ARTICLES
https://doi.org/10.1038/s41588-020-0680-8
nature genetics

Check for updates

Alteration of genome folding via contact domain boundary insertion

Di Zhang[1,2]✉, Peng Huang[1], Malini Sharma[1], Cheryl A. Keller[3], Belinda Giardine[3], Haoyue Zhang[1], Thomas G. Gilgenast[4], Jennifer E. Phillips-Cremins[4,5], Ross C. Hardison[3] and Gerd A. Blobel[1,2]✉

Animal chromosomes are partitioned into contact domains. Pathogenic domain disruptions can result from chromosomal rearrangements or perturbation of architectural factors. However, such broad-scale alterations are insufficient to define the minimal requirements for domain formation. Moreover, to what extent domains can be engineered is just beginning to be explored. In an attempt to create contact domains, we inserted a 2-kb DNA sequence underlying a tissue-invariant domain boundary—containing a CTCF-binding site (CBS) and a transcription start site (TSS)—into 16 ectopic loci across 11 chromosomes, and characterized its architectural impact. Depending on local constraints, this fragment variably formed new domains, partitioned existing ones, altered compartmentalization and initiated contacts reflecting chromatin loop extrusion. Deletions of the CBS or the TSS individually or in combination within inserts revealed its distinct contributions to genome folding. Altogether, short DNA insertions can suffice to shape the spatial genome in a manner influenced by chromatin context.

Whole-genome chromosome conformation capture (Hi-C) studies have described features of animal three-dimensional (3D) genome organization, including compartments and domains[1–5]. Compartments present as plaid-like patterns on Hi-C heatmaps, originally defined as multi-megabase open (A compartment) and closed (B compartment) chromatin regions[1], recently characterized as finer segregations often reflecting transcriptional activities[6–8]. Topologically associating domains (TADs)[2,3], or contact domains[4], are megabase/submegabase squares on Hi-C maps representing regions of enriched interactions. Separating adjacent domains are boundaries, across which interactions are depleted. Among the genomic features frequently co-localized with boundaries are TSSs of housekeeping genes and architectural proteins such as CTCF and cohesin[2,3]. In fact, CTCF or cohesin depletion perturbs, but does not abolish, domain configurations genome wide[6,7,9], whereas transcription inhibition can compromise boundary strength[10]. TADs have been found to remain largely conserved in evolution as integral functional units[8,11–14], and defective domain organization due to chromosomal rearrangements or disrupted binding of architectural factors at targeted genomic loci has been implicated in diseases[11,15–18]. Meanwhile, a new domain (neo-TAD) and regulatory circuitry can result from large-scale genomic duplications or inversions spanning domain boundaries[11]. However, important questions remain as to what features are minimally required for creating a domain[19,20]. Are domains created via megabase-scale genomic rearrangements, or can their formation be driven by kilobase-sized DNA elements? Moreover, as neither all sites bound by CTCF or cohesin nor all housekeeping gene TSSs are at domain boundaries[2,3], is a DNA element demarcating a domain boundary in one context able to delineate a new domain boundary when inserted into other contexts? How is the potentially intrinsic ability to demarcate domains encoded by sequences of boundary-associated DNA, and to what extent is it modulated by genomic context? In the present study, using a gain-of-function approach, we examined whether, and how, a putative boundary element can organize de novo domains in the context of multiple ectopic insertion sites.

Results

Random insertions of a domain boundary-associated DNA fragment. To obtain multiple insertions of a putative boundary sequence genome wide for subsequent multiplexed architectural characterization by Hi-C (Fig. 1a), we used a Sleeping Beauty transposon-based approach[21] in near-haploid human HAP1 cells[22]. The 2-kb DNA fragment we selected resides within a tissue-invariant domain boundary[3,4]. It is bound by CTCF, cohesin subunits and other architectural proteins, and contains a TSS of *PARL*, a housekeeping gene[23,24] (Supplementary Fig. 1; Fig. 1a, inset). After transposition, we established clonal lines and prioritized clones 21 (C21) and 25 (C25), which contained ten and six insertions, respectively—named C21 sites 1–10 (C21S1–C21S10) and C25 sites 1–6 (C25S1–C25S6)—across a total of 11 chromosomes (Extended Data Fig. 1a–f). All inserted DNA fragments retain their TSS function, initiating unidirectional transcripts at levels similar across all integration sites (as measured by quantitative PCR with reverse transcription (RT–qPCR); see Extended Data Fig. 1g), suggesting that any resulting architectural differences among insertions are not attributable to variations in transcription levels.

Ectopic boundary insertions variably demarcate new domains. To characterize how the insertion of a boundary-associated DNA fragment shapes human genome architecture, we performed in situ Hi-C on clones C21 and C25, and on parental (wild-type or WT) HAP1 cells. Overall, the results from the three samples were highly

[1]Division of Hematology, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Biochemistry & Molecular Biology, Pennsylvania State University, University Park, PA, USA. [4]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA. [5]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA.
✉e-mail: dizhang.penn@gmail.com; blobel@email.chop.edu

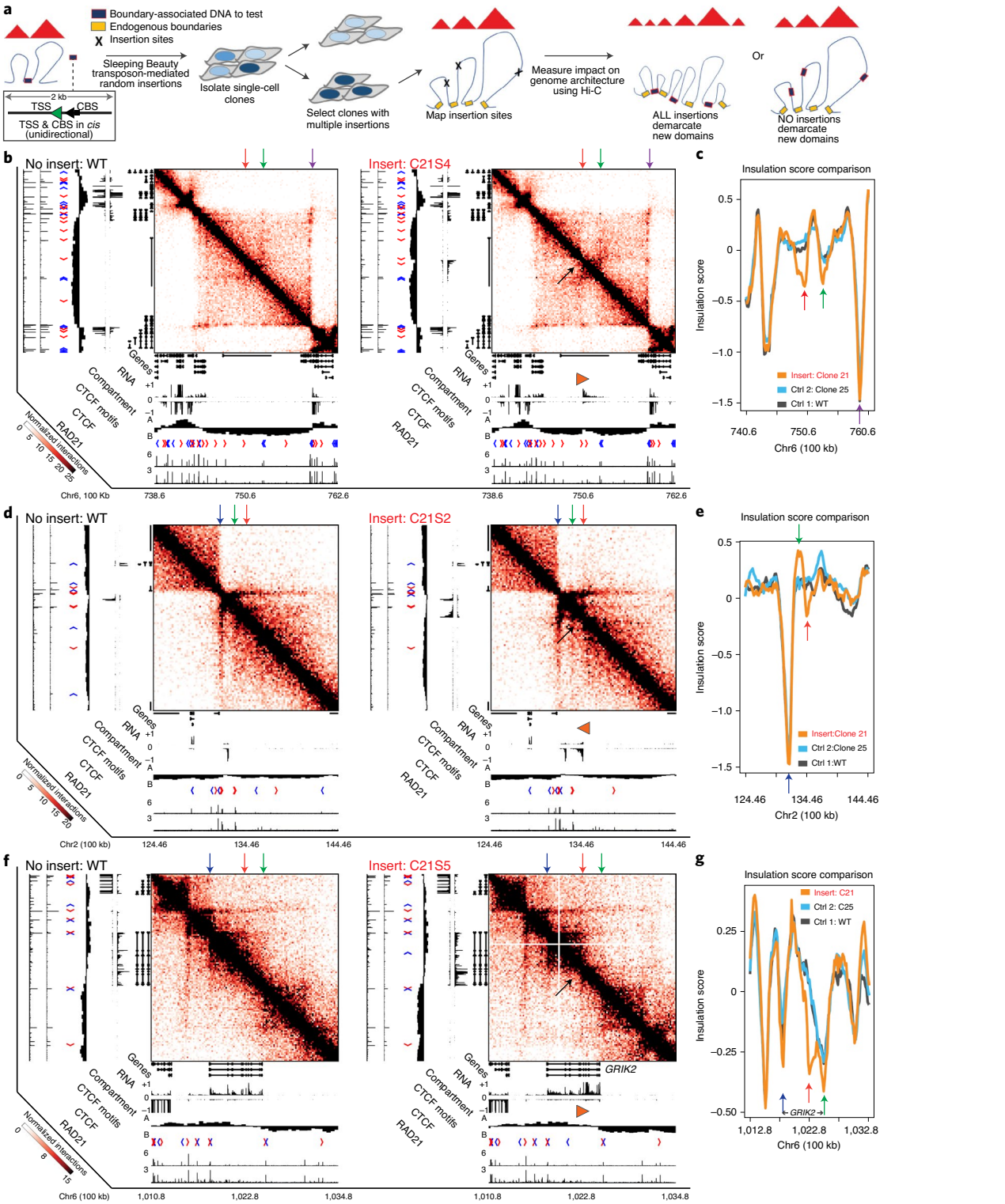


Fig. 1 | Domain boundary insertions create de novo contact domains. a, Schematic of experimental design. Throughout: red arrow: insertion site; green arrow: up- or downstream CBSs; blue/purple arrow: nearby boundaries; black arrow: notable architectural changes; orange arrowhead in the browser tracks: site and orientation of the insertion. **b**, Hi-C contact maps of control (no insertion, left) and C21S4 (right): de novo domain formation on insertion. Additional no-insertion control is shown in Extended Data Fig. 2. **c**, Insulation scores from Hi-C results in **b**, revealing strengthened insulation at the insertion site (C21S4) and the downstream CBSs demarcating a new domain. **d**, Hi-C contact maps of control (left) and C21S2 (right): a small pre-existing domain (between blue and green arrows) appears to coalesce into a larger new domain (between blue and red arrows) on insertion. Additional no-insertion control is shown in Extended Data Fig. 2. **e**, Insulation scores from Hi-C results in **d** showing strengthened insulation at the insertion site**. f**, Hi-C contact maps of C21S5: an insertion creates stripe-shaped contacts. **g**, Insulation scores from Hi-C results in **f** demonstrating strengthened insulation evident at the insertion and at the 3′-end of *GRIK2* (green arrow). Each Hi-C heatmap presents merged data from two independent experiments for each genotype. Two CTCF and RAD21 ChIP–seq and two RNA-seq experiments were performed for each genotype, with one of each displayed.

concordant (Supplementary Fig. 2), ruling out drastic genome-wide architectural perturbations caused by transpositions or substantial biases from clonal variations.

Importantly, examination of integration sites revealed at least four instances of apparent de novo contact domain formation (Fig. 1b,d,f and Extended Data Figs. 2 and 3a–f). These domains (at least 80 kb in size), identified at a resolution of 20-kb bins using an insulation score with a window of 200 kb, are defined as regions demarcated by two boundaries—one created at the insertion site and the other that may or may not be a pre-existing boundary. (Note: we use the term 'contact domains' or 'domains' here to refer to squares on an Hi-C heatmap representing, in general terms, regions of enriched interactions, but not in particular reference to TADs, sub-TADs, compartment domains or transcription domains. However, below we describe the characteristics of newly created domains and the mechanisms by which they might be formed.) Intriguingly, these de novo domains all showed distinct patterns. At C21S4, for example, the inserted 2-kb element, together with endogenous downstream convergent CBSs, demarcated a new domain ~250 kb in size (Fig. 1b and Extended Data Fig. 2a). In this case, the new ~250-kb domain corresponded to the length of the insert-driven transcript, which did not extend beyond the two convergent CBSs downstream—although whether the two downstream CBSs causally contribute to transcription termination here remains unclear[25] (Fig. 1b and Extended Data Fig. 2a). The insert thereby partitioned a ~1.7-Mb genome domain into smaller domains, which is visible as a cross pattern on the Hi-C heatmap (Fig. 1b,c and Extended Data Fig. 2a,b). This partitioning did not extensively alter the expression of adjacent genes (Fig. 1b and Extended Data Fig. 2a: RNA). Insulation scores were diminished (reflective of increased insulation) specifically at both ends of the transcribed region, also corresponding to the inserted and endogenous CTCF sites (Fig. 1c and Extended Data Fig. 2b) and supporting the formation of a new domain. Collectively, these observations suggest that both the act of transcription and the pairing of CTCF sites might contribute to domain formation at the C21S4 locus.

In addition, at C21S2, the insertion delimited a de novo ~300-kb domain with a strong, endogenous domain boundary to its left, while also increasing interactions within this newly formed domain (Fig. 1d,e and Extended Data Fig. 2c,d). This domain enclosed a ~100-kb transcribed region induced by the insert, and a pre-existing unannotated transcribed region, previously a subtle small square by itself on the Hi-C map (Fig. 1d and Extended Data Fig. 2c). This observation demonstrates that a larger domain can appear as a result of one smaller domain emerging immediately adjacent to another. Furthermore, at C21S5, the element inserted into the body of the actively transcribed *GRIK2* gene, forming a subtle but clearly detectable structure similar to what has recently been described as a stripe[26,27] (Fig. 1f,g and Extended Data Fig. 2e,f): a single locus forms enriched interactions with its contiguous chromatin region. This finding supports the sufficiency for forming a stripe, indicative of a loop extrusion process[26,28,29], via insertion of a short CBS- and TSS-containing element.

Together, these results highlight that insertions of a 2-kb putative boundary DNA element can form de novo domains with distinct attributes, potentially altering various aspects of genome folding. Moreover, the observed variability in effects of the insertions of the same sequence at different locations indicates modulation by chromosomal contexts.

In addition to the effects of insertions on domain structure, a closer examination of the new domain at the C21S4 locus revealed a plaid pattern, detectable even >30 Mb downstream, reminiscent of a compartment change (Extended Data Fig. 4). Indeed, compartment analysis showed that the de novo domain formed around the insertion site, originally part of a large B compartment, trended toward an A compartment (Fig. 1b: Compartment, and Extended Data Fig. 4). We observed additional instances supporting a focal B-to-A change as a result of the CBS–TSS insertion. At the previously described C21S2 locus, the transcription unit added by the insertion extended the A-compartment-like region, previously around the existing transcription unit, further toward the insertion site (Fig. 1d: Compartment). Compartment changes have been reported at the multi-megabase scale genome wide, associated with development, differentiation, reprogramming or infection[30–34]. This finding exemplifies that a single 2-kb insertion might trigger a focal B-to-A compartment switch in the absence of global perturbations or cell-state changes.

Insertions into pre-existing boundaries can further strengthen them. Five of the integration events occurred within pre-existing boundaries, enabling us to explore how the addition of a boundary-associated DNA element affects a pre-established boundary. At four of these loci, interactions across existing boundaries further decreased (Fig. 2 and Extended Data Figs. 5 and 3g–l). These insertion-associated reductions in interactions across pre-existing boundaries could implicate either a broader chromatin region (Fig. 2a–c) or a more focal region (Fig. 2d–f). Whereas previous computational analyses indicated that domain boundary strength scales with total CTCF levels[35] and architectural protein occupancy[24], these experimental findings imply that the addition of a boundary-associated DNA element might further strengthen an existing domain boundary.

De novo domain formation is impacted by genomic context. Notably, five insertions of this CTCF–TSS element did not result in considerable domain-level changes (Fig. 3a,b and Extended Data Fig. 6). This prompted us to probe contextual constraints that may limit new domain formation. One particular insertion event, which occurred within a ~350-kb domain at the C25S2 locus, did not demarcate an obvious new domain as measured by Hi-C (Fig. 3a,b). The ~2-Mb surrounding region is conspicuous in its complex architecture denoted by a high density of CBSs and genes (Fig. 3a), in stark contrast to previously discussed genomic contexts where new domains formed (Fig. 1b,d,f). To examine this region at a higher resolution, we performed Capture-C[36], RNA-sequencing (RNA-seq), and CTCF, RAD21 and H3K27ac chromatin immunoprecipitation sequencing (ChIP–seq). Together, these results revealed that this locus—located in a region rich in CTCF, cohesin, H3K27ac and transcribed genes—was not devoid of changes in interactions; rather, these changes were confined to an intradomain, sub-100-kb range, unable to manifest themselves as new domains (Fig. 3c,d). Curiously, gained interactions on insertion were limited to ~<25 kb, as far as transcription elongation from the inserted TSS (Fig. 3c). In contrast, de novo domains formed by the insertions were mostly accompanied by effective transcription elongation approaching or above ~100 kb (Fig. 1b,f). These observations suggest that both existing architectural complexity and restriction of transcription elongation may constrain the contexts in which new domains may be formed. We further scrutinized the genomic contexts of all 16 experimentally introduced insertions: restricted transcription elongation as well as existing CTCF or TSS density, albeit not statistically powered, coincide with a low likelihood of de novo domain formation (Supplementary Table 1 and Extended Data Fig. 6). Conceptually, as Hi-C measures interaction patterns resulting from multiple, sometimes competing modes of genome organization[37], we cannot discern whether different contextual constraints occur together or independently. One possible explanation for the lack of observable domain-scale effects of these insertions might be the low permissibility for transcription elongation, restricting transcription-mediated contacts to an unresolvable range in Hi-C. Another possibility is the high baseline level of pre-existing architectural complexity—probably mediated by the high density of

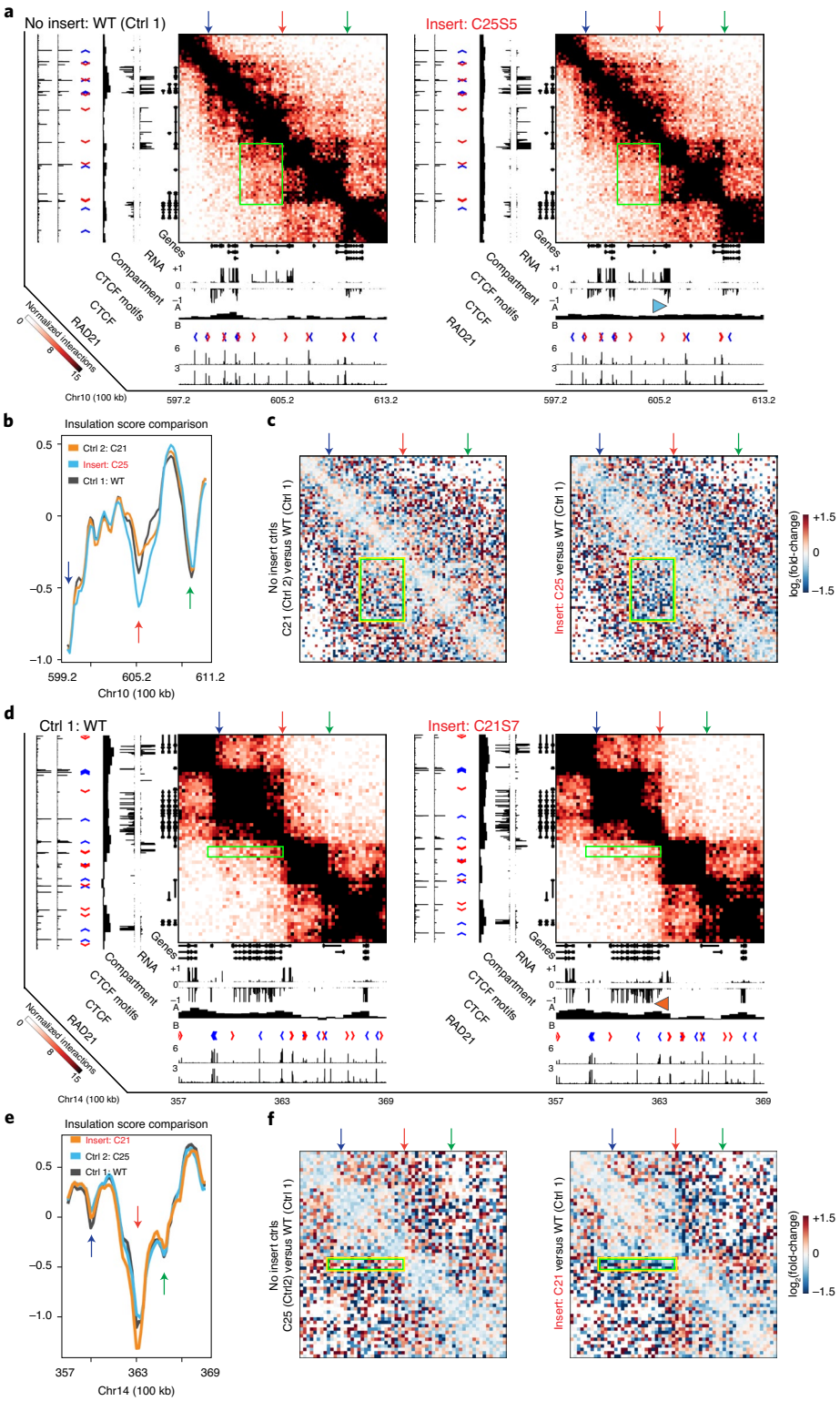


Fig. 2 | Domain boundary insertions can strengthen pre-established boundaries. Throughout: red arrow: insertion site; green/blue arrow: nearby boundaries. The orange (C21)/blue (C25) arrowhead in the browser tracks marks the site and orientation of the insertion. In **a**, **c**, **d** and **f**, yellow/green rectangles denote corresponding regions with overall decreased interactions on insertion. **a**, Hi-C contact maps of control (Ctrl, no insertion, left) and C25S5 (right). The insertion strengthens an existing domain boundary by decreasing interactions across it. Additional no-insertion control is shown in Extended Data Fig. 5. **b**, Insulation scores from Hi-C results in **a** showing strengthened insulation on insertion. **c**, $\log_2$(fold-changes) in interaction frequencies between no-insertion controls (left) and between the insertion clone and no-insertion control (right) for the region in **a**. **d**, Hi-C contact maps of control (no insertion, left) and C21S7. The insertion strengthens an existing boundary by decreasing interactions in its immediate proximity. Additional no-insertion control is shown in Extended Data Fig. 5. **e**, Insulation scores from Hi-C results in **d**. **f**, $\log_2$(fold-changes) in interaction frequencies between no-insertion controls (left) and between the insertion clone and no-insertion control (right) for the region in **d**. Each Hi-C heatmap presents merged data from two independent experiments performed for each genotype. Two CTCF and RAD21 ChIP-seq and two RNA-seq experiments were conducted for each genotype, with one of each shown.

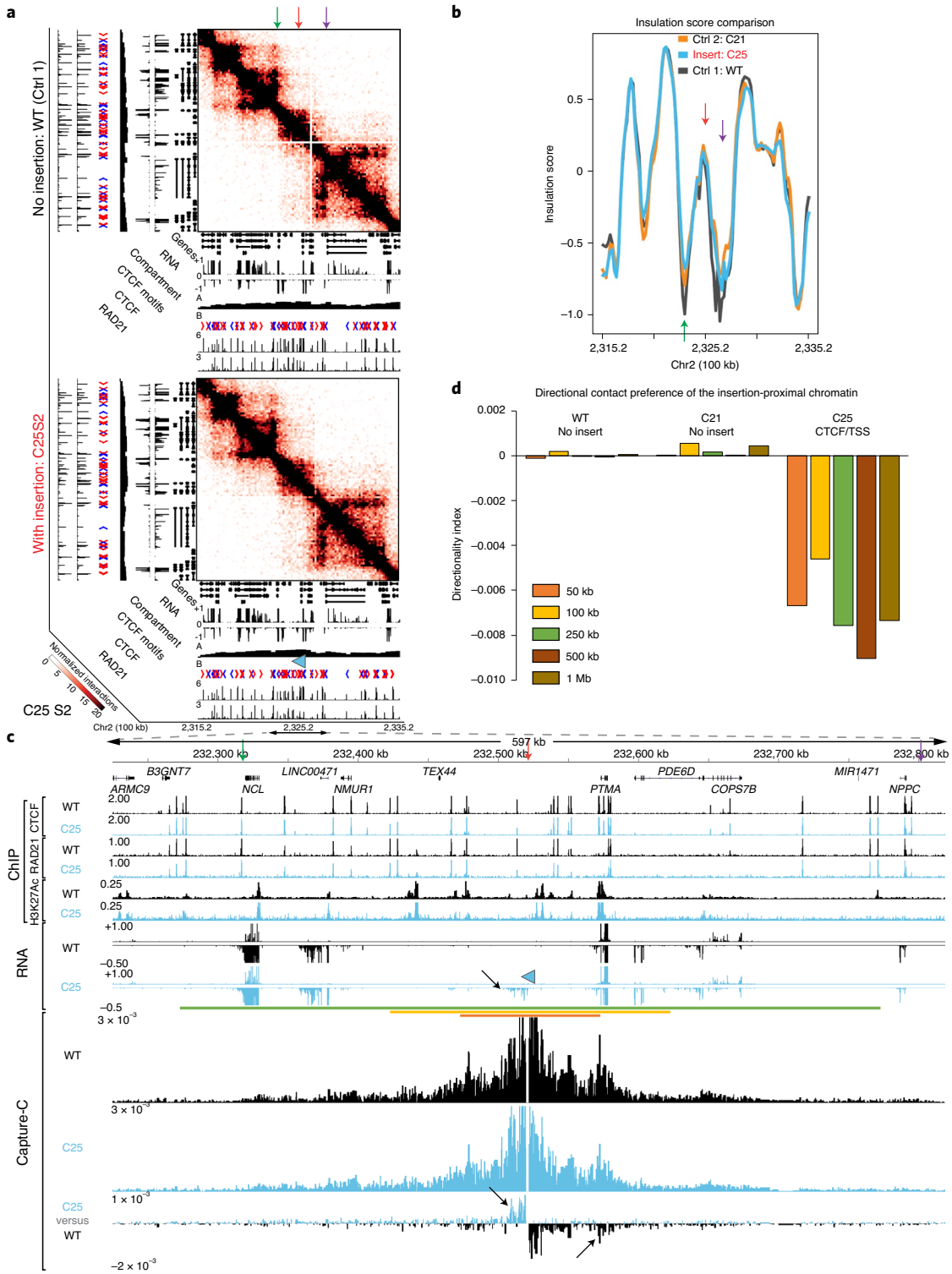


Fig. 3 | An insertion into a complex genomic region modestly changes short-range interactions, without domain-level impact. Throughout: red arrow: insertion site; green/purple arrow: a nearby boundary; blue arrowhead in the browser tracks: site and orientation of the insertion; black arrows: notable transcriptional/architectural changes. **a**, Hi-C maps of control (no insert, top) and C25S2 (bottom) showing no obvious domain-level changes on insertion. **b**, Insulation scores from Hi-C results in **a**, confirming the few apparent changes at the insertion site. Variations at the two boundaries (green and purple arrows) flanking the insertion probably caused by the empty bin on the control Hi-C heatmap in **a**. **c**, Examination of the ~600-kb region surrounding C25S2 reveals modest changes. The insertion coincides with possible reductions in RAD21 and CTCF binding ~8 kb to the left. Insertion-driven de novo transcripts do not elongate beyond ~25 kb. Capture-C anchored at the insertion site shows gained interactions along the transcribed region, and reduced interactions in the opposite direction (Capture-C: C25 versus WT). Colored lines denote distance ranges for measuring DIs in **d**. **d**, DIs revealing that the insertion induces preferential contacts to the left (negative DI). Hi-C/Capture-C results represent merged data from two independent experiments for each genotype. Two CTCF and RAD21 ChIP–seq, one H3K27ac ChIP-–seq and two RNA-seq experiments were performed for each genotype, with one of each shown.

CBSs and genes observed in the surrounding regions—which may prevent the detection of any comparatively modest changes caused by the insertions.

Next, we examined whether CTCF–TSS insertions affect transcription. Transcriptome wide, WT, C21 and C25 were highly congruent: only ~95 and ~160 genes were differentially expressed in C21 and C25, respectively (Extended Data Fig. 7a–e). The only differentially expressed gene near insertions (excluding gene-body insertions) was *MLKL* (Extended Data Fig. 7e,f), a recently characterized gene essential for necroptosis[38,39] and implicated in diseases[40–42]. An insertion event occurred in C25 in or near a putative *cis*-regulatory region[43–45] ~80 kb away from the *MLKL* gene, coinciding with the ~80% reduction in its transcript level (Extended Data Fig. 7g–h). Capture-C showed that the insertion formed strong contacts with the *GLG1* promoter, with which the *MLKL* promoter also interacted, albeit weakly (Extended Data Fig. 7h). However, we cannot ascertain the exact mechanism for the downregulation of *MLKL* on insertion, nor is it clear why *GLG1* expression is unaffected by the new contacts. Intriguingly, all the remaining ~191 genes within 1.5 Mb of any insertion were not differentially expressed (Extended Data Fig. 7e), whereas 4 of 9 genes with gene-body insertions were differentially expressed, suggesting that most of the gene expression changes were limited to gene-body insertions. As transposon insertions are almost random, the influence of insertions on the spatial pairing between genes and their putative *cis*-regulatory elements is difficult to decipher. One possible explanation for the absence of differential expression of insertion-proximal genes is that these insertions might not have occurred between gene promoters and their enhancers[46,47] active in HAP1 cells.

Ectopic CTCF binding and TSS contribute distinctly to domain formation. To interrogate newly formed domains at higher resolution, we carried out ChIP–seq for H3K27ac, CTCF and RAD21, RNA-seq and Capture-C experiments (Figs. 4 and 5). We first examined the C21S4 locus, where an insertion event demarcated a new domain with a long-range chromatin interaction pattern that became more A-compartment like (Figs. 1b,c and 4b and Extended Data Fig. 4). Several H3K27ac peaks emerged within the new domain in the clone bearing the insertion, along with a de novo transcript >200 kb in length (Fig. 4g: ChIP H3K27ac and RNA). These changes are concordant with the active chromatin features expected in a domain with a compartment-A signature[1,6–8]. Elevated RAD21 levels were also detected at several sites in the new domain on insertion (Fig. 4f,g and Extended Data Fig. 8k: R1–R5). In particular, the insertion increased interactions of the immediately neighboring chromatin with a pair of convergent CTCF peaks ~250 kb downstream (Fig. 4f,g: green arrow)—both CBSs gained RAD21 binding on insertion (Fig. 4g and Extended Data Fig. 8k: R4 and R5), whereas only the left CBS had moderately increased CTCF binding (Fig. 4g and Extended Data Fig. 8j: C4). This increased accumulation of cohesin on CBS–TSS insertion probably strengthened insulation at the downstream CTCF sites, which now demarcates the new domain (Fig. 4b,e,g).

To dissect the contributions of CTCF and TSSs to genome folding, we deleted the CBSs and the TSSs that are ~50 bp apart within the insert, alone and in combination via CRISPR[48,49], followed by additional Hi-C and Capture-C (Fig. 4 and Extended Data Fig. 8). Importantly, removal of the CBS, which spared transcription (Extended Data Fig. 8b), did not disrupt the newly formed domain with A-compartment features (Fig. 4c), despite weakened insulation at the insertion locus (Fig. 4e: red arrow). However, it did reduce the interactions between the insertion locus and downstream CBSs (Fig. 4h). Notably, deletion of the CBSs led to the emergence of a loop at the corner of the new domain (Fig. 4c), which coincided with elevated cohesin accumulation at both putative loop anchors (Fig. 4h and Extended Data Fig. 8k: R1, R4, R5). In contrast to CBS deletion, TSS deletion at this locus largely eliminated the de novo domain (Fig. 4d,e and Extended Data Fig. 8c). This was accompanied by reduced cohesin binding at the CTCF sites ~250 kb downstream and other nearby loci[34,50] (Fig. 4i and Extended Data Fig. 8k). Deletion of both the CBSs and the TSSs in one clone restored local chromatin structure to pre-insertion configuration as assayed by Hi-C (Extended Data Fig. 8e,g), coinciding with a further reduction in local cohesin accumulation close to the pre-insertion levels (Extended Data Fig. 8i: C21 ΔCTCF/ΔTSS no. 2, and 8k). In the other clone with the combined CBS and TSS deletion, we noticed a reversion to a diploid state that must have occurred after the initial transposon insertion. However, the process of editing resulted in a large ~27-Mb heterozygous deletion, rendering the region of interest de facto haploid (Extended Data Fig. 8d,h,f). The remaining allele has the desired edits that were subsequently characterized by Capture-C (Extended Data Fig. 8i). We next assessed how CBSs and/or TSSs fold their nearby genomic fragment by using a directionality index (DI)[3] on Capture-C data: computing DI, across increasing distances, uncovers directional contact preferences, while being agnostic to the mechanisms by which directional contacts are formed. The intact 2-kb element preferentially contacts downstream regions (as signaled by a positive DI): such preference is evident within 50 kb, becomes more pronounced at 250 kb and then flattens toward 1 Mb (Fig. 4a,b,f,g,j). CBS deletion did not abate the immediate directional preference at 50 kb, but did decrease preference for downstream contacts at the 1-Mb range (inclusive of the strong CTCF/cohesin-occupied boundary marked by the purple arrow in Fig. 4a,f) and to a lesser degree at the 250-kb range (Fig. 4c,h,j). Conversely, TSS deletion heavily reduced preference for downstream interactions from within 50 kb up to ~250 kb (Fig. 4d,i,j). Deletions of both CBSs and TSSs neutralized the genomic fragment's interaction preference close to the baseline level before insertion (Fig. 4j and Extended Data Fig. 8i). Therefore, at the C21S4 locus, the TSS folds local chromatin into a new domain through transcription elongation of ~250 kb.

Next, we applied this systematic approach to investigate the new domain at C21S2 (Fig. 5 and Extended Data Fig. 9). Disruption of the CBS here moderately reduced transcription (Extended Data Fig. 9b) and diminished the interactions with both CBSs downstream (Fig. 5c and Extended Data Fig. 9i: R6, R7). By contrast, these interactions, probably mediated by CTCF–CTCF pairing, remained largely unaffected on TSS deletion (Extended Data Fig. 9c and Fig. 5d,i). Unlike the new domain at C21S4, where the TSS is more important for its formation, here the CBS and the TSS act more cooperatively to change chromatin folding. Whereas the inserted CBS is responsible for the strengthened insulation (Fig. 5a–e) and long-range interactions (Fig. 5h), the TSS establishes short-range unidirectional contact preference (Fig. 5f–j). Deletions of both the CBS and the TSS restored the contact preference and cohesin levels to its original state (Extended Data Fig. 9d–i and Fig. 5j). Altogether, these results determine that, within the 2-kb insertion element, the ~100-bp sequence comprising the CBS and TSS is the most instructive for genome folding, whereas the relative importance of CBS and TSS to new domain formation can be context specific. Thus, genetic dissections disentangled two components of boundary elements often co-localized and intertwined: CBS forms distal interactions with convergent CBSs between demarcating boundaries, whereas TSS enforces strong directionality bias in genome folding in the orientation of transcription elongation.

Having illustrated how the 2-kb CBS–TSS element is capable of altering genome folding at multiple ectopic loci, we scrutinized its function at its endogenous context. Although the deletion of the 2-kb element at its endogenous *PARL* gene locus did not lead to the fusion of two neighboring domains (Extended Data Fig. 10a,b,d,e), the boundary seemed to have shifted ~60 kb to the left (Extended Data Fig. 10c). This shift is roughly the distance between

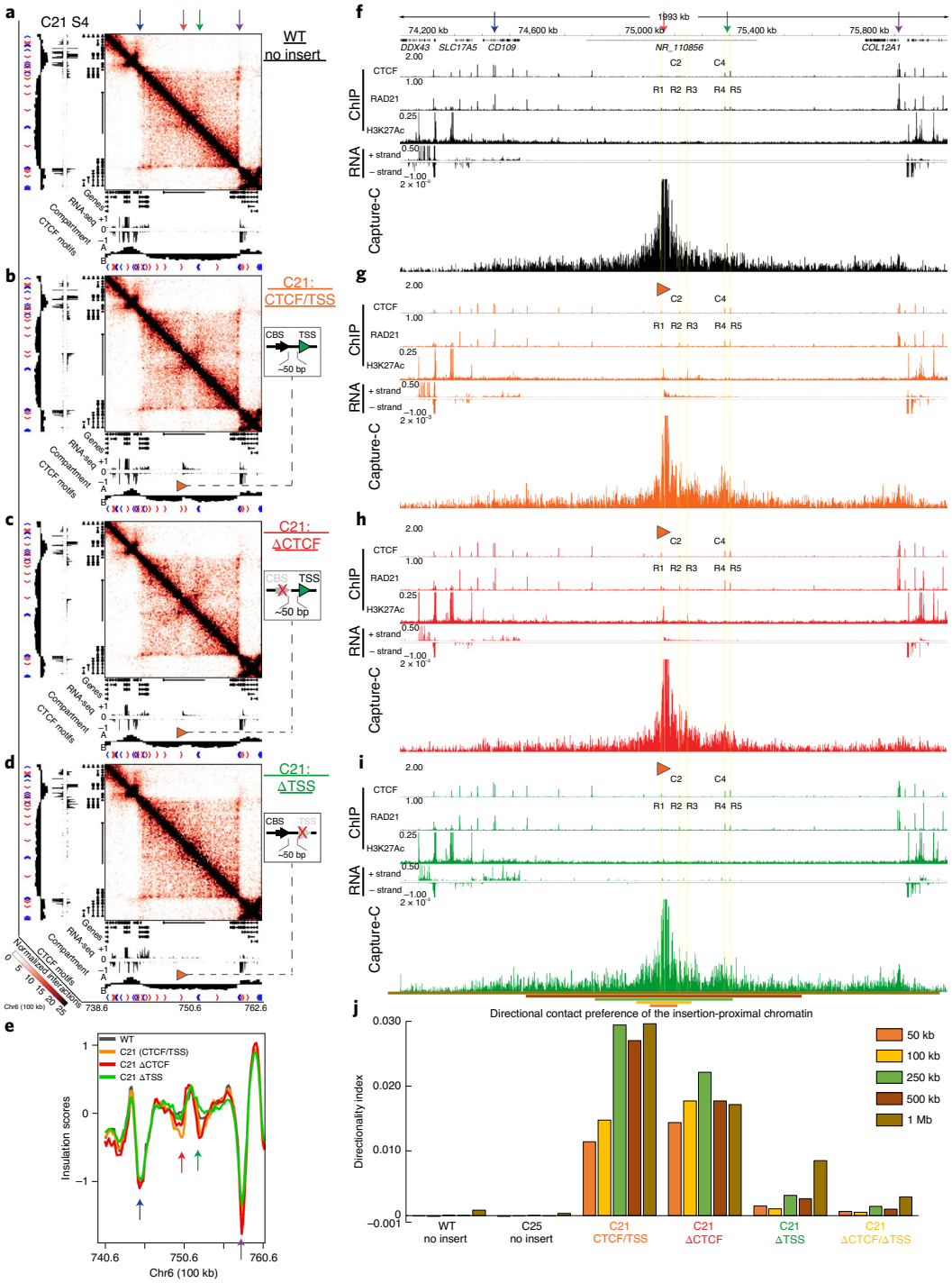


Fig. 4 | TSS can influence domain formation by switching its compartment signature. a–i, Hi-C of each genotype (insets) at C21S4 (**a**–**d**) and corresponding data tracks (**f**–**i**). Colored arrows mark corresponding loci in **a**–**i** (red arrow: insertion site; green arrow: downstream CBSs; blue/purple arrows: strong boundaries nearby). The orange arrowhead in the browser tracks: site and orientation of the insertion. **a**, Hi-C of control (no insert) cells. **b**, Hi-C of C21 with CBS/TSS insertion showing the new domain. **a**,**b**, Same as in Fig. 1b, shown here for comparison. **c**, The new domain persists on CBS deletion. **d**, The new domain diminishes on TSS deletion. **e**, Insulation scores from Hi-C results in **a**–**d**. **f**, WT no-insert ChIP/RNA-seq/Capture-C tracks of C21S4 in **a**. Differentially bound CTCF/RAD21 peaks on CBS–TSS insertion highlighted throughout **f**–**i**. C2, C4: CTCF peaks 2 and 4; R1–R5: RAD21 peaks 1–5. **g**, CBS–TSS insertion as in **b** produces >250-kb transcripts (RNA), spreads active histone marks (H3K27ac), increases interactions with the downstream boundary CBSs and diffusely within the new domain (Capture-C), and coincides with increased local CTCF/RAD21 binding (Extended Data Fig. 8j–k). **h**, CBS deletion as in **c** does not eliminate transcription or the gained H3K27ac marks, but reduces interactions with the downstream boundary CBSs (Capture-C). **i**, TSS deletion as in **d** abolishes transcription and the gained H3K27ac marks, while largely sparing CBS-associated interactions (Capture-C). TSS deletion is accompanied by locally reduced RAD21 (Extended Data Fig. 8k). Colored horizontal lines denote distance ranges for DI analysis in **j**. **j**, DI on Capture-C data (**f**–**i**) uncovers contributions of CBSs and TSSs to local chromatin folding (Extended Data Fig. 8i: ΔCTCF/ΔTSS). Hi-C/Capture-C results depict merged data of at least two independent experiments for each genotype. Two CTCF and RAD21 ChIP–seq, one H3K27ac ChIP–seq and two RNA-seq experiments were performed for each genotype, with one of each shown.

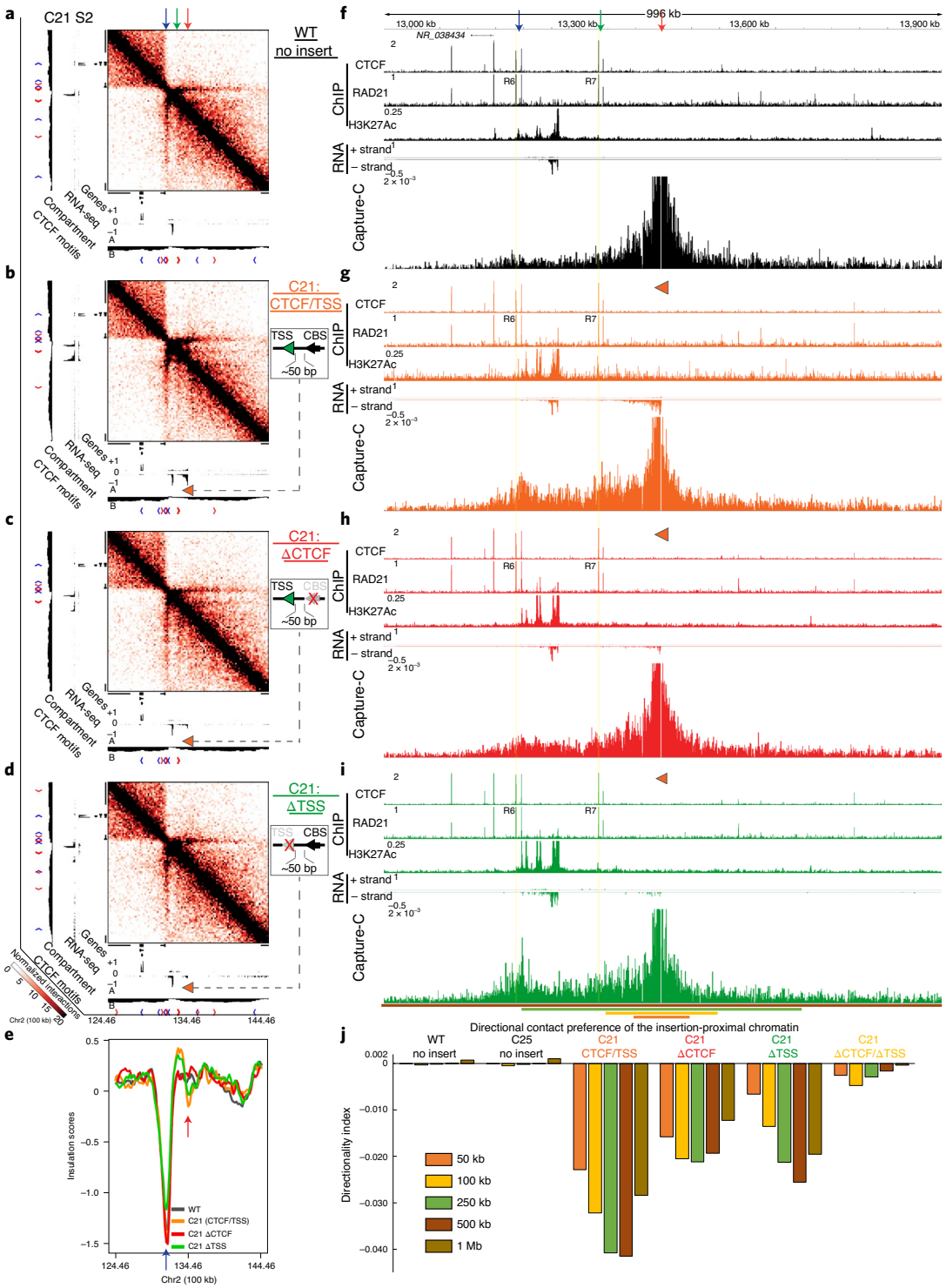


Fig. 5 | TSSs and CTCF cooperatively contribute to new domain formation by driving proximal and distal genome folding, respectively. a–**i**, Hi-C of each genotype (insets) at C21S2 (**a**–**d**), and corresponding data tracks (**f**–**i**). Red arrow: insertion site; green or blue arrow: downstream CBSs; orange arrowhead in the browser tracks: site and orientation of the insertion. **a**, Hi-C of no-insertion control. **b**, CTCF/TSS insertion forms a new domain (between red and blue arrows). **a**,**b**, Same as in Fig. 1d displayed here for ease of comparison**. c**, CBS deletion perturbs the new domain. **d**, TSS deletion partially reduces boundary strength at insertion, as in **e**. **e**, Insulation scores from Hi-C results in **a**–**d**. **f**, WT no-insert ChIP-/RNA-seq/Capture-C tracks of C21S2 in **a**. Differentially bound RAD21 peaks on CBS–TSS insertion highlighted throughout **f**–**i**. R6, R7: RAD21 peaks 6 and 7. **g**, CBS/TSS insertion as in **b** produces ~100-kb transcripts (RNA), increases interactions with the downstream CBSs and diffusely within the new domain (Capture-C), and coincides with increased local RAD21 (Extended Data Fig. 9i). **h**, CBS deletion as in **c** does not eliminate transcription, while diminishing interactions with the downstream CBSs to the left (Capture-C). **i**, TSS deletion as in **d** spares CBS-associated interactions (Capture-C). Colored horizontal lines indicate distance ranges for DI analysis in **j**. **j**, DI on Capture-C data (**f**–**i**): TSS and CTCF drive proximal and distal chromatin folding, respectively (Extended Data Fig. 9h: C21 ΔCTCF/ΔTSS). Hi-C/Capture-C results depict merged data of at least two independent experiments for each genotype. Two CTCF and RAD21 ChIP-seq, one H3K27ac ChIP-seq and two RNA-seq experiments were performed for each genotype, with one of each shown.

the TSSs of *PARL* and its neighboring transcribed gene and a nearby CBS (Extended Data Fig. 10f), suggesting that nearby TSS–CBS elements may assume boundary function in the absence of the 2-kb element[51,52].

Context may modulate how SINE B2 transposons shape the mouse genome. The tissue-invariant boundary element we inserted here drives detectable new domain formation in a context-dependent manner. Intrigued by this finding, we wondered to what extent transposable element insertions during evolution contribute to new genome domain formation, and whether context might affect the likelihood of such an outcome. To this end, we considered a class of transposable elements linked to domain boundary formation[3]: SINE B2, which harbors CBS and Pol III TSSs. SINE B2 elements are often transcribed at low levels, escaping detection by conventional RNA-seq[53,54], in contrast to human endogenous retroviruses, the high transcription level of which is essential for their boundary activity[19]. This suggests a different mechanism of action that potentially relies less on high levels of transcription and more often on CBSs[4,52,55], presenting an opportunity to compare and contrast the contribution of boundary formation by transcription and CTCF separately. SINE B2 elements have been inserted into tens of thousands of loci in the mouse (but not primate) genome since the mouse diverged from its placental ancestor ~75 million years ago[56,57]. This provides a window into recent evolution for exploring how these insertions that expand CBSs may have contributed to the formation of putative new domain boundaries—defined as mouse Hi-C boundaries overlapping with mouse-specific CTCF-bound SINE B2 insertions while lacking ancestral CTCF binding[52,56]. To approach this question, we analyzed CTCF ChIP–seq data[54] in a mouse erythroid cell line and identified ~7,136 SINE B2 elements bound by CTCF[58] (Fig. 6a). We also carried out Hi-C for this cell line, which revealed ~7,508 domain boundaries genome wide[54]. Among these, we identified ~625 putative new boundaries harboring ~701 (~9.8% of the ~7,136) CTCF-bound SINE B2 insertions (Fig. 6b). Next, we explored whether the outcome of CTCF-bound SINE B2 insertions to detectably alter domain structure could be modulated by local context. We used the number of TSSs within 200 kb of a CTCF-bound SINE B2 element (the TSS density) as a proxy for architectural complexity[59–61]. By comparing co-localization rates between CTCF-bound SINE B2 and putative new boundaries across the TSS density spectrum (Fig. 6c), we found that, of the total ~7,136 CTCF-bound SINE B2 insertions, 527 occurred in the most TSS-sparse regions (with no other TSSs within 200 kb of the SINE B2 element). Of these, 10.06% (53/527) co-localized with putative new domain boundaries (Fig. 6c). By contrast, 721 CTCF-bound SINE B2 insertions took place in the most TSS-dense regions (Fig. 6c) (defined as having ≥11 TSSs within 200 kb), and only 5.27% (38/721) of them co-localized with putative new boundaries (Fig. 6c). As it is impossible to ascertain chromatin architecture for the genome of the placental ancestor, we are unable to annotate the list of definitive new mouse boundaries. Nevertheless, based on these findings, we speculate that, in recent mouse genome evolution, CTCF-bound SINE B2 insertion events may be more likely to contribute to the creation of detectable domain boundaries in TSS-sparse regions than in regions with high TSS densities.

Discussion

Our results establish that insertions of a 2-kb element can demarcate new domains of several hundred kilobases and shape chromosome architecture potentially up to tens of megabases away (Extended Data Fig. 4)—with both transcription and CTCF contributing to local changes in domain structure. Different de novo domains formed by the same sequence may manifest distinct facets of genome organization (Fig. 6d). First, a new domain can be repackaged from the B to the A compartment. Although compartments have historically been thought of as much larger than domains (tens of megabases in size), our creation of a de novo domain by changing its compartment signature is consistent with more recent observations that small genomic segments can be autonomous in their ability to compartmentalize, through comparative analysis[8,62] or global cohesin depletion[37]. Second, a larger domain can be formed via confluence of two smaller ones[63], which is also evident in 3D genome reconfiguration on mitotic exit[54]. Third, a stripe, perhaps reflecting cohesin-mediated loop extrusion[26], can be formed with a CTCF–TSS insertion. Moreover, through genetic dissections, we have unraveled functional elements driving genome folding. Specifically, the TSS enforces on adjacent chromatin a strong directional contact preference in the direction of transcription, contributing to directionality index (DI)-based boundary detection[3]. Meanwhile, convergent CBSs form distal interactions to demarcate boundaries (Fig. 6d). Importantly, genomic context not only can modulate effects of the inserted element to display discrete features of de novo domains, but also may pose constraints that limit or mask measurable new domain formation[20]. Under these latter scenarios, the effects of CTCF–TSS insertion—especially the proximal directional bias introduced by a TSS—are not entirely absent, but rather confined to an intradomain, sub-100-kb range (Fig. 6d) that might be further resolved with emerging techniques with subkilobase resolution[64,65]. Through our experimental findings and our analysis of recent genome evolution, we have begun to explore possible contextual constraints—they might include low permissibility for transcription elongation and/or existing architectural complexity marked by a high density of CBSs and/or gene TSSs.

The loop extrusion model states that a domain is formed as cohesin extrudes chromatin until it is stalled by CTCF[28,29]; however, given the fact that many CBSs are not at boundaries, does the sequence underlying a domain boundary, in addition to a CBS alone[20], encode the function of domain demarcation? Only recently have several putative boundary elements with different compositions begun to be tested in mammalian systems, leading to varying results. The integrations of a ~72-kb Firre cDNA consisting of ~15 CBSs with various combinations of convergent CTCF pairs did not lead to measurable chromatin structural changes, regardless of the element's transcription state[20]. Insertions of three CBSs in *cis* formed loops and stripes, although it was not immediately clear whether new boundaries or domains formed[27]. By contrast, a locally repositioned (deleted from its endogenous site and inserted ~1 Mb away) element with two pairs of divergently oriented CBSs still functioned as a boundary[51], although it is unknown whether this element can still form a boundary without the deletion of its endogenous copy or when placed beyond its local context. Without any CBS, a human endogenous retrovirus element has been shown to function as a boundary in a transcription-dependent manner, with its transcripts confined within ~8 kb, the element's length[19]. Our data clearly demonstrate the DNA-encoded ability to alter genome folding—a ~100-bp element spanning the TSS and CBS within the 2-kb insertion is mostly responsible for new domain formation while manifesting itself as a domain boundary. Specifically, the TSS folds its nearby chromatin along the direction of transcription, whereas the CBS forges comparatively focal contacts with endogenous CTCFs at the distal demarcating boundary. Meanwhile, the ability of domain organization/boundary formation is subject to modulation by context, which possibly underlies the observation by colleagues[19,20,27] and by us that not all insertions have resulted in domain-level changes. These observations underscore the importance of using multiplexed edited genomes, beyond local, individual loci, to more comprehensively investigate how domain formation is causally connected with DNA sequences and genomic context (Fig. 6d). In the present study, we have leveraged genome editing and Hi-C for multiplexed characterization of the inserted putative boundary DNA element's effects on the human

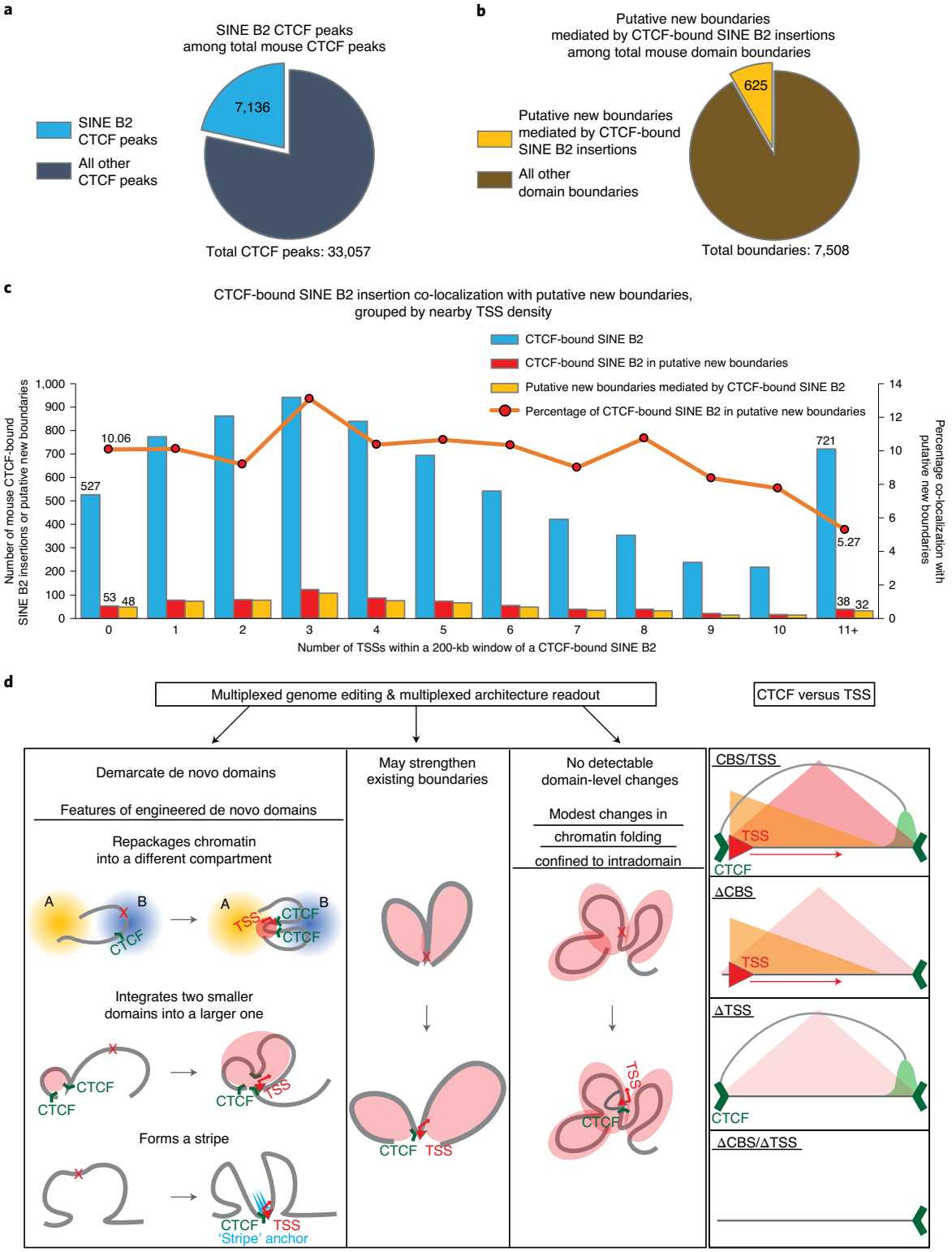


Fig. 6 | Possible context dependency in how SINE B2 elements shape mouse genome architecture in recent evolution, and a graphic summary of the present study. a, SINE B2-derived CTCF peaks (~7,136, sky blue) constitute ~21.6% of all CTCF peaks (~33,057) in the mouse genome[54]. **b**, Putative SINE B2-mediated new domain boundaries (~625, yellow) may constitute ~8.3% of all mouse domain boundaries (~7,508)[54]. **c**, A clustered column–line chart (columns: counts; line: percentage) showing the distribution of all ~7,136 CTCF-bound SINE B2 insertions (blue columns), those that co-localize with putative new boundaries (red columns) and putative new boundaries possibly mediated by CTCF-bound SINE B2 insertions (yellow columns, from the yellow portion in **b**), based on their nearby TSS density (horizontal axis): the number of TSSs within a 200-kb window of each CTCF-bound SINE B2 insertion. Each red dot in the line plot indicates, at each TSS density, the percentage of putative, new boundary, co-localized, CTCF-bound SINE B2 (red column) among all CTCF-bound SINE B2 (blue column). At a TSS density of 0, 10.06% (53/527) CTCF-bound SINE B2 insertions co-localize with putative new boundaries, compared with 5.27% (38/721) at TSS density ≥11 (two-sided Fisher's exact test, $P = 0.0019$). **d**, Graphic summary of the present study.

genome—creating de novo domains that display multiple important features of chromatin architecture. This work demonstrates that it is feasible to harness short DNA insertions to explore a diverse genomic space toward understanding how sequence and context together influence genome architecture, and ultimately toward rationally engineering the 3D genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0680-8.

Received: 25 August 2019; Accepted: 23 July 2020;
Published online: 31 August 2020

References

1. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
2. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
3. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
4. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
5. Phillips-Cremins, J. E. et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
6. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
7. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320.e24 (2017).
8. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852.e7 (2017).
9. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944.e22 (2017).
10. Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* **169**, 216–228.e19 (2017).
11. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
12. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
13. Fudenberg, G. & Pollard, K. S. Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl Acad. Sci. USA* **116**, 2175–2180 (2019).
14. Symmons, O. et al. The shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Dev. Cell* **39**, 529–543 (2016).
15. Lupiáñez, D. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
16. Narendra, V. et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015).
17. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
18. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
19. Zhang, Y. et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* **51**, 1380–1388 (2019).
20. Barutcu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat. Commun.* **9**, 1444 (2018).
21. Mátés, L. et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **41**, 753–761 (2009).
22. Carette, J. E. et al. Ebola virus entry requires the cholesterol transporter Niemann–Pick C1. *Nature* **477**, 340–343 (2011).
23. Haarhuis, J. H. I. et al. The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693–707.e14 (2017).
24. Van Bortle, K. et al. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.* **15**, R82 (2014).
25. Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541–554 (2015).
26. Vian, L. et al. The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178.e20 (2018).
27. Redolfi, J. et al. DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. *Nat. Struct. Mol. Biol.* **26**, 471–480 (2019).
28. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, 6456 (2015).
29. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
30. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
31. Krijger, P. H. L. et al. Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. *Cell Stem Cell* **18**, 597–610 (2016).
32. Ke, Y. et al. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell* **170**, 367–381.e20 (2017).
33. Du, Z. et al. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232–235 (2017).
34. Heinz, S. et al. Transcription elongation can affect genome 3D structure. *Cell* **174**, 1522–1536.e22 (2018).
35. Gong, Y. et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nat. Commun.* **9**, 542 (2018).
36. Hughes, J. R. et al. Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
37. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl Acad. Sci. USA* **115**, E6697–E6706 (2018).
38. Sun, L. et al. Mixed lineage kinase domain-like protein mediates necrosis signaling downstream of RIP3 kinase. *Cell* **148**, 213–227 (2012).
39. Zhao, J. et al. Mixed lineage kinase domain-like is a key receptor interacting protein 3 downstream component of TNF-induced necrosis. *Proc. Natl Acad. Sci. USA* **109**, 5322–5327 (2012).
40. Galluzzi, L., Buqué, A., Kepp, O., Zitvogel, L. & Kroemer, G. Immunogenic cell death in cancer and infectious disease. *Nat. Rev. Immunol.* **17**, 97–111 (2017).
41. Shan, B., Pan, H., Najafov, A. & Yuan, J. Necroptosis in development and diseases. *Genes Dev.* **32**, 327–340 (2018).
42. Yuan, J., Amin, P. & Ofengeim, D. Necroptosis and RIPK1-mediated neuroinflammation in CNS diseases. *Nat. Rev. Neurosci.* **20**, 19–33 (2019).
43. Chung, C. C. et al. Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nat. Genet.* **45**, 680–685 (2013).
44. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
45. Mitchell, J. S. et al. Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nat. Commun.* **7**, 12050 (2016).
46. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc. Natl Acad. Sci. USA* **105**, 20398–20403 (2008).
47. Rawat, P., Jalan, M., Sadhu, A., Kanaujia, A. & Srivastava, M. Chromatin domain organization of the TCRb locus and its perturbation by ectopic CTCF binding. *Mol. Cell Biol.* **37**, e00557–16 (2017).
48. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
49. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
50. Busslinger, G. A. et al. Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* **544**, 503–507 (2017).
51. Despang, A. et al. Functional dissection of the Sox9—Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet* **51**, 1263–1271 (2019).
52. Choudhary, M. N. et al. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* **21**, 16 (2020).
53. Karijolich, J., Zhao, Y., Alla, R. & Glaunsinger, B. Genome-wide mapping of infection-induced SINE RNAs reveals a role in selective mRNA export. *Nucleic Acids Res.* **45**, 6194–6208 (2017).
54. Zhang, H. et al. Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* **576**, 158–162 (2019).

55. Sundaram, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
56. Schmidt, D. et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).
57. Bourque, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18**, 1752–1762 (2008).
58. Thybert, D. et al. Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Res.* **28**, 448–459 (2018).
59. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
60. Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
61. Kentepozidou, E. et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.* **21**, 5 (2020).
62. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
63. Zhan, Y. et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.* **27**, 479–490 (2017).
64. Hsieh, T. S. et al. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell* **78**, 539–553.e8 (2020).
65. Krietenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* **78**, 554–565.e7 (2020).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Experiments. *HAP1 cell culture and maintenance.* HAP1 cells[22] (a kind gift from B. van Steensel through the 4D Nucleome consortium) were cultured in Iscove's modified Dulbecco's medium with 10% fetal bovine serum and 1% penicillin–streptomycin at 37 °C with 5% CO_2. To enrich a near-haploid population, HAP1 cells were routinely stained with a cell-permeant double-stranded DNA dye, followed by FACS. Briefly, HAP1 cells were trypsinized, pelleted, counted using a hemocytometer and resuspended at a density of 1 million cells ml^{-1} of growth medium containing 10 µg ml^{-1} of Hoechst 33342 (Thermo Fisher Scientific, catalog no. H3570) for 30 min. Stained cells were then pelleted, resuspended in 1 ml of MACS buffer (1× phosphate-buffered saline (PBS), pH 7.2, 0.5% bovine serum albumin and 2 mM ethylenediaminetetraacetic acid (EDTA)), and subsequently sorted on a MoFlo Astrios (Beckman Coulter). The sorting gate was stringently set to enrich cells with half the DNA content of a diploid (2*n*) cell line (HUDEP-2 (ref. [66])) at G1. Growth medium was then added to sorted cells for continued culture.

Sleeping Beauty transposon genome editing. The candidate 2-kb DNA (see Boundary-underlying DNA selection) underwent PCR (primer sequences in Supplementary Table 2) with HAP1 genomic DNA (gDNA) as the template, and was cloned into pSA-MCS[67] (a kind gift from A. Recchia and Z. Ivics; this plasmid can be requested from them on ordering Addgene, plasmid no. 26557), to generate pSA-MCS-2kb (plasmid sequence GenBank file; Supplementary Data 1). HAP1 cells were co-transfected with the two components of the Sleeping Beauty transposon system: the transposon vector with the 2-kb insert, pSA-MCS-2kb and the transposase vector, pCMV(CAT)T7-SB100 (ref. [21]) (a kind gift from Z. Ivics and Z. Izsvák, Addgene, plasmid no. 34879), together with pmaxGFP (Lonza), using Nucleofector Kit L (Lonza) and the program X-001 on an Amaxa electroporator (Lonza). Then, 24 h after transfection, the top 1% green fluorescent protein-positive (GFP$^+$) transfected HAP1 cells were sorted on a FACSJazz sorter (BD Biosciences) as single cells into five 24-well plates, as well as a pooled population (~340 cells); 34 single-cell clones recovered after sorting. Genomic DNA was harvested from sorted clonal and pooled populations as they continued to expand.

Insertion copy number estimation. To screen for single-cell clones with higher transposition copy numbers, real-time PCR was performed using genomic DNA from edited single-cell clones, with two primer pairs (IR–qPCR-1 and IR–qPCR-2; Supplementary Table 2) targeting the inversed repeat regions that are part of each transposition flanking the 2-kb insert, but not in the endogenous genome, as well as another primer pair (hCD4; Supplementary Table 2) targeting an endogenous genome locus. The estimated insertion copy number (EICN) was then calculated using the following formula:

$$\mathrm{EICN} = \mathrm{Mean}\left(\frac{2^{Ct_{\mathrm{IR}-1}-Ct_{\mathrm{CD4}}}}{2}, \frac{2^{Ct_{\mathrm{IR}-2}-Ct_{\mathrm{CD4}}}}{2}\right)$$

The insertion copy number for the transfected, sorted and pooled HAP1 population was also estimated using this approach.

Insertion site mapping and validation. Capture using biotinylated oligonucleotides and pull down with streptavidin beads was used to map transposon insertion sites as outlined in Extended Data Fig. 1b. Specifically, for each clone chosen to have its transposition sites mapped, 8 µg of gDNA, prepared with PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific, catalog no. K182001), was sonicated at 100% amplitude, 30 s on:30 s off, for 40 min, in a bath sonicator (QSonica, catalog no. Q800R3). Sonicated DNA was cleaned up using AMPure XP beads (Beckman Coulter), end repaired and dA-tailed (NEBNext Ultra), adapter ligated, indexed (NEBNext Multiplex Oligos for Illumina) and P5/P7 amplified (NEBNext Q5 Hot Start HiFi PCR Master Mix). This library-prepared DNA, dried in a PCR machine with the tube cap open, was resuspended in 7.5 µl of NimbleGen 2× hybridization buffer, 3 µl of NimbleGen Hybridization Component A and 2.5 µl of nuclease-free water, and incubated for 10 min at room temperature. Resuspended DNA was then denatured at 95 °C in a PCR machine for 10 min. Then 2 µl of 1.5 µM 5′-biotinylated hybridization oligonucleotide (IR_Junc_Hyb; Supplementary Table 2), which targets the inversed repeat regions immediately proximal to the endogenous genome, was subsequently added. After vortexing for a few seconds and spinning down, the mixture was incubated in a PCR machine at 47 °C (lid temperature 57 °C) overnight. Each mixed library DNA and biotinylated hybridization oligonucleotide was then added to 40 µl of washed Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific, catalog no. 65001) at 47 °C in a thermomixer for 1 h. With 100 µl of pre-heated 1× wash buffer I added to the beads–DNA mixture, the tubes were vortexed and placed on a magnetic stand, with the supernatant containing unbound DNA subsequently removed. DNA-bound streptavidin beads were then washed twice with the Stringent Wash Buffer, followed by one wash each with Wash Buffers I, II and III (all Wash Buffers, together with NimbleGen 2× Hybridization Buffer and Nimblegen Hybridization Component A, were supplied in the SeqCap EZ Hybridization and Wash Kits, Roche NimbleGen, catalog no. 05634261001). Hybridized DNA was eluted off the beads with 25 µl of 0.125 M NaOH, neutralized with 25 µl of 1 M Tris-HCl, pH 8.8, and followed by AMPure XP bead cleanup. Eluted DNA was then enriched with PCR using P5/P7 primers (NEBNext Q5 Hot Start HiFi PCR Master Mix) for 12 cycles. PCR-enriched pulldown DNA underwent an additional round of biotinylated oligonucleotide hybridization and streptavidin bead pulldown. Pulldown enrichment was confirmed using qPCR before Illumina sequencing. After the insertion sites have been mapped (Insertion-site mapping), the location and orientation of each insertion were validated via targeted PCR.

In situ Hi-C. In situ Hi-C was performed as previously described[4,9,68]. Briefly, ~10 million HAP1 cells, WT, C21, C25 and all the CRISPR-edited subclones were crosslinked in 2% formaldehyde at room temperature for 10 min, and then quenched in 0.125 M glycine for 5 min, on an orbital shaker. Cells were then transferred from a culture flask to a 15-ml tube, pelleted, washed with 1 ml of cold PBS, transferred to a microcentrifuge tube, pelleted, resuspended in 1 ml of cold cell lysis buffer (10 mM Tris pH 8.0, 10 mM NaCl and 0.2% NP-40/Igepal) and incubated on ice for 10 min. Nuclei were then pelleted, washed once using 800 µl of NEBuffer DpnII, pelleted and resuspended in 500 µl of NEBuffer DpnII. A final concentration of 0.3% sodium dodecylsulfate (SDS) was added and samples were incubated at 37 °C. for 1 h in a thermomixer (Eppendorf Thermomixer R or equivalent tabletop thermomixers). A final concentration of 1.8% Triton X-100 was then added to each sample, which was subsequently incubated at 37 °C for 1 h in a thermomixer. Nuclease-free water, 40 µl, as well as 300 U DpnII (New England Biolabs, catalog no. R0543M), was added and samples were digested at 37 °C overnight in a thermomixer. Another 300 U DpnII was added for an additional 4 h of digestion with mixing. The samples were then incubated at 65 °C in a thermomixer for 20 min. Nuclei were then pelleted, resuspended in 1× NEBuffer 2, with biotin-14-dATP, dTTP, dCTP and dGTP, and DNA polymerase I, Large (Klenow) Fragment (New England Biolabs, catalog no. M0210), and incubated at 37 °C in a thermomixer for 90 min. Ligation reaction was subsequently carried out in a total volume of 1.2 ml with 4,000 U T4 DNA Ligase (New England Biolabs) at 16 °C for 4 h, followed by 30 min at room temperature. Then 20 µl of 20 mg ml^{-1} of proteinase K and 120 µl of 10% SDS were added to each sample, followed by crosslinking reversal at 65 °C overnight. An additional 10 µl of proteinase K was added to each sample, which was incubated at 55 °C for 2 h. Then, 2 µl of DNase-free RNase was added to each sample, followed by incubation at 37 °C for 30 min. Phenol–chloroform extraction was performed to purify DNA.

DNA was sonicated to ~200–300 bp in a bath sonicator (QSonica, catalog no. Q800R3), followed by bead cleanup. Biotinylated nucleotide-filled ligation junctions were pulled down with 50 µl of Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific, catalog no. 65001). Sequencing libraries were subsequently prepared as described in Insertion site mapping and validation: end repair, dA-tailing and adapter ligation, followed by six cycles of indexing PCR.

The quality and size of each library were evaluated using the Agilent Bioanalyzer 2100 (Agilent Technologies), followed by quantification using RT-PCR with the KAPA Library Quant Kit for Illumina (KAPA Biosystems, catalog no. KK4835). Libraries were then pooled and sequenced in paired-end mode on the NextSeq 500 to generate 2× 75-bp reads using Illumina-supplied kits as appropriate.

Capture-C. Capture-C was performed as previously described[36,68,69]. Briefly, ~10 million HAP1 cells were crosslinked in 1% formaldehyde at room temperature for 10 min, and then quenched in 1 M glycine for 5 min, on an orbital shaker. Cells were then pelleted, washed with 1 ml of cold PBS, transferred to a microcentrifuge tube, pelleted, resuspended in 1 ml of cold cell lysis buffer (10 mM, Tris pH 8.0, 10 mM NaCl and 0.2% NP-40/Igepal) and incubated on ice for 10 min. Nuclei were then pelleted, washed once using NEBuffer DpnII, pelleted and resuspended in 500 µl of NEBuffer DpnII. Samples were incubated in 0.3% SDS at 37 °C for 1 h in a thermomixer, followed by the addition of a final concentration of 1.8% Triton X-100, for 1 h at 37 °C. After adding 40 µl of water, 300 U DpnII was added for in situ digestion at 37 °C, overnight. An additional 300 U DpnII was then added, and samples were incubated for 4 h more at 37 °C, followed by 20 min at 65 °C. Nuclei were pelleted, resuspended in a total volume of 1.2 ml with 4,000 U T4 DNA ligase (New England Biolabs) and incubated at 16 °C for 4 h, and then at room temperature for 30 min. Then, 20 µl of 20 mg ml^{-1} of proteinase K and 120 µl of 10% SDS were added, and samples were incubated at 65 °C overnight to reverse crosslinking. An additional 10 µl proteinase K was added to each sample for 2 h more at 55 °C. Finally, 2 µl of DNase-free RNase was added to each sample, followed by incubation at 37 °C for 30 min. DNA was then extracted with phenol–chloroform.

DNA was sonicated to ~200–300 bp in a bath sonicator (QSonica, catalog no. Q800R3), followed by a bead cleanup. Sequencing libraries were subsequently prepared similarly as described in Insertion site mapping and validation: end repair, dA-tailing, adapter ligation, six cycles of indexing PCR and six cycles of amplification using P5/P7.

After sequencing library preparation, two rounds of target enrichment/capture were carried out using biotinylated hybridization oligonucleotides (sequences in Supplementary Table 2) similar to the description in Insertion site mapping and validation. Pulldown enrichment was confirmed using qPCR. Target-enriched libraries were quality checked and quantified as in situ Hi-C samples before 2× 75-bp sequencing on the Illumina NextSeq 500.

ChIP. ChIP was performed as previously described[70], using 10 µg per ChIP H3K27ac antibody (Active Motif, catalog no. 39685), 10 µl of stock per ChIP CTCF antibody (Millipore, catalog no. 07-729) and 10 µg per ChIP RAD21 antibody (abcam, catalog no. ab992). Briefly, ~20 million HAP1 cells were fixed in 1% formaldehyde in fresh culture medium at room temperature for 10 min, followed by quenching in 1 M glycine for 5 min. Crosslinked cells were lysed for 10 min in 1 ml of cold cell lysis buffer (10 mM Tris, pH 8.0, 10 mM NaCl and 0.2% NP-40/Igepal), supplied with protease inhibitors (Sigma-Aldrich, catalog no. P8340) and phenylmethylsulfonyl fluoride (PMSF). Nuclei were pelleted, resuspended in 1 ml of room-temperature nuclei lysis buffer (50 mM Tris, pH 8, 10 mM EDTA, 1% SDS), with protease inhibitors and PMSF, and were incubated on ice for 20 min. The samples were sonicated at 100% amplitude, 30 s on:30 s off, for 45 min, in a bath sonicator (QSonica, catalog no. Q800R3). Sonicated materials were centrifuged, with the supernatant subsequently collected, and diluted with 4 ml of IP dilution buffer (20 mM Tris, pH 8, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100, 0.01% SDS), with protease inhibitors and PMSF. Then, 50 µl of protein A/G agarose beads (Thermo Fisher Scientific, catalog nos. 15918014 and 15920010) and 50 µg of isotype-matched immunoglobulin (Ig)G control were added to sonicated chromatin to pre-clear it for >2 h at 4 °C. Beads were then spun down, with 200 µl of supernatant containing pre-cleared chromatin saved as 'input' before immunoprecipitation. The remaining pre-cleared chromatin was split into equal volumes, each incubated with antibody or 10 µg per ChIP isotype-matched control (mouse IgG (Sigma, catalog no. I8140) or rabbit IgG (Sigma, catalog no. I8765)) pre-bound protein A/G beads and rotated overnight at 4 °C.

Chromatin-bound beads were washed on ice, once with IP wash 1 (20 mM Tris, pH 8, 2 mM EDTA, 50 mM NaCl, 1% Triton X-100, 0.1% SDS), twice with high-salt buffer (20 mM Tris, pH 8, 2 mM EDTA, 500 mM NaCl, 1% Triton X-100, 0.01% SDS), once with IP wash 2 (10 mM Tris, pH 8, 1 mM EDTA, 0.25 M LiCl, 1% NP-40/Igepal, 1% sodium deoxycholate) and twice with TE. Beads were then moved to room temperature and eluted twice with a total volume of freshly prepared 200 µl of elution buffer (100 mM $NaHCO_3$, 1% SDS). Into each IP and input, 12 µl of 5 M NaCl and 2 µl of RNase A (10 mg ml^{-1}, Roche through Sigma, catalog no. 10109169001) were added, and samples were incubated at 65 °C overnight. Then, 3 µl of proteinase K (20 mg ml^{-1}, Roche through Sigma, catalog no. 3115879) was added, for an additional 2 h at 65 °C. DNA was column cleaned using a QIAquick PCR Purification Kit (QIAGEN, catalog no. 28106).

For ChIP–seq, library construction was performed using Illumina's TruSeq ChIP sample preparation kit (Illumina IP-202–1012), followed by size selection using SPRIselect beads (Beckman Coulter, catalog no. B23318). Libraries were quality checked, quantified before 1× 75-bp sequencing on the Illumina NextSeq 500.

RNA-seq. HAP1 cells were washed with PBS, and resuspended in 1 ml of TRIzol (Thermo Fisher Scientific), with 200 µl of chloroform then added. RNA was extracted using RNeasy Mini Kit (QIAGEN). For the initial round of WT and transposon-engineered cell lines, sequencing libraries were constructed from 500 ng of DNase-treated total RNA using the ScriptSeq v.2 Complete Kit (Illumina BHMR1224). Briefly, the RNA was depleted of ribosomal (r)RNA using Ribo-Zero removal reagents and fragmented. First-strand cDNA was then synthesized using a 5′-tagged random hexamer and reverse transcription, followed by annealing of a 5′-tagged, 3′-end-blocked, terminal-tagged oligonucleotide and second-strand synthesis. The twice-tagged cDNA fragments were purified, barcoded and PCR amplified for 15 cycles. For the subsequent CRISPR-edited clones, as the ScriptSeq v.2 Complete Kit has been discontinued, TruSeq Stranded Total RNA (Illumina, catalog no. 20020598) was used following the manufacturer's instructions, which rely on a dUTP-based second-strand synthesis to preserve the stranded nature of RNA. Libraries were quality checked and quantified before 2× 76-bp sequencing on the Illumina NextSeq 500.

RT–qPCR. Extracted RNA was reverse transcribed using iScript Reverse Transcription Supermix (Bio-Rad), which contains a combination of oligo(dT) and random primers. Then, qPCR was carried out with Power SYBR Green (Thermo Fisher Scientific). Transcripts were normalized relative to the geometric mean of C_t values of 11 housekeeping genes. Supplementary Table 2 contains all RT–qPCR primer sequences.

CRISPR genome editing. The guide RNA (gRNA) targeting the CBS within the 2-kb insert was designed using the Benchling CRISPR gRNA design tool. Oligos (Supplementary Table 2) encoding this gRNA were annealed and cloned into a plasmid co-expressing Cas9 and gRNA, with GFP, modified from pX330 (Addgene, catalog no. 42230). C21 HAP1 cells were transfected with this pX330 using Nucleofector Kit L (Lonza) and program X-001 on an Amaxa electroporator (Lonza). Then 24 h post-transfection, GFP$^+$ cells were sorted on a FACSJazz sorter (BD Biosciences) as single cells, which were expanded as clonal populations and subsequently subcloned. To genotype genome edits at each transposon insertion locus, PCR, Sanger sequencing and Inference of CRISPR Edits[71] were performed. Clones with CBS disruptions at insertion loci where new domains had formed were identified for downstream characterizations. A CRISPR, ribonucleoprotein (RNP)-based[72], paired-cut approach was used for subsequent rounds of editing. For TSS editing, ΔTSS cells were derived from C21 cells, whereas ΔCTCF/ΔTSS cells were derived from the subclone with CBS disruptions at the C21S2 and C21S4 insertion loci that were previously derived from C21. Cells with deletion of the endogenous 2-kb element (WT 2 kb Del) were derived from WT cells. Specifically, 150 pmol per single guide (sg)RNA (2 sgRNAs, thus 300 pmol in sgRNA) (Synthego), along with 150 pmol spCas9 protein (Synthego), were mixed and incubated at room temperature for 10 min. The RNP mixture was then added to Nucleofector Kit L solution (Lonza) with 2 µg of pmaxGFP plasmid (Lonza), and this transfection solution was then added to ~1 million trypsinized and pelleted HAP1 cells. The resuspended cells were transfected using program X-001 on an Amaxa electroporator. GFP$^+$ (probably transfection-positive) cells were FACS sorted the next day as a pooled population. This transfected population was then FACS sorted into single cells ~4–5 d later. After ~2 weeks of growth, clonal cells were characterized by RNA using RT–qPCR (TSS-edited) and DNA (for 2-kb deleted and selected TSS-edited clones with decreased transcription). Clonal cells with TSSs edited at both C21S2 and C21S4 insertion loci, along with clones with a complete 2-kb deletion, were expanded for further characterizations.

Analyses. *Boundary-underlying DNA selection.* The selection process was summarized in Supplementary Fig. 1a. Briefly, K562 boundary coordinates (hg19) were obtained from the domain calls using Arrowhead[4] (Gene Expression Omnibus (GEO): accession no. GSE63525), with domain start and end coordinates extended 5 kb both upstream and downstream. Human embryonic stem cell (hESC) boundary coordinates were obtained from the domain calls using the DI[3] (http://chromosome.sdsc.edu/mouse/hi-c/download.html), with domain start and end coordinates extended 20 kb both upstream and downstream. K562 and hESC boundaries were intersected using BEDTools[73]. This shared list of boundaries was then intersected with K562 CTCF ChIP–seq peaks[74,75] (DCC accession no.: ENCSR000EGM, Michael Snyder Lab, ENCODE Consortium), narrowing down to a list of shared boundaries with K562 CTCF binding. This list was subsequently intersected with the top 100 K562 CTCF-binding sites ranked by the number of co-bound putative architectural proteins[24]. The boundary at chr3: ~183,600,000 was chosen as the candidate domain boundary, after visual examination of the K562 Hi-C heatmap[4] (GEO, accession no. GSE63525). As shown in Supplementary Fig. 1, this candidate boundary was further confirmed to be present in GM12878, HMEC, HUVEC, IMR90 and NHEK cell lines[4] (GEO, accession no. GSE63525), and by additional analytical methods[76,77] such as Insulation Score[78] and Armatus[79]. The CBS at this boundary, which is among the top 100 K562 CBSs ranked by architectural protein co-occupancy, was verified to have the bindings of CTCF (DCC accession no. ENCSR000AKO, Bradley Bernstein Lab, ENCODE Consortium), SMC3 (DCC accession no. ENCSR000EGW, Michael Snyder Lab, ENCODE Consortium), RAD21 (DCC accession no. ENCSR000FAD, Sherman Weissman Lab, ENCODE Consortium) and Pol2 (DCC accession no. ENCSR000FAY, Sherman Weissman Lab, ENCODE Consortium) in K562 (Supplementary Fig. 1c). In HAP1, similarly, this site was also bound by CTCF and SMC1 (ref. [23]) (GEO, accession nos. GSM2493878 and GSM2493882). The candidate 2-kb DNA fragment was chosen to have the CTCF/Cohesin binding site in the middle; it also contains a TSS of a housekeeping gene: *PARL*[80].

Insertion-site mapping. Demultiplexed Illumina sequencing reads underwent adapter trimming, with read pairs removed if both reads mapped to the inversed repeat. Read1 of the remaining read pairs was further filtered to those that had a partial match to the inversed repeats, with right-trimming performed to remove these inverse repeat sequence fragments. All these steps were carried out using the BBDuk tool (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide). Adapter- and inverse repeat-trimmed sequence fragments were then mapped to hg19 using Bowtie 2 (ref. [81]). The resulting SAM files were converted to BAM files, which were then sorted, both using SAMtools[82]. Genome-wide coverage was then obtained using BEDTools[73]. Genome coordinates with >25× coverage were visualized as peaks, and the insertion sites were identified to be between the two bases with the highest coverage in each peak. Each insertion site was subsequently validated using targeted PCRs.

Hi-C data processing. Two replicates of each sample, WT, C21, C25 and all the CRISPR-edited subclones, underwent a pilot sequencing run, generating at least ~30 million raw reads for each replicate. Reads from each replicate were then mapped to hg19 using Bowtie 2 (ref. [81]). Detection and filtering of valid interaction pairs, assignment to restriction fragment and binning, and interaction matrix balancing with iterative correction and eigenvector decomposition (ICE)[83] were all performed using the HiC-Pro pipeline[84]. With each replicate of a given sample having highly reproducible metrics—>~58% valid interaction pairs:raw read pairs ratio, <~18% *trans* interaction and contact ranges of interaction pairs—two replicates of the sample were subsequently pooled for further analysis.

Deeper sequencing was subsequently performed to yield ~248–300 million raw reads for each sample (with subsampling performed where necessary), generating ~161 million to ~173 million valid interaction pairs per sample, after HiC-Pro processing (Supplementary Table 3). ICE-balanced interaction matrices were binned at 20-kb resolution, and used for downstream analyses, unless otherwise specified. Genome-wide Hi-C contact Pearson's correlations between two samples were calculated on nonempty bins shared between two samples.

Heatmap generation. To generate heatmaps for each insertion site, ICE-balanced matrices for the clone with the insertion and the subclones with edited derivatives, as well as for WT and the other clones without the insertion at this position, were extracted with the insertion bin in the middle and 150 bins (3 million bp) on either side of it. These extracted matrices, adjusted for minor coverage differences via division by a scaling factor reflecting the total number of valid interactions, were plotted using lib5C[85], the dependencies of which include pandas, scipy and numpy, in linear scale using color scheme 'red8' from matplotlib[86], highlighting a megabase-scale region centering around each insertion. Linear features extracted as .BedGraph from BigWig format, using UCSC kentUtils accompanying each Hi-C heatmap included: CTCF motif orientation, obtained through PWMScan[87], using the JASPAR 2018 (ref. [88]) CTCF motif, with a P value cutoff of 5×10^{-5} and overlapped with: CTCF peaks identified from our ChIP–seq data with a signal cutoff of 20; RNA-seq; eigenvectors reflective of compartment states; and CTCF and RAD21 ChIP–seq.

Insulation score. Insulation score computations were implemented in R, mechanistically similar to those in Crane et al.[78]. Given an ICE-balanced interaction matrix with the insertion bin at the center, averaged interactions within a sliding window of 10×10 bins^2 (15×15 bins^2 when the insert was at a pre-established boundary), were recorded for each bin along the diagonal of interaction matrices. The insulation score for each diagonal bin was then normalized to all insulation scores across its nearby 240 bins (4.8 Mb chromosomal region), with the $\log_2$(ratio) subsequently calculated. Positive $\log_2$(values) indicate relatively enriched interactions, whereas negative $\log_2$(values) suggest relatively depleted interactions, with local minima marking domain boundaries. Shared domain boundaries genome wide among WT and insertion clones were identified similarly: averaged interactions within a sliding window of 10×10 bins^2, $\log_2$(values) calculated after being normalized to all scores chromosome wide and local minima determined, with a minimum domain size of 4 bins (80 kb).

Compartment analysis. After HiC-Pro processing, allValidPairs format was converted to .hic format using Juicer Tools[89]. Eigenvectors, the first principal component of the distance-adjusted correlation matrix, were calculated for each of the selected chromosomes in *cis* (intrachromosomal) with insertions with KR normalization at 50-kb resolution using Juicer Tools[89]. The signs were manually adjusted when necessary to reflect active/inactive chromatin states, based on H3K27ac ChIP- and RNA-seq data we generated. Adjusted eigenvalues around the insertions were then plotted.

Capture-C. Two biological replicates were performed for parental WT cells and for most cell lines. In subclones with the deletion of TSSs and the deletion of both CBSs and TSSs, four biological replicates were performed. Raw reads were processed using published scripts[69]. Briefly, Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore), a wrapper for FastQC and Cutadapt[90], was used to access the quality and to trim adapter sequences of raw sequences. Trimmed reads were then merged or interleaved using FLASH[91]. Once concatenated, the reads were in silico DpnII digested, aligned to hg19 and analyzed using CCanalyser3 (ref. [69]). Interactions were then pooled from all replicates for each genotype and were normalized to total interactions. To compare how the insertion element and its edited derivatives alter how the immediate insertion-proximal chromatin folds, we used DIs[3] to quantify directional interaction preference over a series of distance ranges. Specifically, we focus on the restriction fragment targeted by the capture oligonucleotides as the center, and computed its DI for each of the distance ranges 50 kb, 100 kb, 250 kb, 500 kb and 1 Mb, respectively, using the equation:[3]

$$\mathrm{DI} = \left(\frac{(B-A)}{|B-A|}\right)\left(\frac{(A-E)^2}{E}+\frac{(B-E)^2}{E}\right)$$

where B is the sum of the captured fragment's normalized interactions to the right (in increasing genome coordinate) within a given distance range, while A is the sum of the captured fragment's normalized interactions to the left (in decreasing genome coordinate). E is the expected number of contacts under null hypothesis, showing no contact preference: $E=(A+B)/2$. Conceptually, this quantification approach detects at which distance the directional contact preference introduced by the inserted CBS/TSS, or ΔCTCF, ΔTSS or ΔCBS/ΔTSS, elements emerges, propagates or reduces.

ChIP–seq. Raw sequencing reads (Supplementary Table 4) were mapped to hg19 using Bowtie[92]. SAM files were converted to BAM files using SAMtools. For histone marks (H3K27ac), BAM files were converted to Wiggle using model-based analysis of ChIP–seq (MACS)[93] and subsequently converted to bigWig for visualization. BAM files were converted to BED for subsequent peak calling using Sicer[94]. For transcription factors or factors with narrower peaks as in CTCF and RAD21 in our case, MACS was used for peak calling and generating Wiggle files. For the ChIP–seq data tracks visualized in detail (such as those in Figs. 3, 4 and 5), bamCoverage from Deeptools[95] was used to generate counts-per-million (c.p.m.)-normalized bigWig files from bam files. The c.p.m.-normalized bigWig files were converted to bedgraphs to accompany Hi-C heatmaps. DiffBind[96] was used for differential binding analysis of CTCF and RAD21 near insertions. Comparisons between C21 and non-C21, which consists of three cell lines without C21 insertions—WT, WT with deletion of the endogenous 2-kb (WT 2-kb Del) and C25, each with two ChIP–seq replicates—were performed using the consensus peak set, which includes peaks identified in at least two replicates and re-centered to include 250 bp upstream and downstream from consensus summits. Each pairwise comparison between C21 CTCF/TSS and other CRISPR subclones derived from it was based on two ChIP–seq replicates of each cell line (genotype). The consensus peak set for these comparisons includes peaks identified in at least three individual replicates among ChIP–seq data of the same factor for all samples, with each peak again re-centered to include 250 bp upstream and downstream from a consensus summit. P values are derived after performing a negative binomial Wald test through DiffBind[96] (http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf).

RNA-seq. Two biological replicates were performed for each cell line. For differential expression (DE) analysis, reads were mapped to the indexed Ensembl hg19 transcriptome (GRCh37.67) and quantified using Salmon[97]. Transcript-level quantifications were then imported in R, using tximport[98], for gene-level (EnsDb.Hsapiens.v75) measurements with DESeq2 (ref. [99]). Genes with fewer than 10 reads were excluded from downstream DE analysis. DESeq2 was performed to identify DE genes between C21 and non-C21, and between C25 and non-C25, at a false discovery rate < 0.01.

For RNA-seq analysis for genome track visualization, the sequence reads were processed using the ENCODE long RNA-seq pipeline (https://www.encodeproject.org/pipelines/ENCPL002LPE). Briefly, raw sequencing reads were initially accessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Mapping to hg19 was performed using STAR[100], with the default normalization reads per million mapped reads (or c.p.m.) unchanged. Resulting bedGraph files were then converted to the bigWig format for genome browser visualization with strand specificity. Insertion-proximal regions were extracted from bigWig files as bedGraph to accompany Hi-C heatmaps.

Recent evolution of the mouse architectural genome. Ancestral CTCF binding, defined as CTCF peaks shared among humans, macaques, mice, rats and dogs, and presumably in their common placental ancestor, was obtained from Schmidt et al.[56] (http://ftp.ebi.ac.uk/pub/databases/vertebrategenomics/FOG03/). More recent, highly rodent-specific, CTCF–Pol III TSS insertions expanded by SINE B2 elements, were obtained from Thybert et al.[58] (http://ftp.ebi.ac.uk/pub/databases/vertebrategenomics/FOG21/repeatPeaks/mus_musculus_RepeatAssociated.txt). These two CTCF peak lists were intersected with CTCF binding in G1E-ER4 cells[101], a mouse erythroid cell line, for which we have generated both CTCF binding and in situ Hi-C data[54], to obtain ancestral CTCF peaks as well as recently gained, rodent-specific SINE B2 CTCF binding for downstream analysis. Mouse genome domain boundaries (20-kb resolution) were obtained using 3DNetMod[102] on in situ Hi-C data[54] from G1E-ER4 cells. Putative, recently gained genome domain boundaries associated with SINE B2 CTCF expansion were defined as boundaries that co-localize with SINE B2 CTCF binding, and not with ancestral CTCF binding. TSS annotations were obtained from intersecting RefSeq and University of California, Santa Cruz known gene databases, with 1 bp upstream output. All SINE B2 CTCF sites were then grouped by nearby TSS density: the number of TSSs within a 200-kb window. Finally, Fisher's exact test (two-sided) was used to test whether CTCF-bound SINE B2 insertions in regions with 0 TSSs and ≥11 TSSs within a 200-kb range differentially co-localize with putative, recently gained SINE B2 CTCF-associated domain boundaries, based on a 2×2 contingency table.

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

Data availability

All main, extended data and supplementary figures include publicly available data. All Hi-C, Capture-C, RNA-seq, ChIP–seq, and other applicable next-generation sequencing raw data and processed data generated from the present study are available under accession no. GSE137376 (GEO database). Mouse CTCF ChIP–seq and mouse Hi-C domain boundaries (both asynchronous) shown in Fig. 6a–c are derived from Zhang et al.[19] (https://doi.org/10.1038/s41586-019-1778-y), accession no. GSE129997 (GEO database). In Supplementary Fig. 1: Hi-C heatmaps from all cell lines, except for HAP1, are from GEO, accession no. GSE63525 by Rao et al.[4] (https://doi.org/10.1016/j.cell.2014.11.021); K562 ChIP–seq data are from

ENCODE, CTCF (DCC accession no. ENCSR000AKO), SMC3 (DCC accession no. ENCSR000EGW), RAD21 (DCC accession no. ENCSR000FAD) and Pol2 (DCC accession no. ENCSR000FAY). Source data are provided with this paper.

Code availability

Code used in the present study is available upon request as well as on GitHub (https://github.com/dizhmp/boundary-insertion).

References

66. Kurita, R. et al. Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. *PLoS ONE* **8**, e59890 (2013).
67. Zayed, H., Izsvák, Z., Walisko, O. & Ivics, Z. Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Mol. Ther.* **9**, 292–304 (2004).
68. Huang, P. et al. Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes Dev.* **31**, 1704–1713 (2017).
69. Davies, J. O. J. et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
70. Hsiung, C. C.- et al. A hyperactive transcriptional state marks genome reactivation at the mitosis-G1 transition. *Genes Dev.* **30**, 1423–1439 (2016).
71. Hsiau, T. et al. Inference of CRISPR edits from Sanger trace data. Preprint at *bioRxiv* https://doi.org/10.1101/251082 (2019).
72. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).
73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
74. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
75. Sloan, C. A. et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, 726–732 (2016).
76. Kerpedjiev, P. et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
77. Forcato, M. et al. Comparison of computational methods for Hi-C data analysis. *Nat. Methods* **14**, 679–685 (2017).
78. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
79. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).
80. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
81. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
82. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
83. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
84. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
85. Gilgenast, T. G. & Phillips-Cremins, J. E. Systematic evaluation of statistical methods for identifying looping interactions in 5C data. *Cell Syst.* **8**, 197–211.e13 (2019).
86. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
87. Ambrosini, G., Groux, R. & Bucher, P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* **34**, 2483–2484 (2018).
88. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
89. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
90. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
91. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
92. Langmead, B. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinform.* **Chapter 11**, Unit 11.7 (2010).
93. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
94. Xu, S., Grullon, S., Ge, K. & Peng, W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol. Biol.* **1150**, 97–111 (2014).
95. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
96. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
97. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* **14**, 417–419 (2017).
98. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2015).
99. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
100. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
101. Weiss, M. J., Yu, C. & Orkin, S. H. Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line. *Mol. Cell. Biol.* **17**, 1642–1651 (1997).
102. Norton, H. K. et al. Detecting hierarchical genome folding with network modularity. *Nat. Methods* **15**, 119–122 (2018).

Acknowledgements

We thank B. van Steensel (Netherlands Cancer Institute) for providing HAP1 cells; Z. Izsvák (Max Delbrück Center) and Z. Ivics (The Paul Ehrlich Institute) for providing the Sleeping Beauty transposon constructs; A. Raj, O. Symmons and F. Yue for helpful comments on the manuscript. We thank the Flow Cytometry Core at the Children's Hospital of Philadelphia; J. Yano and P. Evans for assistance; and members of the Blobel laboratory for helpful discussions. This work was supported by grants (nos. R01DK054937 and U01HL129998A to G.A.B and R24DK106766 to R.C.H. and G.A.B.). This work was also supported by the Spatial and Functional Genomics program at the Children's Hospital of Philadelphia.

Author contributions

D.Z. and G.A.B. conceived the study and designed the experiments. D.Z. performed a large majority of the experiments, analyzed all datasets and interpreted the results. P.H. conducted Hi-C and Capture-C for half the replicates for transposon-edited and control cell lines, and helped with Hi-C and Capture-C analysis and interpretation. M.S. helped generate and characterize cell lines derived from CRISPR targeting the TSSs and the 2-kb elements. C.A.K., B.G. and R.C.H. prepared ChIP–seq and RNA-seq libraries, performed all sequencing, uploaded sequencing data, and conducted RNA-seq alignment and ChIP–seq peak calling. H.Z. generated mouse ChIP and Hi-C datasets used for recent mouse genome evolution analysis. T.G.G. and J.E.P.-C. helped with Hi-C data visualization and interpretation. D.Z. and G.A.B wrote the paper with input from all authors.

Competing interests

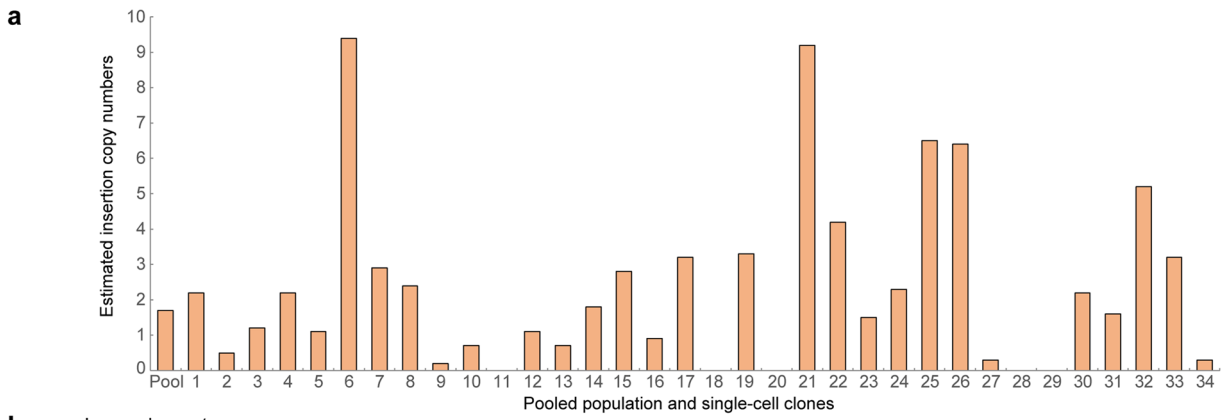
The authors declare no competing interests.

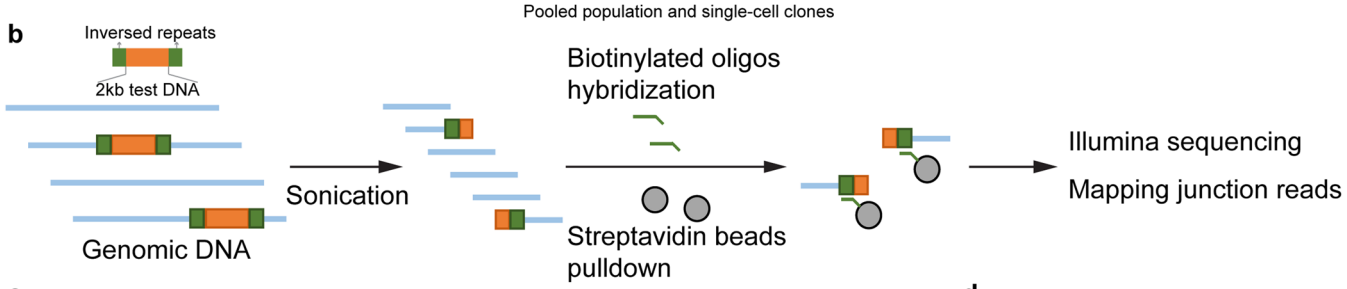
Additional information

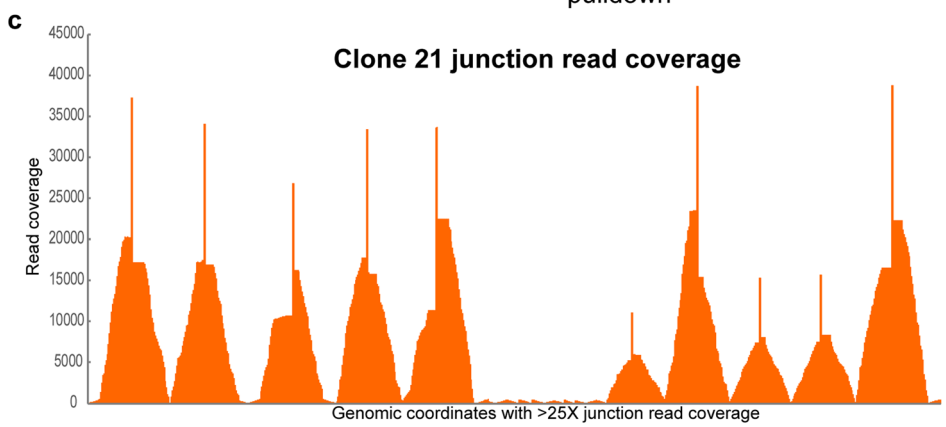
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0680-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0680-8.

Correspondence and requests for materials should be addressed to D.Z. or G.A.B.



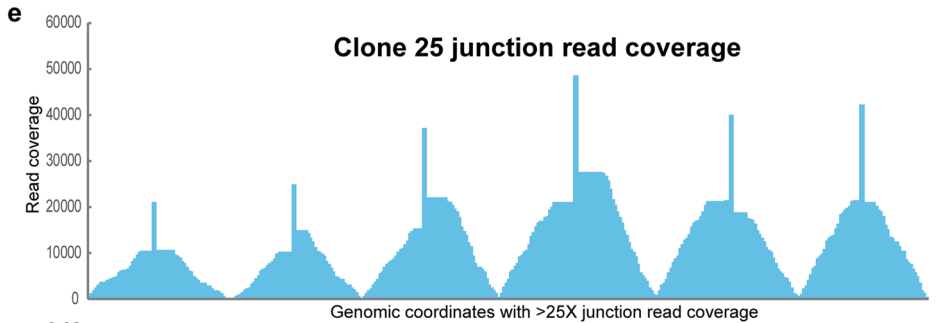




d

Clone 21 insertion sites

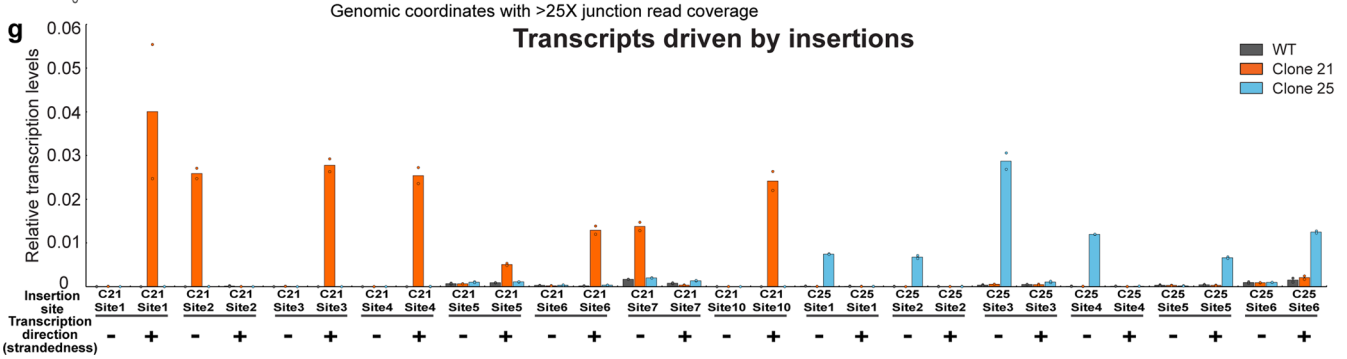
Insertion site	Chromosome	Coordinate	Orientation
C21S1	chr1	160133892	→
C21S2	chr2	13445903	←
C21S3	chr4	34839665	→
C21S4	chr6	75063309	→
C21S5	chr6	102281045	→
C21S6	chr13	67702536	→
C21S7	chr14	36298742	←
C21S8	chr15	20278576	←
C21S9	chr15	21285765	←
C21S10	chr15	68703456	→



f

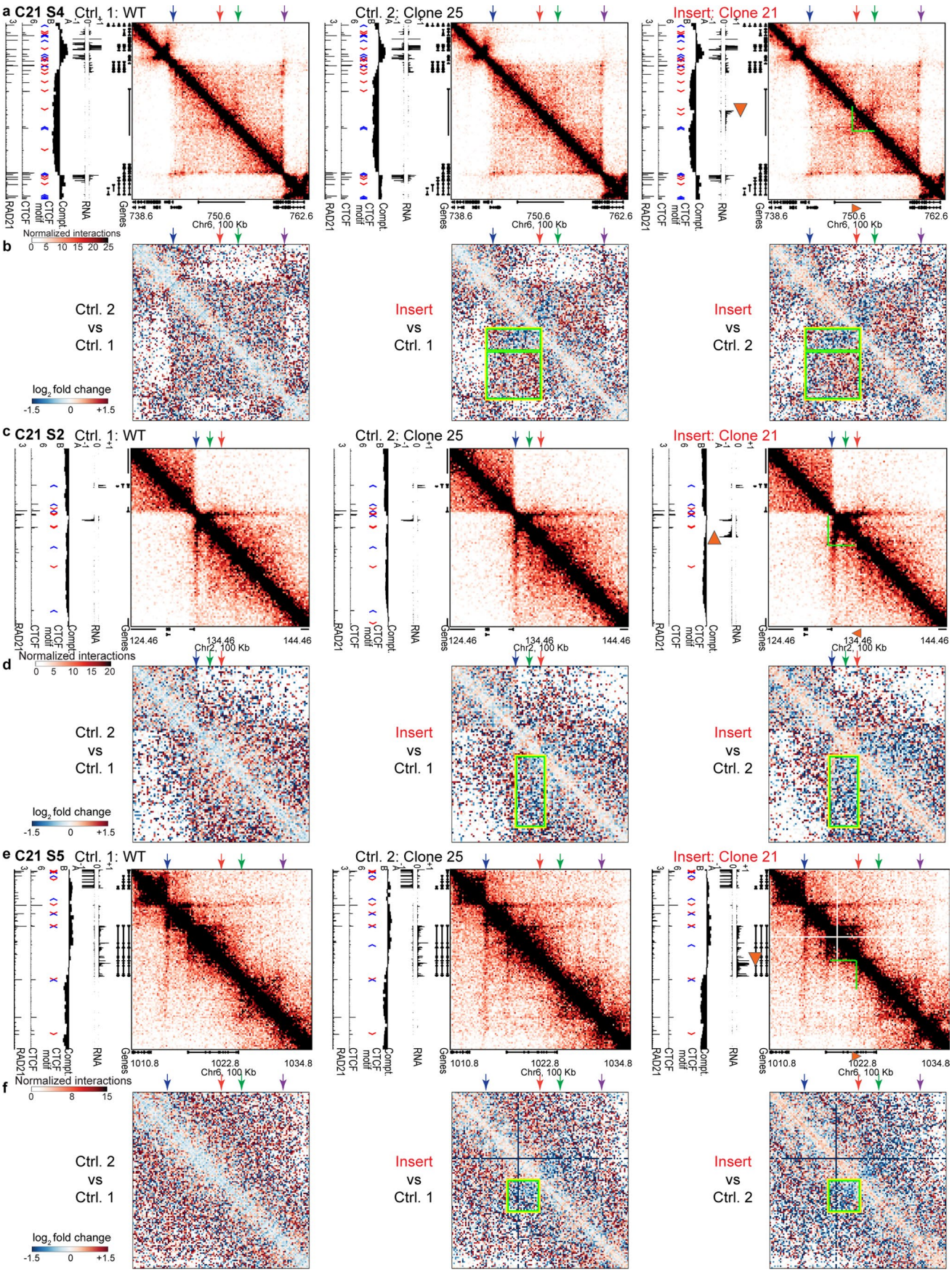
Clone 25 insertion sites

Insertion site	Chromosome	Coordinate	Orientation
C25S1	chr2	213004989	←
C25S2	chr2	232521564	←
C25S3	chr5	173339152	←
C25S4	chr7	20789579	←
C25S5	chr10	60518172	→
C25S6	chr16	74651969	→



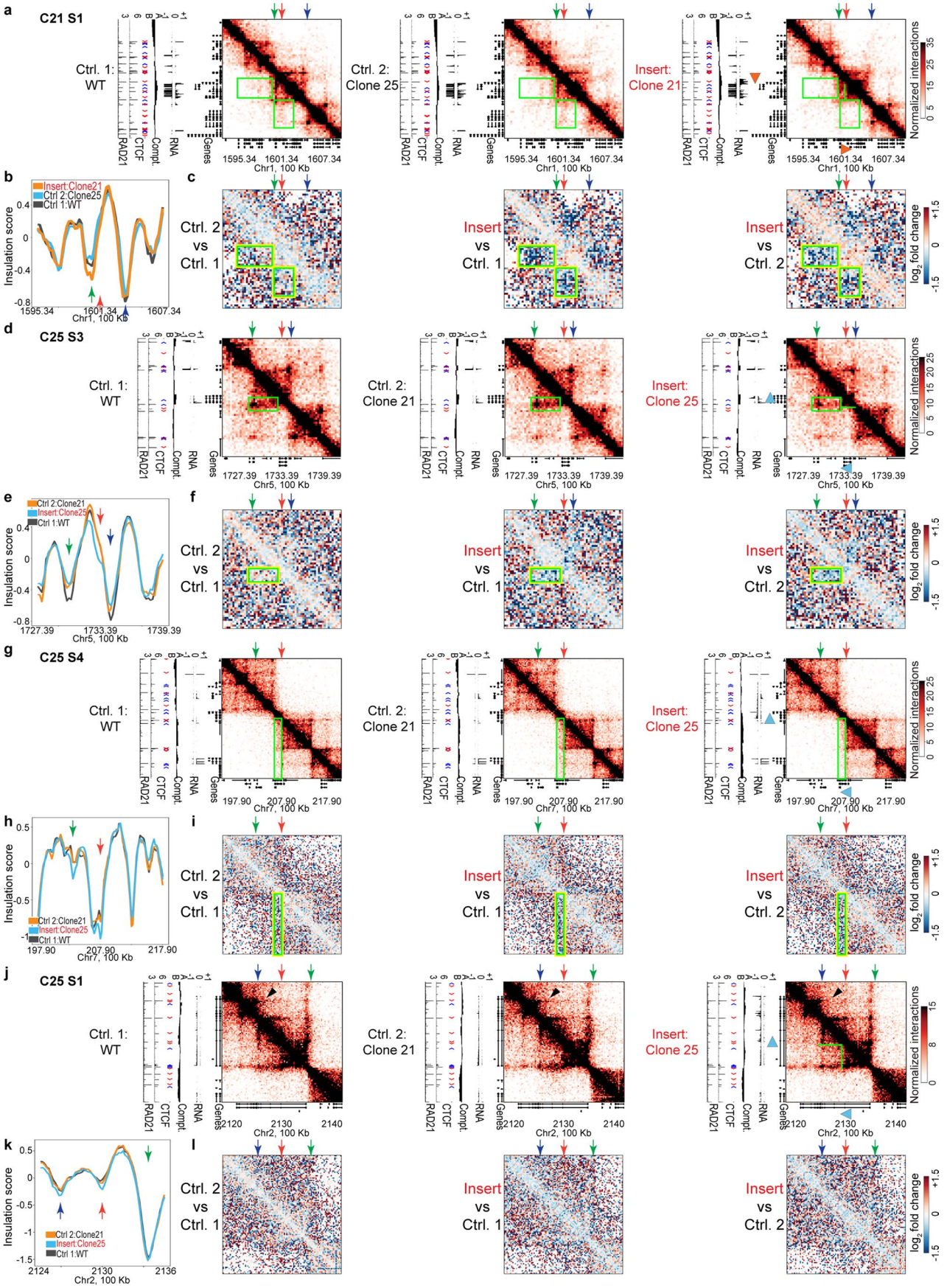
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Generation and characterization of transposon genome-edited clones with multiple insertions. a, Estimated insertion copy numbers using qPCR (see Methods) after transposon insertion in pooled cells and in single-cell-derived clones (numbered). N = 1 qPCR measurement. **b**, Insertion site mapping: fragmented gDNA containing insertions are captured by biotinylated oligos capturing the inversed repeats (green rectangles), which flank the 2 kb element (orange rectangles). Junction reads are mapped to identify insertion sites. **c**, Junction read coverage for Clone 21: horizontal axis denotes genomic coordinates (single nucleotide resolution) with > 25X coverage; vertical axis shows read coverage. The spike in the middle of each peak consists of two neighboring nucleotides between which an insertion is located. Data from N = 1 experiment. **d**, The locations and orientations of Clone 21 insertion sites. The CBS and TSS are in *cis* (Fig. 1a). "→" denotes that the CBS is on the plus strand and that the TSS transcribes from left to right, and vice versa for "←". Each insertion site orientation was confirmed in (**g**). **e**, Junction read coverage for Clone 25, similar to (**c**). Data is from N = 1 experiment. **f**, The locations and orientations of Clone 25 insertion sites, similar to (**d**). **g**, Insertion-driven transcription in both directions/strands measured by quantitative PCR with reverse transcription (RT-qPCR). Transcript levels were normalized relative to the geometric mean of the Ct values of 11 housekeeping genes. N = 2 independent experiments for each genotype.



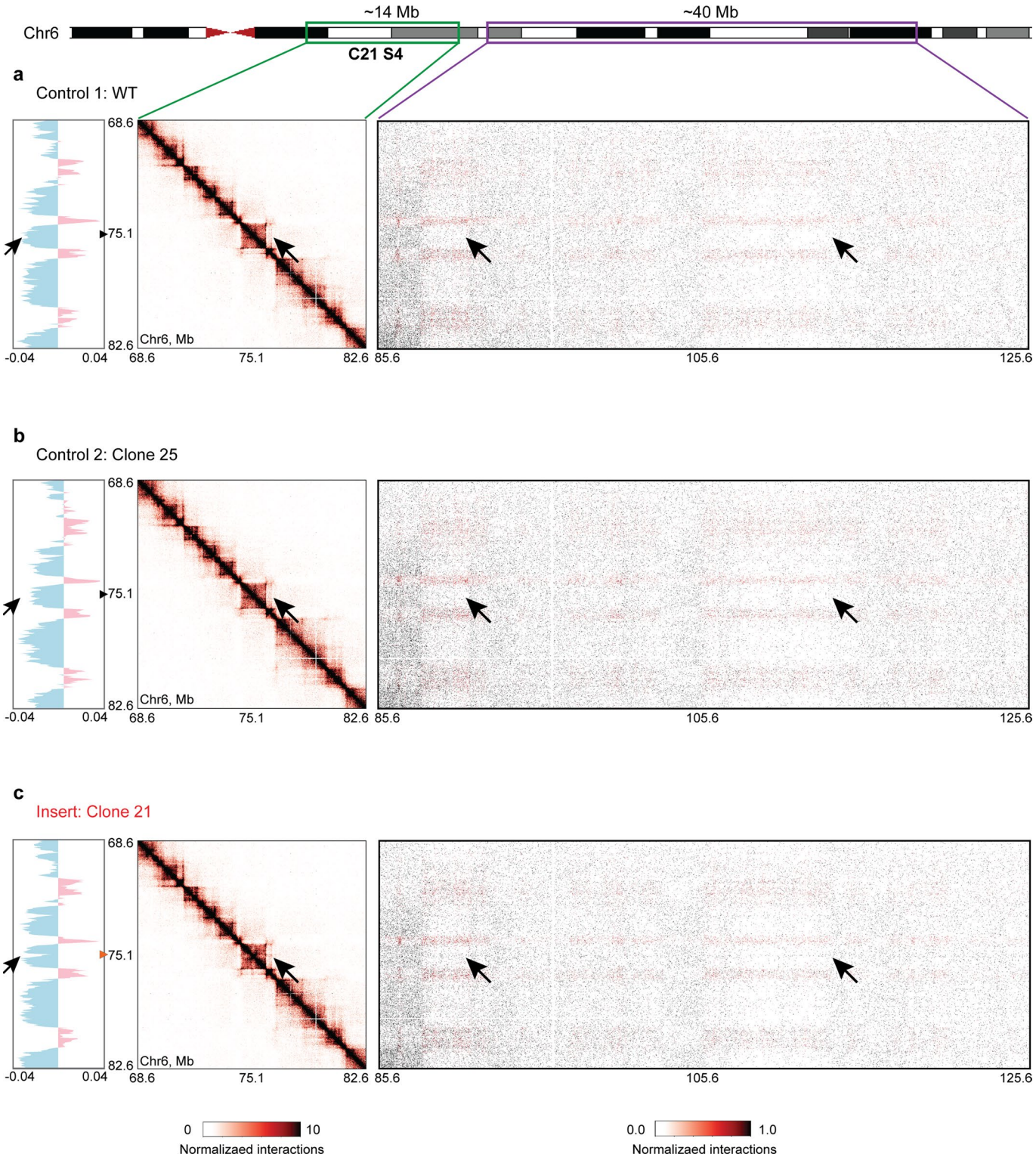
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Insertion-driven new domains: detailed comparisons (an extension to Fig. 1). Throughout, red arrow: insertion site; green arrow: upstream or downstream CBSs; blue/purple arrow: nearby boundaries; orange arrowhead in the browser tracks: site and orientation of the insertion. Green lines demarcate new domains. Yellow/green rectangles (squares) indicate regions with overall depleted (enriched) contacts upon insertion. (**a**, **b**): related to Fig. 1b, c. **a**, An extension to Fig. 1b showing Hi-C maps for both no-insertion controls (left and middle) and the insertion clone (right) at C21S4, each accompanied by corresponding data tracks. **b**, Log2 fold changes in interaction frequencies between two no-insertion controls (left), and between the insertion clone and no-insertion controls (middle and right) for the region in (**a**). Yellow/green rectangles: depleted interactions upon insertion; yellow/green squares: increased interactions between two B-compartment domains partitioned by the new domain with A compartment signature. (**c**, **d**): related to Fig. 1d, e. **c**, An extension to Fig. 1d showing both no-insertion controls at C21S2. **d**, Log2 fold changes in interaction frequencies between no-insertion controls and between insertion and no-insertion controls for the region in (**c**). (**e**, **f**): related to Fig. 1f, g. **e**, An extension to Fig. 1f showing both no-insertion controls at C21S5. **f**, Log2 fold changes between no-insertion controls and between insertion and no-insertion controls for the region in (**e**). Each Hi-C heatmap presents merged data from 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq and 2 RNA-seq experiments were performed for each genotype, with 1 of each displayed.

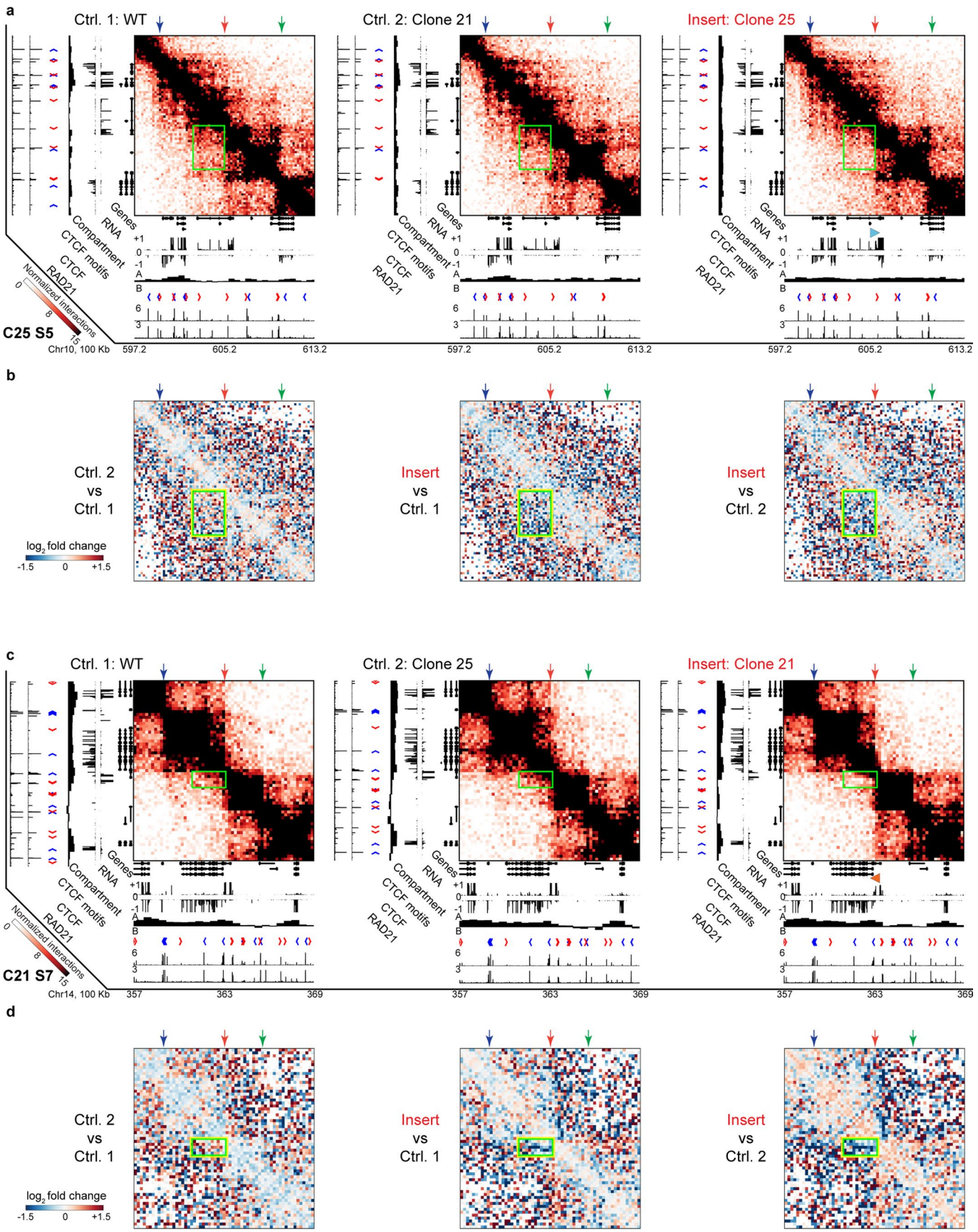


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Additional insertion loci with possible domain-level changes. Throughout, red arrow: insertion site; green or blue arrow: nearby boundaries; orange/blue arrowhead in the browser tracks: site and orientation of insertion. Green lines demarcate (possible) new domains. Yellow/green rectangles indicate regions with overall depleted contacts upon insertion. **a**, *De novo* domain upon insertion at C21S1: Hi-C maps for both no-insertion controls (left and middle) and the insertion clone (right) at C21S1, each accompanied by corresponding data tracks. **b**, Insulation scores for the region in (**a**). **c**, Log2 fold changes in interaction frequencies between the two no-insertion controls (left) and between the insertion clone and no-insertion controls (middle and right) for the region in (**a**). **d**, A small subtle domain forms upon insertion at C25S3 locus. **e**, Insulation scores for the region in (**d**). **f**, Log2 fold changes in interaction frequencies for the region in (**d**). **g**, Modest strengthening of an existing boundary upon insertion at C25S4. **h**, Insulation scores for the region in (**g**). **i**, Log2 fold changes for the region in (**g**). **j**, Subtle strengthening of an existing boundary upon insertion at C25S1. The black arrowheads point at insertion-associated changes. **k**, Insulation scores for the region in (**j**). **l**, Log2 fold changes for the region in (**j**). Each Hi-C heatmap presents merged data from 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq and 2 RNA-seq experiments were performed for each genotype, with 1 of each displayed.

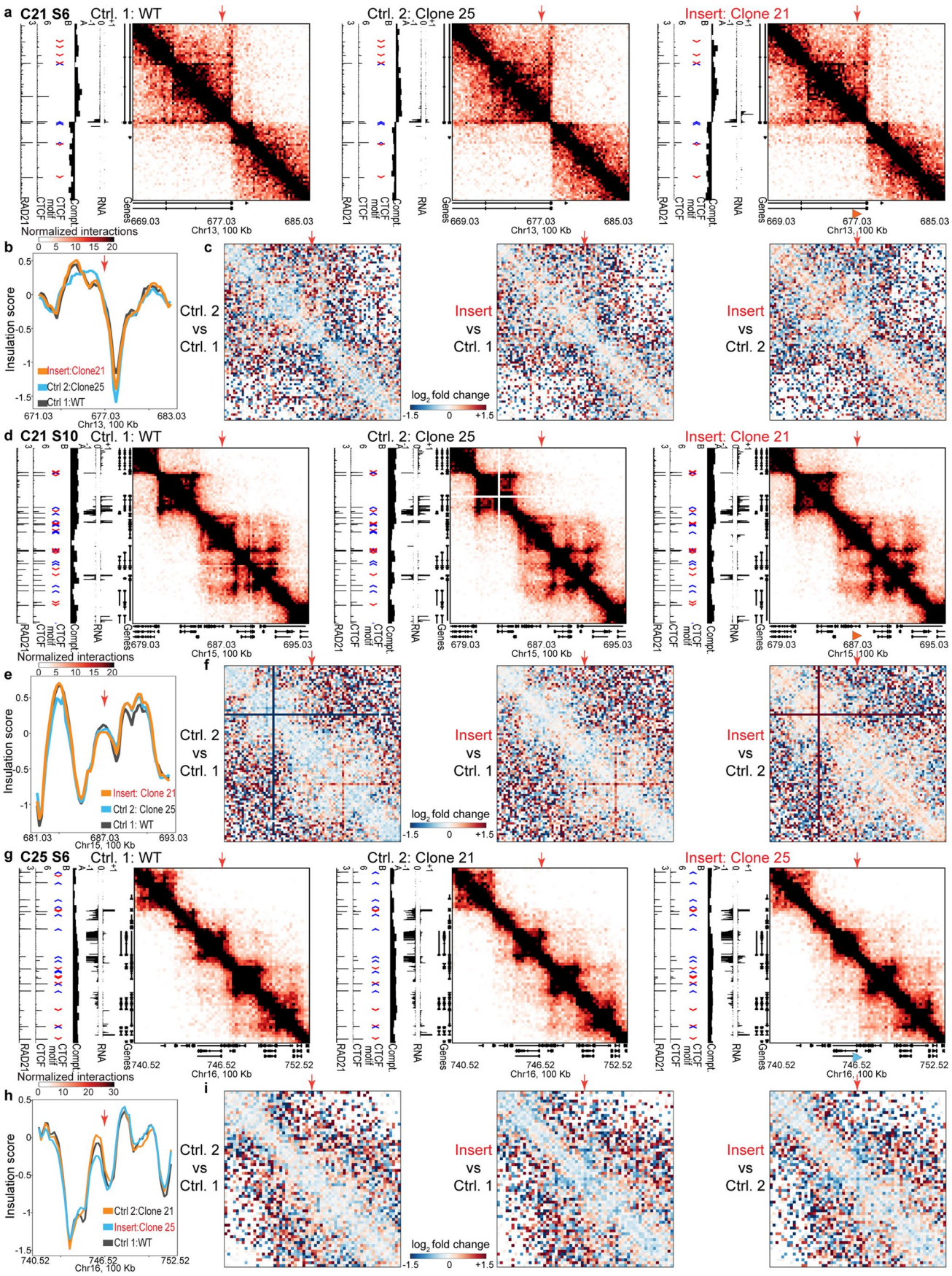


Extended Data Fig. 4 | An ectopic insertion can redirect its local chromatin from B to A compartment. Throughout, left: compartment eigenvectors (cyan denotes B compartment; red denotes A compartment) for the ~14 Mb region marked by the green rectangle on the chromosome diagram; middle: Hi-C heatmaps for this ~14 Mb region surrounding C21S4; right: distal interactions between this ~14 Mb region and a ~40 Mb region downstream marked by the purple rectangle. Black arrows: compartment switch; orange arrowhead: location of the insertion; black arrowhead: corresponding locations in no-insertion controls. **a**, No-insertion control 1 (WT) at C21S4. **b**, No-insertion control 2 (Clone 25) at C21S4. **c**, Insertion clone (Clone 21) at C21S4: compartment eigenvectors demonstrate the insertion locus trending from a strong B compartment towards A as the largest change in the region. The Hi-C heatmap for the ~14 Mb with the insertion at the center shows a plaid like pattern, with gained interactions between the insertion locus and its nearby A compartment regions. Distal interactions (right) shows the insertion locus forming distal interactions with other A-compartment regions (black arrows), which are absent in (**a**, **b**). Each Hi-C result depicts merged data from 2 independent Hi-C experiments for each genotype.



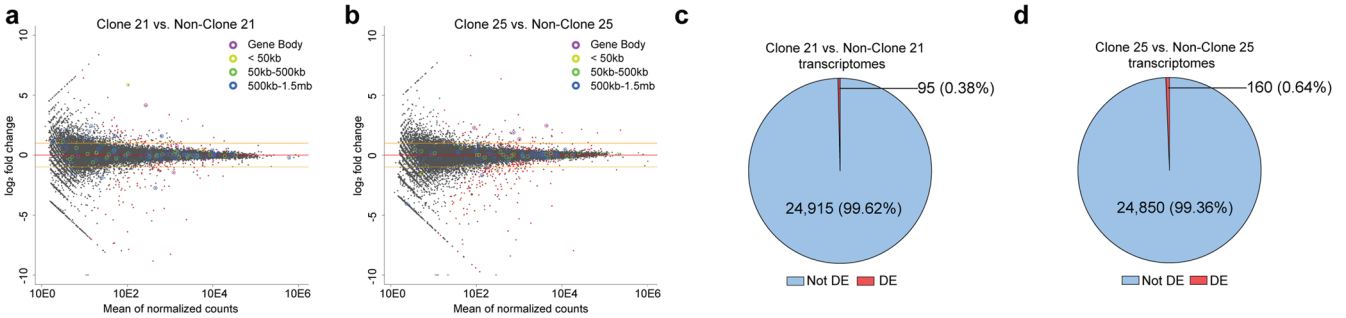
Extended Data Fig. 5 | See next page for caption.

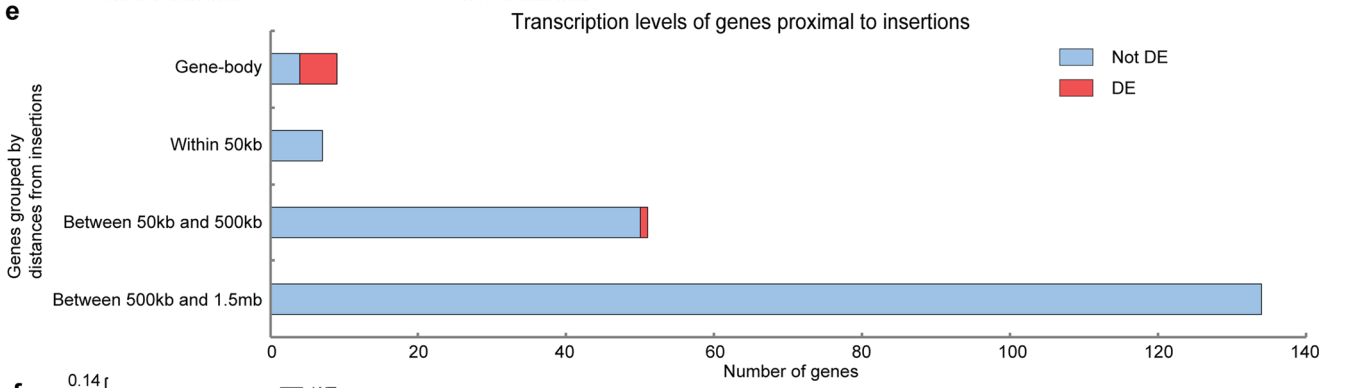
Extended Data Fig. 5 | Boundary-associated DNA insertions can strengthen pre-established boundaries: additional controls (an extension to Fig. 2). Throughout, red arrow: insertion site; green or blue arrow: nearby boundaries; Blue/orange arrowhead in the browser tracks: site and orientation of the insertion. Yellow/green rectangles indicate regions with overall depleted contacts upon insertion. (**a**, **b**) are related to Fig. 2a–c. **a**, An extension to Fig. 2a showing both no-insertion controls (left and middle) and the insertion clone (right) at C25S5, each accompanied by corresponding data tracks. **b**, An extension to Fig. 2c: log2 fold changes in interaction frequencies between two no-insertion controls (left) and between the insertion clone and no-insertion controls (middle and right) for the region in (**a**). (**c**, **d**) are related to Fig. 2d–f. **c**, An extension to Fig. 2d showing both no-insertion controls at C21S7. **d**, An extension to Fig. 2f: log2 fold changes in interaction frequencies between two no-insertion controls and between the insertion clone and no-insertion controls for the region shown in (**d**). Each Hi-C heatmap represents merged data from 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq and 2 RNA-seq experiments were conducted for each genotype, with 1 of each exhibited.

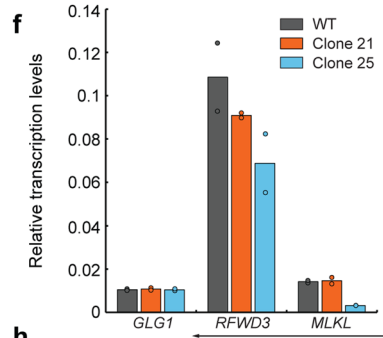


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Insertion loci without apparent detectable domain-level changes. Throughout, red arrow: insertion site; orange/blue arrowhead in the browser tracks: locus/orientation of the insertion. **a**, An insertion at C21S6: Hi-C maps for both no-insertion controls (left and middle) and the insertion clone (right) at C21S6, each accompanied by corresponding data tracks. **b**, Insulation scores for the region in (**a**). **c**, Log2 fold changes in interaction frequencies between two no-insertion controls (left) and between the insertion clone and no-insertion controls (middle and right) for the region in (**a**). **d**, Hi-C contact maps at C21S10. **e**, Insulation score profiles for the region in (**d**). **f**, Log2 fold changes in interaction frequencies between two no-insertion controls and between the insertion clone and no-insertion controls for the region in (**d**). **g**, Hi-C contact maps at C25S6. **h**, Insulation score profiles for the region in (**g**). **i**, Log2 fold changes in interaction frequencies for the region shown in (**g**). Each Hi-C heatmap presents merged data from 2 independent experiments performed for each genotype. 2 CTCF & RAD21 ChIP-seq and 2 RNA-seq experiments were performed for each genotype, with 1 of each displayed.

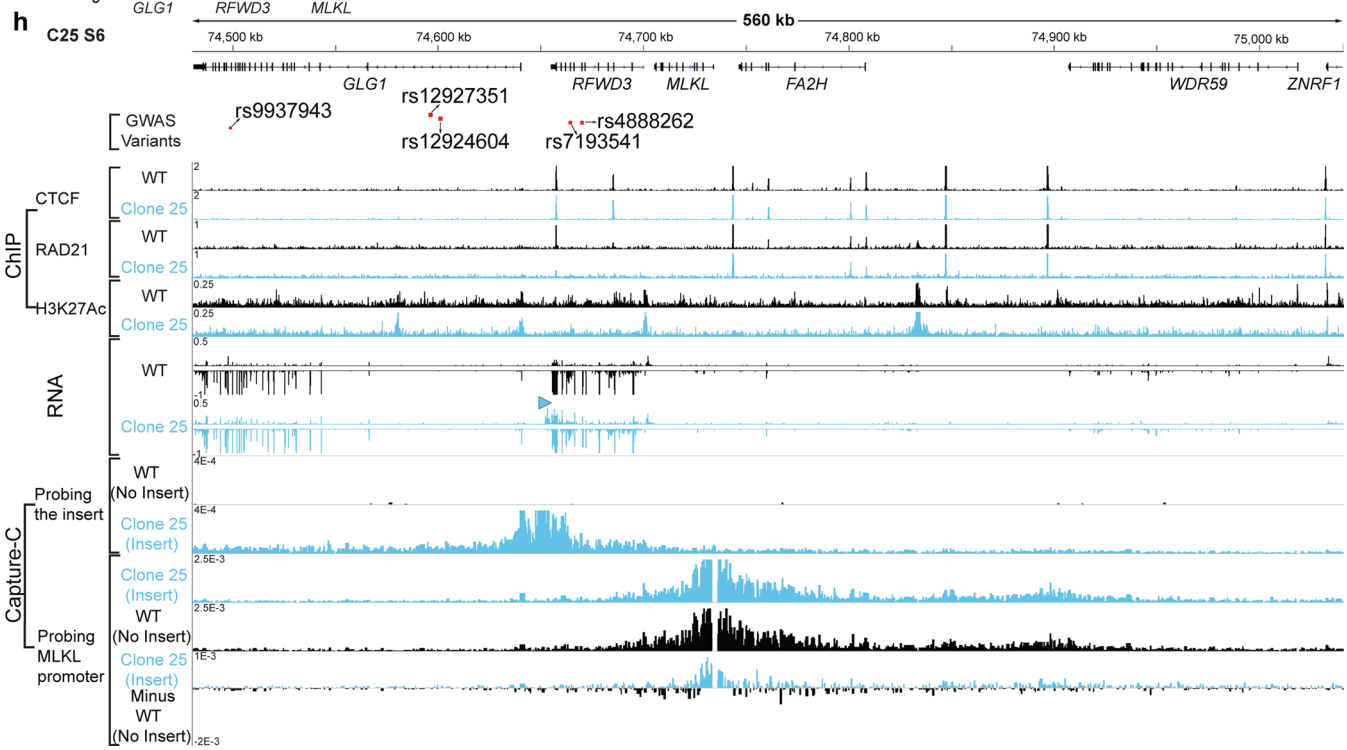






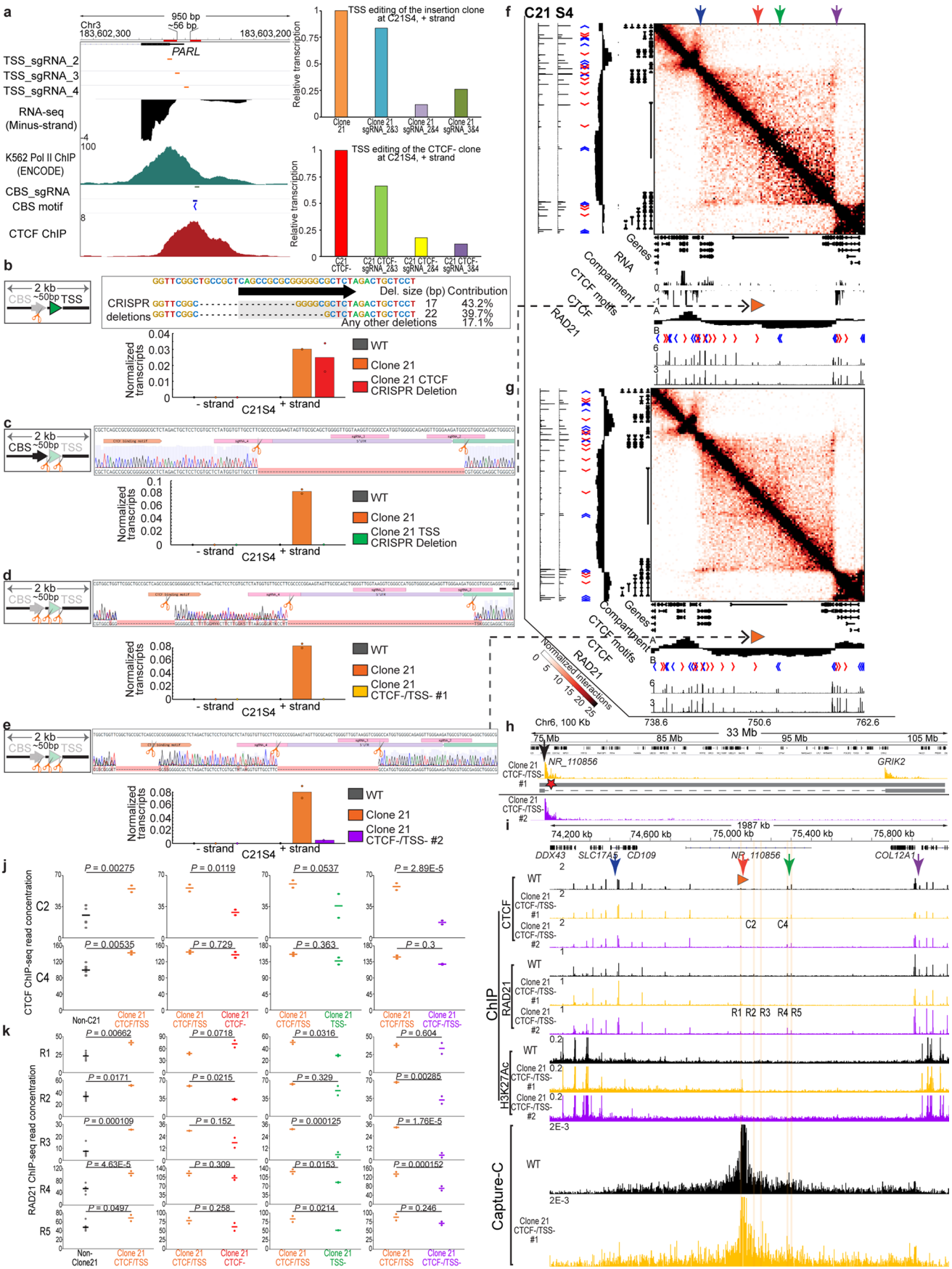
g

SNP	HGVS notation	Publication	Disease/trait
rs9937943	NC_000016.9:g.74499046G>C	Astle WJ, Cell, 2016	Neutrophil percentage
rs12927351	NC_000016.9:g.74596618C>A	Astle WJ, Cell, 2016	Granulocyte percentage
rs12924604	NC_000016.9:g.74601486G>A	Astle WJ, Cell, 2016	Monocyte percentage
rs7193541	NC_000016.9:g.74664743T>C	Mitchell JS, Nat. Comm., 2016	Multiple myeloma
rs4888262	NC_000016.9:g.74670458C>T	Chung CC, Nat. Genetics, 2013	Testicular germ cell tumor



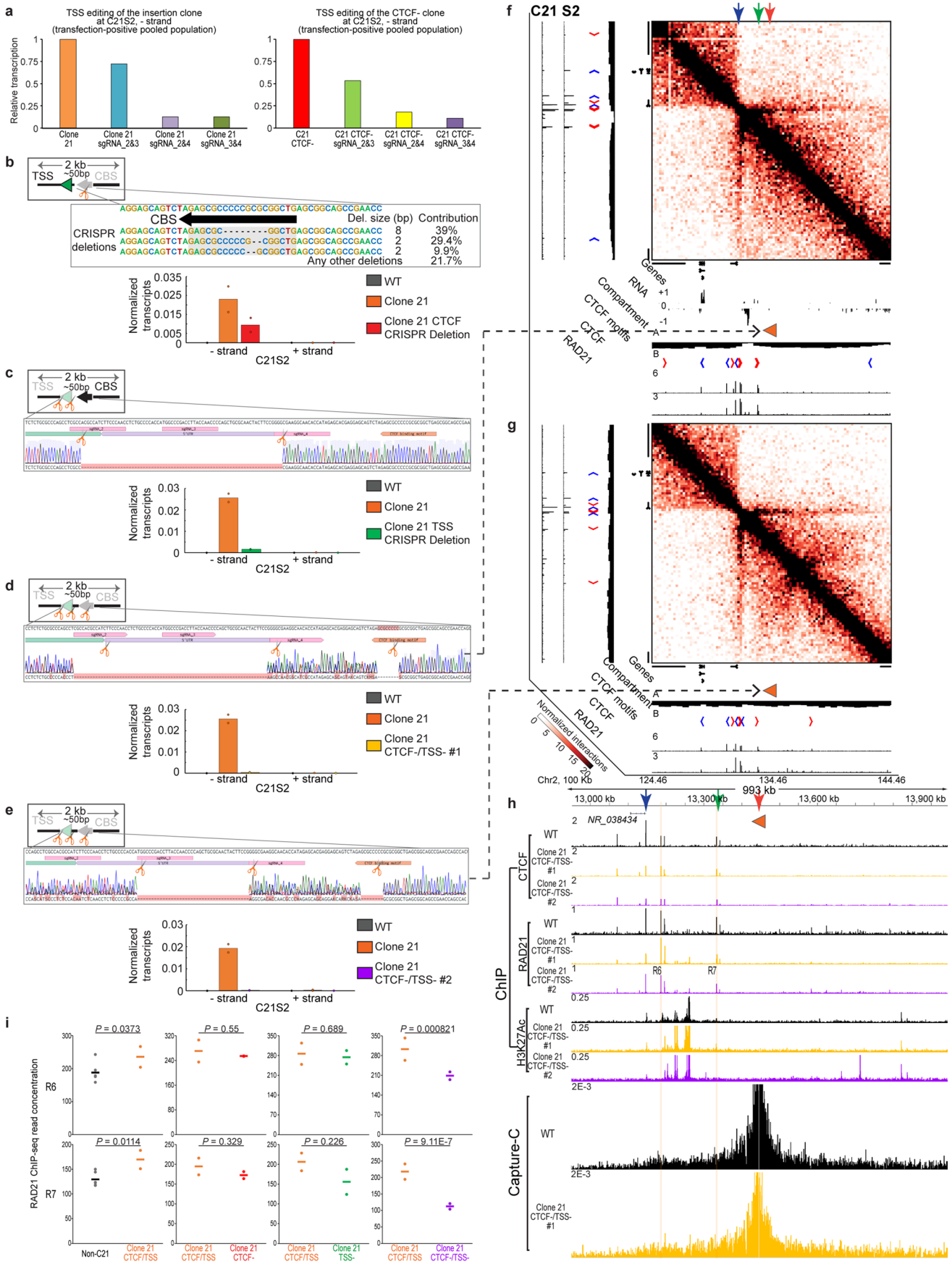
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Transcription of insertion-proximal genes remains mostly stable, with *MLKL* as an exception. a, An MA plot showing Clone 21 vs. non-Clone 21 transcriptomes. Each dot: a gene; red dots: differentially expressed (DE) genes at an FDR < 0.01; color-coded circles: insertion-proximal genes by distance ranges; red line: no-change line; two orange lines: +/− 1 log2 fold change. **b**, Clone 25 vs. non-Clone 25 transcriptomes. **c**, Clone 21 has ~95 DE genes transcriptome-wide (related to (**a**)). **d**, Clone 25 has ~160 DE genes transcriptome-wide (related to (**b**)). **e**, DE status of all insertion-proximal genes. The DE gene between 50 kb and 500 kb to an insertion, *MLKL*, is characterized in (**f**) and (**h**). In (**a**–**e**), 2 RNA-seq experiments were performed for each genotype. DE analysis was conducted with Clone 21 vs. non-Clone 21 (WT and Clone 25) and Clone 25 vs. non-Clone 25 (WT and Clone 21). **f**, RT-qPCR of *MLKL* and *GLG1/RFWD3*, two genes flanking the insertion (see (**h**)). N = 2 independent experiments for each genotype. **g**, GWAS significant variants near *GLG1/RFWD3/MLKL* insertion locus[43–45]. **h**, *GLG1/RFWD3/MLKL* locus (blue arrowhead: location/orientation of the insertion) using ChIP-seq/RNA-seq/Capture-C. The insertion coincides with reduced RAD21 binding at a peak immediately downstream. The insertion contacts the promoter of *GLG1* (Capture-C: Probing the insert). *MLKL* promoter also interacts with *GLG1* promoter (Capture-C: Probing *MLKL* promoter), albeit no apparent changes in interactions of *MLKL* promoter upon insertion. Capture-C presents merged data from 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq, 1 H3K27ac ChIP-seq and 2 RNA-seq experiments were conducted for each genotype, with 1 of each shown.



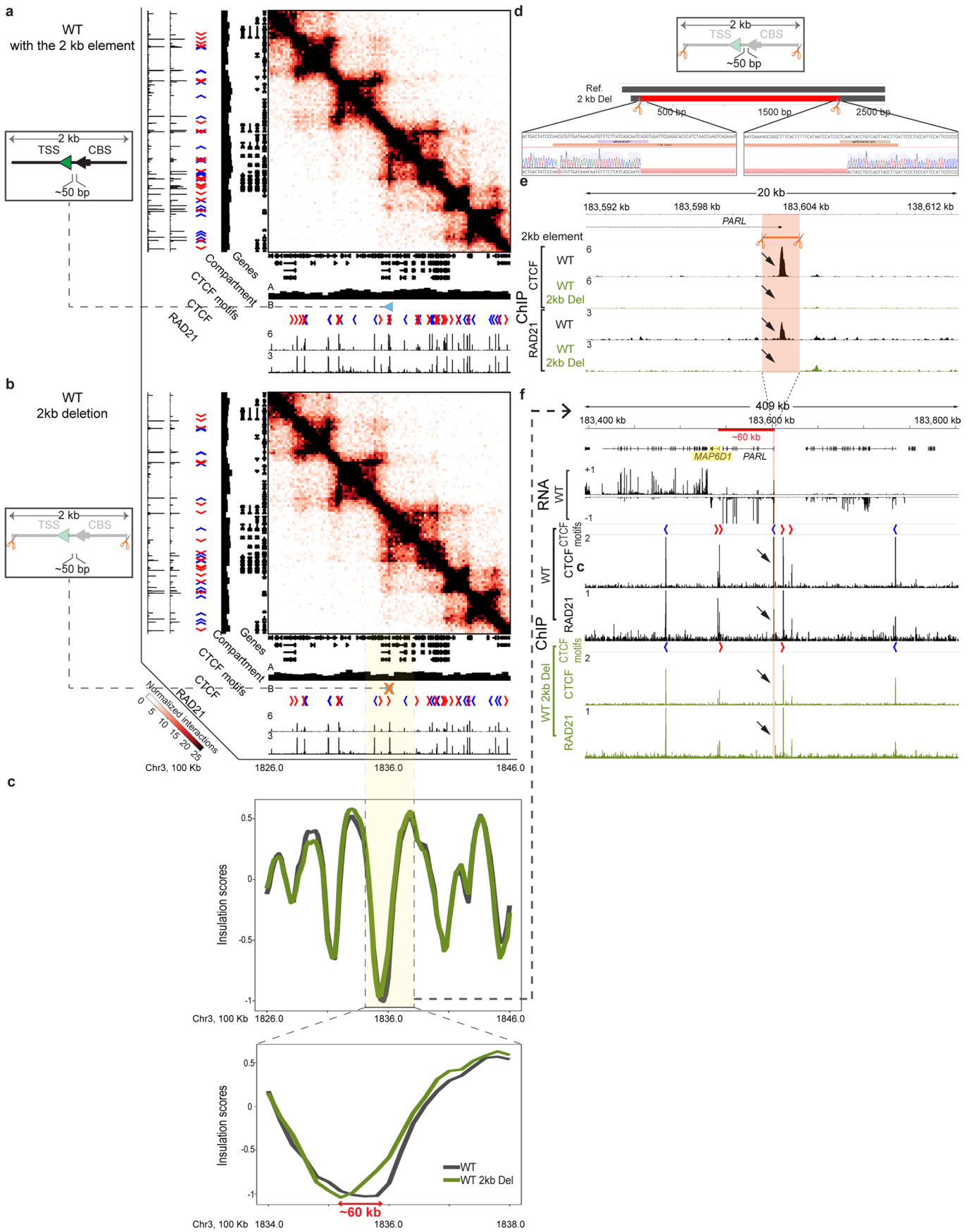
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | CRISPR dissections of insertion, and CTCF/RAD21 at C21S4. a, Left: sgRNAs within the insertion element (red lines: Pol2/CTCF peak centers). Right: TSS_sgRNA_2&4 and TSS_sgRNA_3&4 reduce transcription more effectively at C21S4. N = 1. **b**, CRISPR deletion of the inserted CBS spares transcription. **c**, Clone 21 ΔTSS: TSS_sgRNA_2&4-edited Clone 21 abrogates transcription, with the CBS intact. **d, e**, Clone 21 ΔCTCF/ΔTSS #1&#2: Clone 21 with its CBS already disrupted (**b**) further edited with TSS_sgRNA_2&4 and TSS_sgRNA_3&4, respectively. In (**b**–**e**), N = 2 experiments for each genotype. In (**f**, **g** and **i**), red arrow: insertion site; green arrow: downstream CBSs; blue/purple arrow: strong boundary nearby; orange arrowhead: insertion location/orientation. **f**, Hi-C of Clone 21 ΔCTCF/ΔTSS #1 (**d**) at C21S4: a ~27 Mb heterozygous deletion (**h**) influences heatmap interpretation. **g**, Hi-C of Clone 21 ΔCTCF/ΔTSS #2 (**e**) at C21S4: domain configuration restored close to pre-insertion level (Fig. 4a). **h**, Virtual 4 C (black arrow: viewpoint; red star: C21S4; *GRIK2*: C21S5): Clone 21 ΔCTCF/ΔTSS #1 has both short-range contacts and strong >25-Mb distal contacts, suggesting a heterozygous deletion between C21S4 and C21S5 (grey bars: chromosomes; dashed line: deletion). **i**, ΔCBS/ΔTSS restores nearby chromatin folding pattern to pre-insertion levels. Differentially bound CTCF (C2, C4) and RAD21 peaks (R1-R5) upon insertion highlighted. Directionality Index of Clone 21 ΔCTCF/ΔTSS #1 Capture-C: Fig. 4j. In (**f**–**i**), each Hi-C/Capture-C describes merged data from at least 2 independent experiments for each genotype. 2 CTCF/RAD21 ChIP-seq and 1 H3K27ac ChIP-seq for each genotype, with 1 of each shown. **j**, Pairwise comparisons between genotypes of CTCF binding (C2 and C4: (**i**) and Fig. 4f–i). **k**, Pairwise comparisons between genotypes of RAD21 binding (R1-R5: (**i**) and Fig. 4f–i). In (**j**, **k**), non-Clone 21: 3 genotypes without Clone21 insertions, each with 2 ChIP-seq replicates. Clone21 CTCF/TSS and derived CRISPR clones: 1 genotype, each with 2 ChIP-seq replicates. P-values (not adjusted for multiple comparisons): from a two-sided Wald test.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | CRISPR dissections of insertion, and RAD21 distribution at C21S2. a, TSS_sgRNA_2&4 and TSS_sgRNA_3&4 (as in Extended Data Fig. 8a) reduce transcription more effectively at C21S2 in CRISPR-Cas9 RNP-transfected cells. N = 1 experiment. **b**, Deletion of the inserted CBS reduces but does not abolish transcription at C21S2. **c**, Clone 21 ΔTSS derived from TSS_sgRNA_2&4-edited Clone 21 abrogates transcription, with the CBS intact. **d**, Clone 21 ΔCTCF/ΔTSS #1: derived from CBS-disrupted Clone 21 (**b**) further edited with TSS_sgRNA_2&4. **e**, Clone 21 ΔCTCF/ΔTSS #2: derived from CBS-disrupted Clone 21 (**b**) further edited with TSS_sgRNA_3&4. In (**b**–**e**), N = 2 independent experiments for each genotype. In (**f**–**h**), red arrow: insertion site; green or blue arrow: downstream CBSs; orange arrowhead in the browser tracks: locus/orientation of the insertion. **f**, **g**, Hi-C maps of Clone 21 ΔCTCF/ΔTSS #1 (**d**) and of Clone 21 ΔCTCF/ΔTSS #2 (**e**), respectively, at C21S2: deletions of both the CBS and the TSS restore the domain configuration close to pre-insertion level (Fig. 5a). **h**, Capture-C and corresponding data tracks showing that ΔCTCF/ΔTSS rescues local chromatin contact pattern close to that of WT. Differentially bound RAD21 peaks (R6, R7) upon CBS-TSS insertion highlighted. Directionality Index of Clone 21 ΔCTCF/ΔTSS #1 Capture-C: Fig. 5j. In (**f**–**h**), each Hi-C/Capture-C depicts merged data from at least 2 independent experiments for each genotype. 2 CTCF/RAD21 ChIP-seq and 1 H3K27ac ChIP-seq for each genotype, with 1 of each shown. **i**, Pairwise comparisons between genotypes of RAD21 binding at two RAD21 peaks (R6 and R7, as in (**h**) and Fig. 5f–i). Non-Clone 21: 3 genotypes without Clone 21 insertions, each with 2 ChIP-seq replicates. All others: 1 genotype, each with 2 ChIP-seq replicates. P-values (not adjusted for multiple comparisons) are derived from a two-sided Wald test through DiffBind.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Deletion of the endogenous 2 kb element leads to a boundary shift, while local domain organization is stable. a, Hi-C of no-deletion control showing the endogenous boundary where the 2 kb element (blue arrowhead) is derived, accompanied by corresponding data tracks. **b**, Deletion of the 2 kb (crossed-out blue arrowhead) leaves the overall domain configuration largely intact. The highlighted ~400 kb region is further examined in (**c**) and (**f**). **c**, Insulation scores show overall concordance, with a possible shift in boundary by ~60 kb to the left upon deletion. **d**, Genotyping confirms the desired deletion between sgRNAs flanking the 2 kb. **e**, ChIP-seq further verifies the deletion, as reflected in lack of signal (black arrows) within the 2 kb element (highlighted). **f**, Upon 2 kb deletion (highlighted in red), the point of local maximal insulation shifts ~60 kb to the left (**c**), coinciding with the distance between the TSSs of *PARL* and its nearest transcribed gene: *MAP6D1* (highlighted in yellow). This shift (red line) also corresponds to the distance between the deleted CBS and its nearest CTCF peak to the left, which now has reduced CTCF/RAD21 binding. Each Hi-C result presents merged data from 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq experiments for each genotype, with 1 of each shown.

nature research

Corresponding author(s): Di Zhang, Gerd Blobel

Last updated by author(s): Jul 13, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	For processing and analyzing next-gen sequencing data, we used the following tools and packages: BBDuk tool 37.68 bowtie 2: 2.3.3.1 SAMtools: 1.5 BEDtools: v2.17.0 HiC-Pro_2.9 and HiC-Pro 2.11.3-beta Lib5C 0.5.3 matplotlib 2.2.3 pandas 0.22.0, scipy 0.19.1, numpy 1.13.3, juicer_tools.1.7.6 FastQC 0.11.5 trim galore: 0.4.1-0 cutadapt: 1.18 FLASH: 1.2.8 CCAnalyzer v3 Salmon v0.9.1. DESeq2 1.26.0 3D netmod domain calling (v3.0_development) as described in Norton et. al., 2018 (DOI: 10.1038/nmeth.4560) and in Zhang et. al., 2019 (DOI: 10.1038/s41586-019-1778-y) FastQC v0.10.1 STAR_2.5.1b

kentUtils UCSC Genome Browser v369
MACS 1.3.7.1
Sicer V1.1
bamCoverage 3.1.3
DiffBind v2.14.0
tximport_1.14.2
For statistical analyses and implementation of insulation score calculations, we used R (3.5.0)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All main, extended data and supplementary figures include publicly available data. All Hi-C, Capture-C, RNA-seq, ChIP-seq, and other applicable next-gen sequencing raw data and processed data generated from this study are now available as GSE137376 on GEO database. Mouse CTCF ChIP-seq and mouse Hi-C domain boundaries (both asynchronous) shown in Figure 6a-c are derived from Zhang et. al., 2019 (DOI: 10.1038/s41586-019-1778-y), the updated GEO database for which has the accession number GSE129997. In Supplementary Figure 1: Hi-C heatmaps from all cell lines, except for HAP1, are from from GEO: GSE63525 by Rao et. al., 2014 (DOI: 10.1016/j.cell.2014.11.021); K562 ChIP-seq data are from ENCODE, CTCF (DCC accession: ENCSR000AKO), SMC3 (DCC accession: ENCSR000EGW), RAD21 (DCC accession: ENCSR000FAD) and Pol2 (DCC accession: ENCSR000FAY).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was not pre-determined. We used sample sizes commonly accepted for next-gen sequencing based genome-wide experiments (Huang et. al., 2017 (DOI: 10.1101/gad.303461.117); Zhang et. al., 2019 (DOI: 10.1038/s41586-019-1778-y); Despang et. al., 2019 (DOI: 10.1038/s41588-019-0466-z)). For each genotype (wildtype or genome-edited cell line) whose Hi-C results are shown, we performed 2 independent Hi-C experiments. For each genotype (wildtype or genome-edited cell line) whose Capture-C results are shown, we performed at least 2 independent Capture-C experiments. Hi-C and Capture-C data were pooled for down-stream analyses. For each genotype (wildtype or genome-edited cell line) whose RNA-seq and CTCF/RAD21 ChIP-seq results are shown, we also performed 2 independent RNA-seq experiments and ChIP-seq of CTCF and RAD21. For each genotype whose target-enriched insertion mapping and H3K27Ac ChIP-seq results are shown, we performed 1 experiment.
Data exclusions	No data were excluded from analyses.
Replication	2 biological replicates (2 independent experiments per genotype) were performed for Hi-C, RNA-seq, and ChIP-seq of CTCF and RAD21. At least 2 biological replicates (2 independent experiments per genotype) were performed for Capture-C. Data from all replication attempts were included. Mapping and other quality metrics were compared between replicates for Hi-C and Capture-C before data from replicates were pooled. Mapping metrics were examined, and data tracks were visualized in multiple genomic regions for RNA-seq replicates prior to downstream differential expression analysis. ChIP-seq peak call concordance and data tracks were examined for each ChIP-seq replicate.
Randomization	Experiments were not randomized. No animal or human subjects were involved in this study.
Blinding	Investigators were not blinded to allocation during experiments and outcome assessment. Blind was not relevant to our study as no human subjects were involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Histone H3K27ac antibody (Mouse Monoclonal), Active Motif, catalog number 39685, lot numbers 33417016 and 5919019. Dilution: 10ug/ChIP.

CTCF antibody (Rabbit Polyclonal), EMD Millipore, catalog number 07-729, lot number 3273150. Dilution: 10uL stock/ChIP.

RAD21 antibody (Rabbit Polyclonal), abcam, catalog number ab992, lot numbers GR3253930-6, GR3310168-1, and GR3310168-2. Dilution: 10ug/ChIP.

Mouse IgG, Sigma, catalog number I8765, lot number SLBW2188. Dilution: 50ug for pre-clearing and 10ug/ChIP.

Rabbit IgG, Sigma, catalog number I8140, lot number SLBT3686. Dilution: 50ug for pre-clearing and 10ug/ChIP.

Validation

The Histone H3K27ac antibody has been claimed to react with human and to be ChIP-grade per manufacturer. This antibody has also been previously used in our lab (Behera et al. 2019, DOI: 10.1016/j.celrep.2019.03.057), as well as 15 other publications that used and cited this antibody on manufacturer's website.

The CTCF and RAD21 antibodies have been claimed to react with human and to be ChIP-grade per the manufacturers. These two antibodies have also been previously used in our lab (Zhang et. al. 2019, DOI: 10.1038/s41586-019-1778-y), and in other published studies.

The mouse and rabbit IgGs have been previously used in our lab for pre-clearing and for isotype-matched control (Zhang et. al. 2019, DOI: 10.1038/s41586-019-1778-y). They are commonly used as control for ChIP per the manufacturer, and have been cited in at least 17 and 25 publications listed on the manufacturer's website.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HAP1 cells are near-haploid cells derived from KBM7 cells (Carette et al., Nature 477, 340-343 (2011)).
Authentication	HAP1 cells were routinely stained with a cell-permeant dsDNA dye, followed by sorting.
Mycoplasma contamination	HAP1 cells were tested for Mycoplasma.
Commonly misidentified lines (See ICLAC register)	The cell line used (HAP1) is not in the ICLAC database.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	All Hi-C, Capture-C, RNA-seq, ChIP-seq, and other applicable next-gen sequencing raw data and processed data generated from this study are now uploaded into the GEO database, under super series GSE137376. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137376
Files in database submission	Please see the GEO template sheet: https://upenn.box.com/s/2qqvmsmcdd5f90egoro049zdcfwtngg3
Genome browser session (e.g. UCSC)	

1. Open Integrated Genomics Viewer (IGV)
2. Download these IGV session files to local drive:
   -Fig. 4f: https://upenn.box.com/s/kpslkten6052qblo0hsig58kg0q06nda
   -Fig. 4g: https://upenn.box.com/s/1xz5i2xv1oxuuvn7p2rn3dllox599xjr
   -Fig. 4h: https://upenn.box.com/s/x1upxfdjhlyp8roww3kcnr7mhx5ph1dp
   -Fig. 4i: https://upenn.box.com/s/tklurgo5oj4qecm8uukz5krpxbs4m02k
   -Fig. 5f: https://upenn.box.com/s/s8m1ua7ed7rqg8y73l6ghozvw5cmtjod
   -Fig. 5g: https://upenn.box.com/s/knzg0sik3tohylox351ijjjs91a2qod4
   -Fig. 5h: https://upenn.box.com/s/yu4mng8qhdt62obq2451btql78arm6ol
   -Fig. 5i: https://upenn.box.com/s/idps6i0gpz0ba8pf32bcfrqg6620yfxm
3. In IGV, choose file-->open session-->select the session file downloaded from the link
4. Depending on network speed, it might take some time to fully load.
5. Alternatively, data can be downloaded to a local drive using the links in the IGV session files (.xml). Session files can be modified ("Resources" and "Tracks") for visualizing the data tracks locally.

Methodology

Replicates	As noted previously, for each genotype (wildtype or genome-edited cell line) whose Capture-C results are shown, we performed at least 2 independent Capture-C experiments. Hi-C and Capture-C data were pooled for down-stream analyses. For each genotype (wildtype or genome-edited cell line) whose RNA-seq and CTCF/RAD21 ChIP-seq results are shown, we also performed 2 independent RNA-seq experiments and ChIP-seq of CTCF and RAD21. For each genotype whose target-enriched insertion mapping and H3K27Ac ChIP-seq results are shown, we performed 1 experiment.
Sequencing depth	Hi-C for each clonal/parental populations was sequenced to generate ~248-300 million raw reads for each genotype (clonal or parental populations), with biological replicates pooled and with sub-sampling performed where necessary. This leads to ~161-173 million valid interaction pairs per genotype. Capture-C for each biological replicate was sequenced to ~5-15M raw reads per captured locus. RNA-seq for each replicate was sequenced to ~70-120M reads. H3K27Ac ChIP was sequenced to ~50-70M reads, with the input sequenced to ~46M reads. CTCF and RAD21 ChIP samples were sequenced to ~20-40M reads per replicate. Please refer to Supplementary Tables 3 and 4 for more details.
Antibodies	Histone H3K27ac antibody (Mouse Monoclonal), Active Motif, catalog number 39685, lot numbers 33417016 and 5919019. CTCF antibody (Rabbit Polyclonal), EMD Millipore, catalog number 07-729, lot number 3273150. RAD21 antibody (Rabbit Polyclonal), abcam, catalog number ab992, lot numbers GR3253930-6, GR3310168-1, and GR3310168-2.
Peak calling parameters	Basecalls using bcl2fastq2 v2.15.0.4, and parameters --barcode-mismatches 1 Mapping to reference genome hg19 canon with Bowtie 1.0.0 using parameters --chunkmbs 1024 -y -n 2 --best -k 1 --maxbts 800 -l 28 -e 80 --sam-nohead --sam Wiggle generated with MACS using parameters --nomodel --shiftsize=120--format BAM --gsize 2615371906 --tsize 36 --bw 120 --mfold 12 --wig --space 1 For H3K27Ac: Peaks called with Sicer V1.1 using parameters hg19 1 200 150 0.74 600 0.01 For CTCF and RAD21: Peaks were called with MACS using MFOLD=12, PVALUE="" --format BAM --gsize $EFFGENSIZE --tsize 36 --bw 120 --mfold $MFOLD --wig --space 1 $PVALUE Filter blacklist regions from peaks, and convert to broadpeak format (see UCSC Genome Browser for format specs)
Data quality	ChIP-qPCRs of positive control loci confirmed enrichment before the library was sent for sequencing. Raw fastq files were assessed with FastQC (v0.10.1) prior to processing. Peaks were called using input controls. Please see methods section for details with regards to Hi-C, Capture-C, and RNA-seq.
Software	We used Bowtie 0.12.8, SAMtools 0.1.18, MACS 1.3.7.1, BEDTools 2.16.2, and Sicer V1.1.