Published in final edited form as:
Nat Genet. 2020 October ; 52(10): 1076–1087. doi:10.1038/s41588-020-0680-8.

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], **Gerd A. Blobel**[1,2,*]

[1]Division of Hematology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[3]Department of Biochemistry & Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA

[4]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104, USA

[5]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

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 only 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 reflective of chromatin loop extrusion. Deletions of the CBS or the TSS individually or in combination within inserts revealed their distinct contributions to genome folding. Altogether, short DNA insertions can suffice to shape the spatial genome in a manner influenced by chromatin context.

[*]Corresponding authors: Di Zhang (dizhang.penn@gmail.com), Gerd A. Blobel (blobel@email.chop.edu).

Author Contributions
D.Z. and G.A.B. conceived the study and designed 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 of 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 TSS and the 2-kb element. C.A.K., B.G. and R.C.H. prepared ChIP-seq and RNA-seq sequencing libraries, performed all sequencing, uploaded seq 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 statement
The authors declare no competing financial interests, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

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 reflective of transcriptional activities[6–8]. Topologically associating domains (TADs)[2, 3], or contact domains[4], are megabase/sub-megabase 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 have 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, because 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? Here, 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). Following transposition, we established clonal lines and prioritized Clones 21 (C21) and 25 (C25), which contained ten and six insertions, respectively—named C21 Sites1 – 10 (C21S1 – C21S10) and C25 Sites1 – 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 RT-qPCR, 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 (WT) HAP1 cells. Overall, the results from the 3 samples were highly concordant (Supplementary Fig. 2), ruling out drastic genome-wide architectural perturbations caused by transpositions or significant biases from clonal variations.

Significantly, examination of integration sites revealed at least four instances of apparent *de novo* contact domain formation (Fig. 1b, d, f; Extended Data Fig. 2; Extended Data Fig. 3a-f). These domains (at least 80 kb in size), identified at a resolution of 20-kb bins using 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 a Hi-C heatmap representing in general terms regions of enriched interactions, but not in particular refence 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 Clone 21 Site 4 (C21S4), for example, the inserted 2-kb element, together with endogenous downstream convergent CBSs, demarcated a new domain ~250 kb in size (Fig. 1b, 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—though whether the two downstream CBSs causally contribute to transcription termination here remains unclear[25] (Fig. 1b, 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, 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, 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 Clone 21 Site 2 (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, Extended Data Fig. 2c-d). This newly formed 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, 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 Clone 21 Site 5 (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, 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 towards an A compartment (Fig. 1b: Compartment, 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 towards 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, Extended Data Fig. 5, and Extended Data Fig. 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 with 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 significant domain-level changes (Fig. 3a-b, 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 CBS 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-seq, and CTCF, RAD21 and H3K27ac 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 intra-domain, sub-100-kb range, unable to manifest themselves as new domains (Fig. 3c, d). Curiously, gained interactions upon 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, d, 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, Extended Data Fig. 6). Conceptually, since 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 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—likely mediated by the high density of 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, Clone 21, and Clone 25 were highly congruent: only ~95 and ~160 genes were differentially expressed in C21 and in 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 *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* upon insertion, nor is it clear why *GLG1* expression is unaffected by the new contacts. Intriguingly, all the remaining ~191 genes within 1.5 Mb from any insertion were not differentially expressed (Extended Data Fig. 7e), while 4 out of 9 genes with gene-body insertions were differentially expressed, suggesting that most of the gene expression changes were limited to gene-body insertions. Since transposon insertions are nearly 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 (Fig. 4 and Fig. 5). We first examined the C21S4 locus, where an insertion event demarcated a new domain whose long-range chromatin interaction pattern became more A compartment-like (Fig. 1b, c; Fig.

4b; 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 compartment A signature[1, 6–8]. Elevated RAD21 levels were also detected at several sites in the new domain upon 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 upon insertion (Fig. 4g and Extended Data Fig. 8k: R4 and R5), while only the left CBS had moderately increased CTCF binding (Fig. 4g and Extended Data Fig. 8j: C4). This increased accumulation of cohesin upon CBS-TSS insertion likely strengthened insulation at the downstream CTCF sites, which now demarcates the new domain (Fig. 4b, e, g).

To dissect the contributions of CTCF and TSS to genome folding, we deleted the CBS and the TSS 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 CBS 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; 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, Extended Data Fig. 8k). Deletion of both the CBS and the TSS in one clone restored local chromatin structure to pre-insertion configuration as assayed by Hi-C (Extended Data Fig. 8e, g), coinciding with further reduction in local cohesin accumulation close to the pre-insertion levels (Extended Data Fig. 8i: Clone21 CTCF-/TSS- #2, Extended Data Fig. 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 desired edits that were subsequently characterized with Capture-C (Extended Data Fig. 8i). We next assessed how CBS and/or TSS folds its nearby genomic fragment by using 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 towards 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 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 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 CBS and TSS neutralized the genomic fragment's interaction preference close to the baseline level prior to insertion (Fig. 4j, 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, 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, Extended Data Fig. 9i: R6, R7). By contrast, these interactions, likely mediated by CTCF-CTCF pairing, remained largely unaffected upon TSS deletion (Extended Data Fig. 9c; 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 for 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, 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, while 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, while 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. While 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 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].

Chromatin 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 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 (HERV), whose high level of transcription 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 inserted into tens of thousands of loci in the mouse (but not primate) genome since mouse diverged from its placental ancestor ~75 million years ago[56, 57]. This provides a window into recent evolution for exploring how these insertions expanding 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 out of the total ~7,136 CTCF-bound SINE B2 insertions, 527 occurred in the most TSS-spare 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 TSS within 200 kb), and only 5.27% (38/721) of them co-localized with putative new boundaries (Fig. 6c). Since 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. While compartments have historically been thought of as much larger than domains (tens of Mb 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 re-configuration upon mitotic exit[54]. Third, a stripe, perhaps reflective of 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-based boundary detection[3]. Meanwhile, convergent CBSs form distal interactions to demarcate boundaries (Fig. 6d). Importantly, genomic contexts can not only modulate effects of the inserted element to display discrete features of *de novo* domains, but may also pose constraints that limit or mask measurable new domain formation[64]. 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 intra-domain, sub-100-kb range (Fig. 6d) that might be further resolved with emerging techniques with sub-kilobase resolution[65, 66]. 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[64], 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, though 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], though 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 HERV 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 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 contexts (Fig. 6d). Here, we have leveraged genome editing and Hi-C for multiplexed characterization of the inserted putative boundary DNA element's effects on the human 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 towards understanding how sequence and context together influence genome architecture, and ultimately towards rationally engineering the 3D genome.

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.

Methods (Experiments)

HAP1 cell culture and maintenance—HAP1 cells[22] (a kind gift from Bas van Steensel through the 4D Nucleome consortium) were cultured in IMDM with 10% FBS and 1% penicillin/streptomycin at 37°C with 5% CO_2. To enrich near-haploid population, HAP1 cells were routinely stained with a cell-permeant dsDNA dye, followed by fluorescence assisted cell sorting (FACS). Briefly, HAP1 cells were trypsinized, pelleted, counted using a hemocytometer, and re-suspended at a density of 1 million cells/ml growth medium containing 10 μg/ml Hoechst 33342 (Thermo Fisher, H3570) for 30 minutes. Stained cells were then pelleted, resuspended in 1 ml MACS buffer (1× PBS pH 7.2, 0.5% BSA and 2 mM EDTA), and subsequently sorted on a MoFlo Astrios (Beckman Coulter). The sorting gate was stringently set to enrich cells with half of the DNA content of a diploid (2n) cell line (HUDEP-2[68]) at G1. Growth medium was then added to sorted cells for continued culture.

Sleeping Beauty transposon genome editing—The candidate 2-kb DNA (see Methods section "**Boundary-underlying DNA selection**") was PCR-ed (primer sequences in Supplementary Table 2) with HAP1 gDNA as the template, and was cloned into pSA-MCS[69] (a kind gift from Alessandra Recchia and Zoltán Ivics, this plasmid can be requested from them upon ordering Addgene plasmid 26557 and finishing the MTA for it), 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[21] (a kind gift from Zoltán Ivics and Zsuzsanna Izsvák, Addgene plasmid 34879), together with pmaxGFP (Lonza), using Nucleofector Kit L (Lonza) and the program X-001 on an Amaxa electroporator (Lonza). 24 hours after transfection, the top 1% 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. gDNA 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 gDNA 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 was then calculated using the following formula:

$$Est.Ins.Copy.Num = mean\left(\frac{2^{Ct_{IR-1} - Ct_{CD4}}}{2}, \frac{2^{Ct_{IR-2} - Ct_{CD4}}}{2}\right)$$

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

Insertion site mapping and validation—Capture using biotinylated oligonucleotides and pull-down with streptavidin beads were used to map transposon insertion sites as outlined in Extended Data Figure 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, K182001), was sonicated at 100% amplitude, 30 seconds on/30 seconds off, for 40 minutes, in a bath sonicator (QSonica Q800R3). Sonicated DNA was cleaned up using AMPure XP beads (Beckman Coulter), End Repaired, dA-Tailed (NEBNext® Ultra™), adaptor 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 NimbleGen 2× Hybridization buffer, 3 μl NimbleGen Hybridization Component A, and 2.5 μl nuclease-free water, and incubated for 10 minutes at room temperature. Resuspended DNA was then denatured at 95°C in a PCR machine for 10 minutes. 2 μl 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 washed Dynabeads™ MyOne™ Streptavidin C1 (Thermo Fisher 65001) at 47°C in a thermomixer for 1 hour. With 100 μl 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 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, 05634261001). Hybridized DNA was eluted off the beads with 25 μl 0.125 M NaOH, neutralized with 25 μl 1 M Tris-HCl, pH 8.8, followed by AMPure XP beads 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 beads pulldown as described above. Pulldown enrichment was confirmed using qPCR prior to Illumina sequencing. After the insertion sites are mapped (see "Insertion-site mapping" in Methods (Analyses) section), the location and orientation of each insertion was validated via targeted PCR.

***In situ* Hi-C—**In situ Hi-C was performed as previously described[4, 9, 70]. Briefly, ~10 million HAP1 cells, WT, Clone 21, Clone 25 and all the CRISPR-edited subclones were crosslinked in 2% formaldehyde at room temperature for 10 minutes, and then quenched in

0.125 M glycine for 5 minutes, on an orbital shaker. Cells were then transferred from a culture flask to a 15 ml tube, pelleted, washed with 1 ml cold PBS, transferred to an Eppendorf tube, pelleted, resuspended in 1 ml 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 minutes. Nuclei were then pelleted, washed once using 800 µl NEBuffer™ DpnII, pelleted, and resuspended in 500 µl NEBuffer™ DpnII. A final concentration of 0.3% SDS was added, and samples were incubated at 37°C, 950 rpm for 1 hour in a thermomixer. A final concentration of 1.8% Triton X-100 was then added to each sample, which was subsequently incubated at 37°C, 950 rpm for 1 hour. 40 µl nuclease-free water, as well as 300 U DpnII (NEB R0543M), was added, and samples were digested at 37°C, at 950 rpm overnight in a thermomixer. Another 300 U DpnII were added for an additional 4 hours of digestion with mixing. The samples were then incubated at 65°C, 950 rpm for 20 minutes. Nuclei were then pelleted, resuspended in 1× NEBuffer™ 2, with biotin-14-dATP, dTTP, dCTP and dGTP and DNA polymerase I, Large (Klenow) Fragment (NEB M0210), and incubated at 37°C, 950 rpm in a thermomixer for 90 minutes. Ligation reaction was subsequently carried out in a total volume of 1.2 ml with 4,000 U T4 DNA Ligase (NEB) at 16°C for 4 hours, followed by 30 minutes at room temperature. 20 µl of 20 mg/ml Proteinase K and 120 µl of 10% SDS were then added to each sample, followed by crosslinking reversal at 65°C overnight. An additional 10 µl proteinase K was added to each sample, which was incubated at 55°C for 2 hours. 2 µl DNase-free RNase was added to each sample, followed by incubation at 37°C for 30 minutes. Phenol-chloroform extraction was performed to purify DNA.

DNA was sonicated to 200–300 bp in a bath sonicator (QSonica Q800R3), followed by beads clean-up. Biotinylated nucleotides-filled ligation junctions were pulled down with 50 µl Dynabeads™ MyOne™ Streptavidin C1 (Thermo Fisher 65001). Sequencing libraries were subsequently prepared as described above in the "**Insertion site mapping and validation"** section: End Repair, dA-Tailing, adaptor ligation, followed by 6 cycles of indexing PCR.

The quality and size of each library was evaluated using the Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA), followed by quantitation using real-time PCR using the KAPA Library Quant Kit for Illumina (KAPA Biosystems, 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, 70, 71]. Briefly, ~10 million cells HAP1 cells were crosslinked in 1% formaldehyde at room temperature for 10 minutes, and then quenched in 1 M glycine for 5 minutes, on an orbital shaker. Cells were then pelleted, washed with 1 ml cold PBS, transferred to an Eppendorf tube, pelleted, resuspended in 1ml 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 minutes. Nuclei were then pelleted, washed once using NEBuffer™ DpnII, pelleted, and resuspended in 500 µl NEBuffer™ DpnII. Samples were incubated in 0.3% SDS at 37°C, 950 rpm for 1 hour in a thermomixer, followed by the addition of a final concentration of 1.8% Triton X-100, for 1 hour at 37°C, 950 rpm. After adding 40 µl 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 more hours at

37°C, followed by 20 minutes at 65°C, 950 rpm. Nuclei were pelleted, resuspended in a total volume of 1.2 ml with 4,000 U T4 DNA Ligase (NEB), and incubated at 16°C for 4 hours and then at room temperature for 30 minutes. 20 μl of 20 mg/ml 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 more hours at 55°C. 2 μl DNase-free RNase was added to each sample, followed by incubation at 37°C for 30 minutes. DNA was then extracted with phenol-chloroform.

DNA was sonicated to 200–300 bp in a bath sonicator (QSonica Q800R3), followed by beads clean-up. Sequencing libraries were subsequently prepared similarly as described in "**Insertion site mapping and validation"** section: End Repair, dA-Tailing, adaptor ligation, 6 cycles of indexing PCR, and 6 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) similarly as described in "**Insertion site mapping and validation"** section. Pulldown enrichment was confirmed using qPCR. Target enriched libraries were quality checked and quantified as *in situ* Hi-C samples prior to 2 × 75 bp sequencing on the Illumina NextSeq 500.

ChIP—Chromatin immunoprecipitation (ChIP) was performed as previously described[72], using the 10 μg/ChIP H3K27ac antibody (Active Motif 39685), 10 μL stock/ChIP CTCF antibody (Millipore 07–729) and 10 μg/ChIP Rad21 antibody (abcam ab992). Briefly, ~20 million HAP1 cells were fixed in 1% formaldehyde in fresh culture medium at room temperature for 10 minutes, followed by quenching in 1 M glycine for 5 minutes. Crosslinked cells were lysed for 10 minutes in 1 ml 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 P8340) and phenylmethylsulfonyl fluoride (PMSF). Nuclei were pelleted, resuspended in 1ml room temperature Nuclei Lysis Buffer (50mM Tris pH 8, 10 mM EDTA, 1% SDS), with Protease Inhibitors and PMSF, and were incubated on ice for 20 minutes. The samples were sonicated at 100% amplitude, 30 seconds on/30 seconds off, for 45 minutes, in a bath sonicator (QSonica Q800R3). Sonicated materials were centrifuged, with the supernatant subsequently collected, and diluted with 4ml 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. 50 μl protein A/G agarose beads (Thermo Fisher 15918014, Thermo Fisher 15920010) and 50 μg isotype-matched IgG control were added to sonicated chromatin to pre-clear it for >2 hours at 4°C. Beads were then spun down, with 200 μl 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/ChIP isotype-matched control (mouse IgG (Sigma I8140) or rabbit IgG (Sigma 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 were eluted twice with a total volume of freshly prepared 200 μl Elution Buffer (100 mM $NaHCO_3$, 1% SDS). Into each IP and input, 12 μl 5 M NaCl and 2 μl RNase A (10 mg/ml, Roche through Sigma 10109169001) were added, and samples were incubated at 65°C overnight. 3 μl proteinase K (20 mg/ml, Roche through Sigma 3115879) was then added, for an additional 2 hours at 65°C. DNA was column cleaned up using a QIAquick PCR Purification Kit (QIAGEN 28106).

For ChIP-sequencing, 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, B23318). Libraries were quality checked, quantified prior to 1 × 75 bp sequencing on the Illumina NextSeq 500.

RNA-seq—HAP1 cells were washed with PBS, and resuspended in 1 ml TRIzol (Thermo Fisher), with 200 μl 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 ScriptSeqv2 Complete Kit (Illumina BHMR1224). Briefly, the RNA was depleted of rRNA using the 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 Di-tagged cDNA fragments were purified, barcoded, and PCR-amplified for 15 cycles. For the subsequent CRISPR-edited clones, as the ScriptSeqv2 Complete Kit was discontinued, TruSeq Stranded Total RNA (Illumina: 20020598) was used per manufacturer's instructions, which relies on a dUTP-based second strand synthesis to preserve the strandedness of RNA. Libraries were quality checked and quantified prior to 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 oligodT and random primers. qPCR was carried out with Power SYBR Green (Thermo Fisher). Transcripts were normalized relative to the geometric mean of Cts 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, 42230). Clone 21 HAP1 cells were transfected with this pX330 using Nucleofector Kit L (Lonza) and program X-001 on an Amaxa electroporator (Lonza). 24 hours 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[73] were performed. Clones with CBS disruptions at insertion loci where new domains had formed were identified for downstream characterizations. A CRISPR RNP-based[74] paired-cut approach was used for subsequent rounds of editing. For TSS editing, TSS- cells were derived from

Clone 21 cells, whereas CTCF-/TSS- cells were derived from the subclone with CBS disruptions at the C21S2 and C21S4 insertion loci that was previously derived from Clone 21. Cells with the deletion of the endogenous 2-kb element (WT 2kb Del) were derived from WT cells. Specifically, 150 pmol per sgRNA (2 sgRNAs, thus 300 pmol in sgRNA) (Synthego) along with 150 pmol spCas9 protein (Synthego) were mixed and incubated at room temperature for 10 minutes. The RNP mixture was then added to Nucleofector Kit L solution (Lonza) with 2 μg pmaxGFP plasmid (Lonza), this transfection solution was then added to ~1 million trypsinized and pelleted HAP1 cells. The resuspended cells were transfected using program X-001on an Amaxa electroporator. GFP+ (likely 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 days 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.

Methods (Analyses)

Boundary-underlying DNA selection—The selection process was summarized in Supplementary Figure 1a. Briefly, K562 boundary coordinates (hg19) were obtained from the domain calls using Arrowhead[4] (GEO: GSE63525), with domain start and end coordinates extended 5 kb both upstream and downstream. Human Embryonic Stem Cells (hESC) boundary coordinates were obtained from the domain calls using the Directionality Index (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[75]. This shared list of boundaries was then intersected with K562 CTCF ChIP-seq peaks[67, 76] (DCC accession: 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, following visual examinations of K562 Hi-C heatmap[4] (GEO: GSE63525). As shown in Supplementary Figure 1, this candidate boundary was further confirmed to be present in GM12878, HMEC, HUVEC, IMR90 and NHEK cell lines[4] (GEO: GSE63525), and by additional analytical methods[77, 78] such as Insulation Score[79] and Armatus[80]. The top 100 K562 CTCF binding site ranked by architectural proteins co-occupancy at this boundary was verified to have the bindings of CTCF (DCC accession: ENCSR000AKO, Bradley Bernstein Lab, ENCODE Consortium), SMC3 (DCC accession: ENCSR000EGW, Michael Snyder Lab, ENCODE Consortium), RAD21 (DCC accession: ENCSR000FAD, Sherman Weissman Lab, ENCODE Consortium) and Pol2 (DCC accession: ENCSR000FAY, Sherman Weissman Lab, ENCODE Consortium) in K562 (Supplementary Fig. 1c). In HAP1, similarly, this site was also bound by CTCF and SMC1[23] (GEO: 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*[81].

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 were further filtered to those that had a partial match to the Inversed Repeats, with right-trimming performed to remove these Inverse Repeats 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 Repeats-trimmed sequence fragments were then mapped to hg19 using Bowtie 2[82]. Resulting SAM files were converted to BAM files, which were then sorted, both using SAMtools[83]. Genome-wide coverage was then obtained using BEDTools[75]. 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, Clone 21, Clone 25 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[82]. Detection and filtering of valid interaction pairs, assignment to restriction fragment and binning, and interaction matrix balancing with iterative correction and eigenvector decomposition (ICE)[84] were all performed using the HiC-Pro pipeline[85]. 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 ~161–173 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 contacts Pearson correlations between two samples were calculated on non-empty bins shared between two samples.

Heatmap generation—To generate heatmaps for each insertion site, ICE balanced matrices for the clone with the insertion, the subclones with edited derivatives, as well as for WT and the other clone without the insertion at this position, were extracted with the insertion bin in the middle and with 150 bins (3 million bps) on either side of it. These extracted matrices, adjusted for minor coverage differences via division by a scaling factor reflective of the total number of valid interactions, were plotted using lib5C[86], whose dependencies include pandas, scipy and numpy, in linear scale using color scheme “red8” from matplotlib[87], 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[88], using the JASPAR 2018[89] CTCF motif, with p-value cutoff 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 & RAD21 ChIP-seq.

Insulation score—Insulation score computations were implemented in R mechanistically similar to those in Crane et al., 2015[79]. Given an ICE balanced interaction matrix with the insertion bin at the center, averaged interactions within a sliding window of 10 bins by 10 bins (15 bins by 15 bins 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 bins by 10 bins, $\log_2$ after 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[90]. Eigenvectors, the first principal component of the distance-adjusted correlation matrix, were calculated for the each of the selected chromosomes in *cis* (intrachromosomal) with insertions with KR normalization at 50-kb resolution using Juicer Tools[90]. 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 for were performed for parental WT cells and for most cell lines. In subclones with the deletion of TSS and the deletions of both CBS and TSS, four biological replicates were performed. Raw reads were processed using published scripts[71]. Briefly, Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), a wrapper for FastQC and Cutadapt[91], was used to access the quality and to trim adapter sequences of raw sequences. Trimmed reads were then merged or interleaved using FLASH[92]. Once concatenated, the reads were *in silico* DpnII digested, aligned to hg19, and analyzed using CCanalyser3[71]. 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 Directionality Indexs[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 Directionality Index for each of the distance range of 50 kb, 100 kb, 250 kb, 500 kb, and 1 Mb, respectively, using the equation[3]:

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

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 did the directional contact preference introduced by the inserted CBS/TSS, CTCF-, TSS-, or CBS-/TSS- elements emerges, propagates or reduces.

ChIP-seq—Raw sequencing reads (Supplementary Table 4) were mapped to hg19 using Bowtie[93]. SAM files were converted to BAM files using SAMtools. For histone marks (H3K27ac), BAM files were converted to Wiggle using MACS[94], and subsequently converted to bigWig for visualization. BAM files were converted to BED for subsequent peak calling using Sicer[95]. For transcription factors or factors with narrower peaks as in CTCF and RAD21 in our case, MACS was used for peak calling and for generating Wiggle files. For the ChIP-seq data tracks visualized in detail (such as those in Figs. 3, 4 and 5), bamCoverage from Deeptools[96] was used to generate counts-per-million (CPM)-normalized bigWig files from bam files. The CPM-normalized bigWig files were converted to bedgraphs to accompany Hi-C heatmaps. DiffBind[97] was used for differential binding analysis of CTCF and RAD21 near insertions. Comparisons between Clone 21 and Non-Clone 21, which consists of 3 cell lines without Clone 21 insertions: WT, WT with the deletion of the endogenous 2 kb (WT 2 kb Del), and Clone 25, each with 2 ChIP-seq replicates, were performed using the consensus peak set which includes peaks identified in at least 2 replicates and re-centered to include 250 bp upstream and downstream from consensus summits. Each pairwise comparison between Clone 21 CTCF/TSS and other CRISPR subclones derived from it was based on 2 ChIP-seq replicates of each cell line (genotype). The consensus peak set for these comparisons includes peaks identified in at least 3 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[97] (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[98]. Transcript-level quantifications were then imported in R, using tximport[99], for gene-level (EnsDb.Hsapiens.v75) measurements with DESeq2[100]. Genes with less than 10 reads were excluded from downstream DE analysis. DESeq2 was performed to identify DE genes between Clone 21 and non-Clone 21, and between Clone 25 and non-Clone 25, at an FDR < 0.01.

For RNA-seq analysis for genome tracks 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[101], with the default normalization reads per million mapped reads (RPM, or CPM) unchanged. Resulting bedGraph files were then converted to bigWig format for genome browser visualization with strand specificity. Insertion-proximal regions were extracted from bigWig files as bedGraph to accompany Hi-C heatmaps.

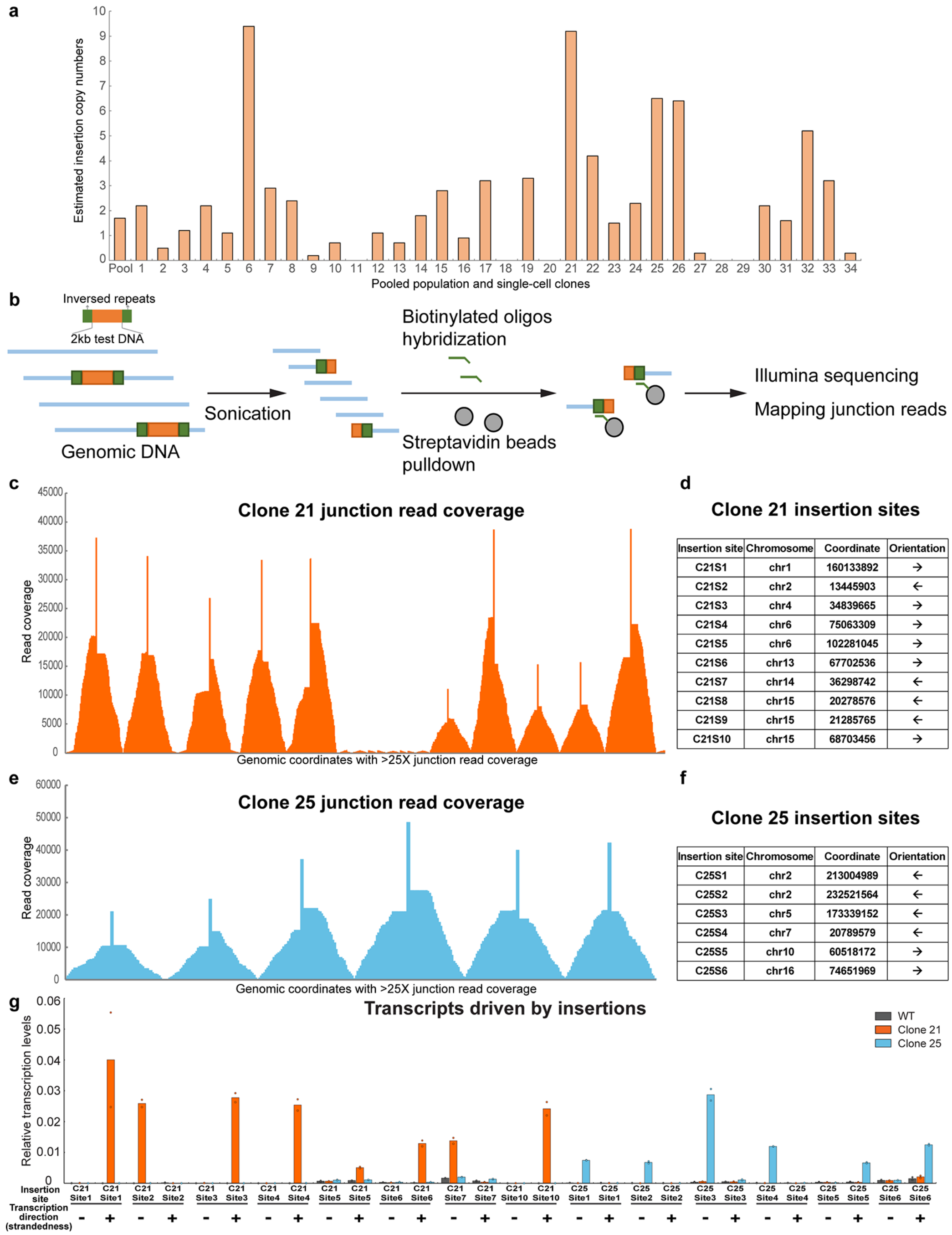
Recent evolution of the mouse architectural genome—Ancestral CTCF binding, defined as CTCF peaks shared among humans, macaques, mice, rats and dogs, and presumably in their common placental ancestor, was obtained from Schmidt et al., 2012[56] (ftp://ftp.ebi.ac.uk/pub/databases/vertebrategenomics/FOG03/calls/CTCF_mouse_5way.gff). More recent, highly rodent-specific, CTCF-Pol III TSS insertions expanded by SINE B2 elements, was obtained from Thybert et al., 2018[58] (ftp://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[102], 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[103] 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 colocalize with SINE B2 CTCF binding, and not with ancestral CTCF binding. TSS annotations were obtained from intersecting RefSeq and UCSC known genes 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 into regions with 0 TSSs and with ≥ 11 TSSs within a 200-kb range differentially colocalize with putative recently gained, SINE B2 CTCF-associated domain boundaries, based on a 2-by-2 contingency table.

Reporting Summary—Further information on research design is available in the Nature Research Life Sciences 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 this study are available under accession GSE137376 (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), accession number GSE129997 (GEO database). In Supplementary Figure 1: Hi-C heatmaps from all cell lines, except for HAP1, are 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).

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

Extended Data



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	→

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:

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.

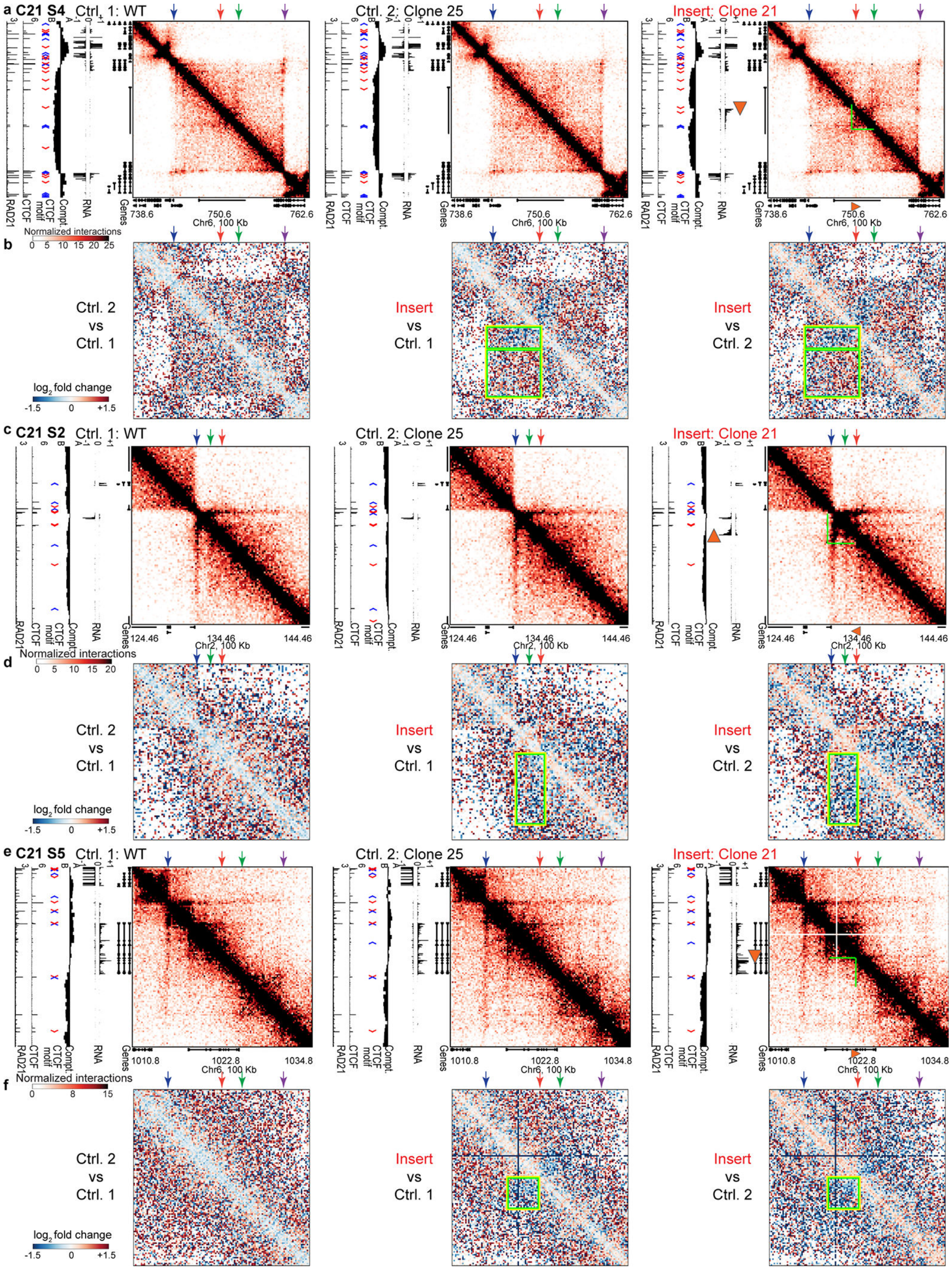
c, Junction read coverage for Clone 21: horizontal axis denotes genomic coordinates (single nucleotide resolution) with > 25X coverage; vertical axis shows read coverage. The spike in the middle of each peak consists of two neighboring nucleotides between which an insertion is located. Data from N=1 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:

Insertion-driven new domains: detailed comparisons (an extension to Figure 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) and (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) and (d): related to Fig. 1d-e.

c, An extension to Fig. 1d showing both no-insertion controls at C21S2.

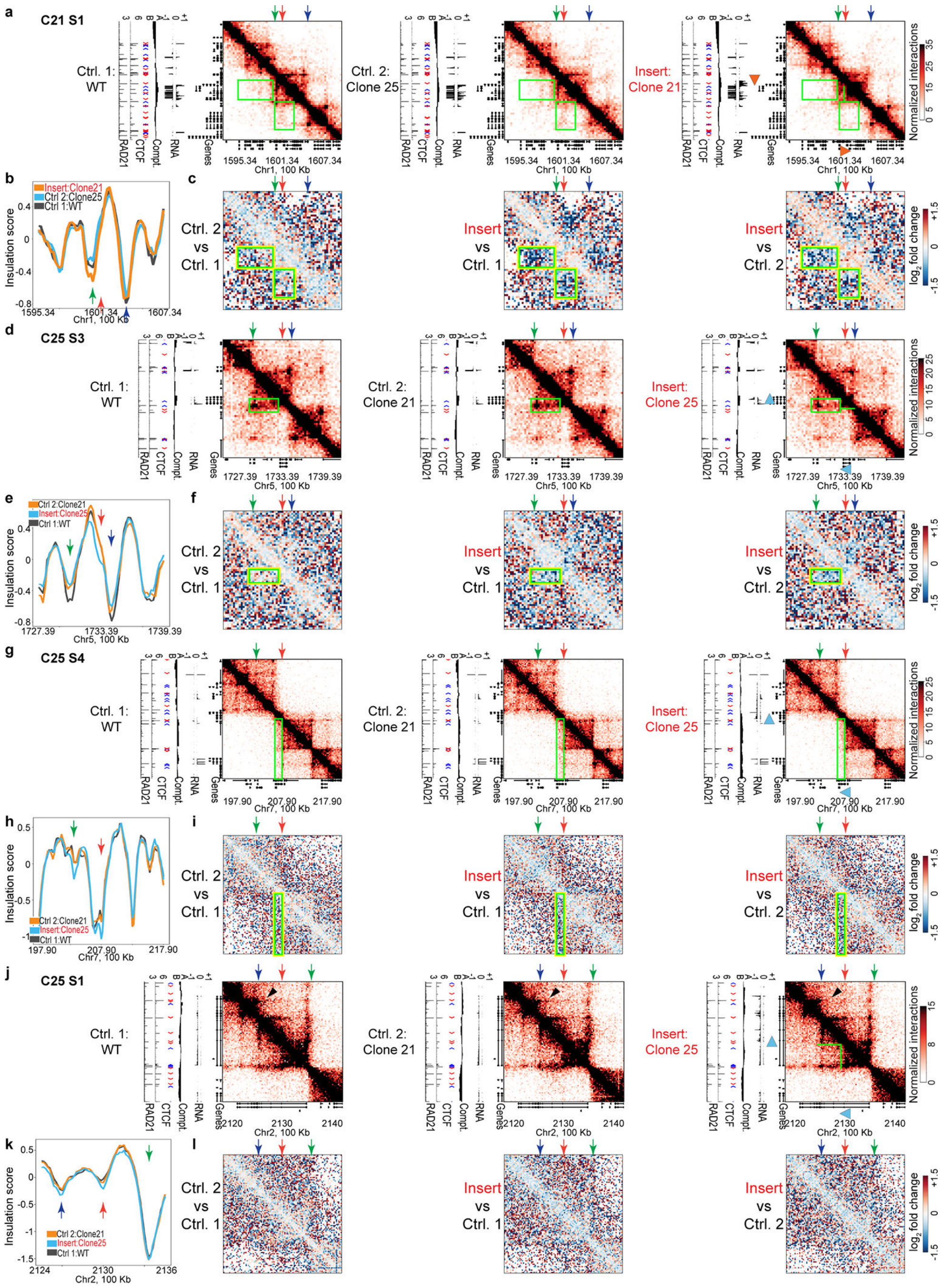
d, Log2 fold changes in interaction frequencies between no-insertion controls and between insertion and no-insertion controls for the region in (c).

(e) and (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:

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).

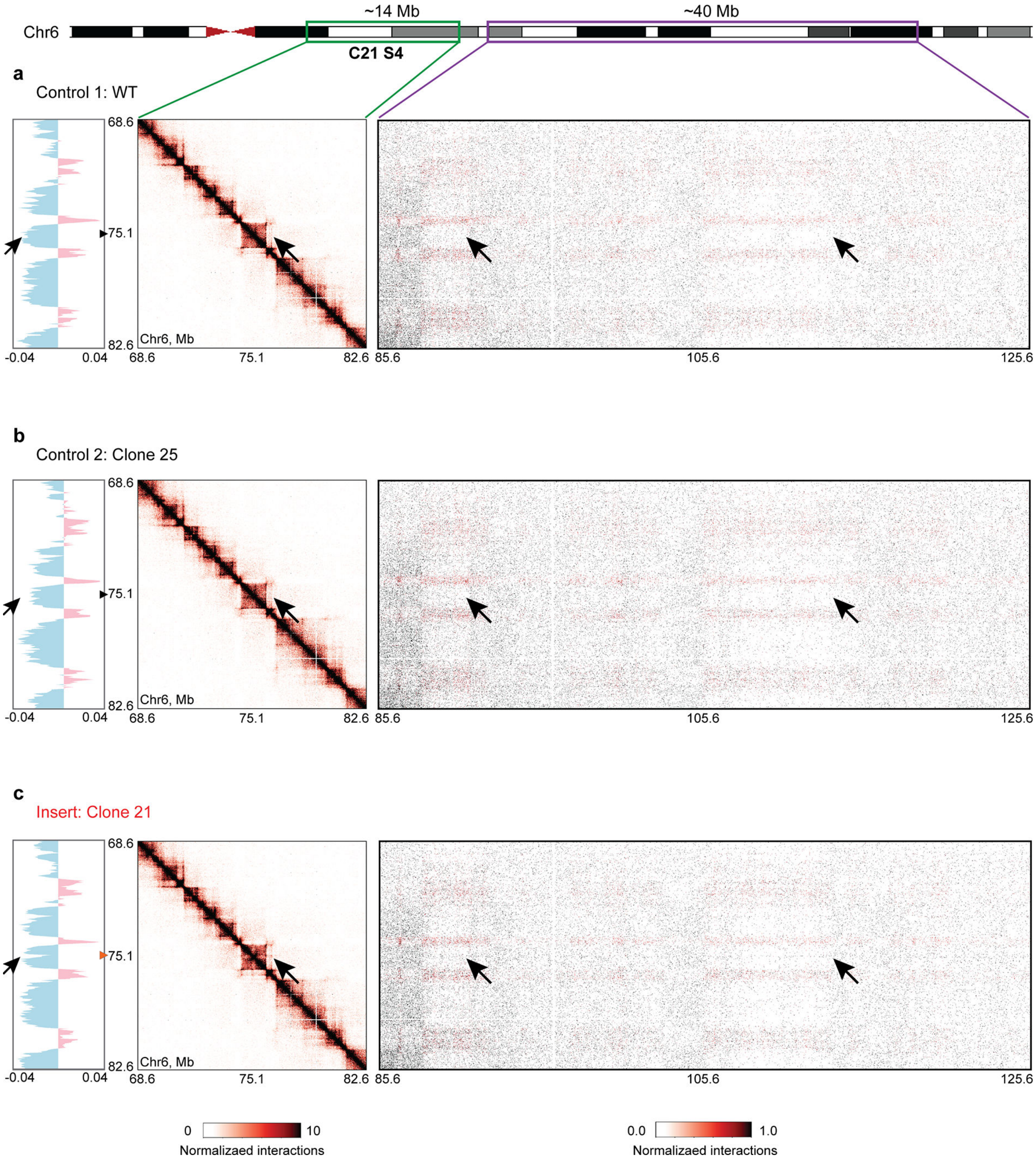
i, Log2 fold changes for the region in (g).

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:

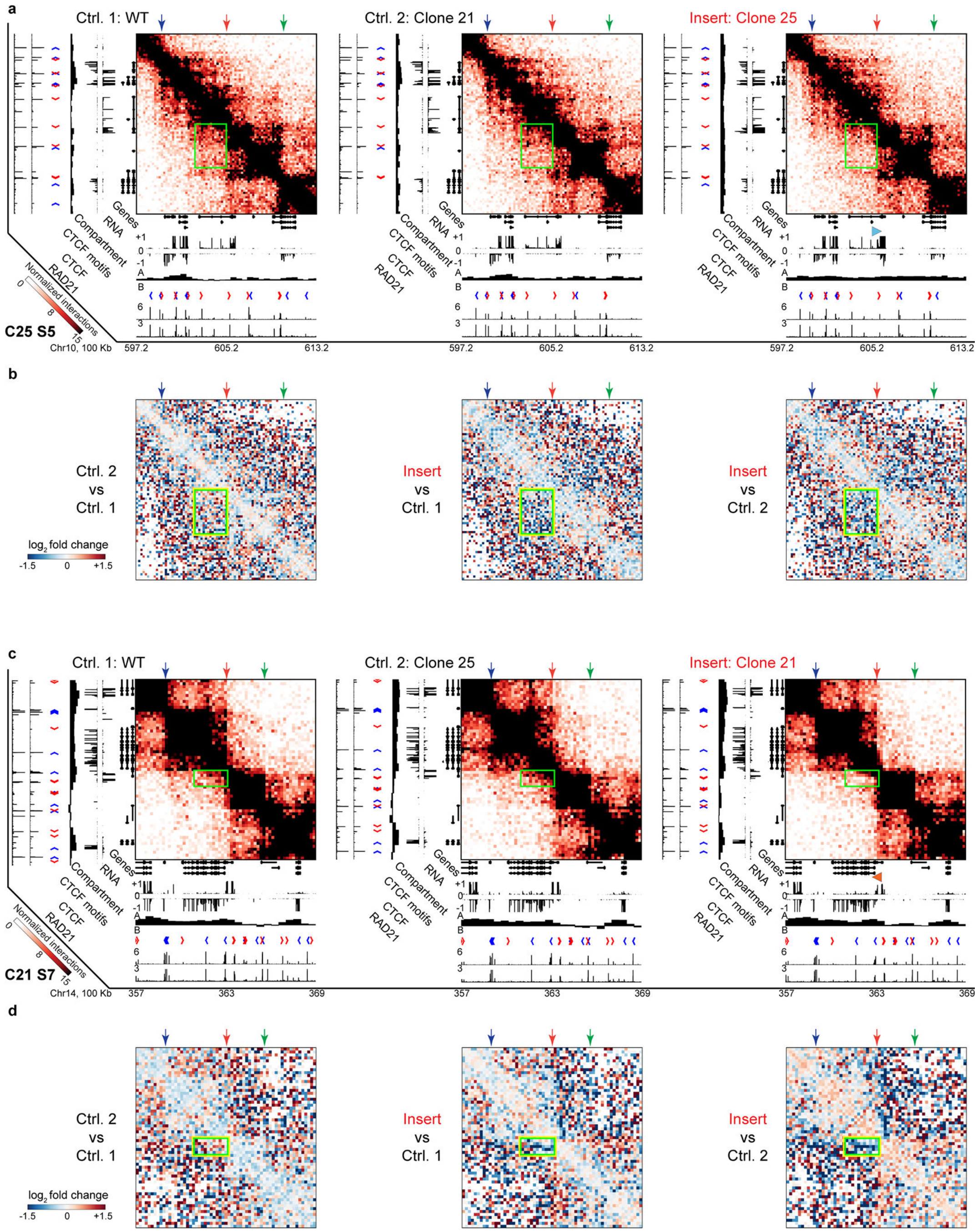
An ectopic insertion can redirect its local chromatin from B to A compartment. Throughout, left: compartment eigenvectors (cyan denotes B compartment; red denotes A compartment) for the ~14 Mb region marked by the green rectangle on the chromosome diagram; middle: Hi-C heatmaps for this ~14 Mb region 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) and (b).

Each Hi-C result depicts merged data from 2 independent Hi-C experiments for each genotype.



Extended Data Fig. 5:

Boundary-associated DNA insertions can strengthen pre-established boundaries: additional controls (an extension to Figure 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) and (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.

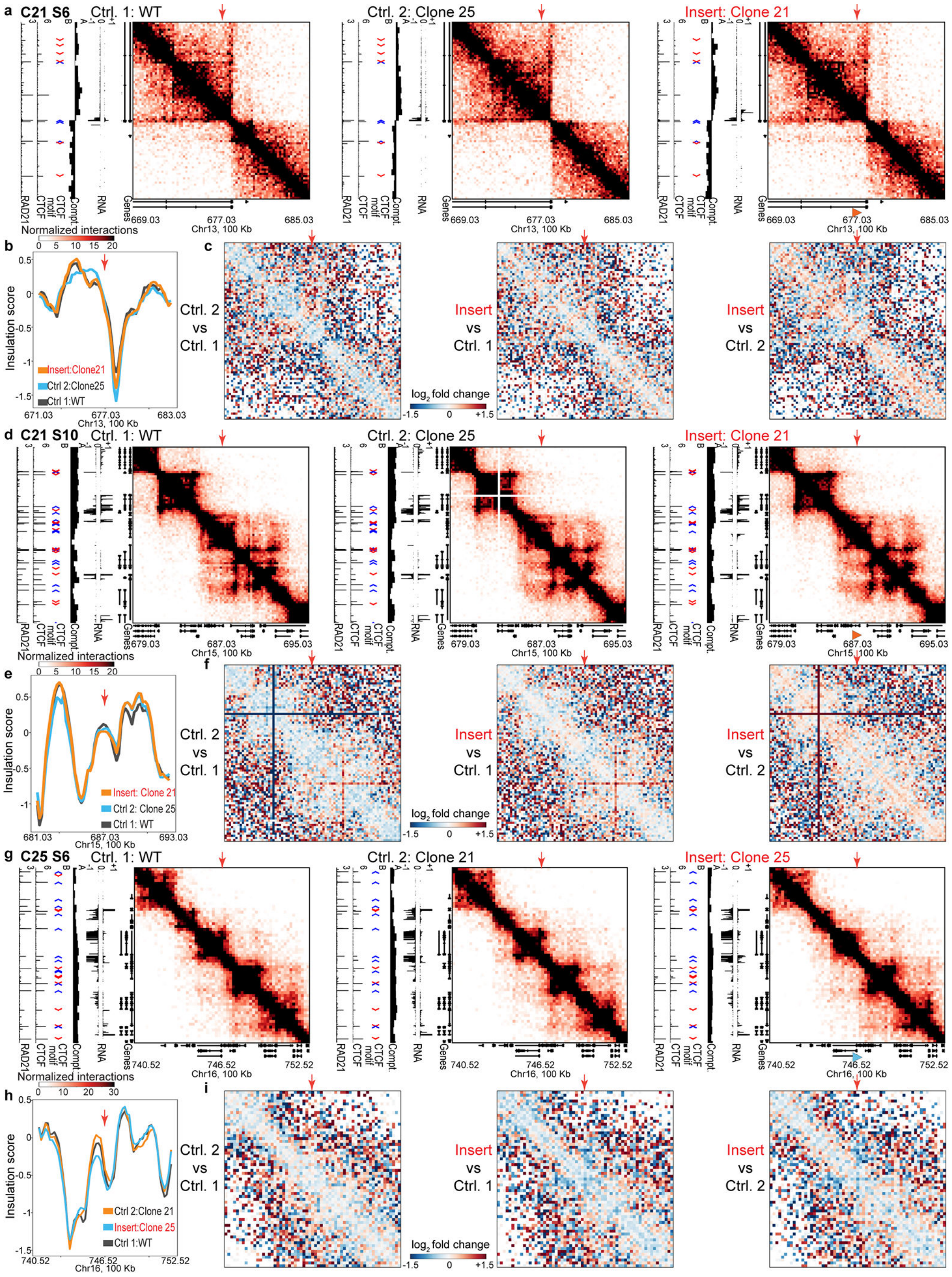
b, An extension to Fig. 2c: log2 fold changes in interaction frequencies between two no-insertion controls (left) and between the insertion clone and no-insertion controls (middle and right) for the region in (a).

(c) and (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:

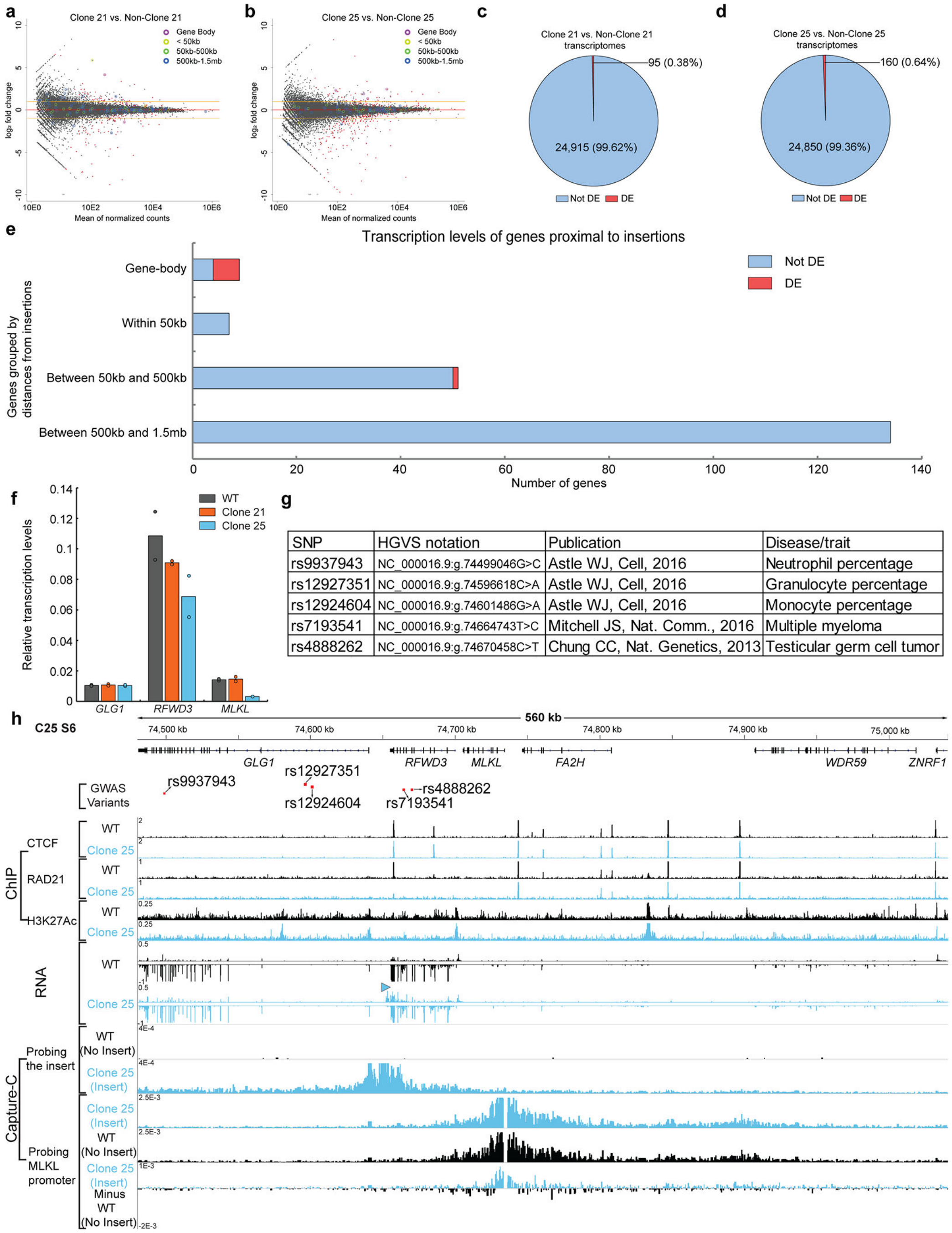
Insertion loci without apparent detectable domain-level changes.

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 the 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.



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:

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)).

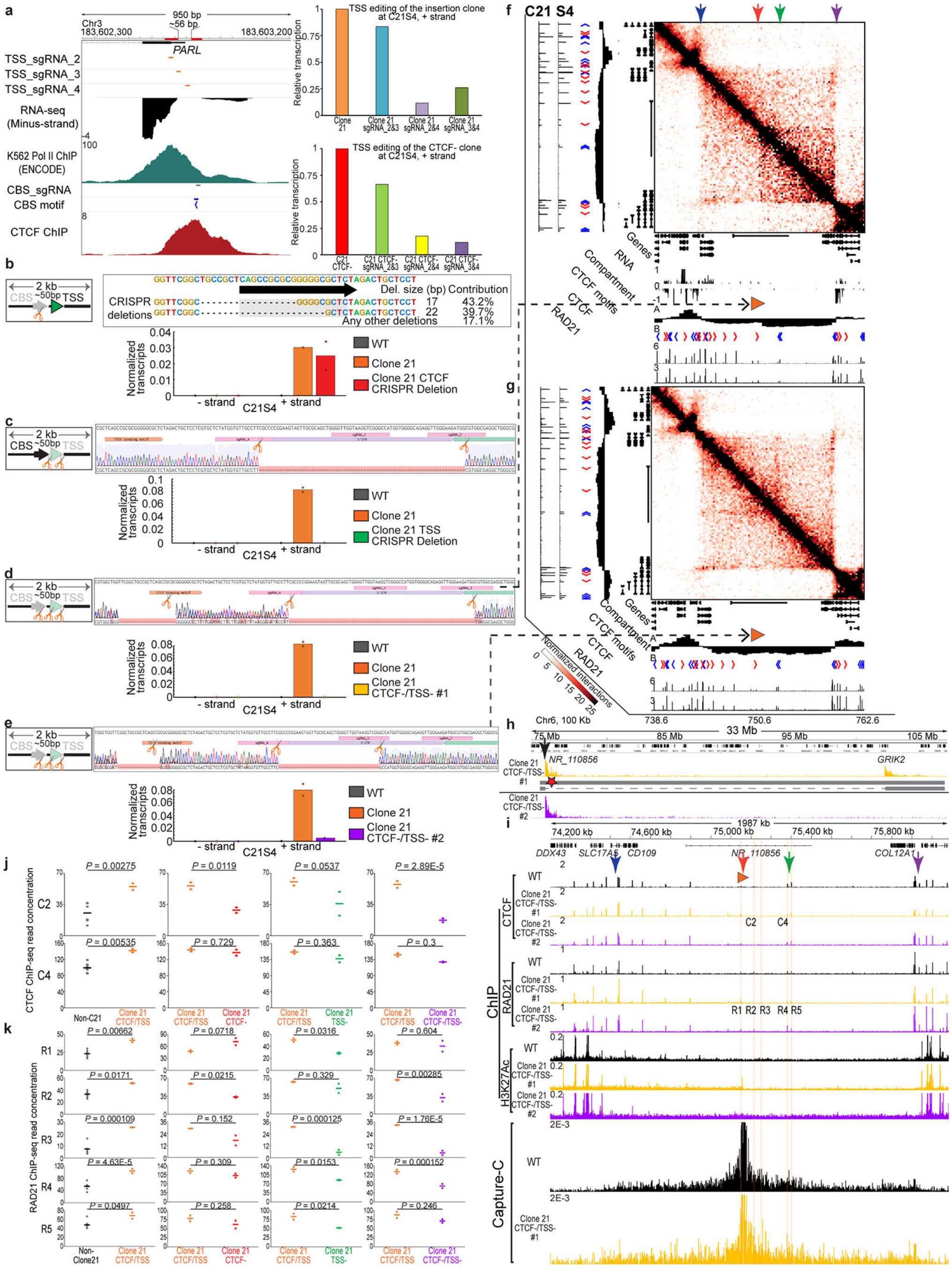
e, DE status of all insertion-proximal genes. The DE gene between 50 kb and 500 kb to an insertion, *MLKL*, is characterized in (f) and (h).

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:

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 TSS-: TSS_sgRNA_2&4-edited Clone 21 abrogates transcription, with the CBS intact.

d, e, Clone 21 CTCF-/TSS- #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 CTCF-/TSS- #1 (d) at C21S4: a ~27 Mb heterozygous deletion (h) influences heatmap interpretation.

g, Hi-C of Clone 21 CTCF-/TSS- #2 (e) at C21S4: domain configuration restored close to pre-insertion level (Fig. 4a).

h, Virtual 4C (black arrow: viewpoint; red star: C21S4; *GRIK2*: C21S5): Clone 21 CTCF-/TSS- #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).

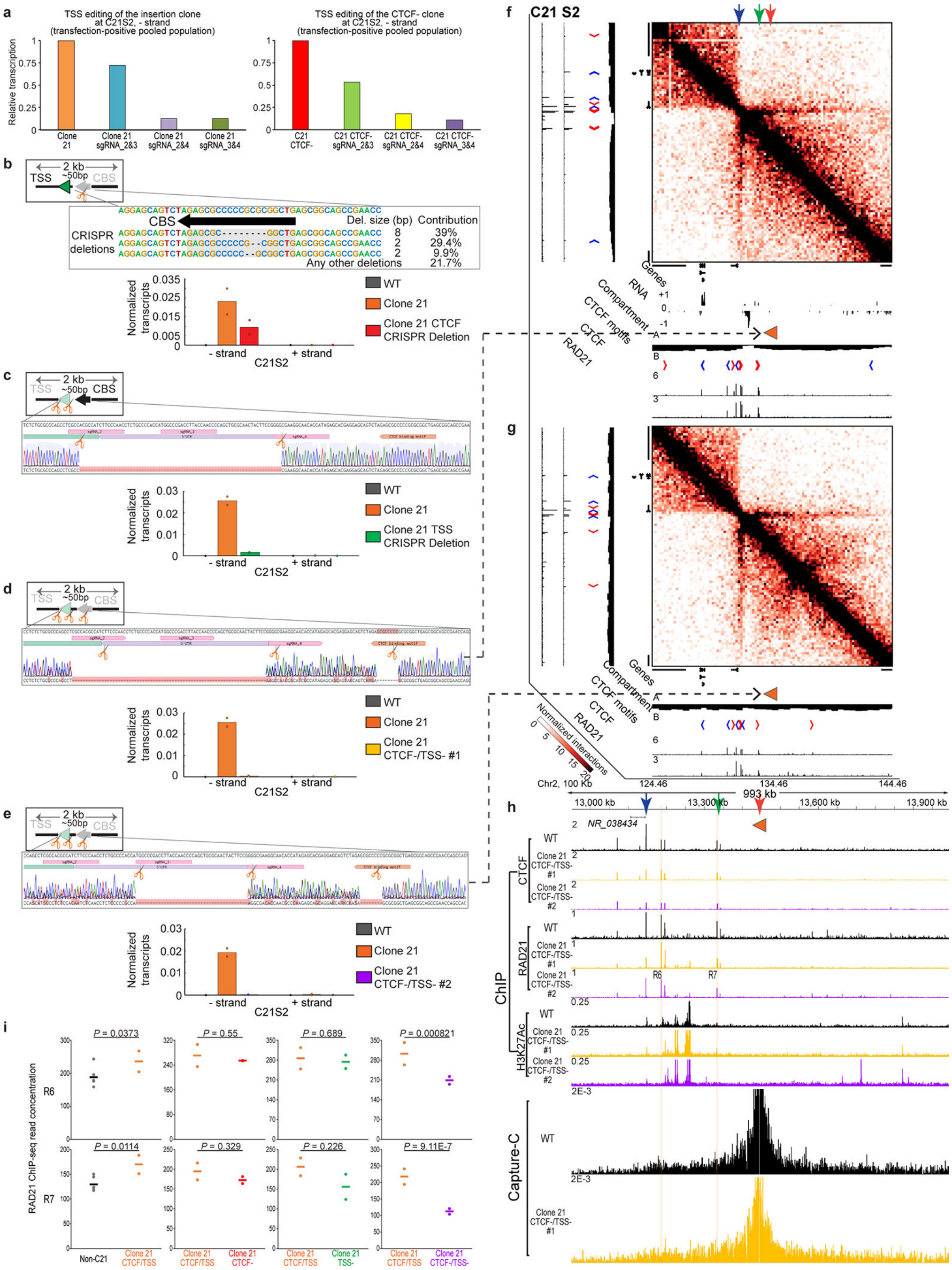
i, CBS-/TSS- restores nearby chromatin folding pattern to pre-insertion levels. Differentially bound CTCF (C2, C4) and RAD21 peaks (R1-R5) upon insertion highlighted. Directionality Index of Clone 21 CTCF-/TSS- #1 Capture-C: Fig. 4j.

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:

CRISPR dissections of insertion, and RAD21 distribution at C21S2.

a, TSS_sgRNA_2&4 and TSS_sgRNA_3&4 (as in Ext 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 TSS- derived from TSS_sgRNA_2&4-edited Clone 21 abrogates transcription, with the CBS intact.

d, Clone 21 CTCF-/TSS- #1: derived from CBS-disrupted Clone 21 (b) further edited with TSS_sgRNA_2&4.

e, Clone 21 CTCF-/TSS- #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.

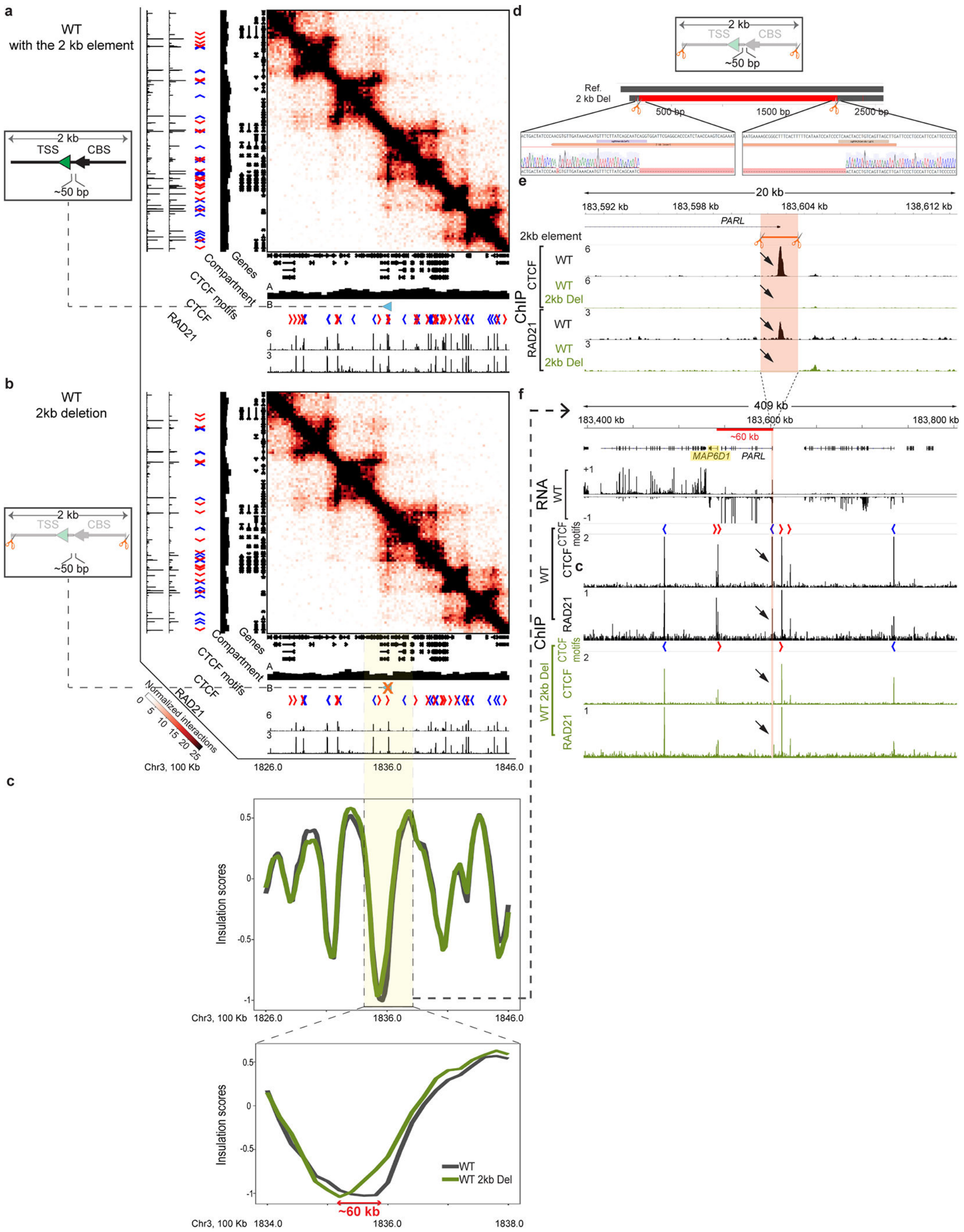
In (f)-(h), red arrow: insertion site; green or blue arrow: downstream CBSs; orange arrowhead in the browser tracks: locus/orientation of the insertion.

f, g, Hi-C maps of Clone 21 CTCF-/TSS- #1 (d) and of Clone 21 CTCF-/TSS- #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 CTCF-/TSS- 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 CTCF-/TSS- #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). **N**on-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:

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank B. van Steensel (Netherlands Cancer Institute) for providing HAP1 cells. We thank Z. Izsvák (Max Delbrück Center) and Z. Ivics (The Paul Ehrlich Institute) for providing the Sleeping Beauty Transposon constructs. We are grateful for 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 their assistance. We thank J. Yano and P. Evans for assistance. We thank members of the Blobel laboratory for helpful discussions. This work was supported by grants 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.

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). [PubMed: 19815776]
2. Nora EP et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385 (2012). [PubMed: 22495304]
3. Dixon JR et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012). [PubMed: 22495300]
4. Rao SSP et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014). [PubMed: 25497547]
5. Phillips-Cremins JE et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153, 1281–1295 (2013). [PubMed: 23706625]
6. Schwarzer W et al. Two independent modes of chromatin organization revealed by cohesin removal. Nature 551, 51–56 (2017). [PubMed: 29094699]
7. Rao SSP et al. Cohesin Loss Eliminates All Loop Domains. Cell 171, 305–320.e24 (2017). [PubMed: 28985562]
8. Rowley MJ et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Molecular Cell 67, 837–852.e7 (2017). [PubMed: 28826674]
9. Nora EP et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930–944.e22 (2017). [PubMed: 28525758]
10. Hug CB, Grimaldi AG, Kruse K & Vaquerizas JM Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. Cell 169, 216–228.e19 (2017). [PubMed: 28388407]
11. Franke M et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature 538, 265–269 (2016). [PubMed: 27706140]

12. Vietri Rudan M et al. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Reports 10, 1297–1309 (2015). [PubMed: 25732821]

13. Fudenberg G & Pollard KS Chromatin features constrain structural variation across evolutionary timescales. PNAS 116, 2175–2180 (2019). [PubMed: 30659153]

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). [PubMed: 27867070]

15. Lupiáñez D et al. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell 161, 1012–1025 (2015). [PubMed: 25959774]

16. Narendra V et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science 347, 1017–1021 (2015). [PubMed: 25722416]

17. Flavahan WA et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 529, 110–114 (2016). [PubMed: 26700815]

18. Hnisz D et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 351, 1454–1458 (2016). [PubMed: 26940867]

19. Zhang Y et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat. Genet. (2019).

20. Barutcu AR, Maass PG, Lewandowski JP, Weiner CL & Rinn JL A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. Nat Commun 9, 1444 (2018). [PubMed: 29654311]

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). [PubMed: 19412179]

22. Carette JE et al. Ebola virus entry requires the cholesterol transporter Niemann–Pick C1. Nature 477, 340–343 (2011). [PubMed: 21866103]

23. Haarhuis JHI et al. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell 169, 693–707.e14 (2017). [PubMed: 28475897]

24. Van Bortle K et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. 15, R82 (2014). [PubMed: 24981874]

25. Mayer A et al. Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. Cell 161, 541–554 (2015). [PubMed: 25910208]

26. Vian L et al. The Energetics and Physiological Impact of Cohesin Extrusion. Cell 173, 1165–1178.e20 (2018). [PubMed: 29706548]

27. Redolfi J et al. DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. Nat. Struct. Mol. Biol. (2019).

28. Sanborn AL et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U. S. A. 112, 6456 (2015).

29. Fudenberg G et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep 15, 2038–2049 (2016). [PubMed: 27210764]

30. Dixon JR et al. Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336 (2015). [PubMed: 25693564]

31. Krijger PHL et al. Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. Cell Stem Cell 18, 597–610 (2016). [PubMed: 26971819]

32. Ke Y et al. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. Cell 170, 367–381.e20 (2017). [PubMed: 28709003]

33. Du Z et al. Allelic reprogramming of 3D chromatin architecture during early mammalian development. Nature 547, 232–235 (2017). [PubMed: 28703188]

34. Heinz S et al. Transcription Elongation Can Affect Genome 3D Structure. Cell 174, 1522–1536.e22 (2018). [PubMed: 30146161]

35. Gong Y et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nat Commun 9, 542 (2018). [PubMed: 29416042]

36. Hughes JR et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet. 46, 205–212 (2014). [PubMed: 24413732]

37. Nuebler J, Fudenberg G, Imakaev M, Abdennur N & Mirny LA Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc. Natl. Acad. Sci. U. S. A. 115, E6697–E6706 (2018). [PubMed: 29967174]

38. Sun L et al. Mixed lineage kinase domain-like protein mediates necrosis signaling downstream of RIP3 kinase. Cell 148, 213–227 (2012). [PubMed: 22265413]

39. Zhao J et al. Mixed lineage kinase domain-like is a key receptor interacting protein 3 downstream component of TNF-induced necrosis. PNAS 109, 5322–5327 (2012). [PubMed: 22421439]

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). [PubMed: 27748397]

41. Shan B, Pan H, Najafov A & Yuan J Necroptosis in development and diseases. Genes Dev. 32, 327–340 (2018). [PubMed: 29593066]

42. Yuan J, Amin P & Ofengeim D Necroptosis and RIPK1-mediated neuroinflammation in CNS diseases. Nat. Rev. Neurosci. 20, 19–33 (2019). [PubMed: 30467385]

43. Chung CC et al. Meta-analysis identifies four new loci associated with testicular germ cell tumor. Nat. Genet. 45, 680–685 (2013). [PubMed: 23666239]

44. Astle WJ et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell 167, 1415–1429.e19 (2016). [PubMed: 27863252]

45. Mitchell JS et al. Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. Nat Commun 7, 12050 (2016). [PubMed: 27363682]

46. Hou C, Zhao H, Tanimoto K & Dean A CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc. Natl. Acad. Sci. U. S. A. 105, 20398–20403 (2008). [PubMed: 19074263]

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 (2017).

48. Cong L et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819–823 (2013). [PubMed: 23287718]

49. Mali P et al. RNA-guided human genome engineering via Cas9. Science 339, 823–826 (2013). [PubMed: 23287722]

50. Busslinger GA 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). [PubMed: 31358994]

52. Choudhary MN et al. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol. 21, 16 (2020). [PubMed: 31973766]

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). [PubMed: 28334904]

54. Zhang H et al. Chromatin structure dynamics during the mitosis-to-G1 phase transition. Nature 576, 158–162 (2019). [PubMed: 31776509]

55. Sundaram V et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome research 24, 1963–1976 (2014). [PubMed: 25319995]

56. Schmidt D et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148, 335–348 (2012). [PubMed: 22244452]

57. Bourque G et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome research 18, 1752–1762 (2008). [PubMed: 18682548]

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). [PubMed: 29563166]

59. Jin F et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290–294 (2013). [PubMed: 24141950]

60. Zhang Y et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504, 306–310 (2013). [PubMed: 24213634]

61. Kentepozidou E et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. Genome Biol. 21, 5 (2020). [PubMed: 31910870]

62. Rowley MJ & Corces VG Organizational principles of 3D genome architecture. Nat. Rev. Genet. 19, 789–800 (2018). [PubMed: 30367165]

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). [PubMed: 28057745]

64. Barutcu AR, Maass PG, Lewandowski JP, Weiner CL & Rinn JL A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. Nat Commun 9, 1444 (2018). [PubMed: 29654311]

65. Hsieh TS et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Molecular Cell 78, 539–553.e8 (2020). [PubMed: 32213323]

66. Krietenstein N et al. Ultrastructural Details of Mammalian Chromosome Architecture. Molecular Cell 78, 554–565.e7 (2020). [PubMed: 32213324]

67. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]

68. Kurita R et al. Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. PLoS ONE 8, e59890 (2013). [PubMed: 23533656]

69. Zayed H, Izsvák Z, Walisko O & Ivics Z Development of hyperactive sleeping beauty transposon vectors by mutational analysis. Mol. Ther. 9, 292–304 (2004). [PubMed: 14759813]

70. Huang P et al. Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. Genes Dev. 31, 1704–1713 (2017). [PubMed: 28916711]

71. Davies JOJ et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat. Methods 13, 74–80 (2016). [PubMed: 26595209]

72. Hsiung CC −. et al. A hyperactive transcriptional state marks genome reactivation at the mitosis-G1 transition. Genes Dev. 30, 1423–1439 (2016). [PubMed: 27340175]

73. Hsiau T et al. Inference of CRISPR Edits from Sanger Trace Data. bioRxiv, 251082 (2019).

74. Kim S, Kim D, Cho SW, 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). [PubMed: 24696461]

75. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). [PubMed: 20110278]

76. Sloan CA et al. ENCODE data at the ENCODE portal. Nucleic Acids Res. 44, 726 (2016).

77. Kerpedjiev P et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 19, 125 (2018). [PubMed: 30143029]

78. Forcato M et al. Comparison of computational methods for Hi-C data analysis. Nat. Methods 14, 679–685 (2017). [PubMed: 28604721]

79. Crane E et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523, 240–244 (2015). [PubMed: 26030525]

80. Filippova D, Patro R, Duggal G & Kingsford C Identification of alternative topological domains in chromatin. Algorithms Mol Biol 9, 14 (2014). [PubMed: 24868242]

81. Eisenberg E & Levanon EY Human housekeeping genes, revisited. Trends Genet. 29, 569–574 (2013). [PubMed: 23810203]

82. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). [PubMed: 22388286]

83. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009). [PubMed: 19505943]

84. Imakaev M et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999–1003 (2012). [PubMed: 22941365]

85. Servant N et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16, 259 (2015). [PubMed: 26619908]

86. Gilgenast TG & Phillips-Cremins JE Systematic Evaluation of Statistical Methods for Identifying Looping Interactions in 5C Data. Cell Syst 8, 197–211.e13 (2019). [PubMed: 30904376]

87. Hunter JD Matplotlib: A 2D Graphics Environment. Computing in Science Engineering 9, 90–95 (2007).

88. 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). [PubMed: 29514181]

89. 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). [PubMed: 29140473]

90. Durand NC et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst 3, 95–98 (2016). [PubMed: 27467249]

91. Martin M Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).

92. Magoč T & Salzberg SL FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963 (2011). [PubMed: 21903629]

93. Langmead B Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11, Unit 11.7 (2010).

94. Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008). [PubMed: 18798982]

95. 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). [PubMed: 24743992]

96. Ramírez F et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, 160 (2016).

97. Ross-Innes CS et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481, 389–393 (2012). [PubMed: 22217937]

98. Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 14, 417–419 (2017). [PubMed: 28263959]

99. Soneson C, Love MI & Robinson MD Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4, 1521 (2015). [PubMed: 26925227]

100. Love MI, Huber W & Anders S Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014). [PubMed: 25516281]

101. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]

102. Weiss MJ, Yu C & Orkin SH 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). [PubMed: 9032291]

103. Norton HK et al. Detecting hierarchical genome folding with network modularity. Nat Methods 15, 119–122 (2018). [PubMed: 29334377]

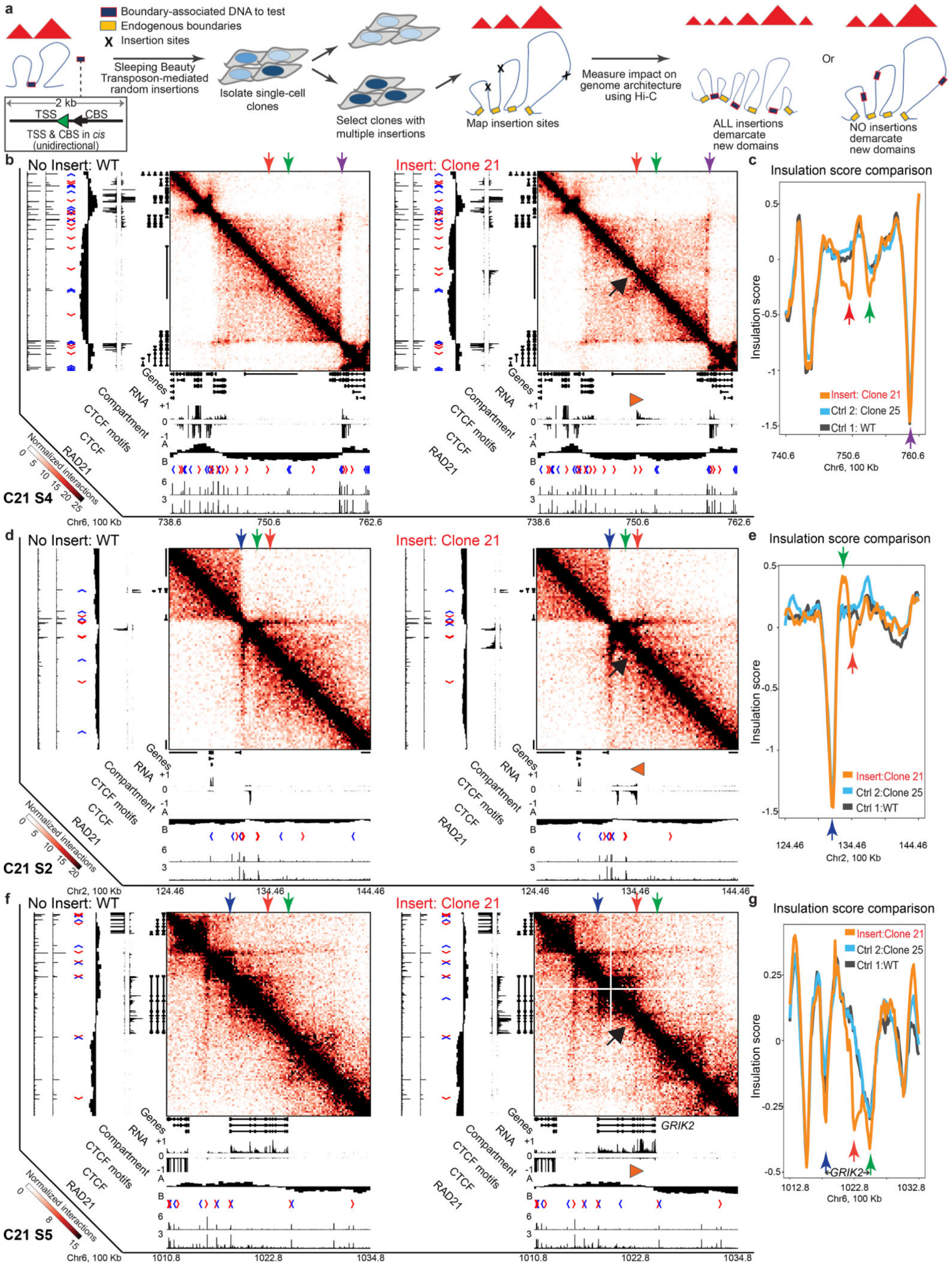


Figure 1 |. Domain boundary insertions create *de novo* contact domains.

a, Schematic of experimental design.

Throughout, red arrow: insertion site; green arrow: up- or down-stream 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 Clone 21 Site 4 (C21S4, right): *de novo* domain formation upon insertion. Additional no insertion control in Extended Data Figure 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 Clone 21 site 2 (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) upon insertion. Additional no insertion control in Extended Data Figure 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 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.

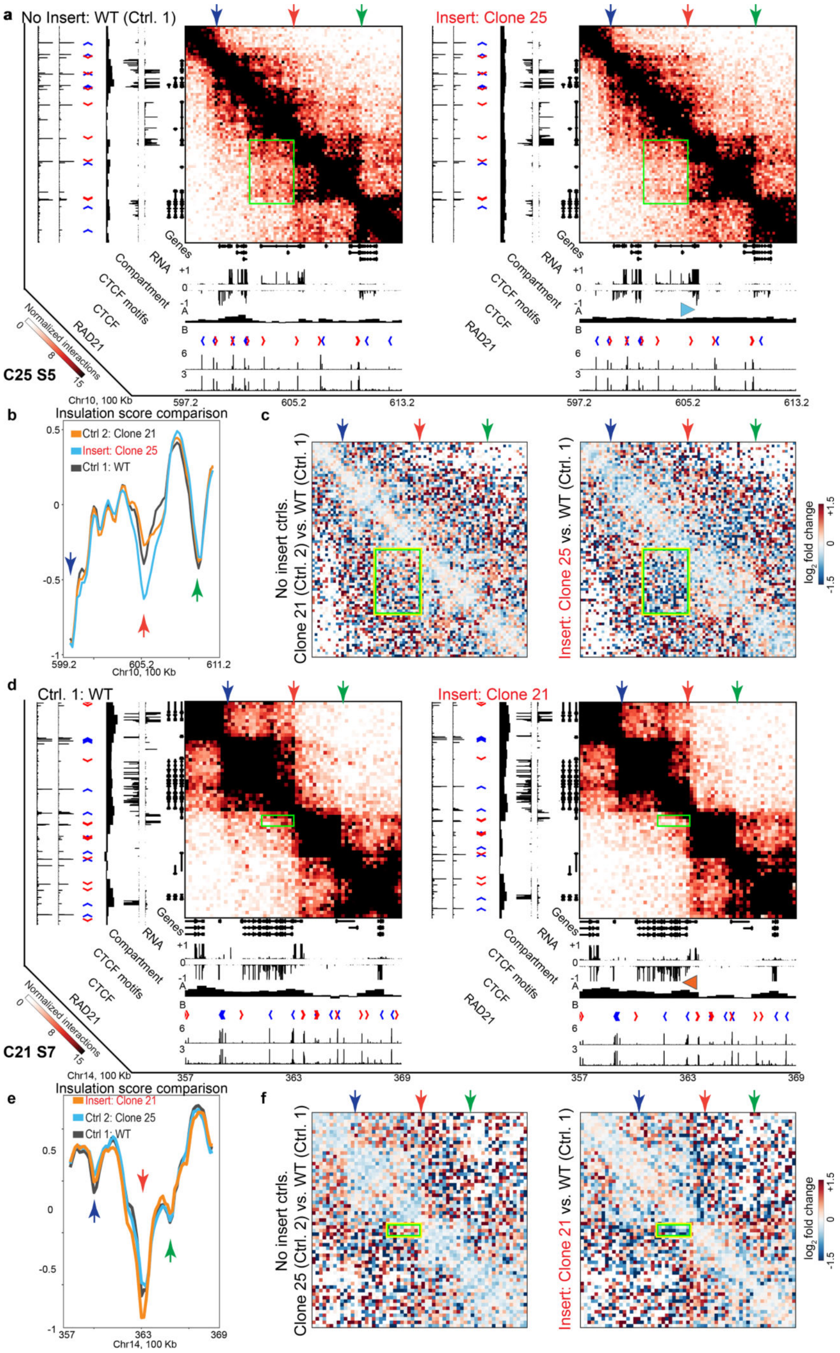


Figure 2 |. Domain boundary insertions can strengthen pre-established boundaries.

Throughout, red arrow: insertion site; green/blue arrow: nearby boundaries. The orange (Clone 21)/blue (Clone 25) arrowhead in the browser tracks marks the site and orientation of the insertion.

In **a**, **c**, **d**, **f**, yellow/green rectangles denote corresponding regions with overall decreased interactions upon insertion.

a, Hi-C contact maps of control (no insertion, left) and Clone 25 Site 5 (C25S5, right). The insertion strengthens an existing domain boundary by decreasing interactions across it. Additional no insertion control in Extended Data Figure 5.

b, Insulation scores from Hi-C results in (a) showing strengthened insulation upon insertion.

c, Log2 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 Clone 21 Site C21S7. The insertion strengthens an existing boundary by decreasing interactions in its immediate proximity. Additional no insertion control in Extended Data Figure 5.

e, Insulation scores from Hi-C results in (d).

f, Log2 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 2 independent experiments performed for each genotype. 2 CTCF & RAD21 ChIP-seq and 2 RNA-seq experiments were conducted for each genotype, with 1 of each shown.

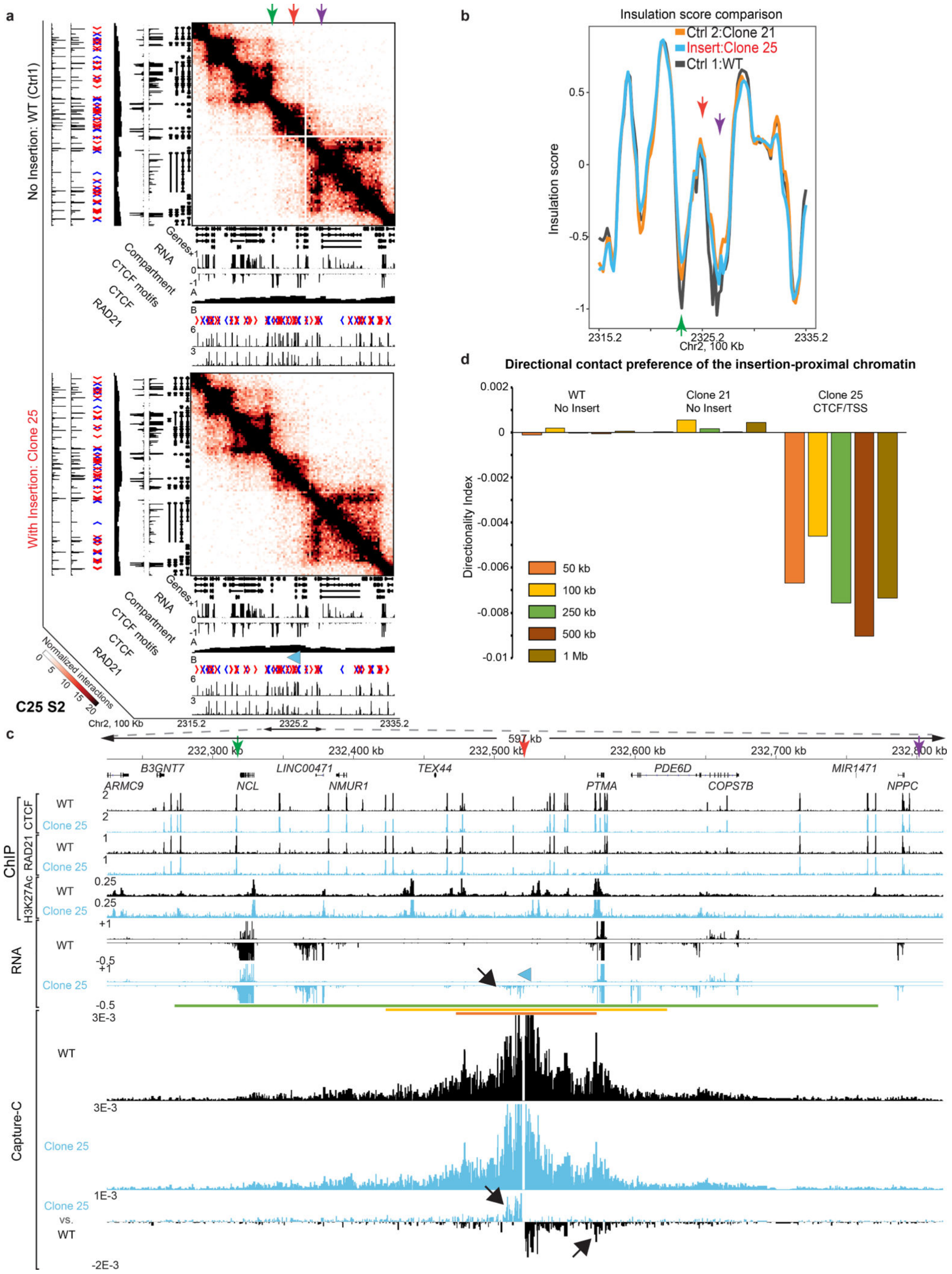


Figure 3 |. An insertion into a complex genomic region modestly changes short-range interactions, without domain-level impact.

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 clone 25 site 2 (C25S2, bottom) showing no obvious domain-level changes upon insertion.

b, Insulation scores from Hi-C results in (a), confirming the little apparent changes at the insertion site. Variations at the two boundaries (green and purple arrows) flanking the insertion likely 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 transcribed region, and reduced interactions in the opposite direction (Capture-C: Clone 25 vs. WT). Colored lines denote distance ranges for measuring Directionality Indices in (d).

d, Directionality Indices (DI) revealing that the insertion induces preferential contacts to the left (negative DI).

Hi-C/Capture-C results represent merged data from 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq, 1 H3K27ac ChIP-seq and 2 RNA-seq experiments were performed for each genotype, with 1 of each shown.

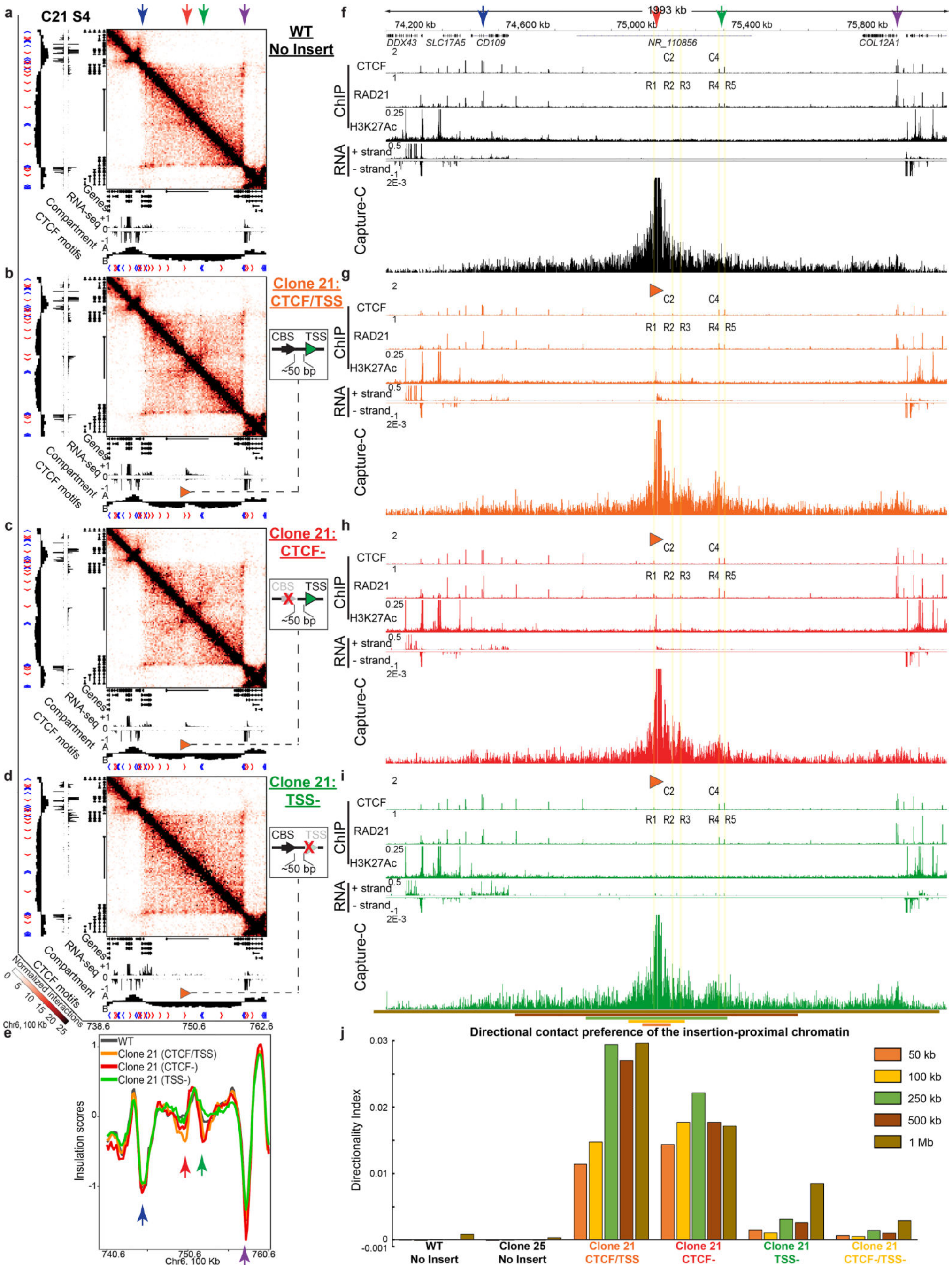


Figure 4 |. TSS can influence domain formation by switching its compartment signature.

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 Clone 21 with CBS/TSS insertion showing the new domain.

(a) and (b) same as in Fig. 1b shown here for comparison.

c, The new domain persists upon CBS deletion.

d, The new domain diminishes upon 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 upon 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, Directionality Index on Capture-C data (f)-(i) uncovers contributions of CBS and TSS to local chromatin folding (Extended Data Fig. 8i: CTCF-/TSS-).

Hi-C/Capture-C results depict merged data of at least 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq, 1 H3K27ac ChIP-seq and 2 RNA-seq experiments were performed for each genotype, with 1 of each shown.

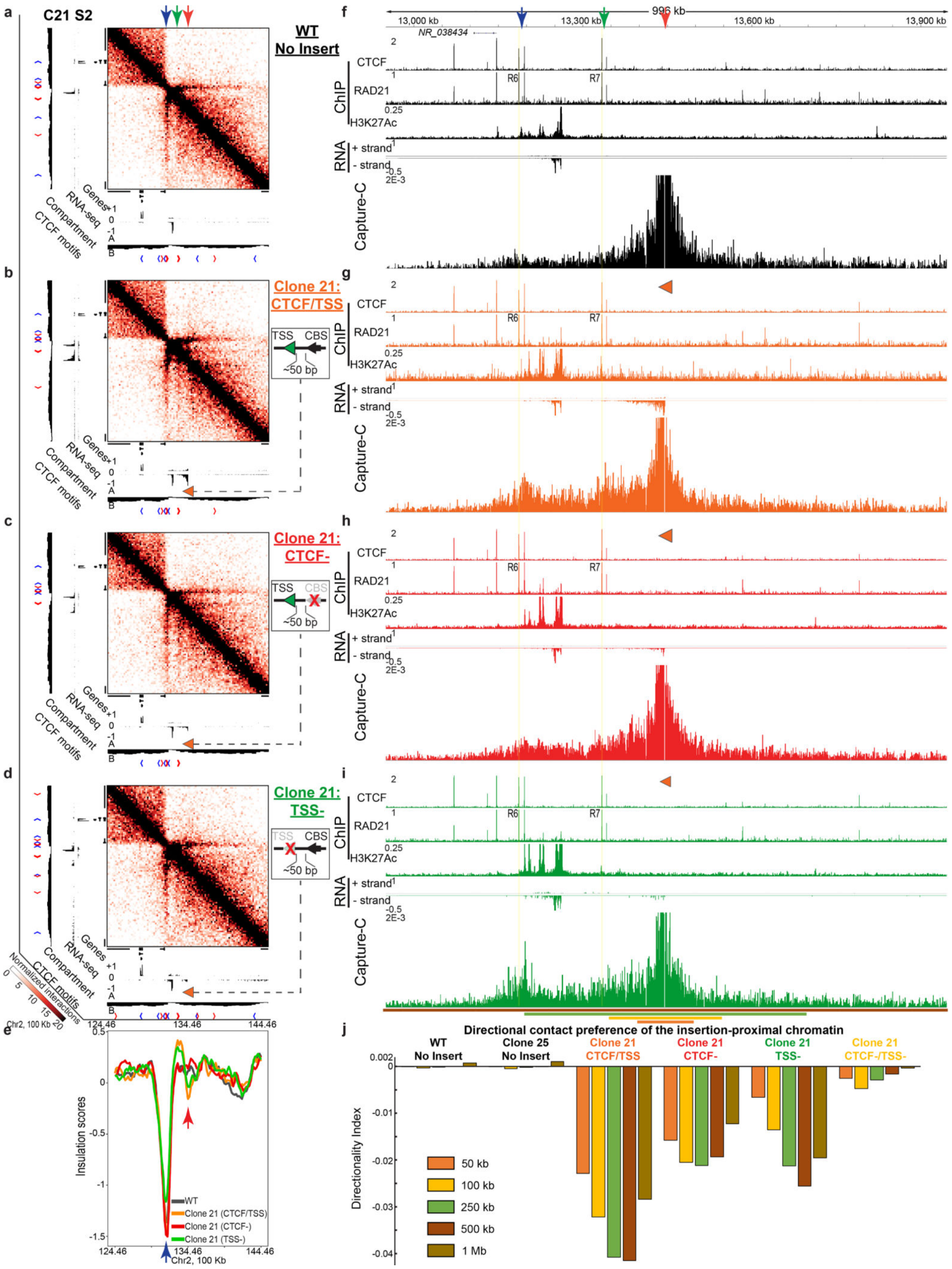


Figure 5 |. TSS and CTCF cooperatively contribute to new domain formation by driving proximal and distal genome folding, respectively.

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) and (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 upon 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, Directionality Index on Capture-C data (f)-(i): TSS and CTCF drives proximal and distal chromatin folding, respectively (Extended Data Fig. 9h: Clone 21 CTCF-/TSS-).

Hi-C/Capture-C results depict merged data of at least 2 independent experiments for each genotype. 2 CTCF & RAD21 ChIP-seq, 1 H3K27ac ChIP-seq and 2 RNA-seq experiments were performed for each genotype, with 1 of each shown.

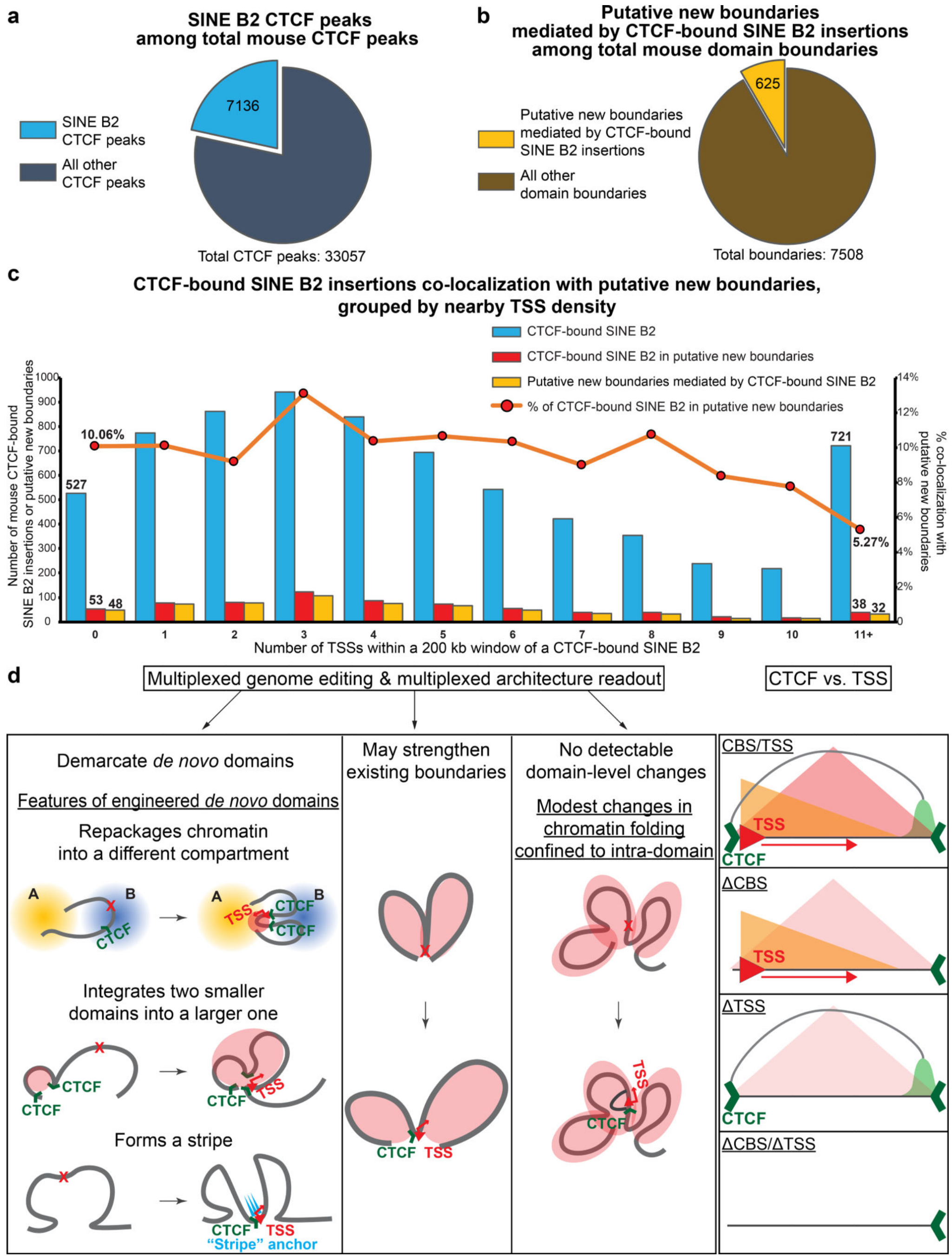


Figure 6 |. Possible context dependency in how SINE B2 elements shape mouse genome architecture in recent evolution, and graphic summary of this study.

a, SINE B2-derived CTCF peaks (~7,136, sky blue) constitute ~21.6% of all CTCF peaks (~33,057) in 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-colocalized CTCF-bound SINE B2 (red column) among all CTCF-bound SINE B2 (blue column). At TSS density = 0, 10.06% (53/527) CTCF-bound SINE B2 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 this study.