nature cell biology

LETTERS

https://doi.org/10.1038/s41556-019-0273-x

Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations

Katerina Kraft[1,2,7,8], Andreas Magg[1,8], Verena Heinrich[3,8], Christina Riemenschneider[1], Robert Schöpflin[1,2], Julia Markowski[3], Daniel M. Ibrahim[1,2,4], Rocío Acuna-Hidalgo[1,2], Alexandra Despang[1,4], Guillaume Andrey[1,4], Lars Wittler[5], Bernd Timmermann[6], Martin Vingron[3] and Stefan Mundlos[1,2,4*]

Balanced chromosomal rearrangements such as inversions and translocations can cause congenital disease or cancer by inappropriately rewiring promoter–enhancer contacts[1,2]. To study the potentially pathogenic consequences of balanced chromosomal rearrangements, we generated a series of genomic inversions by placing an active limb enhancer cluster from the *Epha4* regulatory domain at different positions within a neighbouring gene-dense region and investigated their effects on gene regulation in vivo in mice. Expression studies and high-throughput chromosome conformation capture from embryonic limb buds showed that the enhancer cluster activated several genes downstream that are located within asymmetric regions of contact, the so-called architectural stripes[3]. The ectopic activation of genes led to a limb phenotype that could be rescued by deleting the CCCTC-binding factor (CTCF) anchor of the stripe. Architectural stripes appear to be driven by enhancer activity, because they do not form in mouse embryonic stem cells. Furthermore, we show that architectural stripes are a frequent feature of developmental three-dimensional genome architecture often associated with active enhancers. Therefore, balanced chromosomal rearrangements can induce ectopic gene expression and the formation of asymmetric chromatin contact patterns that are dependent on CTCF anchors and enhancer activity.

Precise spatiotemporal gene expression is essential for normal development and homeostasis. Cell-type- and time-specific gene expression is driven by *cis*-regulatory elements known as enhancers through long-range chromatin contacts with their cognate promoters. The specificity of enhancer–promoter contacts is, in part, controlled by the three-dimensional organization of the genome into topologically associated domains (TADs), megabase-sized chromatin domains that are visually and computationally identified in high-throughput chromosome conformation capture (Hi-C) and chromosome conformation capture carbon-copy (5C) data[4–6]. TADs are frequently flanked by convergently oriented CCCTC-binding factor (CTCF) sites that insulate the regulatory activity within a TAD, defining the genes that an enhancer can act on[7,8]. Although TADs constrain the interactions within the domains, TAD substructures seem to direct enhancers to their specific target genes and/or promoters[9,10]. Rearrangements that break this architectural configuration can result in contact between previously separated units (TAD fusion or TAD shuffling), gene misexpression and disease[2,11,12]. Although this principle provides a plausible explanation for enhancer–promoter contacts and gene activation in regions in which clear contact boundaries exist, it remains unclear whether it is also valid for gene-rich regions in which no clear TAD structures can be discerned. Moreover, the identification of TADs in Hi-C maps is highly dependent on the resolution of the data and the bioinformatic algorithms used[13], leading to differing definitions of the architectural units and structures. The functional and biological relevance of the various concepts of chromatin structures, however, remains to be shown.

To systematically probe the relationship between TAD boundaries, their disruption and gene expression, we investigated the *Epha4* locus. Genomic rearrangements at this locus have been implicated in a number of conditions, including limb and craniofacial malformations[2,14]. Furthermore, the region contains several developmentally important genes such as *Pax3*, *Epha4*, *Ihh*, *Wnt6* and *Wnt10a*. Genes of the hedgehog and WNT pathways have also been implicated in the pathogenesis of malignant processes such as prostate and colon cancers[15]. We studied the three-dimensional configuration of this locus by using capture Hi-C (cHi-C)[11,15] in embryonic day E11.5 mouse limb buds, a tissue and developmental stage at which *Epha4* and many other genes in the locus are active. Capture Hi-C shows that the genes *Epha4*, *Pax3* and *Pinc* reside in clearly defined TADs, bordered by strong loop-forming CTCF sites (Fig. 1a). By contrast, the gene-dense region located between the *Epha4* and *Pinc* TADs shows no clear structure. The gene density, number of genes and lack of defined boundaries suggest that this region is not a single regulatory unit. Such gene-dense regions appear to follow specific characteristics that are different from TADs that contain few or only one gene.

To study gene regulation in more detail in this region, we used serial inversions to relocate a previously described *Epha4* enhancer

[1]RG Development & Disease, Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, Berlin, Germany. [3]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [4]Berlin-Brandenburg Center for Regenerative Therapies, Charité Universitätsmedizin Berlin, Berlin, Germany. [5]Department of Developmental Genetics, Max Planck Institute for Molecular Genetics, Berlin, Germany. [6]Sequencing Core Facility, Max Planck Institute for Molecular Genetics, Berlin, Germany. [7]Present address: Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA. [8]These authors contributed equally: Katerina Kraft, Andreas Magg, Verena Heinrich *e-mail: mundlos@molgen.mpg.de

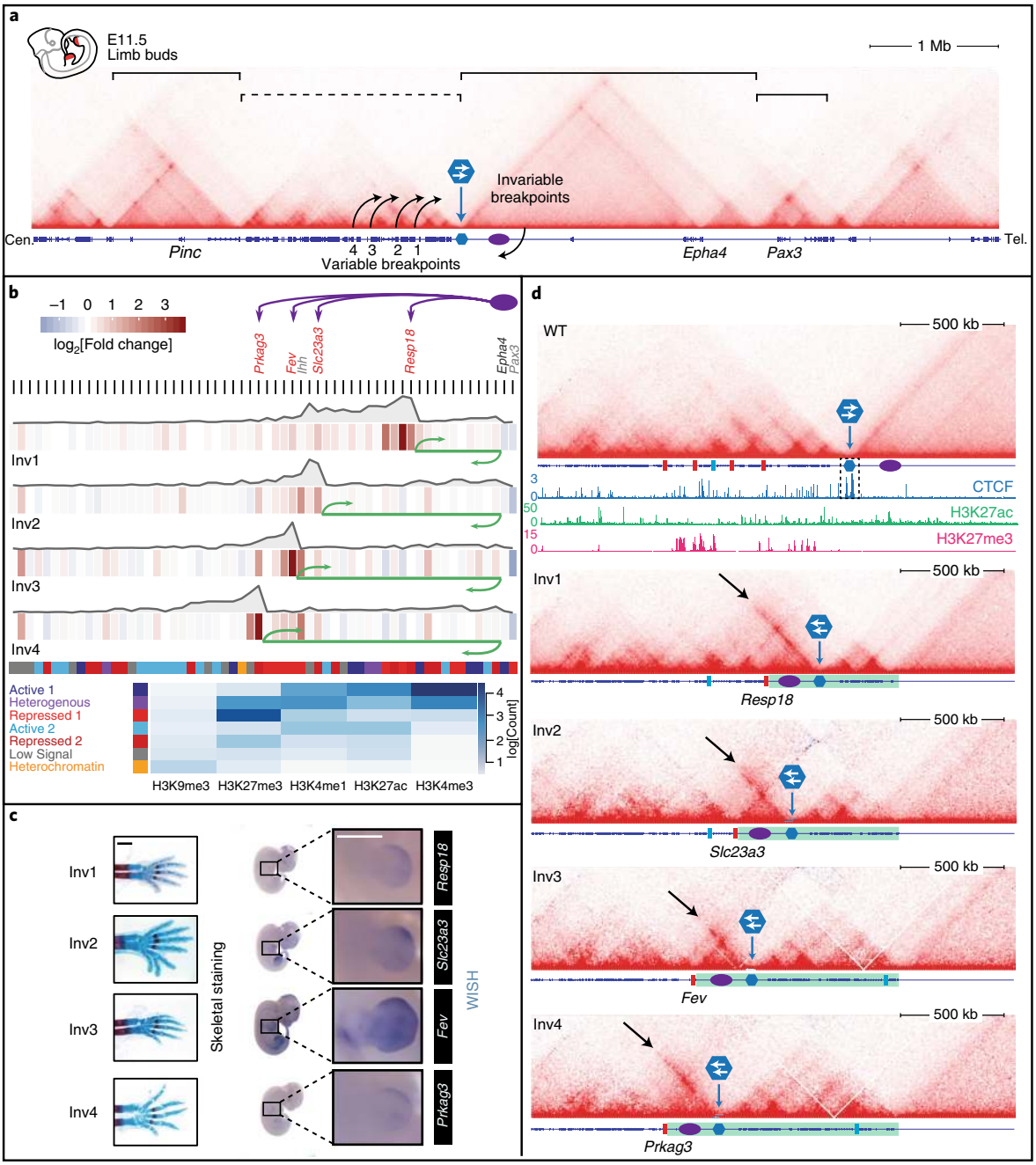

Fig. 1 | Serial inversions induce architectural stripes leading to changes in gene expression. a, Overview of the locus and inversions. Breakpoint inversions (1, 2, 3 and 4) at different promoters (variable) and the invariable breakpoint at the telomeric side of the enhancer cluster (purple oval) are shown in a cHi-C map from wild-type (WT) distal limb buds of E11.5 mice. The blue hexagon indicates the boundary between the *Epha4* TAD and the gene-dense region; the white arrows indicate the CTCF motif orientation. Note the clear TAD formation with CTCF-based loops and boundaries in the gene-poor regions (solid lines) in contrast to the gene-dense region (dashed line). Cen., centromere; Tel., telomere. **b**, Several layers of information per gene (black ticks) from distal limb buds of E11.5 mice showing expression and contact changes. The inverted regions are indicated by green lines and the breakpoint genes are highlighted in red. The change in gene expression in inversions compared to WT is represented by the $\log_2$ fold change. Promoter–enhancer contact frequencies are shown as grey curves. Genome-wide segmentation was performed on ChIP-seq data for histone modifications (average levels of histone marks; see blue colour scale). **c**, Skeletal staining from E16.5 mouse embryos carrying inversions, and WISH from the first gene at the breakpoint. Inv2 mice show polydactyly and Inv1 mice have a duplicated thumb, whereas Inv3 and Inv4 mice show no abnormality. WISH shows a gain in expression corresponding to the activity pattern of the *Epha4* enhancer. WISH and skeletal staining were performed three times with similar results. Scale bars, 1 mm. **d**, cHi-C and ChIP-seq from E11.5 distal limb buds. Top, WT cHi-C (magnification from **a**) with ChIP-seq for CTCF (boundary-forming CTCF sites are displayed enlarged in the black-dotted box), H3K27ac and H3K27me3. Bottom, cHi-C maps of inversions 1–4. Green bars indicate inverted regions; blue and red rectangles denote the position of *Ihh* or the first gene at the breakpoint, respectively; architectural stripes are indicated by black arrows. Capture Hi-C and ChIP-seq for H3K9me3 were performed once with limb buds collected from 10 embryos per condition. ChIP-seq data in **b** and **d**, except for H3K9me3, were obtained from a published data set (accession code: GSE84795).

cluster with strong activity in the distal limb bud[2] from the *Epha4* TAD into the gene-dense region (Fig. 1a). The inversions consisted of an invariable breakpoint that was located telomeric of the enhancer cluster, and a variable centromeric breakpoint located at the promoter of the genes *Resp18* (Inv1), *Slc23a3* (Inv2), *Fev* (Inv3), and *Prkag3* (Inv4) in the gene-dense region. To standardize our comparison across breakpoints, we chose genes that are coded from the same strand and are not expressed in WT E11.5 limb buds (based on RNA-sequencing (RNA-seq) data), the stage and tissue in which the enhancer cluster is active. These genes were also selected because of their different histone modification patterns (H3K27ac, H3K27me3, H3K4me1 and H3K4me3) to compare the effects of the histone landscape on the inversion phenotype (Supplementary Fig. 1a). Specifically, *Resp18* and *Slc23a3* show low to no signal in H3K27me3 chromatin immunoprecipitation sequencing (ChIP-seq) data and no CpG islands, whereas *Prkag3* and *Fev* show repressive (H3K27me3) and active (H3K4me3) chromatin signatures, respectively, indicative of bivalency or heterogeneous behaviour[16,17]. *Fev* also exhibits these marks at the promoter region and at a CpG island at the end of the gene. The inversions include two strong CTCF sites that constitute the *Epha4* boundary with the motifs oriented towards the *Epha4* gene. Thus, when inverted, these CTCF sites are now positioned on the telomeric side of the enhancer, pointing away from *Epha4* and towards the enhancer cluster, into the gene-dense region.

To study the effect of these rearrangements on gene expression during development in vivo, we produced homozygous (Inv1) and heterozygous (Inv2,-3 and -4) mutant embryos and performed gene expression analysis in E11.5 limb buds using RNA-seq (Fig. 1b).

The activating effect of the enhancer on the closest gene varied across individual inversions with *Fev* exhibiting the strongest upregulation. Notably, gene activation was not confined to the gene closest to the enhancers on the linear genome but extended several genes further towards the centromere. Using the chromatin segmentation software EpiCSeg[18], we characterized the epigenomic landscape at the promoter regions in WT limb buds by integrating histone ChIP-seq signals (Fig. 1b, Supplementary Fig. 1b). We observed a positive correlation between the abundance of H3K27me3 marks at the promoter and gene activation, indicating that polycomb-repressed genes that had a H3K27me3 ChIP-seq signal responded more strongly to the enhancer. These genes require less contact with the enhancer to reach the same $\log_2$ fold change in expression than genes without H3K27me3 (Supplementary Fig. 1c). Gene activation beyond the direct enhancer vicinity was most pronounced in Inv1 and most restricted in Inv4. Interestingly, the propagation of enhancer-gene activation appeared to decrease for all inversions with distance from the breakpoint, resulting in a total gene activation stop at around the same genomic position.

We used whole-mount in situ hybridization (WISH) to analyse the spatial distribution of gene expression in limb buds (Fig. 1c). WISH from the first gene at the breakpoint showed a gain in expression in an *Epha4*-like pattern in the anterior distal mesenchyme. These genes are not expressed in the WT embyros. To understand how these structural alterations contribute to developmental abnormalities, we analysed the phenotypes of each inversion by skeletal preparations (Fig. 1c). Inv2 mice developed severe polydactyly, presumably because *Ihh*, the third gene after the breakpoint, is ectopically activated. Similarly, Inv1 mice developed preaxial polydactyly despite the fact that 13 other genes are located between the enhancers and *Ihh*. Normally, *Ihh* is not expressed in limb buds before E12.5 but it is known to cause polydactyly when expressed earlier in development[14]. We observed ectopic, asymmetric expression of *Ihh* in the anterior portion of the limb bud in the Inv2 and, to a lesser degree, in the Inv1 inversion mutants (Supplementary Fig. 1d). This was reminiscent of the doublefoot (*Dbf*) mutant, in which the same enhancer cluster is positioned close to *Ihh*, but less pronounced[2]. The magnitude of ectopic misexpression decreased with the increase in linear distance between the enhancers and *Ihh*, which also corresponded to the number of extra digits. Interestingly, *Ihh* is also activated in Inv3 and Inv4, despite the gene being located distally at the other side of the breakpoint in the remaining *Epha4* TAD. Therefore, it is likely that other enhancers with different limb activities are located in this residual *Epha4* TAD fragment, leading to a more diffuse expression pattern (Supplementary Fig. 1d) that is insufficient to produce a phenotype.

Next, to better understand aberrant gene activation, we studied the three-dimensional architecture induced by the inversions (Fig. 1d). Around the inversion breakpoints, we observed an asymmetric pattern of contacts that was formed between a single locus and a contiguous genomic interval spanning over 20 genes up to 0.5 megabases (Mb), leading to a stripe-like structure in cHi-C maps. Similar structures were predicted by a loop extrusion model and called 'flames' or 'tracks'[19] and were recently described in vitro and called 'architectural stripes'[3]. The anchor point of these interactions localized at the CTCF sites of the inverted boundary with the motif oriented towards the stripe. We observed an extension of interactions beyond the next convergent CTCF sites with many loci on the centromeric side bypassing several active and repressed promoters, skipping several differentially oriented CTCF sites and Rad21 binding peaks, an example for Inv1 is shown in Supplementary Fig. 2. This specific pattern was observed in all inversions but was most pronounced in Inv1, Inv2 and Inv4 (Fig. 1d). We used the cHi-C data to quantify the enhancer–promoter contact frequency within this region (Fig. 1b, Supplementary Fig. 3). These profiles show that the increased contact frequency spreads beyond the first gene, correlating with the observed gene activation. We also determined the interaction frequency of the stripe anchor with promoters and compared it to the enhancer–promoter contact frequency (Supplementary Fig. 3). We found that enhancer–promoter contacts occur within the stripe region.

We characterized the architectural stripe in more detail in Inv1 (Fig. 2); several genes located underneath the stripe were found to be significantly upregulated (four differentially expressed genes with adjusted $P<0.005$, Supplementary Fig. 2). In agreement with previous studies showing that CTCF motif orientation is important for loop formation, we observed that the base of the stripe was located at the inverted CTCF sites[7,8]. To investigate the role of CTCF in the stripe formation, we deleted both CTCF sites in the homozygous inversion (Inv1Δ). Capture Hi-C analysis of the limb buds of Inv1Δ mice compared to Inv1 showed a complete loss of stripes (Fig. 2a,b). Furthermore, the loss of stripe formation was accompanied by the downregulation of gene expression compared to the Inv1 mice with intact CTCF sites, and a decrease in Rad21 binding underneath the stripe (Fig. 2c). However, the gene expression in this region did not completely revert to the WT levels, suggesting that the enhancer cluster was still able to contact and activate the genes. Nevertheless, the deletion of two CTCF sites was sufficient to rescue the polydactyly phenotype (Fig. 2d). In addition, the CTCF deletion led to a slight upregulation of genes on the other side of the boundary, indicating that the CTCF site not only links the limb bud enhancer towards *Ihh*, but also insulates the genes on the other side from enhancer activity. Similar observations had been made in a study showing that the disruption of isolated neighbourhoods can lead to oncogene activation[1].

Next, we asked whether enhancer activity was required for stripe formation. We thus performed cHi-C, CTCF and Rad21 ChIP-seq and RNA-seq in Inv1 mouse embryonic stem cells (mESCs), a cell type in which the *Epha4* enhancer cluster is not active (Fig. 3). Capture Hi-C showed that the chromatin structure at this locus in mESCs differs from that in limb buds. Subtraction of cHi-C in mESCs from cHi-C in Inv1 limb buds shows that the stripe formation is much weaker in mESCs and that the overall structure

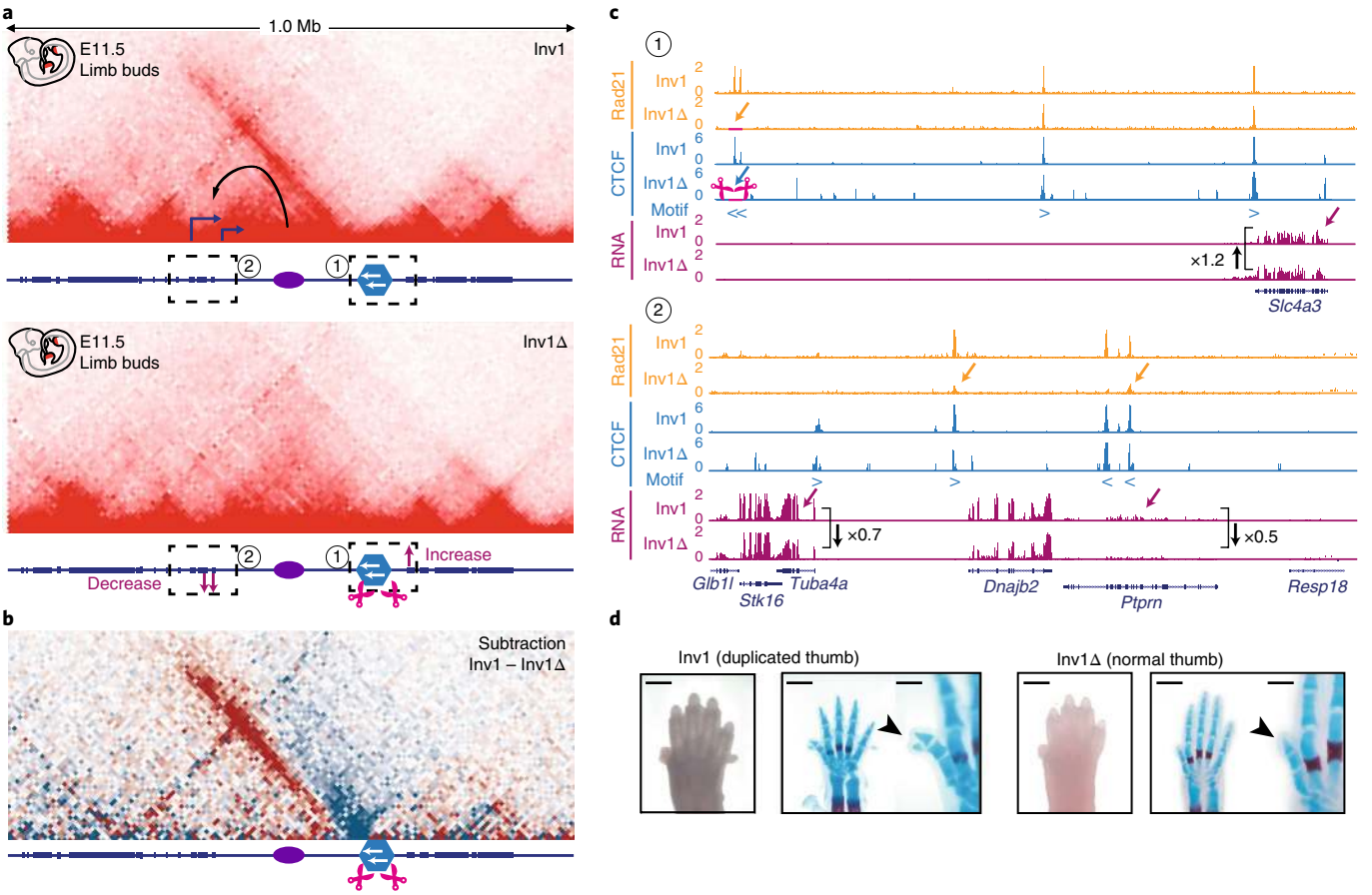

Fig. 2 | Anchor point deletion of the architectural stripe. a, Capture Hi-C of Inv1 (top) and Inv1Δ (bottom). Capture Hi-C of Inv1 shows that interaction of the enhancer cluster with neighbouring genes (arrow) results in gene upregulation. Capture Hi-C of Inv1Δ shows that deletion of the CTCF sites (indicated by pink scissors) results in a loss of the stripe and the downregulation of genes (arrows). Capture Hi-C was performed once with limb buds collected from 10 embryos per condition. The enhancer cluster is indicated by the purple oval; the boundary is indicated by the blue hexagon. **b**, Subtraction of the Inv1Δ mutant cHi-C signal from the Inv1 mutant cHi-C signal shows a complete loss of stripe architecture. Lost contacts in Inv1Δ are shown in red; Gained contacts in Inv1Δ are shown in blue. **c**, Magnification of the regions containing boundary CTCF sites (top) and differentially regulated gene (bottom). Rad21, CTCF ChIP-seq (blue arrows indicate motif orientation) and RNA-seq comparisons of both mutants and regions, respectively. Top, deletion of CTCF sites (pink scissors) results in reduced Rad21 binding at this site (arrows) and downregulation of gene expression in the region contacted by the enhancers below the stripe (bottom). Note the upregulation of gene expression on the telomeric side. Expression changes are shown as fold change. Numbers correspond to regions highlighted in **a**. Inv1, $n=3$ biologically independent samples; Inv1Δ, $n=2$ biologically independent samples. Each biologically independent sample refers to one pair of limb buds collected from a single embryo. ChIP-seq for Rad21 and CTCF was performed once with limb buds pooled from 10 embryos per condition. **d**, Skeletal staining shows the rescue of the duplicated thumb phenotype (arrowhead) by deletion of the stripe anchor. Skeletal staining was performed three times with similar results. Scale bars, 1 mm (2 mm in magnified images).

becomes more symmetrical, indicating that enhancer activity reinforces stripe formation (Fig. 3a,b). However, the stripe did not disappear completely, indicating that other factors besides enhancer activity might play a role.

Subsequently, we searched the whole genome for the presence of architectural stripes in developing E11.5 mouse limb buds through the identification of asymmetric stripes in normalized Hi-C data (Fig. 4). Three imbalance measurements were computed for each chromatin domain on the $\log_2$[observed/expected] (O/E) transformed Hi-C maps between the left and right boundary region, and were used as a feature set of a binary random forest classifier. Active enhancer regions found by EpiCSeg (best described by Active state 2, Supplementary Figs. 1b) were found to aggregate under stripes close to the stripe anchor (left or right), indicating that active enhancers correlate with architectural stripes (Fig. 4c). Interestingly, other chromatin states did not show a similar correlation, demonstrating a diffuse position pattern of chromatin marks relative to the stripe anchor (Supplementary Fig. 4). These results strengthened the hypothesis that enhancer activity contributes to asymmetric stripe formation.

Concomitant with the inversions, we observed the occurrence of asymmetric contact patterns with anchor points formed by two CTCF sites with motif orientation towards the stripe extension. Similar architectural structures were recently described in vitro, and attributed to cohesin loading at a super enhancer and cohesin sliding over long distances involving a loop extrusion process[19,20]. Our inversions generated such a configuration by placing a limb bud enhancer cluster, along with its nearby CTCF sites, into a gene-dense region. In the WT genome, these CTCF sites participate in TAD boundary formation and, when relocated from the classical TAD context, play a role in directing enhancer activity towards several promoters in a gene-dense region. Accordingly, the removal of the stripe anchor leads to stripe loss, reduced cohesin binding and downregulation of gene expression, which is not restored to the WT level. This suggests several layers of gene regulation, as well as CTCF-dependent and CTCF-independent enhancer–promoter

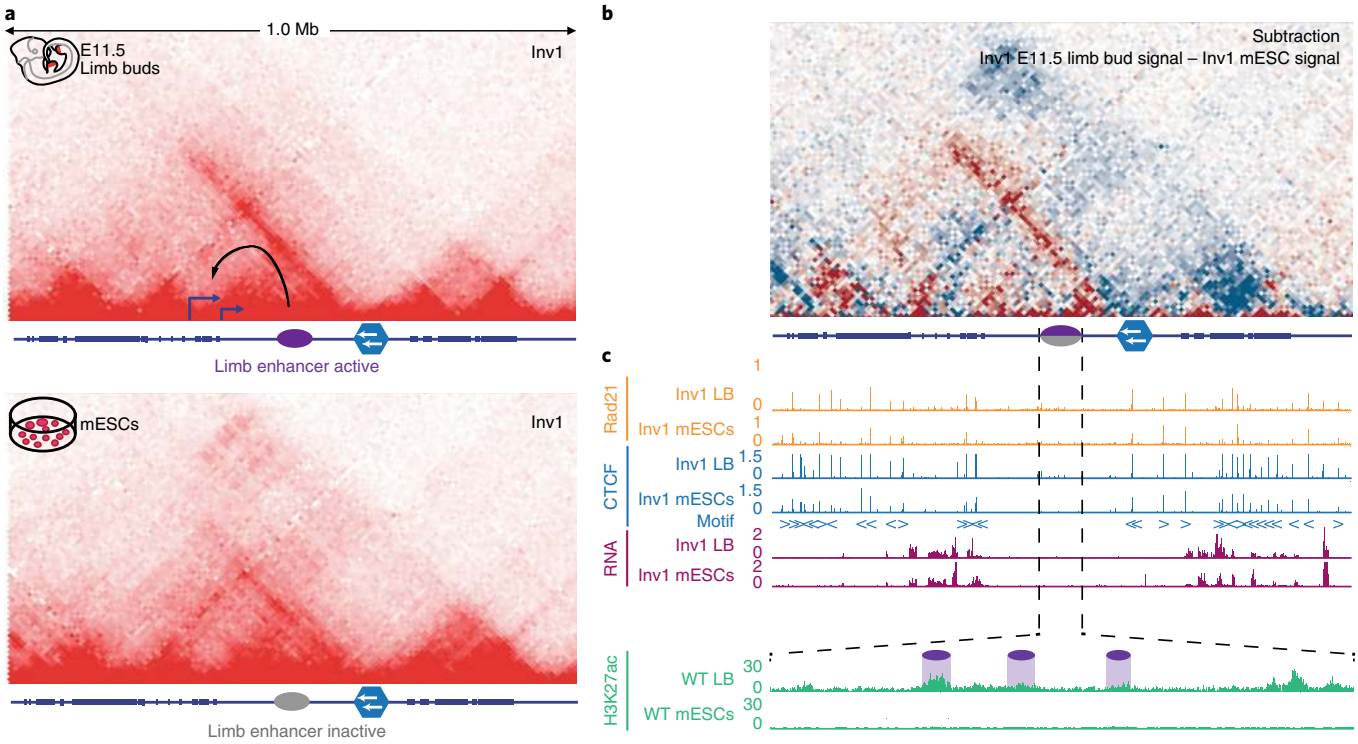

Fig. 3 | Tissue specificity of architectural stripes. a, Capture Hi-C showing an architectural stripe in limb bud tissue (top) and the same region in mESCs where the limb enhancer is not active (bottom). The enhancer cluster is indicated by the purple oval; the boundary is indicated by the blue hexagon. Capture Hi-C was performed once with limb buds from 10 embryos and once with 10^7 mESCs per condition. **b**, Subtraction of the mESC cHi-C signal from the E11.5 limb bud cHi-C signal. Note the reduced interaction in the region of the stripe in mESCs in which the limb enhancer is not active. **c**, Rad21 and CTCF ChIP-seq and RNA-seq of the region. Bottom, enlargement of the enhancer cluster region with H3K27ac ChIP-seq from WT limb buds and mESCs. Rad21 and CTCF ChIP-seq was performed once with limb buds pooled from 10 embryos or 10^7 mESCs per condition. Note that active marks in limbs are absent in mESCs. $n=3$ biologically independent samples. Each biologically independent sample refers to one pair of limb buds collected from a single embryo or to one plate of mESCs that was cultured independently. H3K27ac ChIP-seq for limb bud (LB; accession code: GSE84795) and mESCs (ES E14; accession code: GSE31039) were obtained from published data sets.

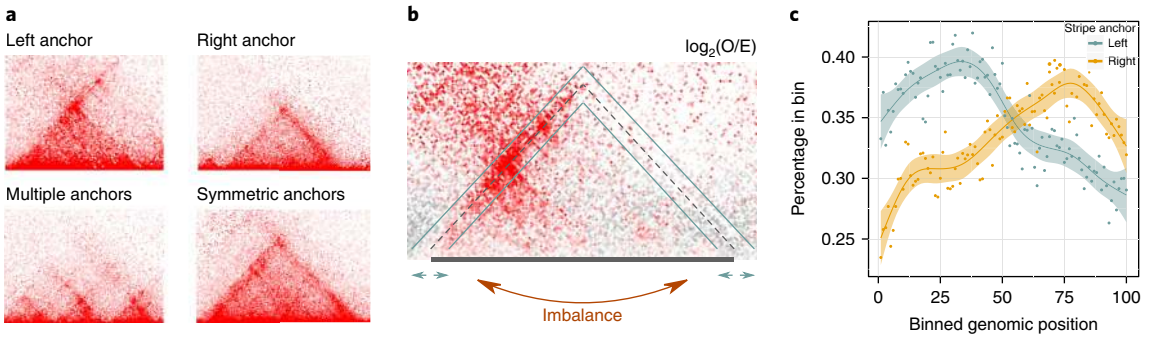

Fig. 4 | Genome-wide identification and analysis of asymmetric stripes in limb tissue. a, Examples of stripes in Hi-C maps obtained from E11.5 limb buds showing a left-anchored stripe, a right-anchored stripe, multiple right-anchored stripes and a symmetrical stripe. Limb buds of 10 embryos were pooled and processed to generate a Hi-C library. From this Hi-C library, three sequencing libraries were generated by three independent PCRs, each of which was sequenced once. **b**, For the genome-wide identification of asymmetric stripes, chromatin domains were first identified. The location of the initial domain boundaries (dashed line) was further refined within a search window (solid lines) in maps of $\log_2$[O/E] signal. For each domain, three different imbalance parameters were computed considering the left border versus the right border region: the difference between the average signals; $\log_2$ ratio of the average signals; and cosine similarity between the signals. The absolute values of all three parameters were used jointly for a binary random forest classifier to identify asymmetric domains. **c**, Aggregated analysis of chromatin domains. Each individual domain was subdivided into 100 intervals. In each bin, the fraction of active enhancers (indicated by Active 2 state from genome-wide segmentation) was calculated and the average profiles were generated within the group of left- and right-anchored domains, respectively. Smoothed curves (solid lines) are fitted using LOESS regression. Shaded envelopes represent 95% confidence intervals.

communication. Our results show that stripe extension occurs beyond the first convergent CTCF site, crossing several CTCF sites with Rad21 or cohesin binding. It is possible that the strength of CTCF binding at the boundaries and promoters is different. Another explanation is that gene-dense regions are organized in a different way to the gene-poor TADs. In the gene-poor TAD regions, a single gene (such as *Epha4* TAD) or few genes (such as *Pinc* TAD) are surrounded by large regulatory regions that are delimited by strong CTCF-binding sites that constitute their boundaries. By contrast, the gene-dense region consists of multiple small interacting domains and a large number of CTCF sites, which are located at promoters. The activity of a strong enhancer that is introduced into a gene-poor TAD through genomic rearrangements is restricted by the boundary of the TAD. In the gene-dense region, however, the activity can spread over many genes induced and confined by architectural stripes. Therefore, ectopic activation of candidate genes can occur through distant promoters, even across several other genes and CTCF sites, if the region is within a stripe and the gene(s) are primed for and/or susceptible to activation. Polycomb repressed genes that displayed a H3K27me3 ChIP-seq signal located within the stripe in the WT limb responded stronger to the enhancer than genes that were not marked by polycomb. Distinct enhancer–promoter specificities have been described for developmental and housekeeping genes using self-transcribing active regulatory region sequencing (STARRseq) in *Drosophila*[21] but not during embryonic development in a mammalian system. The importance of architectural stripes during development and their functional association with active enhancers is supported by our genome-wide study, in which we found active enhancers close to stripe anchor points.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41556-019-0273-x.

Received: 19 July 2018; Accepted: 2 January 2019;
Published online: 11 February 2019

References

1. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
2. Lupiáñez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
3. Vian, L. et al. The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178 (2018).
4. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
5. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
6. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
7. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
8. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
9. Andrey, G. et al. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Res.* **27**, 223–233 (2017).
10. Mumbach, M. R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).
11. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
12. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
13. Dali, R. & Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.* **45**, 2994–3005 (2017).
14. Will, A. J. et al. Composition and dosage of a multipartite enhancer cluster control developmental expression of *Ihh* (Indian hedgehog). *Nat. Genet.* **49**, 1539–1545 (2017).
15. Jäger, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6**, 6178 (2015).
16. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* **128**, 735–745 (2007).
17. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
18. Mammana, A. & Chung, H.-R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.* **16**, 151 (2015).
19. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
20. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
21. Zabidi, M. A. et al. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).

Acknowledgements

This study was supported by grants from Deutsche Forschungsgemeinschaft to S.M. and K.K. We thank J. Fiedler and K. Macura from the transgenic facility; M. Hochradel from the sequencing facility MPIMG; I. Jerkovic for the Rad21 protocol; N. Brieske for WISH; the Mundlos group and B. Flensburger in Berlin; members of the Chang lab and B. Tekila in Stanford; and E. Nora and G. Fudenberg for scientific discussions and support.

Author contributions

K.K. and S.M. conceived the project. K.K. and A.M. performed cHi-C. V.H., J.M and R.S. performed the computational analysis with input from M.V. K.K., C.R. and A.M. produced transgenics, carried out transgenic validation and expressional/phenotypical analysis. D.M.I., R.A.H. and A.D. performed Hi-C. G.A. and A.M. performed ChIP-seq. B.T. sequenced the cHi-C samples. L.W. performed morula aggregation. K.K., S.M. and A.M. wrote the manuscript with input from the other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41556-019-0273-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell culture and mice. *CRISPR–Cas9-engineered allelic series.* Mouse mutants with deletions and inversions were created using the previously described CRISPR–Cas-induced structural variants (CRISVar) protocol[22]. In brief, two single guide RNAs (sgRNAs) were designed for each structural variation using Benchling (https://benchling.com/) and (https://zlab.bio/guide-design-resources), and were selected on the basis of minimal off-target scores (Supplementary Table 1). The sgRNA oligos were annealed and cloned into the pX459 vector (Addgene). G4 mESCs (129/Sv × C57BL/6 F_1 hybrid background)[23] were cultured on mouse embryonic fibroblast (MEF) feeder layers according to standard mESC culture conditions. Cells were co-transfected with pX459 plasmids containing the respective sgRNAs using the FuGENE HD transfection reagent (Promega) in accordance with the manufacturer's conditions. After 12 h, cells were replated on MEF DR4 (mouse embryonic fibroblasts, 4-drug resistant) feeder layers. The next day, puromycin selection (final concentration of 2 μg ml^{-1}) of 48 h was initiated, followed by a cell recovery for 4–6 d. Individual mESC clones (approximately 300–500 per construct) were picked from the plate, transferred to 96-well plates with CD-1 feeders and, after 3 d of culture, split in triplicates (two for freezing or expansion and one for growth and DNA harvesting). Genotypes of the picked clones were ascertained by PCR and quantitative PCR analyses.

qPCR analysis. To determine the genotype of Inv1Δ mESC clones, copy number variation analyses were carried out using the $\Delta\Delta C_t$ method to calculate the fold changes between samples. Two primer pairs each were used that flanked the deletion-spanning region and control regions outside the deletion. qPCR primers of related experiments include: q_Inv1CTCF1_F, CAGGAGACCAGCACACACCA; q_Inv1CTCF1_R, GCCAGACTACTTCACACTCAGAAAC; q_InvCTCF2_F, GGAAGTGCCTGGGAAAGTTGAG; q_Inv1CTCF2_R, GAAGGAGGAGATGGGATGGAAGAG; q_672,AGCTAGATTACCCTGAGTCCA; q_673,TTCAAGTAGGCTCGGTCACC; q_71.440k_F, TGATGTTGGTGGAAGGAAGCA; q_71.440k_R, GGGTATTGGGAGGATGTGGG.

Aggregation of mESCs. Embryos of the desired stage were generated from mESCs by tetraploid aggregation[24]. For each clone, one frozen mESC vial was thawed, seeded on CD-1 feeders, and grown for 2 d. Female CD-1 mice were used as foster mothers.

Animal procedures. All animal procedures were performed according to institutional, state and government regulations. Animal research was approved by the ethics committee of the LAGeSo (Landesamt für Gesundheit und Soziales) Berlin (G0247/13).

WISH. The messenger RNA expression of the analysed genes *Resp18, Slc23a3, Fev, Prkag3* (Supplementary Table 3) and *Ihh*, in E11.5 mouse embryos, was investigated by WISH, according to a previously describedprotocol[25]. The WISH probes designed for related experiments include: Resp18_F1_in_situ, CCCCCTGGCTATTGTTGCT; Resp18_R1_in_situ, GCCTTTGGGATTACTTTGGTG; Slc23a3_F1_in_situ, GTTGTTCTGCATGGGGCTTG; Slc23a3_R1_in_situ, GCAACAAGTCGGCTCTCTCT; Fev_F1_in_situ, ACGCCTACCGCTTTGACTT; Fev_R1_in_situ, GGTTTCCCATCCTTTCTTCC; Prkag_F1_in_situ, GCCACCATTGCATCACCATC; Prkag_R1_in_situ, ACTTGCAGCTTTTCCCTCCA.

Skeletal tissue preparation. E16.5 mouse embryos were treated and stained for bone and cartilage markers as follows. Embryos were kept in H_2O for 1–2 h at room temperature and then heat shocked at 65 °C for 1 min. After carefully taking off the skin, abdominal and thoracic viscera were removed using forceps. The embryos were transferred to glass vials and then fixed in 100% ethanol overnight. On the second day, cartilage was stained using alcian blue staining solution (150 mg l^{-1} alcian blue 8GX in 80% ethanol and 20% acetic acid) overnight. On the third day, embryos were post-fixed and washed in 100% ethanol overnight. Embryos were initially cleared by incubation for 20 min in 1% KOH in H_2O, followed by alizarin red (50 mg l^{-1} alizarin red S in 0.2% KOH) staining of bones overnight. From the fifth day onwards, rinsing and clearing was done using low concentrations (0.2%) of KOH. Limbs of the stained embryos were dissected in 80% glycerol and imaged using a Zeiss Discovery V.12 microscope and Leica DFC420 digital camera.

Hi-C. *Chromosome conformation capture library preparation and sequencing.* The Hi-C protocol was adapted from a previous publication[7]. Hi-C was performed as technical triplicates from limb buds of 10 embryos in total. 3C-libraries were prepared from E11.5 limb buds as follows. After pooling the dissected tissue, it was turned into a single-cell suspension by digestion with trypsin-EDTA 0.05% (Gibco) for 10 min at 37 °C, disrupting the tissue by pipetting every 2 min. The cells were diluted in 10% FCS and PBS and homogenized using a 40-μm cell strainer (Falcon). Cells were pelleted by centrifugation (260*g*, 5 min), resuspended in 10% FCS and PBS and fixed by adding 37% formaldehyde (Sigma-Aldrich) with a final concentration of 2% in 10 ml total volume and put on a roller machine for 10 min at room temperature. By adding 1 ml 1.425 M glycine (Merck) on ice, fixation was quenched and the samples were immediately centrifuged at 260*g* for 8 min at 4 °C. The supernatant was aspirated and the pellet was resuspended in Hi-C lysis buffer and incubated for 10 min on ice. Nuclei were pelleted, snap-frozen in liquid nitrogen and stored at −80 °C. Frozen pellets were washed with 500 μl DpnII buffer (NEB) and centrifuged for 5 min at 4 °C. Then, the Hi-C protocol was conducted as previously described[7], except the size selection of the fragments was performed after the the Hi-C library was directly amplified from the T1 beads according to Illumina protocols (Illumina, 2007). After amplification was complete, the volume of the library was adjusted to 250 μl with 10 mM Tris-HCl (pH 7.5). Following to addition of 137.5 μl of AMPure XP beads, the reaction mix was incubated 5 min at room temperature. Beads were reclaimed through a magnet and the supernatant was again subjected to 37.5 μl of AMPure XP beads with 5 min incubation at room temperature. Beads were captured using a magnet and washed twice with 700 μl of 70% ethanol. After air-drying the beads, the bound DNA was eluted by addition of 25 μl of Tris buffer (10 mM, pH 7.5) and incubation for 5 min at room temperature. Once again, the beads were separated with a magnet and the supernatant transferred to a new tube. The libraries were sequenced on an Illumina HiSeq4000 with approximately 200 million reads (75 bp, paired-end) per sample.

Processing of Hi-C experiments. Raw reads were mapped to reference genome mm9, and then filtered and processed further using the Juicer pipeline v.1.5.6 that incorporates the alignment tool bwa (v.0.7.17). Valid read pairs from all three replicates were merged, filtered for a mapping quality ≥30, binned to 10 kilobases (kb) windows and Knight–Ruiz (KR) normalized[26] using the Juicer tool kit[27] (v.1.7.5). Chromatin domains were identified by applying TopDom v.0.0.2[28] on 25-kb binned and KR-normalized maps using a window size of 250 kb for the TopDom algorithm. For the detection of imbalanced structures in chromatin domains, maps with the (O/E) signal for 10-kb and KR-normalized maps were generated using Juicer tools after taking the $\log_2$[O/E] of the signal. Only positive values were considered, and values above a threshold of the 98th percentile were truncated.

Prediction of imbalanced structures between TAD boundaries. A training set consisting of high-confidence domains with balanced (negative set, 16 domains) and imbalanced (positive set, 15 domains) domain boundaries was hand-picked after visual inspection of Hi-C maps. On the basis of $\log_2$[O/E]-transformed Hi-C matrices, the location of the domain boundaries were first refined by shifting the initial boundary location x to the bin $\in [x-5, \ldots, x, \ldots, x+5]$ with the maximum mean signal along the boundary. For all of the identified chromatin domains, three imbalance measurements were calculated: the absolute difference between the averaged signals along the left and right boundary, the absolute $\log_2$ ratio of the averaged signals along the boundaries and the cosine similarity of the signal along the two boundaries. The sign of the $\log_2$ ratio between the left and right side indicates the orientation of the imbalanced structure, and therefore the location of the stripe anchor. A random forest with 100 trees was grown using the training set and the three described imbalance measurements (using R package randomForest v.4.6–12). Next, the trained model was applied to all chromatin domains identified with TopDom v.0.0.2. The domains were ranked according to their imbalance probabilities (fraction of trees that voted for imbalanced structures) and the top 500 regions were used for further analysis of active enhancer regions at the chromatin domain boundaries.

Analysis of active enhancer regions at imbalanced TAD boundaries. The top 500 chromatin domains with imbalanced boundary regions were selected and further divided into left- and right-anchored stripes (according to the $\log_2$ ratio of the signal averaged along the boundaries). Domains were divided into 100 bins and the fraction of each bin covered by the Active 2 state defined by the EpiCSeg segmentation was calculated. This state is characterized by an increased signal of H3K4me1 and H3K27ac, which are both marks for active enhancers. The results were additionally smoothed using a LOESS regression (stat_smooth function in the ggplot2 R library, v.2.2.1).

Capture HI-C. *Sureselect design.* Using the SureDesign tool from Agilent, the capture Hi-C SureSelect library was designed over the genomic interval (mm9, chr1:71000001–81000000).

Chromosome conformation capture library preparation and sequencing. Preparation and fixation of mESCs (10^7 cells per mutant or WT) and limb tissue (limb buds from 10 embryos per mutant or WT) was conducted as previously described[11]. In brief, fixed cells were lysed, digested with DpnII, ligated and de-crosslinked. Next, religated fragments were sheared using a Covaris sonicator with the conditions as for Hi-C (duty cycle: 10%; intensity: 5; cycles per burst: 200; time: 6 cycles of 60 s each; set mode: frequency sweeping; temperature: 4–7 °C). After adaptors were added to the sheared DNA, amplification was performed according to Agilent instructions for Illumina sequencing. The library was hybridized to the custom-designed sure-select probes and indexed for sequencing (50 bp, paired-end) as instructed by Agilent protocols. Capture Hi-C experiments were performed once. As an internal control, we compared the results from seven experiments

for two regions outside the region of interest (chr1:73100000–73800000 and chr1:77900000–78500000) and calculated pairwise Spearman correlation coefficients, yielding coefficients that ranged from 0.91 to 0.99.

Processing of cHi-C experiments. Preprocessing and mapping of paired-end raw sequencing reads, and filtering of mapped di-tags was performed with the HiCUP pipeline v.0.5.8[29] (Nofill: 1, no size selection, Format: Sanger). The pipeline incorporates the alignment tool Bowtie2 v.2.2.6[30] for mapping short reads to the reference genome (mm9). Aligned and filtered reads were combined from all replicates and further reduced to a minimum mapping quality (MAPQ) = 30. Filtered di-tags were further processed with Juicer tools[27] (v.1.4) to bin di-tags (10-kb bins) and to normalize the interaction matrix by the KR matrix-balancing method[26,27,31]. The DNA-capturing step enriches the genomic region chr1:71000001–81000000 on mm9 leading to three different regimes in the cHi-C map: (1) enriched versus enriched, (2) enriched versus non-enriched, and (3) non-enriched versus non-enriched. For binning and normalization, only di-tags in regime (1) were considered. All cHi-C data sets from mutants were mapped to a customized reference genome that incorporates the respective inversion. To remove the contribution of the WT allele from maps of heterozygous mutants, we applied the following procedure. The cHi-C data from the WT sample was mapped to each customized reference genome. Interaction matrices were generated and scaled to obtain half of the total number of contacts observed in the corresponding map of the heterozygous mutant. Next, each scaled map from the WT sample was subtracted from the corresponding map of the heterozygous mutant to generate virtual homozygous maps of the allele with the inversion.

For the subtraction of interaction maps in visual comparisons, maps were first normalized pairwise by the sum of their subdiagonals to improve comparability. To this end, each subdiagonal vector in one map was divided by its sum and multiplied by the average of the sums of both maps. Additionally, all entries were divided by the total sum of the map and multiplied with 10^6 (similar to reads per million).

RNA-seq. *RNA-seq library preparation and sequencing.* For the analysis of differential gene expression, mESCs or mouse E11.5 distal limb buds were directly lysed or micro dissected, and homogenized using a syringe, respectively. The RNeasy Mini Kit from Qiagen was used according to manufacturer's instructions to isolate RNA. Each condition for WT (five replicates) or mutant samples (Inv1: three replicates; Inv2: four replicates; Inv3: three replicates; Inv4: three replicates; Inv1Δ: two replicates; and, in mESCs, Inv1: three replicates) was sequenced in at least biological duplicates on an Illumina HiSeq 2500 (single read, 50 bp read length). For each sample, 1,600 ng of total RNA were used, and 20 million reads were generated on average.

Processing of RNA-seq experiments. Single-end, 50 bp reads from Illumina sequencing were mapped to the reference genome (mm9) using the STAR mapper (splice junctions based on RefSeq; options: --alignIntronMin20 --alignIntronMax500000 --outFilterMismatchNmax 10). Differential gene expression was ascertained using the DESeq2 package[32]. The cut-off for significantly altered gene expression was an adjusted *P* value of 0.05.

ChIP-seq. *ChIP–seq library preparation and sequencing.* As described for cHi-C and Hi-C experiments, mESCs (10^7 cells per mutant or WT) or limb tissue cells (limb buds of 10 embryos per mutant or WT) were turned into a single-cell suspension and then fixed by adding 37% formaldehyde (Sigma-Aldrich) to a final concentration of 1%, and put on a roller for 10 min at 4 °C. Fixation was quenched by the addition of 1 ml 1.425 M glycine and immediate centrifugation at 400*g* for 8 min. Cells were washed twice with cold PBS and snap-frozen in liquid nitrogen if not further processed immediately. The applied protocol was modified from a previously described ChIP-protocol[33]. In brief, chromatin was sonicated after cell lysis with a Bioruptor NextGen to a fragment size of 200–500 bp. Following sonication, the cell debris was removed by adding 150 µl 10% Triton X-100 to the sample and centrifugation at 16,000*g* and 4 °C. The supernatant was transferred to a new tube and an aliquot was taken for quality control of the sonicated chromatin. Initially, 30 µg (CTCF or Rad21) or 20 µg (H3K9me3) of sonicated chromatin dissolved in 1.2 ml of lysis buffer 3 (+Triton X-100; 1% final concentration) was mixed with 5 µg (CTCF or Rad21) or 1 µg (H3K9me3) of antibody and incubated overnight, with gentle rocking, at 4 °C. For each sample, 30 µl of blocked protein G beads were subjected to the immunoprecipitated chromatin and incubated overnight, with gentle rocking, at 4 °C. The next day, beads were washed seven times with RIPA buffer and once with TE buffer. DNA was eluted from the beads, reverse-crosslinked and ethanol-precipitated. The purified samples were sequenced on an Illumina HiSeq 2500 (single end, 50 bp) with approximately 25 million reads per sample, of which on average 13.5 million reads mapped uniquely to mm9. ChIP-seq experiments were conducted once. To assess the reproducibility of experiments we calculated pairwise Spearman correlation coefficients between the individual samples. We excluded the 4-kb region in the Inv1Δ mutant in all of the samples. We summarized raw read counts in 10-kb bins, yielding coefficients that ranged from 0.73 to 0.87 for CTCF and from 0.78 to 0.86 for Rad21.

Processing of ChIP-seq experiments. The data were processed as previously described[9]. In brief, single-end reads from ChIP-seq experiments were mapped with Bowtie[30] (v.2.2.6) to the reference genome mm9. Mapped reads were filtered for mapping quality MAPQ ≥ 10, and duplicates were removed using samtools[34] (v.1.8). Reads were extended (Rad21 and CTCF: 200 bp; histone modifications: 300 bp) and scaled by the total number of unique reads (total count of reads per 10^6) to produce coverage tracks that could be visualized in the University of California, Santa Cruz (UCSC) genome browser. ChIP-seq data for WT forelimbs of H3K4me1, H3K4me3, H3K27ac, H3K27me3 and CTCF was taken from a previous study[9](GEO accession number: GSE84795). ChIP-seq experiments for WT forelimbs of H3K9me3 were generated and performed as described above. Because H3K9me3 often targets repetitive regions, no mapping quality cut-off was applied.

CTCF motif analysis. The analysis of CTCF motif orientation in ChIP-seq peaks was performed using the FIMO algorithm with default parameters of the MEME suite. The genomic sequence underlying a CTCF peak was used as input to determine the motif orientation. The CTCF motif matrix used as input corresponds to the position weight matrix that has been reported previously[35].

Genome-wide segmentation. Genome-wide segmentation was performed by applying the hidden Markov model (HMM)-based approach EpiCSeg[18] to ChIP-seq experiments. In this study two different segmentations were performed. One incorporates H3K9me3 (no mapping quality filter was applied) and the other was performed solely on the ChIP-seq experiments taken from a previous study[9] (with MAPQ ≥ 10). For the first strategy, seven states were chosen to characterize the epigenomic landscape, including heterochromatic sites that are described by H3K9me3. The second segmentation was performed for six states and was used for the genome-wide analysis of active enhancer regions (state Active 2) at imbalanced TAD boundaries.

Statistics and reproducibility. WISH and skeletal tissue staining experiments were performed three times independently with similar results. RNA-seq analyses involved at least two biologically independent samples with each sample referring to one pair of limb buds collected from a single embryo. Statistical analyses for differential gene expression was conducted with DESeq2[32] using 3–5 biologically independent replicates for RNA-seq to allow basic statistical inference. For Hi-C, limb buds of 10 embryos were pooled and processed to generate a Hi-C library. From this Hi-C library, three sequencing libraries were generated by three independent PCRs, each of which was sequenced once. Capture Hi-C and ChIP-seq analyses were performed once with limb buds pooled from 10 embryos per condition. To assess the reproducibility of cHi-C and ChIP-seq experiments, we calculated pairwise Spearman correlation coefficients between the individual samples. For cHi-C, we compared the results from seven experiments for two regions outside the region of interest (chr1:73100000–73800000 and chr1:77900000–78500000) and calculated pairwise Spearman correlation coefficients, yielding coefficients from 0.91 to 0.99. For ChIP-seq, we summarized raw read counts in 10-kb bins genome-wide except for the region under investigation (that is, the 4-kb region featuring two CTCF sites of the stripe anchor that was deleted using CRISPR–Cas9 in the Inv1Δ mutant), yielding coefficients that ranged from 0.73 to 0.87 for CTCF and from 0.78 to 0.86 for Rad21.

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

Code availability

Custom code to predict domains with asymmetric stripes is available at https://github.com/VerenaHeinrich/HiC2Imbalance.

Data availability

Hi-C, cHi-C, ChIP-seq and RNA-seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE116794. Previously published ChIP-seq data that were reanalysed here are available under accession code GSE84795[9] and GSE31039[36]. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

References

22. Kraft, K. et al. Deletions, inversions, duplications: engineering ofstructural variants using CRISPR/Cas in mice. *Cell Rep.* **10**, 833–839 (2015).
23. George, S. H. L. et al. Developmental and adult phenotyping directly from mutant embryonic stem cells. *Proc. Natl Acad. Sci. USA* **104**, 4455–4460 (2007).
24. Artus, J. & Hadjantonakis, A.-K. Generation of chimeras by aggregation of embryonic stem cells with diploid or tetraploid mouse embryos. *Methods Mol. Biol.* **693**, 37–56 (2011).

25. Kragesteen, B. K. et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* **50**, 1463–1473 (2018).
26. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).
27. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
28. Shin, H. et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).
29. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000 Res.* **4**, 1310 (2015).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
31. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
32. Love, M. I., Huber, W. & Anders, M. I. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
33. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* **1**, 729–748 (2006).
34. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
36. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

nature research

Corresponding author(s): Stefan Mundlos

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data in this study.
Data analysis	cHi-C Fastq files were processed with the HiCUP pipeline v.0.5.8 performing the mapping as well as the filtering for valid and unique di-tags. The pipeline was set up with Bowtie2 v2.2.6. Filtered di-tags were further processed with Juicer tools to bin di-tags (10 kb bins) and to normalize the map by Knight-Ruiz (KR) matrix balancing. RNA-seq Fastq files are mapped to the reference genome using the STAR mapper v2.4.2a. Differential gene expression analysis was performed using the R package DEseq2. ChIP-seq Fastq files were mapped to the reference genome using Bowtie2 v2.2.6 and duplicates were removed using samtools v1.8.

Analysis of CTCF motif orientation was performed using the FIMO algorithm as part of the MEME suite.

Genome-wide segmentation was performed using EpicSeg.

Hi-C
Fastq files were mapped to the reference genome using the Juicer pipeline v1.5.6.

The pipeline was set up with bwa v0.7.17.

Filtered di-tags were further processed with Juicer tools to bin di-tags (10 kb bins) and to normalize the map by Knight-Ruiz (KR) matrix balancing.

Chromatin domains were identified using TopDom v0.0.2 using a 25kb window.

The classification of imbalanced domain structures were done using a customized method.
This incorporates a random forest approach utilizing the R package randomForest v4.6-12.

Smoothing of the average fraction of chromatin domains covering the Active 2 state was performed utilizing LOESS regression which is part of the stat_smooth function in the ggplot2 library v2.2.1.

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

Data

Policy information about availability of data

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

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

Hi-C, Capture Hi-C, ChIP-seq, and RNA–seq data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE116794.
Previously published ChIP-seq data that were re-analysed here are available under accession code GSE84795 or GSE31039, respectively. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	No sample size calculation was performed. For RNA-seq, cHi-C and Hi-C experiments we used commonly accepted numbers of replicates. For ChIP-seq experiments we were limited by the number of animals being used. For Hi-C, limb buds of 10 embryos were pooled and processed to generate a Hi-C library. From this Hi-C library, three sequencing libraries were generated by three independent PCRs , each of which was sequenced once. For expression analyses (RNA-seq) sample size contained at least 2 biologically independent samples (each biologically independent sample refers to one pair of limb buds collected from a single embryo; WT: n=5, Inv1: n=3, Inv2: n=4, Inv3: n=3, Inv4: n=3, Inv1Δ: n=2 embryos, mESCs: n=3 different plates of mESCs that were cultured separately). Capture Hi-C, as in similar studies (see Franke et al., Nature 2016 or Bianco et al., Nat. Gen. 2018), was performed once with limb buds pooled from 10 embryos per condition. ChIP-seq was performed once with limb buds pooled from 10 embryos per condition.
Data exclusions	There was no exclusion of data in the analysis.
Replication	Statistical analyses for differential gene expression was conducted with DESEQ2 using 3 to 5 biologically independent replicates for RNA-seq to allow basic statistical inference. WISH and skeletal staining experiments were performed three times independently with similar results. To assess the reproducibility of cHi-C and ChIP-seq experiments, we calculated pair-wise Spearman correlation coefficients between the individual samples.
Randomization	There was no randomization of samples, as they were allocated according to geno- and/or cell-type.

Blinding

Investigators were not blinded during experiments and outcome assessment. The genotyping of tissues had to occur before the preparation of RNA, ChIP-seq and cHi-C libraries. As a result investigators knew what samples they were handling.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

All cell lines that were used to generate mutant embryos via ESC complementation are available on request.

Antibodies

Antibodies used

We used Anti-Digoxigenin-AP antibody to perform WISH (Roche, reference number: 11093274910, lot number: 11266027, dilution: 1:100)
For ChIP-seq, we used antibodies against H3K9me3 (Abcam, reference number: ab8898, lot number: 837566, dilution: 1:240), Rad21 (Abcam, reference number: ab992, lot number: 221348, dilution: 1:150) and CTCF (Active Motif, reference number: 61311, lot number: 34614003, dilution: 1:150)

Validation

Digoxigenin-AP:
Manufacturer's statement: "The polyclonal antibody from sheep is specific to digoxigenin and digoxin and shows no cross-reactivity with other steroids, such as human estrogens and androgens."
- Weiszmann et al., Determination of gene expression patterns using high-throughput RNA in situ hybridization to whole-mount Drosophila embryos. Nat Protoc. 2009;4(5):605-18.
- Jimenez-Mateos et al., Silencing microRNA-134 produces neuroprotective and prolonged seizure-suppressive effects. Nature Medicine 2012-6-12.

CTCF:
Validated by the manufacturer by ChIP-qPCR, ChIP-seq and Western Blot. Species reactivity: human, mouse
- Weischenfeldt et al., Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet. 2017 Jan;49(1):65-74.
- Thormann et al., Genomic dissection of enhancers uncovers principles of combinatorial regulation and cell type-specific wiring of enhancer-promoter contacts. Nucleic Acids Res. 2018 Apr 6;46(6):2868-2882.

Rad21:
Manufacturer's statement: "The epitope recognized by ab992 maps to a region between residue 575 and the C-terminus (residue 631) human Rad21 homolog using the numbering given in entry NP_006256.1 (GeneID 5885)." Validated for primary applications by the manufacturer (ChIP-seq, IF, IHC/P, WB, IP). Species reactivity: mouse, human, Xenopus laevis, Indian muntjac
- Vian L et al., The Energetics and Physiological Impact of Cohesin Extrusion. Cell 173:1165-1178.e20 (2018)
- Schwarzer W et al., Two independent modes of chromatin organization revealed by cohesin removal. Nature 551:51-56 (2017)

H3K9me3:
Manufacturer's statement: "Specific for Histone H3 tri methyl Lysine 9. Shows slight cross-reactivity with tri methyl K27, which shares a similar epitope . Does not react with mono or di methylated K9." Validated for primary applications by manufacturer (IHC-Fr, IHC-P, ICC/IF, ChIP, WB, ChIP/ChIP, Flow Cyt, CHIPseq).Species reactivity: Mouse, Rat, Chicken, Human, Saccharomyces cerevisiae, Xenopus laevis, Drosophila melanogaster, Indian muntjac, Xenopus tropicalis, Cyanidioschyzon merolae
- Tchasovnikarova IA et al., Hyperactivation of HUSH complex function by Charcot-Marie-Tooth disease mutation in MORC2. Nat Genet 49:1035-1044 (2017).
- Zhao M et al., IL-6/STAT3 pathway induced deficiency of RFX1 contributes to Th17-dependent autoimmune diseases via epigenetic regulation. Nat Commun 9:583 (2018).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	We used G4 Embryonic Stem cells (G4 ESCs, George et al., 2007; obtained from Andreas Nagy from the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto). CD1 and DR4 feeder cell lines produced from CD1 and DR4 transgenic embryos were used to culture the G4 cells.
Authentication	Genetically modified ESCs were used to generate embryos through tetraploid aggregation. Genotyping confirmed that the ES cells featured the desired mutations. DR4 and CD1 feeder cell lines were directly produced from mouse embryos originating from DR4 and CD1 mice crosses, respectively. DR4 and CD1 mouse lines are regularly genotyped in the animal facility of the MPI for Molecular Genetics.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No cell lines used in this study were found in the database of commonly misidentified cell lines that is maintained by ICLAC and NCBI Biosample.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Used species: Mus musculus Embryos and live animals were generated from wildtype or genetically engineered male G4 ESCs (129/Sv×C57BL/6 F1 Hybrid ES Cell) by tetraploid complementation. Female mice of CD1 strain were used as foster mothers (age: 3 - 6 months). All animal procedures were done according to institutional, state, and government regulations (Berlin: LAGeSo G0247/13).
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.

ChIP-seq

Data deposition

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

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

Data access links *May remain private before publication.*	Datasets are available through the Gene Expression Omnibus (GEO) under accession number GSE116794.
Files in database submission	CTCF-ESC-Inv1-Mm-R1-L11722.bw CTCF-FL-E115-Inv1-Mm-R1-L10133.bw CTCF-FL-E115-Inv1DelCTCF-Mm-R1-L11158.bw H3K9me3-FL-E115-Wt-Mm-R1-L9490.fastq.gz.F4.sort.rmdup.bw Rad21-ESC-Inv1-Mm-R1-L11728.bw Rad21-FL-E115-Inv1-Mm-R1-L10134.bw Rad21-FL-E115-Inv1DelCTCF-Mm-R1-L11727.bw CTCF-ESC-Inv1-Mm-R1-L11722_S90_R1_001.fastq.gz CTCF-FL-E115-Inv1-Mm-R1-L10133_S1_R1_001.fastq.gz CTCF-FL-E115-Inv1DelCTCF-Mm-R1-L11158_S34_R1_001.fastq.gz H3K9me3-FL-E115-Wt-Mm-R1-L9490.fastq.gz Rad21-ESC-Inv1-Mm-R1-L11728_S86_R1_001.fastq.gz Rad21-FL-E115-Inv1-Mm-R1-L10134_S1_R1_001.fastq.gz Rad21-FL-E115-Inv1DelCTCF-Mm-R1-L11727_S85_R1_001.fastq.gz
Genome browser session (e.g. UCSC)	no longer applicable

Methodology

Replicates	ChIP-seq experiments were performed as singleton.
Sequencing depth	ample Read depth Unique Rad21-FL-E115-Inv1-Mm-R1-L10134 17034409 13897807 CTCF-FL-E115-Inv1-Mm-R1-L10133 19576546 15818801 CTCF-FL-E115-Inv1DelCTCF-Mm-R1-L11158 17508333 625814 Rad21-FL-E115-Inv1DelCTCF-Mm-R1-L11727 19878945 13649213 CTCF-ESC-Inv1-Mm-R1-L11722 26684283 18530478 Rad21-ESC-Inv1-Mm-R1-L11728 21143440 11575452 H3K9me3-FL-E115-Wt-Mm-R1-L9490 3449035 9422648

	All ChIP-seq experiments were performed as single-end experiments with a read length of 50 bp.
Antibodies	For ChIP-seq, we used antibodies against H3K9me3 (Abcam, reference number: ab8898, lot number: 837566), Rad21 (Abcam, reference number: ab992, lot number: 221348) and CTCF (Active Motif, reference number: 61311, lot number: 34614003)
Peak calling parameters	No peak calling was performed in this study.
Data quality	ChIP-seq experiments (for CTCF and Rad21) were checked for similiar global enrichment folds between the different conditions and also visually inspected for reproducibility by comparing binding profiles of inspected proteins at loci/regions that were not potentially affected by the inserted mutations.
Software	Analysis of CTCF motif orientation was performed using the FIMO algorithm as part of the MEME suite. Genome-wide segmentation was performed using EpicSeg.