-CTCF/cohesin interactions that were found overlapping in at least two of the three cell types.
This resulted in 10,624 constitutive CTCF-CTCF loops (i.e. “constitutive neighborhoods”) (Fig. S11C, Table S8).

ICGC mutations analysis


To compare mutation frequencies at CTCF binding sites in the genome, we defined the set of constitutive neighborhood boundary CTCF sites, and two sets of control CTCF sites (“non-boundary CTCF sites”). To define constitutive neighborhood boundary CTCF sites, we used the ChIP-Seq peaks that form the boundaries of the constitutive neighborhoods described above. This resulted in a set of 16,637 CTCF binding sites. We then defined two sets of non-boundary CTCF sites. Ideally, an appropriate control set would include CTCF sites that do not, or are less likely to be involved in looping interactions than the boundary sites of constitutive neighborhoods, and we used two known features of CTCF-CTCF/cohesin looping interactions to define such control sets. Previous studies have established that CTCF is bound at the anchor sites of chromosome loop structures, but not every CTCF-bound site in the genome is involved in looping interactions, and the CTCF-bound sites that are detected at the boundaries at looping interactions tend to be co-bound by cohesin (8, 9, 11, 24, 40, 42, 63-66). Based on this, we first defined a set of non-boundary CTCF sites (“control set 1”) the following way: we identified the set of interactions in Jurkat, GM12878 and K562 cohesin ChIA-PET data that had an FDR>0.9, which represent interactions of very low statistical significance. We then identified a set of low-confidence interactions that occur in all three cell types (with a reciprocal overlap of at least 80% of the length of the interaction). Only the interactions that overlapped a CTCF ChIP-Seq peak at both ends were used for the analysis. The CTCF ChIP-Seq peaks at the anchors of these interactions were defined as “non-boundary CTCF sites (control set 1)”. The control set 1 contains 44,461 CTCF binding sites. This set of CTCF sites thus represents a set of sites that appear to be involved in looping interactions at a much lower statistical significance than the constitutive neighborhood boundary CTCF sites (the latter being defined using the high-confidence ChIA-PET interactions).

To define a second control set of non-boundary CTCF sites (“control set 2”), we used previous evidence that the CTCF-bound sites that are detected at the boundaries at looping interactions tend to be co-bound by cohesin (8, 9, 11, 24, 40, 42, 63-66). The corollary of this is that CTCF-sites that are not bound by cohesin are less likely to be involved in looping interactions. Therefore, we first identified the CTCF ChIP-Seq peaks in Jurkat, GM12878 and K562 cells, and the cohesin binding sites in Jurkat, GM12878 and K562 cells, and created a union of CTCF binding sites and a union of cohesin binding sites in the three cell types. We then
identified the set of CTCF binding sites in the union that did not overlap a cohesin (SMC1 in Jurkat; RAD21 in K562 and GM12878) binding site in the union of cohesin binding sites, and these were defined as “non-boundary CTCF sites (control set 2)”. The control set 2 contains 4,699 CTCF binding sites. This set of CTCF sites thus represents a set of sites that are not bound by cohesin in any of the three cell types. Analysis of somatic mutations in constitutive neighborhood boundary CTCF sites, and the two sets non-boundary CTCF sites (control set 1 and 2) revealed a striking enrichment of mutations in constitutive neighborhood boundary CTCF sites, and a much more moderate enrichment of mutations in the non-boundary CTCF sites (control set 1), and no enrichment of mutations in the non-boundary CTCF sites (control set 2) (see analytical details below, and Fig. 4B, Fig. S12D, S14). We note that the moderate enrichment of mutations in the non-boundary CTCF sites (control set 1) is likely caused by the presence of CTCF sites in this control set that may be involved in looping interactions. This is supported by the finding that a significant fraction of interactions of very low statistical significance (FDR>0.9) that were used to define this set, connect genomic features such as enhancers, promoters and CTCF sites, therefore potentially represent bona fide looping interactions (Fig. S2G).

CTCF motifs under the CTCF binding sites were identified as follows. We used a position weight matrix (PWM) for CTCF from JASPAR CORE vertebrates 2014 database (MA139.1) and the Biostrings Bioconductor R package (Pages H, Aboyoun P, Gentleman R and DebRoy S. Biostrings: String objects representing biological sequences, and matching algorithms. R package version 2.38.1.) using the matchPWM function with a min.score parameter of 80%. Only the CTCF sites that had at least one CTCF motif detected within 100 bp of the MACS peak call summit were used for further analysis. If multiple motifs were detected at the same CTCF peak, then only the strongest one was used based on the motif score assigned by matchPWM. As an alternative method, we also identified CTCF motifs by an independent approach. For this analysis, we used the CTCF motif PWM identified using HT-SELEX from a previous study (67) (Fig. S14B). Overall, both approaches yielded similar results (Fig. S14). To rule out that the enrichment of mutations was not dependent on particular motif calling tool, we also identified CTCF motifs within the CTCF peaks using FIMO (43) with default parameters and the same JASPAR CTCF motif (Fig. S14C). Overall, both motif identification tools yielded similar results (Fig. S14).

To determine the relative enrichment of somatic mutations at constitutive neighborhood boundary CTCF sites, we adapted an approach used in a previous study (23); we counted the number of somatic mutations +/-5bp around the CTCF motif identified under the constitutive neighborhood boundary CTCF sites and the two non-boundary CTCF sites control sets. The mutation count was normalized by the number of CTCF peaks that contained at least one CTCF motif within each set. The mutations identified in the constitutive neighborhood boundary sites are listed in Table S10.

To assess the significance of the observed enrichment of mutations at CTCF sites, we performed permutation tests. We randomly permuted the mutations detected within +/-1 kb around the CTCF binding motif of the CTCF sites, and scored the frequency at which the number of mutations around (+/-5bp) the CTCF motifs was greater than or equal to the number of mutations observed around (+/-5bp) the CTCF motifs in the experimental data. The permutation was performed 10,000 times (Fig. 4B, Fig. S12A-B, S12D, S14, S15).
To calculate the frequency of recurrent mutations in constitutive neighborhood boundary CTCF sites and non-boundary CTCF sites, we counted the number of recurrent somatic mutations +/-5bp around the CTCF motif identified under the constitutive neighborhood boundary CTCF sites and the two non-boundary CTCF site control sets. Recurrent mutations were defined as the same base substitution that occurs in at least two patient samples of the same Genome Project (or Pan-cancer set). The mutation count was normalized by the number of CTCF peaks that contained at least one CTCF motif within each set (Fig. S16).

To calculate the enrichment of mutations in the constitutive neighborhood boundary CTCF sites versus non-boundary CTCF sites, we calculated the ratio of mutations that occur in the constitutive neighborhood boundary CTCF sites and the mutations that occur in each non-boundary CTCF site control set (described above). This was performed using the mutations annotated in individual cancer Genome Projects, and a “pan-cancer set” created by aggregating all the mutations seen across all the analyzed Genome Projects in the ICGC datasets. Only Genome Projects that contained at least 5 mutations for each of the three CTCF site sets were included in this analysis (Fig. S12E). We then estimated the 95% confidence interval using a bootstrap procedure. For each Genome Project, we recalculated the enrichment ratio by resampling the number of mutations that fell within the CTCF motif and +/-1kb window around it in CTCF site set, and then recalculated the enrichment ratio as described above. Resampling was performed over 1,000 iterations. From this distribution of enrichment ratios, we took the .025 and .975 quantiles as the low and high ends of the confidence interval (Fig. S12E).

To calculate the enrichment of mutations in transcription factor binding sites in liver cancer (Fig. S15), we first identified the ChIP-Seq binding peaks of the transcription factors (CTCF, MAX, MYC, FOXA1, FOSL2, JUND, NR2F2) in a liver cancer cell line (HepG2) using publicly available ChIP-Seq data (Table S13). We used a position weight matrix (PWM) for each transcription factor from the JASPAR CORE vertebrates 2014 database and use the Biostrings Bioconductor R package as described above to identify the binding motifs under the ChIP-Seq peaks. PWMs used: CTCF: MA0139.1, MAX: MA0058.2, cMYC: MA0147.2, FOXA1: MA0148.3, FOSL2: MA0478.1, JUND: MA0491.1, and NR2F1/2: MA0017.1. Only the transcription factor ChIP-Seq peaks that contained at least one motif detected within 100 bp of the MACS peak call summit were used for further analysis. If multiple motifs were called at the same peak, then only the strongest was used based on the motif score assigned by Bioconductor.

To compare the distribution of somatic mutations in constitutive neighborhood boundary CTCF sites and coding regions we counted the number of mutations in the constitutive neighborhood boundary CTCF sites and coding regions, respectively; and normalized the counts to the size of genome covered by constitutive neighborhood boundary CTCF sites and coding regions, respectively (Fig. S13).

To investigate the distribution of HapMap SNPs in constitutive neighborhood boundary CTCF sites, we downloaded data from the GATK resource bundle that included the 1000 Genomes Project data (ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle). Simple and indel SNPs were merged into one set for further analysis. The SNPs were then plotted within +/-1 kb around the CTCF binding motif of the constitutive neighborhood boundary CTCF sites (Fig. S12A).

Estimating differences in CTCF motif strength caused by somatic mutations
To estimate the effect of the observed somatic mutations in the ICGC data on CTCF binding, we calculated the score based on the position weight matrix (PWM) of the CTCF motif consisting of the reference genome sequence before and after inserting the somatic mutation into the motif sequence (Fig. S12C). The motifs in constitutive neighborhood boundary sites were included for this analysis if they contained at least one somatic mutation in the “pan cancer” dataset. The PWM scores were estimated using the \textit{PWMscoreStartingAt} function from the R Bioconductor \textit{Biostrings} package. The CTCF consensus motif used was from the JASPAR motif database (MA0139.1).

Notes on the data types in the ICGC dataset

The individual Genome Projects in the ICGC database contain varying numbers of whole genome (WGS) and whole exome (WES) sequence datasets (Table S9). While we used the total number of mutations annotated in each Genome Project regardless of whether the mutation originated from WGS and WES data, we note that the WES data alone potentially limits our ability to observe enrichment of mutations of constitutive neighborhood boundary CTCF sites, because the vast majority of these are intergenic (Table S2), and thus are not captured in WES data. Therefore, the enrichment of such mutations in Genome Projects that predominantly contain WES data is considered an estimate of the genomic distribution that will be further refined in the future as more WGS data is collected in ICGC. We note that the two Genome Projects in which we observed the highest level of enrichment for mutations in constitutive neighborhood boundary CTCF sites contain >100 WGS data each (ESAD-UK and LIRI-JP). Furthermore, a recent study that used >200 WGS data in colorectal cancer genomes has also reported frequent somatic mutations in CTCF bound sites (23).

Gene Ontology (GO) -term enrichment analysis

To characterize the genes located in constitutive neighborhoods whose boundaries are most frequently mutated we performed Gene Ontology (GO) -term enrichment analysis as follows. We identified all protein-coding genes within the GENCODE version 19 annotation, and the genes whose TSS +/-500bp was contained within the constitutive neighborhoods were identified. We then filtered for genes found within constitutive neighborhoods whose boundary CTCF sites contained at least 3 mutations This step was performed using 1) the complete ICGC dataset (i.e. Pan-cancer), 2) mutations in the ESAD-UK Genome Project, and 3) mutations in the LIRI-JP Genome Project. Genes that appeared multiple times were listed only once. We then estimated the enrichment of GO terms (biological process, complex component or molecular function) for this set of genes using the Bioconductor topGO package. We filtered out all GO terms with less than 10 genes assigned to that term, and we used the Fisher test to estimate the p-value per GO-term. To correct for multiple hypothesis testing, we converted all the p-values into q-values via pFDR (68) using the Bioconductor qvalue package. The list of GO-terms and q-values are listed in Table S11A-C (Pan-cancer), Table S11D-F (ESAD-UK), and Table S11G-I (LIRI-JP).

Candidate proto-oncogenes

Candidate proto-oncogenes were identified as follows. We first downloaded the genes listed in the Cancer Gene Census on 2015.04.01 from \url{www.cancer.sanger.ac.uk/cosmic}. This list
contains the genes whose mutations have been causally linked to cancer (i.e. both candidate proto-oncogenes and tumor suppressor genes). Proto-oncogenes are generally activated by mutations that result in a dominant phenotype and tumor suppressor genes are de-activated by mutations that have a recessive phenotype \(69\), so we filtered for the genes whose mutations are annotated as dominant in the Cancer Gene Census. This resulted in 329 candidate proto-oncogenes \(\text{Table S12}\).

For the example proto-oncogenes whose neighborhood boundary is disrupted by recurrent somatic mutations \(\text{Fig. 4G, 4H}\) the following evidence indicated activation of the genes in the cancer types the mutations were documented in: \(FGFRI\) was found overexpressed in liver carcinoma in previous studies \(70, 71\). \(LMO1\) was found overexpressed in esophageal adenocarcinoma in the ESCA-US clinical cohort \(7.15\)-fold over corresponding normal tissue; \url{http://firebrowse.org/viewGene.html?gene=lmo1}\).

**Abbreviations**

Abbreviations in \text{Fig. 1A} include: NSCLC: non-small cell lung cancer; CML: chronic myelogenous leukemia; AML: acute myeloid leukemia; T-ALL: T-cell acute lymphoblastic leukemia; GBM: glioblastoma multiforme; SCLC: small cell lung cancer

**Accession numbers**

Datasets generated in this study have been deposited in the Gene Expression Omnibus under the Accession number GSE68978. The GEO Accession numbers of the datasets used in this study are listed in \text{Table S13}.
**Supplementary Figures**

(A) Model of the hierarchical organization of chromosome structures.

(B) The majority of cohesin-bound sites are co-bound by CTCF or H3K27Ac-marked enhancers in Jurkat cells. Displayed is a heatmap representation of ChIP-seq data for SMC1 (cohesin), MYB, RUNX1, GATA3, TAL1, RNAPII, H3K27Ac and CTCF at 79,976 SMC1-bound sites in Jurkat cells. The regions are centered on the summit of the binding peak, and read density is displayed within a 10kb window. Color scale intensities are shown below the heatmaps in rpm/bp units.

**Fig S1. Cohesin ChIA-PET processing and analysis using the Mango pipeline**

(A) Model of the hierarchical organization of chromosome structures.

(B) The majority of cohesin-bound sites are co-bound by CTCF or H3K27Ac-marked enhancers in Jurkat cells. Displayed is a heatmap representation of ChIP-seq data for SMC1 (cohesin), MYB, RUNX1, GATA3, TAL1, RNAPII, H3K27Ac and CTCF at 79,976 SMC1-bound sites in Jurkat cells. The regions are centered on the summit of the binding peak, and read density is displayed within a 10kb window. Color scale intensities are shown below the heatmaps in rpm/bp units.
Overview of the cohesin ChIA-PET data analysis at the RUNX1 locus using the Mango pipeline. The algorithm used to identify paired-end tags (PETs) is described in detail in the Materials and Methods section. PETs and interactions involving enhancers, promoters and CTCF-bound sites within the window are displayed at each step in the analysis pipeline: unique PETs, PET peaks, interactions between PET peaks, and high confidence interactions (FDR ≤ 0.2).

Summary of the major classes of interactions identified in the cohesin ChIA-PET data. Enhancers, promoters, and CTCF sites between which interactions occurred are displayed as blue circles, and the size of the circle is proportional to the number of regions. The interactions between two sites are displayed as gray lines, and the thickness of the gray line is proportional to the number of interactions. Note that in this analysis the CTCF sites displayed include only the non-enhancer, non-promoter CTCF sites.
**Fig. S2. Cohesin ChIA-PET interactions**

(A) The ChIA-PET replicate datasets display high correlation. Scatter plot of the number of uniquely mapped PETs per 50kb bins of the genome in each replicate dataset.

(B) (left) Cumulative percentage of interactions in the Replicate 2 ChIA-PET dataset that are overlapped by an interaction in the Replicate 1 ChIA-PET dataset (regardless of the FDR value of the interaction in Replicate 1). The interactions in Replicate 2 were ranked by their FDR values on the x-axis. (right) Cumulative percentage of interactions in the Replicate 1 ChIA-PET dataset that are overlapped by an interaction in the other cell type. The interactions in Replicate 2 were ranked by their FDR values on the x-axis. The dashed lines indicate that the set of interactions that have an FDR of less than or equal to the displayed values are found to the left of the line, and serve to aid visual orientation.

(C) Calculation of the frequency of self-circularization among the PETs, and estimation of a minimum distance cutoff for the PETs used in the downstream analyses. Plotted is the log2 ratio of PETs whose mate pairs mapped to same DNA strand versus PETs whose mate pairs mapped to opposite DNA strands, and the log2 ratio is plotted at various PET distance cutoffs. When fragments self-circularize during the ChIA-PET processing, the mate pairs map to the opposite DNA strand, and these do not represent PETs that arise from bona fide chromatin interactions (detailed in the Materials and Methods). The minimum PET distance cutoff is estimated as the distance above which the ratio is ~1. For this analysis, the merged ChIA-PET dataset was used (Rep. 1 and 2).
(D) Number of interactions in Jurkat cells that are overlapped by an interaction identified using Hi-C in the indicated eight human cell lines. The high confidence Jurkat interactions (FDR≤0.2) were used for this analysis.

(E) Percentage of interactions that cross or do not cross TAD boundaries (defined in H1 human ESCs). The vast majority of interactions do not cross TAD boundaries (P-value<10-3, permutation test).

(F) The orientation of CTCF motifs at pairs of CTCF sites connected by cohesin ChIA-PET interaction is mostly convergent. The high confidence Jurkat interactions (FDR≤0.2) that are overlapped by CTCF-bound sites on both ends were used for this analysis.

(G) Number of interactions that overlap genomic features (enhancer, or promoter or CTCF-bound site) on both ends, at different FDR cutoffs of the interactions. At FDR≤0.2, 94% of ChIA-PET interactions overlap genomic features on both ends.

(H) Heat map of the density of ChIA-PET interactions around the 9,038 CTCF-CTCF interactions. The CTCF-CTCF interactions were length normalized.
Fig. S3. Cohesin ChIA-PET processing and analysis using the Dowen at al. pipeline

(A) Cohesin ChIA-PET processing and analysis at the RUNX1 locus using the Dowen et al. pipeline. The algorithm used to identify paired-end tags (PETs) is described in detail in the Materials and Methods section (“ChIA-PET data processing and analysis using the Dowen et al. pipeline” section). PETs and interactions involving enhancers, promoters and CTCF-bound sites within the window are displayed at each step in the analysis pipeline: unique PETs, PET peaks, interactions between PET peaks supported by at least three independent PETs and with a false positive likelihood of <1%.

(B) Summary of the major classes of interactions identified in the cohesin ChIA-PET data using the Dowen et al. pipeline. Enhancers, promoters, and CTCF sites where interactions occur are displayed as blue circles, and the size of the circle is proportional to the number of regions. The interactions between two sites are displayed as gray lines, and the thickness of the gray line is proportional to the number of interactions. Note that in this analysis the CTCF sites displayed include only the non-enhancer, non-promoter CTCF sites.

(C) Number of interactions in Jurkat cells that are overlapped by an interaction identified using Hi-C in the indicated eight human cell lines. The high confidence Jurkat interactions (at least 3 PETs) were used for this analysis.
(D) Percentage of interactions that cross or do not cross TAD boundaries (defined in H1 human ESCs). The vast majority of interactions do not cross TAD boundaries (P-value<0.001, permutation test).

(E) The orientation of CTCF motifs at pairs of CTCF sites connected by cohesin ChIA-PET interaction is mostly convergent.
Fig. S4. Comparison of ChIA-PET interactions identified using Mango and the Dowen et al. pipelines

(A) Cohesin ChIA-PET interactions at the RUNX1 locus. The interactions identified using Mango are shown on the left, and the interactions identified using the Dowen et al. pipeline are shown on the right panel. Note that all Mango interactions are displayed (the high confidence interactions (FDR≤0.2) are displayed separately on Fig. S1C), while the high confidence interactions (at least 3 PETs) are displayed from the output of the Dowen et al. pipeline.

(B) Percentage of interactions that overlap between the interactions identified using the Mango and the Dowen et al. pipelines. Only the respective high confidence interactions were used for this analysis.

(C) Summary of types of interactions in the Jurkat ChIA-PET data using (left) Mango and (right) the Dowen et al. pipeline. The high confidence interactions of each pipeline were used for this analysis.
Fig. S5. Active oncogenes and silent proto-oncogenes in insulated neighborhoods in T-ALL

(A) Examples of insulated neighborhoods containing active oncogenes at the \textit{NOTCH1}, \textit{MYB} and \textit{ETV6} loci in Jurkat cells. The cohesin ChIA-PET interactions are displayed above the binding profiles of CTCF, SMC1 (cohesin), H3K27Ac, and RNA-Seq track. Gene models are displayed below the binding profiles.

(B) Examples of insulated neighborhoods containing silent proto-oncogenes at the \textit{TLX3}, \textit{OLIG2} and \textit{TCL6} loci in Jurkat cells. The cohesin ChIA-PET interactions are displayed above the binding profiles of CTCF, SMC1 (cohesin), H3K27Ac, and RNA-Seq track. Gene models are displayed below the binding profiles.
Fig. S6. Chromosomal microdeletions in T-ALL genomes

(A) Distribution plot of the lengths of recurrent genomic deletions found in T-ALL genomes. Only deletions <500 kb in size are plotted.

(B) Number of deletions found in T-ALL genomes that overlap insulated neighborhood boundaries. Significance of the observed overlaps is $P=0.647$ for all insulated neighborhoods, and $P=0.0013$ for insulated neighborhoods containing T-ALL Pathogenesis Genes (permutation test).
Fig. S7. Expression of TAL1 in patient samples harboring deletions that disrupt the TAL1 insulated neighborhood

(A) Insulated neighborhood at the TAL1 locus in Jurkat T-ALL cells. Cohesin ChIA-PET interactions are displayed above the ChIP-Seq binding profiles of CTCF and cohesin (SMC1). Patient deletions described in (22) are shown as bars below the gene models.

(B) Expression level of TAL1 in all patient samples with matched gene expression and genotype information in the clinical cohort described in (22). The expression data was downloaded from (22). The red bars correspond to the samples harboring the deletions denoted as red bars on panel (A). The grey bars correspond to the samples harboring the deletions denoted as grey bars on panel (A). The white bars correspond to the samples where no deletion was detected in the genome. The right side of the panel is the contingency table reflecting the occurrence of each indicated event. A Fisher’s test was used to test whether TAL1 expression above the median value was associated with the presence of a deletion (P=0.003276 two-sided Fisher exact test).
**Fig. S8. Disruption of an insulated neighborhood boundary is linked to TAL1 activation**

(A) ChIP-Seq binding profiles of CTCF, H3K27Ac, p300 and CBP, and RNA-Seq at the TAL1 locus in HEK-293T cells. Arrows point to two candidate enhancer regions occupied by H3K27Ac, p300 and CBP.

(B) The region around the CMPK1 promoter has enhancer activity. Regions 1 and 2 (displayed on panel (A)) were cloned into luciferase reporter vectors, and the luciferase activity was measured in HEK-293T and Jurkat cells. Data from n=3 replicate experiments are shown as mean+SD. Asterisk indicates P<0.05 (Student’s t-test).

(C) ChIP-Seq binding profiles of CTCF, H3K27Ac, and RNA-Seq at the TAL1 locus in primary human T-cells. The region deleted using a CRISPR/Cas9-based approach is highlighted in a grey box.

(D) (left) Experimental scheme of CRISPR/Cas9-mediated editing experiments in primary human T-cells. (right) qRT-PCR analysis of TAL1 expression. A total of 24 experiments were performed, and the difference in Ct values between TAL1 and GAPDH is displayed in the samples with detectable signal. P-value=0.04 between the control deletion and TAL1-neighborhood boundary -deletion samples (two-tailed t-test).
(E) Genotyping of the T-cell transfections described on panel (D). Genomic DNA was extracted 4-10 days after electroporation and the genomic loci containing the target sites for TAL1 were PCR-amplified.

(F) Single molecule RNA FISH of TAL1 and RUNX1 in primary human T-cells after CRISPR/Cas9-mediated editing. Cells were transfected with sgRNAs targeting a control region (control deletion) or sgRNAs targeting the TAL1 neighborhood CTCF boundary site (see C, D). Arrowheads point to the respective gene transcripts. RUNX1 was included as a positive control for the hybridization, and Jurkat T-ALL cells were also imaged as a positive control (for TAL1 and RUNX1 expression). Nuclei are counterstained with Hoechst 33342. Magnification is 40X.

(G) Quantification of the single molecule RNA FISH in the CRISPR-edited primary human T-cells described on panel (F). TAL1 and RUNX1 transcripts were counted in 100 cells in each condition, and the number of transcripts per cell is plotted for the 100 cells. The P-values between the control deletion and TAL1 neighborhood boundary deletion cells are: P<0.0001 (two-tailed t-test) for TAL1, and P=0.6 (not significant; two-tailed t-test) for RUNX1.
Fig. S9. Expression of *LMO2* in patient samples harboring deletions that disrupt the *LMO2* insulated neighborhood

(A) Insulated neighborhood at the *LMO2* locus in Jurkat T-ALL cells. Cohesin ChIA-PET interactions are displayed above the ChIP-Seq profiles of CTCF and cohesin (SMC1). Patient deletions described in (22) are shown as bars below the gene models.

(B) Expression level of *LMO2* in the patient samples in the clinical cohort described in (22). The expression data was downloaded from (22). The red bars correspond to the samples harboring the deletions denoted as red bars on panel (A). The right side of the panel is the contingency table reflecting the occurrence of each indicated event. A Fisher’s test was used to test whether *LMO2* expression above the median value was associated with the presence of a deletion (P=1 Fisher exact test).
Fig. S10. Disruption of insulated neighborhood boundaries leads to defects in chromosome structure

(A) 5C double alternating design. Four restriction fragments are represented with the double alternating primers (F for Forward, R for Reverse, LF for Left Forward and LR for Left Reverse). The universal tails are represented in green. Only the reverse primers are phosphorylated. All possible interactions for only fragment B are represented by arrows (for simplicity). The orange arrow represents the self-ligation of the restriction fragment and is removed during step 1 of the 5C processing.

(B) The neighborhood boundary deletion leads to an increase in contact frequencies across the deleted boundary at the TAL1 locus. 5C contact matrices at the TAL1 locus (chr1:46,740,122-48,740,121) in wild type HEK-293T cells, in cells in which the region containing the candidate LMO2 neighborhood boundary was deleted (LMO2-ΔCTCF HEK-293T), and in cells in which the TAL1 neighborhood boundary site were deleted (TAL1-ΔCTCF HEK-293T). The left and right panels are identical to Fig. 3E. Note that the LMO2-ΔCTCF HEK-293T cells have an intact TAL1 locus, and serve as a control that CRISPR/Cas9 perturbation of an unrelated neighborhood boundary does not lead to changes in chromosome structure at the TAL1 locus. The position of TAL1 neighborhood boundary site highlighted with an arrow. See also Fig. 3E.

(C) z-score difference (5C) map at the TAL1 locus (LMO2-ΔCTCF HEK-293T - wild type HEK-293T or TAL1-ΔCTCF HEK-293T - wild type HEK-293T). Increase in signal is colored red. Note the increase in the 5C signal adjacent to the deleted region in the TAL1-ΔCTCF HEK-293T cells. The position of the region removed in the TAL1-boundary deletion mutant cells is highlighted with an arrow. The right panel is identical to Fig. 3F.
(D) The neighborhood boundary deletion leads to an increase in contact frequencies across the deleted boundary at the \textit{LMO2} locus. 5C contact matrices at the \textit{LMO2} locus (chr11:33,003,550-35,003,549) in wild type HEK-293T cells, in cells where the \textit{TAL1} neighborhood boundary site was deleted (TAL1-\textDelta CTCHF HEK-293T), and in cells where the region containing the candidate \textit{LMO2} neighborhood boundary sites were deleted (LMO2-\textDelta CTCHF HEK-293T). The left and right panels are identical to Fig. 3K. Note that the TAL1-\textDelta CTCHF HEK-293T cells have an intact \textit{LMO2} locus, and serve as a control that CRISPR/Cas9 perturbation of an unrelated neighborhood boundary does not lead to changes in chromosome structure at the \textit{LMO2} locus. The position of the region removed in the LMO2-boundary deletion mutant cells is highlighted with an arrow.

(E) z-score difference (5C) map at the \textit{LMO2} locus (TAL1-\textDelta CTCHF HEK-293T - wild type HEK-293T or LMO2-\textDelta CTCHF HEK-293T - wild type HEK-293T). Increase in signal is colored red. Note the increase in the 5C signal adjacent to the deleted region in the LMO2-\textDelta CTCHF HEK-293T cells. The position of the region removed in the LMO2-boundary deletion mutant cells is highlighted with an arrow. The right panel is identical to Fig. 3L.
Fig. S11. Comparison of CTCF and SMC1 binding and cohesin ChIA-PET interactions in Jurkat, GM12878 and K562 cells

(A) Overlap analysis of CTCF ChIP-Seq binding peaks in Jurkat, GM12878 and K562 cells.

(B) Overlap analysis of Cohesin (SMC1 in Jurkat or RAD21 in GM12878 and K562) ChIP-Seq binding peaks in Jurkat, GM12878 and K562 cells.

(C) Overlap analysis of CTCF-CTCF/cohesin ChIA-PET interactions in Jurkat, GM12878 and K562 cells. The CTCF-CTCF/cohesin interactions that are found in at least two of the three cell types are classified as “constitutive neighborhoods.”
Fig. S12. Enrichment of mutations at the constitutive neighborhood boundary CTCF sites in many cancers

(A) (left) Frequency of somatic mutations and (right) Hapmap SNPs at CTCF sites that form constitutive neighborhood boundaries. The plots are centered on the CTCF motif identified under the CTCF binding sites, and regions 1 kb up- and downstream of the motif position are shown.

(B) The observed number of somatic mutations within constitutive neighborhood boundary CTCF sites (i.e. the data displayed on the left side of panel (A)), is significantly greater than the number of randomly permuted background mutations within +/-1 kb around the CTCF binding motif. Permutation was performed 10,000 times, and the y-axis shows the number of permutations in which the number of mutations around (+/-5bp) the CTCF motifs in the constitutive neighborhood boundary sites occurs at the number shown on the x-axis. The red line indicates the observed mutation frequency.

(C) Position weight matrix (PWM) scores of the sequence motifs containing somatic mutations at the constitutive neighborhood boundary CTCF sites plotted against the PWM score of the same motif consisting of the reference genome sequence.
(D) Frequency of somatic mutations at CTCF sites that (left) form constitutive neighborhood boundaries, and (middle and right) do not form neighborhood boundaries. The number of mutations is normalized to the number of CTCF sites found in each set. The plots are centered on the CTCF motif identified under the CTCF binding sites, and regions 1 kb up- and downstream of the motif position are shown. The two different control sets (i.e. CTCF sites that do not form neighborhood boundaries) are described in the Supplemental Materials and Methods. In the top panel, all mutations in the ICGC database were used (“Pan-cancer”). In the middle panel, the mutations in the ESAD-UK Genome Project were used. In the bottom panel, the mutations from the LIRI-JP Genome Project were used.

(E) Ratio of the normalized mutation rates at CTCF sites that form constitutive neighborhood boundaries versus CTCF sites that do not form neighborhood boundaries (two different control sets). The ratios were calculated using the mutations annotated in the indicated Genome Projects. “Pan-cancer” includes all the mutations in the ICGC database. Whiskers mark the estimated 95% confidence interval which was calculated using a bootstrap procedure. Only Genome Projects that contained at least 5 mutations for each of the three CTCF site sets were included in the analysis.
Fig. S13. Mutations in constitutive neighborhood boundary CTCF sites and in protein coding regions in the ICGC Genome Projects’ datasets

The numbers of mutations in constitutive neighborhood boundary CTCF sites and in protein coding regions were normalized to the size of the genome covered by these genomic elements.
Fig. S14. The enrichment of mutations at the constitutive neighborhood boundary CTCF sites is observed when different sources and methods for the identification of CTCF motifs are used

(A) Frequency of somatic mutations at CTCF sites that (left) form constitutive neighborhood boundaries, and (middle and right) do not form neighborhood boundaries. The number of mutations is normalized to the number of CTCF sites found in each set. The plots are centered on the CTCF motif identified under the CTCF binding sites, and regions 1 kb up- and downstream of the motif position are shown. The two different control sets (i.e. CTCF sites that do not form neighborhood boundaries) are described in the Materials and Methods. The CTCF motif used was downloaded from the JASPAR CORE 2014 vertebrate database (MA139.1), and Bioconductor was used for motif discovery.

(B) Frequency of somatic mutations at CTCF sites that (left) form constitutive neighborhood boundaries, and (middle and right) do not form neighborhood boundaries. The analysis was done as in (A) except the CTCF motif used was derived from genomic SELEX data.

(C) Frequency of somatic mutations at CTCF sites that (left) form constitutive neighborhood boundaries, and (middle and right) do not form neighborhood boundaries. The analysis was done as in (A) except FIMO was used for motif discovery.

(A-C) The enrichment of mutations at the constitutive neighborhood boundary sites compared to regions flanking the binding sites has a P-value <0.0001 (permutation test).
Fig. S15. Somatic mutations occur frequently at constitutive neighborhood boundary CTCF sites, but not at binding sites of other transcription factors in liver cancer

Frequency of somatic mutations at CTCF sites that form constitutive neighborhood boundaries, and at the binding sites of CTCF, MAX, MYC, FOXA1, FOSL2, JUND and NRFR2 in liver cancer cells. The numbers of mutations are normalized to the number of sites bound by the respective transcription factor in each set. The plots are centered on the transcription factor binding motif identified under the binding sites, and regions 1 kb up- and downstream of the motif position are shown.
Fig. S16. Recurrent base substitutions occur frequently in constitutive neighborhood boundary CTCF sites

(A-C) Frequencies of recurrent mutations (at least two occurrence of the same base substitution) in the ICGC dataset (“pan-cancer”) in neighborhood boundary CTCF sites and sites that do not form neighborhood boundaries. The numbers of recurrent mutations are normalized to the number of CTCF sites found in each set. The two control sets (i.e. CTCF sites that do not form insulated neighborhood boundaries) are described in the Supplemental Materials and Methods.

(A) The plot was generated using all mutations in the ICGC database. Corrected P-values (proportion test): constitutive neighborhood boundary CTCF sites vs. non-boundary CTCF sites (control set 1): P=2.11E-14; and constitutive neighborhood boundary CTCF sites vs. non-boundary CTCF sites (control set 2): P=2.09E-7

(B) The plot was generated using the mutations in the ESAD-UK Genome Project. Corrected P-values: constitutive neighborhood boundary CTCF sites vs. non-boundary CTCF sites (control set 1): P=1.16E-16; and constitutive neighborhood boundary CTCF sites vs. non-boundary CTCF sites (control set 2): P=6.05E-13

(C) The plot was generated using the mutations in the LIRI-JP Genome Project. Corrected P-values: constitutive neighborhood boundary CTCF sites vs. non-boundary CTCF sites (control set 1): P=6.21E-3; and constitutive neighborhood boundary CTCF sites vs. non-boundary CTCF sites (control set 2): not significant

All P-values listed are Bonferroni-corrected.
Supplementary Tables

Table S1. Summary statistics of the Jurkat SMC1 ChIA-PET data
Table S2. SMC1 ChIA-PET interactions
Table S3. T-ALL Pathogenesis Genes
Table S4. Gene expression (RPKM) values in Jurkat cells
Table S5. Deletions in T-ALL genomes overlapping insulated neighborhood boundaries
Table S6. 5C primers
Table S7. 5C summary statistics
Table S8. Constitutive neighborhoods across three cell types
Table S9. ICGC datasets
Table S10. Somatic mutations in constitutive neighborhood boundary sites
Table S11. GO-analysis of genes with neighborhood boundary mutations
Table S12. Candidate proto-oncogenes extracted from the Cancer Gene Census
Table S13. GEO accession IDs of the datasets used in this study
References


