Demarcation of Topologically Associating Domains Is Uncoupled from Enriched CTCF Binding in Developing Zebrafish

Yuvia A Perèz Rico, Emmanuel Barillot, Alena Shkumatava

To cite this version:

Yuvia A Perèz Rico, Emmanuel Barillot, Alena Shkumatava. Demarcation of Topologically Associating Domains Is Uncoupled from Enriched CTCF Binding in Developing Zebrafish. iScience, 2020, 23 (5), pp.101046. 10.1016/j.isci. hal-02872297

HAL Id: hal-02872297
https://hal.sorbonne-universite.fr/hal-02872297
Submitted on 17 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License
Demarcation of Topologically Associating Domains Is Uncoupled from Enriched CTCF Binding in Developing Zebrafish

Yuvia A. Pérez-Rico, Emmanuel Barillot, Alena Shkumatava

HIGHLIGHTS
- Identification of CTCF occupancy in zebrafish embryos using a tagged ctcf allele
- CTCF binding at promoters correlates with gene expression levels
- No general CTCF enrichment at topological domain boundaries in zebrafish embryos

Pérez-Rico et al., iScience 23, 101046
May 22, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.101046
Demarcation of Topologically Associating Domains Is Uncoupled from Enriched CTCF Binding in Developing Zebrafish

Yuvia A. Pérez-Rico, Emmanuel Barillot, and Alena Shkumatava

SUMMARY

CCCTC-binding factor (CTCF) is a conserved architectural protein that plays crucial roles in gene regulation and three-dimensional (3D) chromatin organization. To better understand mechanisms and evolution of vertebrate genome organization, we analyzed genome occupancy of CTCF in zebrafish utilizing an endogenously epitope-tagged CTCF knock-in allele. Zebrafish CTCF shares similar facets with its mammalian counterparts, including binding to enhancers, active promoters and repeat elements, and bipartite sequence motifs of its binding sites. However, we found that in vivo CTCF binding is not enriched at boundaries of topologically associating domains (TADs) in developing zebrafish, whereas TAD demarcation by chromatin marks did not differ from mammals. Our data suggest that general mechanisms underlying 3D chromatin organization, and in particular the involvement of CTCF in this process, differ between distant vertebrate species.

INTRODUCTION

CCCTC-binding factor (CTCF) is a key regulator of gene expression and plays a central role in 3D organization of mammalian genomes (Dixon et al., 2012; Guo et al., 2015; Nora et al., 2017). In mammals, CTCF demarcates topologically associating domain (TAD) boundaries, and disruption of the CTCF sites in these regions results in the formation of ectopic contacts between neighboring domains (Despang et al., 2019; Dowen et al., 2014; Lupiáñez et al., 2015). CTCF and its role in gene regulation are conserved throughout bilaterians (Heger et al., 2012), whereas CTCF functions in 3D genome organization have diverged between invertebrates and vertebrates. In contrast to mammals, Drosophila CTCF is not essential for embryogenesis and its binding is not enriched at TAD boundaries (Gambetta and Furlong, 2018; Rowley et al., 2017). Given the functional divergence of CTCF in Drosophila and mammals, CTCF analyses in other non-mammalian vertebrate species are key for understanding the evolution and regulation of the 3D chromatin organization. Although the functions of CTCF in zebrafish development have previously been explored (Carmona-Aldana et al., 2018; Delgado-Olguín et al., 2011; Meier et al., 2018; Rhodes et al., 2010), no genome-wide CTCF in vivo binding data have been achieved in zebrafish so far. Similar to mammals, predicted CTCF binding motifs are distributed in divergent orientation at TAD boundaries in zebrafish (Gómez-Marín et al., 2015; Kaaj et al., 2018), suggesting the conserved role of CTCF in TAD demarcation. Here, we identified and characterize CTCF occupancy in developing zebrafish embryos using an epitope-tagged allele of ctcf. Although several gene regulatory features of zebrafish CTCF are similar to mammals, no enrichment of the in vivo CTCF occupancy was detected at TAD boundaries in zebrafish embryos, suggesting functional differences of CTCF in 3D genome architecture between vertebrates.

RESULTS AND DISCUSSION

Identification of In Vivo CTCF Binding Sites Using the ctcf HPSH Zebrafish Allele

To determine CTCF occupancy in the zebrafish genome, we generated a tagged allele of ctcf, where a tripartite HA-PreScission-His tag was inserted in frame after the start codon of ctcf resulting in N-terminally endogenous CTCF tagged by HPSH (ctcf HPSH allele) (Figures 1A and S1A and Transparent Methods). We confirmed the expression of the tagged protein in ctcf HPSH/HPSH zebrafish (Figure 1B). Homozygous ctcf HPSH/HPSH zebrafish developed normally and were viable and fertile, indicating that the function of CTCF was not affected by the tag (Figures S1B and S1C). Chromatin immunoprecipitation sequencing (ChiP-seq) analyses of CTCF binding in 24 hours postfertilization (hpf) ctcf HPSH embryos showed high correlation between biological replicates (Figures 1C and S1D) and confirmed a previously reported autoregulatory binding of CTCF to its promoter (Figure S1E).
In the ChIP-seq data merged from both replicates, we identified 36,540 CTCF peaks that showed higher phastCons sequence conservation than random control regions (Figure 1D and Transparent Methods); the same trend was observed when considering only CTCF peaks that do not overlap exons (Figure S1F). Notably, the number of CTCF peaks identified in the zebrafish genome roughly corresponds to the number of CTCF sites in mammalian genomes (Pugacheva et al., 2020). Therefore, the ctcf HPSH zebrafish allele enables reliable and reproducible detection of the in vivo CTCF occupancy in the zebrafish genome.

Common Features of CTCF Binding Sites in Vertebrates

Because of the function of mammalian CTCF in establishing enhancer-promoter interactions (Sanyal et al., 2012), we analyzed zebrafish CTCF binding with respect to histone modifications and DNA accessibility in 24-hpf zebrafish embryos (Aday et al., 2011; Bogdanović et al., 2012; Gehlke et al., 2015; Irimia et al., 2012; Ulitsky et al., 2011) (Tables S1 and S2 and Transparent Methods). We found that zebrafish CTCF peaks were enriched for poised (H3K4me1), active (H3K4me3, H3K27ac), and accessible chromatin (ATAC-seq), but not for inactive chromatin (H3K27me3) (Figure 2A). Furthermore, de novo motif discovery identified a 20-bp core motif that was present in 78% of CTCF peaks and showed more similarity to the human CTCF motif than to the human CTCFL or Drosophila CTCF motifs (Figure 2B and Data S1). As reported for other vertebrate species (Boyle et al., 2011; Filippova et al., 1996; Kadota et al., 2017; Rhee and Pugh, 2011; Schmidt et al., 2012), we also identified enriched CTCF upstream motifs, separated from the core motif by 8- or 12-bp spacers (Figures 2C and S2A). Similar to mammalian CTCF binding sites, a fraction of which propagated in the genome through retrotransposition of repeat elements (Schmidt et al., 2012), non-autonomous DNA transposons were significantly enriched on CTCF binding sites (Figures 2D, S2B, S2C, and Table S3). Taken together, our analyses show that zebrafish and mammalian CTCF binding sites share similar features.

CTCF Abundance at Promoters Correlates with the Gene Expression Levels

Although CTCF peaks were mainly located in intronic and intergenic regions, ~6% of the peaks were found within promoters (Figure 3A). In human, a fraction of CTCF sites in promoter and intragenic regions is engaged in loops

(Pugacheva et al., 2006). In the ChIP-seq data merged from both replicates, we identified 36,540 CTCF peaks that showed higher phastCons sequence conservation than random control regions (Figure 1D and Transparent Methods); the same trend was observed when considering only CTCF peaks that do not overlap exons (Figure S1F). Notably, the number of CTCF peaks identified in the zebrafish genome roughly corresponds to the number of CTCF sites in mammalian genomes (Pugacheva et al., 2020). Therefore, the ctcf HPSH zebrafish allele enables reliable and reproducible detection of the in vivo CTCF occupancy in the zebrafish genome.
Figure 2. Characterization of CTCF Peaks and Binding Sites

(A) Heatmap profiles of four histone marks and ATAC-seq data at CTCF peaks ranked by decreasing CTCF ChIP-seq signal over the displayed region. Normalized signal is shown in FPM (fragments per million mapped fragments) for ATAC-seq and in RPKM (reads per kilobase million) for histone marks. All datasets correspond to the 24-hpf zebrafish embryonic stage.

(B) Dendrogram representing hierarchical clustering results of CTCF and CTCFL motifs. In total, 28,538 of 36,540 CTCF ChIP-seq peaks contained at least one matching site to the zebrafish motif, compared with 7,360 of 36,540 control sequences. Information content of each position on the x axis is expressed in bits on the y axis. Ncor, normalized Pearson correlation.

(C) Histogram showing the number of co-occurrences of the CTCF core and upstream motifs at different spacing distances (6–25 bp). Non-significant enriched spacing distances are shown in gray, enrichments are shown in pink, and the highest enrichments are shown in red. Bottom, inferred upstream motifs using sequences matching to the reference motif at the indicated distances from the CTCF core motif. Information content of each position on the x axis is expressed in bits on the y axis.

(D) Top five DNA transposon types enriched for CTCF binding sites. The fraction of repeats overlapping at least one CTCF motif is shown on the y axis. For control regions, the mean and the standard deviation (error bars) calculated by bootstrap analyses are shown.

See also Figure S2, Tables S1–S3 and Data S1.
that impact exon inclusion (Ruiz-Velasco et al., 2017), raising the possibility of an equivalent mechanism in zebrafish. Because the resolution of the available zebrafish high-throughput chromosome conformation capture (Hi-C) data (~20 kb) (Kaaij et al., 2018) does not allow investigation of this type of looping interactions, we sought to determine distinctive features of genes with CTCF-bound promoters. All genes that showed CTCF binding at their promoters were classified in three categories based on the signal strength of CTCF peaks ranging from high (top 10% percentile) to low (bottom 10% percentile) (Figure 3B and Transparent Methods). We found a positive correlation between the presence of CTCF motifs and CTCF occupancy regardless of the site orientation relative to transcription (Figure S3A) (χ² tests of independence, p value < 1.5 x 10⁻³). The increased CTCF occupancy at promoter also positively correlated with increased gene expression (White et al., 2017) (Figure 3C) and DNA accessibility (Figure 3D). Notably, CTCF binding at promoters is not a mere reflection of permissive chromatin, as we also identified promoters with high ATAC-seq signal but no CTCF binding (Figure S3B). Genes with high CTCF promoter occupancy had high signals for histone marks associated with enhancers (H3K4me1, H3K27ac), active promoters (H3K4me3, H3K27ac), and transcriptional elongation (Figures S3C–S3F). By contrast, no correlation between repressive chromatin (H3K27me3) and CTCF abundance was found (Figure S3G), whereas enrichment of the H3K27me3 repressive mark was overall higher at CTCF-bound promoters than at promoters without CTCF peaks (Figure S3H). In summary, our findings suggest that CTCF binding at promoters generally correlates with chromatin states that favor transcription. This observation could be explained by CTCF playing a role in the generation of nucleosome-depleted regions (Nora et al., 2017) or reflect CTCF binding in specific cell types to prevent ectopic gene expression, as previously reported for cis-regulatory elements of runx1 in zebrafish (Marsman et al., 2014).

Figure 3. High Abundance of CTCF Binding at Promoters Associates with High Gene Expression Levels

(A) Distribution of CTCF peaks across different zebrafish genomic regions. Percentages represent the number of CTCF peaks for each category.

(B) Average CTCF ChIP-seq signal profiles over promoters. Each line represents one of the three gene categories defined by CTCF abundance at promoters (low, medium, high) or promoters with no CTCF peaks (no peak).

(C) Expression of the stratified gene categories and genes without CTCF peaks at promoters. Differences in distribution are denoted as significant (*) and non-significant (n.s.) according to two-sided Wilcoxon rank-sum test (p value < 1.5 x 10⁻³).

(D) Average ATAC-seq signal profiles over promoters of gene categories defined by CTCF abundance as explained in (B). See also Figure S3.
No Enrichment of CTCF Binding at TAD Boundaries in Zebrafish Embryos

Next, we sought to investigate the role of CTCF in zebrafish 3D genome organization by analyzing its distribution at TAD boundaries. Visualization of the available 24-hpf Hi-C data (Kaaij et al., 2018) showed enriched interactions of centromeres and telomeres and an uneven distribution of the signal along chromosomes (Figure S4A), which was even more pronounced at earlier developmental stages (Figure S4B). Although this signal distribution reflects the Rabl organization of chromosomes with continuous arm pairing over the cell cycle characteristic for dividing cells (Stadler et al., 2017) and the cell cycle heterogeneity of 24-hpf embryos (Figures S4B and S4C), we, nevertheless, identified 1,307 TADs (median size = 580 kb) using insulation scores (Crane et al., 2015) (Figure S4D and Table S4 and Transparent Methods). The latter was possible given that the 24-hpf Hi-C maps are a composite of the interactions occurring in dividing and interphase cells, which are characterized by lack and presence of TADs, respectively.

Although the annotated CTCF motif was enriched within accessible chromatin sites at TAD boundaries, this enrichment was not boundary specific, as it was also found enriched at accessible sites within TADs (enrichment p values < 1 x 10^-20) (Figure S5A). Our analysis showed a moderate enrichment of in silico predicted CTCF sites and their divergent orientation bias within accessible chromatin at TAD boundaries, consistent with previous reports (Gómez-Marín et al., 2015; Kaaij et al., 2018) (Figures 4A and S5B). Similar to mouse CTCF (Dixon et al., 2012), only a small fraction of zebrafish CTCF peaks was located at TAD boundaries (Figure S5C). However, unlike enriched CTCF binding at TAD boundaries in mammals (Figure 4B), neither CTCF peaks nor the in vivo identified CTCF motifs were enriched at TAD boundaries in 24-hpf zebrafish embryos (Figures 4C and S5D). To exclude analysis bias, we applied the reciprocal insulation method.
and identified hierarchical domains (Zhan et al., 2017). In contrast to mammalian CTCF and in agreement with our zebrafish above-mentioned results, no reciprocal insulation value at which zebrafish CTCF enrichment was clearly maximized was identified (Figures S5E and SSF and Transparent Methods), potentially reflecting low variability in sequence composition of the genome (Costantini et al., 2007). Likewise, we found no zebrafish CTCF enrichment at boundaries of domains identified at 72.5% reciprocal insulation (i.e., the value at which the highest percentage of boundaries overlaps with CTCF peak summits) (Figure S5G). In agreement with the previous report (Kaaij et al., 2018), we found that TAD boundaries in zebrafish were enriched for chromatin marks associated with active transcription (Figures 4D and 4E), depleted for the H3K27me3 repressive mark (Figure 4F), and showed no enrichment for H3K27ac (Figure S5H) suggesting that the biochemical features of zebrafish TAD boundaries are similar to mammals. Our results do not imply that CTCF is dispensable for TAD establishment in the zebrafish genome, as we found moderate enrichment of predicted CTCF motifs within accessible chromatin at TAD boundaries and motif orientation biases, but they rather indicate that, in contrast to mammals, there is no strong correlation between TAD boundaries and high enrichment of CTCF. Importantly, it will require further investigations to determine if this moderate enrichment of CTCF is sufficient to establish TAD boundaries in zebrafish. It is also reasonable to propose that additional architectural proteins or the active chromatin state may play a role in TAD establishment in zebrafish, similar to Drosophila and other eukaryotes lacking this architectural protein. Interestingly, replication timing that is tightly associated with TAD distribution correlates with transcriptional status in zebrafish (Siefert et al., 2017) supporting the latter hypothesis. Moreover, it will be important to investigate colocalization of CTCF and cohesin binding in zebrafish, as the N-terminal CTCF region mediating the interaction with cohesin in mammals differs in its amino acid composition in zebrafish (Li et al., 2020; Pugacheva et al., 2006, 2020). Indeed, this N-terminal region is highly conserved in organisms, in which CTCF is enriched at TAD boundaries including mammals and chicken (Fishman et al., 2018), but it is not conserved in Drosophila, in which CTCF does not delineate TAD boundaries (Moon et al., 2005; Rowley et al., 2017). Therefore, possible differences in the interaction between CTCF and cohesin in zebrafish may explain lack of CTCF enrichment at TAD boundaries. CTCF/cohesin ChIP-seq and Hi-C analyses in specific cell types and using single-cell approaches will be required to further investigate the functions of CTCF in the higher-order organization of the zebrafish genome.

Limitations of the Study

Although we found a positive correlation between CTCF binding at promoters and elevated gene expression, the cellular heterogeneity of 24-hpf zebrafish embryos does not allow to distinguish between CTCF-facilitating gene expression in specific cell types while acting as an insulator in other cells. Future cell-type-specific depletion of CTCF followed by purification of these cells will be required to interrogate CTCF binding and gene expression changes. In addition, it will also be important to analyze the relationship between CTCF enrichment and TAD boundaries in specific cell types to establish whether lack of strong CTCF enrichment at boundaries is maintained across different cell types or is cell specific.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Alena Shkumatava (alena.shkumatava@curie.fr).

Materials Availability

The cctf^{HPSH/HPSH} zebrafish line generated in this study is available from the Lead Contact without restriction.

Data and Code Availability

CTCF ChIP-seq sequencing data generated in this study are available in the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). The accession number for the sequencing data reported in this study is NCBI GEO: GSE133437. No previously unreported algorithms were used to generate the results.
SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101046.

ACKNOWLEDGMENTS

We thank members of the Shkumatava laboratory for discussions. ChIP-seq libraries generated in this study were sequenced at the ICgex NGS platform of Institut Curie funded by grants from “L’Agence Nationale de la Recherche”, France (ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09-08 (France Génomique Consortium) from the “Investissements d’Avenir” program) and by the Canceropole Ile-de-France, France. This research was funded by grants from the European Research Council (FLAME-337440), “Fondation pour la Recherche Médicale”, France (DBI/20132285578), and LabEx DEEP, France (ANR-11-LABX-0044, ANR-10-IDEX-0001-02).

AUTHOR CONTRIBUTIONS


DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 24, 2019

Revised: March 19, 2020

Accepted: April 3, 2020

Published: May 22, 2020

REFERENCES


Supplemental Information

Demarcation of Topologically Associating Domains
Is Uncoupled from Enriched CTCF Binding
in Developing Zebrafish

Yuvia A. Pérez-Rico, Emmanuel Barillot, and Alena Shkumatava
Figure S1. Characterization of the ctcf<sup>HPSH</sup> allele and identification of CTCF binding in zebrafish, Related to Figure 1.
(A) Schematic representation of the *ctcf*<sup>HPSH</sup> zebrafish locus showing exon 2 that was targeted to generate the tagged allele (top panel). Middle and bottom panels show DNA sequencing chromatographs and nucleotide composition of wild type (WT) *ctcf* and *ctcf*<sup>HPSH</sup> zebrafish. The positions of the short guide RNA and PAM sequences are indicated with blue and magenta blocks, respectively. In the bottom panel, the HPSH tag is indicated with orange blocks and homology arms are indicated with black lines. (B) Wild type (WT) and *ctcf*<sup>HPSH</sup>/HPSH embryos during early development. Scale bars for each stage are displayed. Hfp, hours post-fertilization. (C) Wild type (WT) and *ctcf*<sup>HPSH</sup>/HPSH adult fish. Representative fish of each sex and genotype are shown. Scale bars for each genotype are displayed. (D) Correlation between CTCF ChIP-seq biological replicates. In the scatter plot, each dot represents one 1 Mb bin and the axes correspond to the average signals in replicate 1 (y axis) and replicate 2 (x axis). (E) Genome track showing CTCF signal distribution at the *ctcf* locus in two biological replicates. (F) Distribution of average sequence conservation, measured by phastCons scores on the y axis, of non-exonic CTCF peaks and control regions using as reference point the center of peaks and controls.
Figure S2. Analysis of CTCF binding sites, Related to Figure 2.

(A) Extended motif logos for the sequences containing matches to the CTCF core and upstream motifs at spacing distances of 8 (top) and 12 bp (bottom). The number of sequences used to generate the logos is shown. Information content of each position on the x axis is expressed in bits on the y axis. (B) Tracks showing examples of CTCF peaks (purple bars) and CTCF motifs (red bars) at DNA transposons. Displayed signal distributions and peaks correspond to the combined analysis of biological replicates. Signal is represented on the y axis as $-\log_{10} (p\text{-value})$ of the CTCF ChIP-seq signal. (C) Left, color chart representing all sequences with CTCF motifs centered on the region matching to the zebrafish CTCF motif. Right, color chart with the same sequences
clustered by edit distances. Clustered sequences showing enrichment on TDR13B repeats (the DNA transposon with the highest enrichment of sites) are indicated with triangles. Nucleotides are represented as follows: A = green, T = red, C = blue, G = yellow.
Figure S3. Analysis of promoters with CTCF peaks, Related to Figure 3.

(A) Contingency tables showing the number of gene promoters for each of the three defined categories (color-coded according to ChIP-seq signals, lighter colors represent lower abundances) that contain a CTCF motif irrespective of its orientation (top) and in
the same orientation as transcription (bottom). (B) Profiles and heat maps of ATAC-seq signal in promoters that do not contain a CTCF peak. Promoters were assigned to eight clusters (c_1-8) using k-means. (C-G) Average ChIP-seq signal profiles of (C) H3K4me1, (D) H3K4me3, (E) H3K27ac, (F) H3K36me3 and (G) H3K27me3 over gene bodies and flanking sequence of genes with CTCF bound at promoters. Normalized signal is shown in RPKM (reads per kilobase million). Genes were classified based on CTCF signal strength at promoters as High (strongest peaks), Medium and Low (weakest peaks). (H) Average H3K27me3 ChIP-seq signal profiles over gene bodies and flanking sequence of genes without CTCF peaks at promoters.
Figure S4. Visualization of zebrafish Hi-C maps shows Rabl configuration of chromosomes, Related to Figure 4.

(A) Hi-C maps of zebrafish at 24 hpf showing intra-chromosomal and inter-chromosomal normalized interaction signal for all chromosomes. (B) Zoom into Hi-C maps showing intra-chromosomal and inter-chromosomal interaction signal of 5 zebrafish chromosomes at 24 hpf (top) and 4 hpf (bottom). Arrowheads point to representative enriched
interactions between centromeres (purple) and telomeres (green). (C) Karyotype of zebrafish chromosomes showing the locations of centromeres (blue) and the intersections between the two strong diagonals visualized on the Hi-C maps (red). (D) Hi-C maps of zebrafish at 24 hpf displaying examples of TADs annotated using the insulation score approach (green) at genomic regions in chromosomes 7 (top) and 24 (bottom).
Figure S5. Relationship between CTCF peaks and domain boundaries, Related to Figure 4.

(A) CTCF motifs enriched at TAD boundaries (top) and within TADs (bottom). Ncor, normalized Pearson correlation. (B) Distribution of predicted CTCF motifs within ATAC-seq peaks present in forward (green) and reverse (red) orientation along 200 kb regions centered on TAD boundaries (x axis). The y axis shows the number of motifs overlapping each bin. (C) Classification of CTCF peaks relative to TAD boundaries. (D) Distribution of CTCF motifs within ChIP-seq peaks (orange) along 600 kb regions centered on TAD boundaries (x axis). The y axis shows the percentage of total peak counts in the 600 kb region located at each genomic position. The dash line represents the mean background distribution and the grey ribbon depicts the ± 1 standard deviation range from the mean. Differences in mean percentages (central 60 kb) were assessed by Z-scores, n.s.: non-significant. (E) Percentage of domain boundaries identified at different reciprocal insulation scores (x axis) that overlap at least one CTCF peak (y axis) in zebrafish (red) and mouse (green). (F) Average number of CTCF peaks (y axis) within boundaries of domains identified at different reciprocal insulation scores (x axis) in zebrafish (red) and mouse (green). (G) Distribution of CTCF peaks (orange) along 600 kb regions centered on domain boundaries identified at 72.5 % reciprocal insulation (x axis) as described for D. (H) Distribution of H3K27ac-enriched peaks along TAD boundaries as described for D.
Data S1: Zebrafish CTCF motif (homer format) and upstream inferred motifs (meme format), Related to Figure 2.

> GCCWG CAGGGGGCGCTGSDG  CTCF_zebrafish 3.643980 -616.688268 0

T:1993.0(42.57%), B:7136.0(20.02%), P:1e-267

0.125 0.078 0.550 0.247
0.047 0.576 0.309 0.068
0.046 0.908 0.015 0.031
0.453 0.046 0.171 0.330
0.015 0.108 0.876 0.001
0.061 0.892 0.046 0.001
0.953 0.001 0.015 0.031
0.001 0.001 0.997 0.001
0.123 0.001 0.861 0.015
0.078 0.123 0.732 0.067
0.001 0.001 0.997 0.001
0.015 0.001 0.969 0.015
0.001 0.997 0.001 0.001
0.202 0.015 0.782 0.001
0.001 0.860 0.124 0.015
0.217 0.281 0.062 0.440
0.092 0.092 0.815 0.001
0.123 0.396 0.357 0.124
0.283 0.157 0.219 0.341
0.062 0.219 0.611 0.108

MEME version 4

ALPHABET = ACGT
strands: + -

Background letter frequencies (from unknown source):
A 0.322 C 0.178 G 0.178 T 0.322

MOTIF upstream_near_CTCF_gap_8_orientation_0

letter-probability matrix: alength= 4 w= 9 nsites= 20 E= 0e+0
0.039106 0.584730 0.106145 0.270019
0.007449 0.011173 0.981378 0.000000
0.001862 0.960894 0.014898 0.022346
0.204842 0.011173 0.001862 0.782123
0.013035 0.063315 0.918063 0.005587
0.011173 0.109870 0.011173 0.867784
0.033520 0.011173 0.096834 0.858473
0.020484 0.953445 0.009311 0.016760
0.024209 0.186220 0.007449 0.782123

MEME version 4

ALPHABET= ACGT

strands: + -

Background letter frequencies (from unknown source):
A 0.322 C 0.178 G 0.178 T 0.322

MOTIF upstream_near_CTCF_gap_12_orientation_0

letter-probability matrix: alength= 4 w= 9 nsites= 20 E= 0e+0
Table S1: Data sets generated or analyzed in this study, Related to Figures 1 and 2.

<table>
<thead>
<tr>
<th>Data</th>
<th>NCBI GEO serie</th>
<th>Sample</th>
<th>Input sample</th>
<th>Approximate fragment size (bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAC-seq (24 hpf)</td>
<td>GSE61065</td>
<td>GSM1496130</td>
<td>-</td>
<td>-</td>
<td>Gehrke et al., 2015</td>
</tr>
<tr>
<td>H3K4me1 ChIP-seq (24 hpf)</td>
<td>GSE20600</td>
<td>GSM520620, GSM668660</td>
<td>GSM520622, GSM668662</td>
<td>150</td>
<td>Aday et al., 2011</td>
</tr>
<tr>
<td>H3K4me3 ChIP-seq (24 hpf)</td>
<td>GSE20600</td>
<td>GSM520621, GSM668661</td>
<td>GSM520622, GSM668662</td>
<td>150</td>
<td>Aday et al., 2011</td>
</tr>
<tr>
<td>H3K27ac ChIP-seq (24 hpf)</td>
<td>GSE32483</td>
<td>GSM803832</td>
<td>NA</td>
<td>200</td>
<td>Bogdanović et al., 2012</td>
</tr>
<tr>
<td>H3K27me3 ChIP-seq (24 hpf)</td>
<td>GSE35050</td>
<td>GSM861348</td>
<td>NA</td>
<td>200</td>
<td>Irimia et al., 2012</td>
</tr>
<tr>
<td>H3K36me3 ChIP-seq (24 hpf)</td>
<td>GSE32680</td>
<td>GSM813752</td>
<td>GSM813756</td>
<td>150</td>
<td>Ulitsky et al., 2011</td>
</tr>
<tr>
<td>CTCF ChIP-seq (24 hpf)</td>
<td>GSE133437</td>
<td>GSM3908626, GSM3908627</td>
<td>GSM3908628, GSM3908629</td>
<td>185</td>
<td>This report</td>
</tr>
<tr>
<td>Hi-C (4 and 24 hpf)</td>
<td>GSE105013</td>
<td>allValidPairs files</td>
<td>-</td>
<td>-</td>
<td>Kaaij et al., 2018</td>
</tr>
<tr>
<td>mouse CTCF ChIP-seq (mESCs)</td>
<td>GSE29184</td>
<td>GSM723015</td>
<td>GSM723020</td>
<td>300</td>
<td>Shen et al., 2012</td>
</tr>
<tr>
<td>mouse Hi-C (mESCs)</td>
<td>GSE35156</td>
<td>GSM862720, GSM862721</td>
<td>-</td>
<td>-</td>
<td>Dixon et al., 2012</td>
</tr>
</tbody>
</table>
Table S2: ATAC-seq and ChIP-seq mapping statistics, Related to Figures 1 and 2.

<table>
<thead>
<tr>
<th>Data</th>
<th>Species</th>
<th>Total reads</th>
<th>Mapped reads</th>
<th>% of mapped reads</th>
<th>Clean reads</th>
<th>% of clean reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAC-seq</td>
<td>zebrafish</td>
<td>179831298</td>
<td>144502903</td>
<td>80.4</td>
<td>77697950</td>
<td>43.2</td>
</tr>
<tr>
<td>CTCF ChIP-seq 1</td>
<td>zebrafish</td>
<td>15793189</td>
<td>12861051</td>
<td>81.4</td>
<td>5732512</td>
<td>36.3</td>
</tr>
<tr>
<td>CTCF ChIP-seq 2</td>
<td>zebrafish</td>
<td>23303996</td>
<td>18968756</td>
<td>81.4</td>
<td>8700853</td>
<td>37.3</td>
</tr>
<tr>
<td>input (CTCF ChIP-seq 1)</td>
<td>zebrafish</td>
<td>30132254</td>
<td>21301146</td>
<td>70.7</td>
<td>12757021</td>
<td>42.3</td>
</tr>
<tr>
<td>input (CTCF ChIP-seq 2)</td>
<td>zebrafish</td>
<td>35265251</td>
<td>25364354</td>
<td>71.9</td>
<td>15361366</td>
<td>43.6</td>
</tr>
<tr>
<td>CTCF ChIP-seq</td>
<td>mouse</td>
<td>19433603</td>
<td>15591650</td>
<td>80.2</td>
<td>10343560</td>
<td>53.2</td>
</tr>
<tr>
<td>input (CTCF ChIP-seq)</td>
<td>mouse</td>
<td>33137002</td>
<td>29424170</td>
<td>88.8</td>
<td>23494267</td>
<td>70.9</td>
</tr>
<tr>
<td>H3K4me1 ChIP-seq</td>
<td>zebrafish</td>
<td>54206950</td>
<td>43782567</td>
<td>80.8</td>
<td>27424455</td>
<td>50.6</td>
</tr>
<tr>
<td>H3K4me3 ChIP-seq</td>
<td>zebrafish</td>
<td>41419284</td>
<td>34804137</td>
<td>84.0</td>
<td>19159968</td>
<td>46.3</td>
</tr>
<tr>
<td>input (H3K4me1 and H3K4me3 ChIP-seq)</td>
<td>zebrafish</td>
<td>44315679</td>
<td>39178459</td>
<td>88.4</td>
<td>24704265</td>
<td>55.7</td>
</tr>
<tr>
<td>H3K27ac ChIP-seq</td>
<td>zebrafish</td>
<td>11440822</td>
<td>10804838</td>
<td>94.4</td>
<td>7679223</td>
<td>67.1</td>
</tr>
<tr>
<td>H3K27me3 ChIP-seq</td>
<td>zebrafish</td>
<td>11989124</td>
<td>11236840</td>
<td>93.7</td>
<td>7788005</td>
<td>65.0</td>
</tr>
<tr>
<td>H3K36me3 ChIP-seq</td>
<td>zebrafish</td>
<td>15019635</td>
<td>12565631</td>
<td>83.7</td>
<td>7814849</td>
<td>52.0</td>
</tr>
<tr>
<td>input (H3K36me3 ChIP-seq)</td>
<td>zebrafish</td>
<td>12541421</td>
<td>10338391</td>
<td>82.4</td>
<td>6456010</td>
<td>51.5</td>
</tr>
</tbody>
</table>
### Table S3: Zebrafish repeat types enriched with CTCF sites, Related to Figure 2.

<table>
<thead>
<tr>
<th>Repeat type</th>
<th>Number of repeats with CTCF site</th>
<th>p-value (Z-scores)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDR13B</td>
<td>2874</td>
<td>0</td>
</tr>
<tr>
<td>HATN11_DR</td>
<td>562</td>
<td>0</td>
</tr>
<tr>
<td>DNA25TWA1_DR</td>
<td>1310</td>
<td>0</td>
</tr>
<tr>
<td>TDR4</td>
<td>334</td>
<td>0</td>
</tr>
<tr>
<td>DNA8-3_DR</td>
<td>308</td>
<td>0</td>
</tr>
<tr>
<td>DNA-1-3_DR</td>
<td>305</td>
<td>0</td>
</tr>
<tr>
<td>LTR1_DR</td>
<td>290</td>
<td>0</td>
</tr>
<tr>
<td>hAT-N45_DR</td>
<td>178</td>
<td>6.36E-283</td>
</tr>
<tr>
<td>DNA-1-4_DR</td>
<td>258</td>
<td>7.83E-241</td>
</tr>
<tr>
<td>Copia-7-I_DR</td>
<td>107</td>
<td>4.94E-201</td>
</tr>
<tr>
<td>DNA-1-5_DR</td>
<td>217</td>
<td>4.45E-194</td>
</tr>
<tr>
<td>Copia-6-I DR</td>
<td>128</td>
<td>4.66E-178</td>
</tr>
<tr>
<td>DNA-1-3B_DR</td>
<td>235</td>
<td>1.93E-167</td>
</tr>
<tr>
<td>HATN3_DR</td>
<td>153</td>
<td>2.00E-154</td>
</tr>
<tr>
<td>DNA-8-1_DR</td>
<td>107</td>
<td>7.62E-148</td>
</tr>
<tr>
<td>DIRS1a_DR</td>
<td>538</td>
<td>9.87E-144</td>
</tr>
<tr>
<td>hAT-N31_DR</td>
<td>94</td>
<td>9.05E-127</td>
</tr>
<tr>
<td>Nimb-1_DR</td>
<td>150</td>
<td>1.69E-120</td>
</tr>
<tr>
<td>DIRS1_DR</td>
<td>520</td>
<td>1.11E-101</td>
</tr>
<tr>
<td>DNA9NNN1_DR</td>
<td>655</td>
<td>1.42E-75</td>
</tr>
<tr>
<td>Looper-N8_DR</td>
<td>96</td>
<td>1.10E-69</td>
</tr>
<tr>
<td>HATN5_DR</td>
<td>221</td>
<td>2.78E-68</td>
</tr>
<tr>
<td>Harbinger-N11_DR</td>
<td>96</td>
<td>1.43E-65</td>
</tr>
<tr>
<td>Gypsy-169-I_DR</td>
<td>76</td>
<td>4.31E-53</td>
</tr>
<tr>
<td>DIRS-1 DR</td>
<td>114</td>
<td>2.62E-51</td>
</tr>
<tr>
<td>DIRS-10_DR</td>
<td>95</td>
<td>1.52E-48</td>
</tr>
<tr>
<td>Harbinger-N13_DR</td>
<td>100</td>
<td>1.13E-45</td>
</tr>
<tr>
<td>HATN3B_DR</td>
<td>92</td>
<td>1.08E-41</td>
</tr>
<tr>
<td>HarbingerN1_DR</td>
<td>217</td>
<td>1.10E-40</td>
</tr>
<tr>
<td>LRS_DR</td>
<td>65</td>
<td>2.10E-36</td>
</tr>
<tr>
<td>Kolobok-1N1_DR</td>
<td>100</td>
<td>1.38E-32</td>
</tr>
<tr>
<td>hAT-N76_DR</td>
<td>100</td>
<td>3.79E-28</td>
</tr>
<tr>
<td>DIRS-8 DR</td>
<td>59</td>
<td>6.50E-22</td>
</tr>
<tr>
<td>Harbinger-N9_DR</td>
<td>71</td>
<td>2.56E-20</td>
</tr>
<tr>
<td>DNA-8-23_DR</td>
<td>57</td>
<td>1.07E-19</td>
</tr>
<tr>
<td>SAT-1_DR</td>
<td>57</td>
<td>2.97E-16</td>
</tr>
<tr>
<td>DNA-8-9_DR</td>
<td>123</td>
<td>3.62E-06</td>
</tr>
<tr>
<td>HATN9_DR</td>
<td>71</td>
<td>3.62E-05</td>
</tr>
<tr>
<td>EXPANDER1_DR</td>
<td>110</td>
<td>3.80E-03</td>
</tr>
<tr>
<td>hAT-N25_DR</td>
<td>52</td>
<td>5.13E-03</td>
</tr>
</tbody>
</table>
TRANSPARENT METHODS

Generation of the epitope-tagged ctcf<sub>HPSH</sub> zebrafish allele

The AB zebrafish strain was used to generate the ctcf<sub>HPSH</sub> allele using a CRISPR-Cas9-based knock-in protocol (Lavalou et al., 2019). The HPSH tag was introduced into the 5'UTR of ctcf by using one gRNA (5'-AGGGG GACCG ACTGA GGCG-3`) designed to target 3 bp down-stream of the ctcf ATG. A 163-nt single-stranded DNA oligo (ssDNA oligo) thiolated at 5'- and 3'-end nucleotides with 33- and 54-bp homology arms, respectively, flanking both sides of the HPSH tag was designed and manufactured by Ulramer IDT (ssDNA oligo: 5'-TTTGTT ATTTT AACCA ATGTA GGTAT TACCC ATGTA CCTTT ACGAC GTGCC TGTGGA AGATGG CAGGG GATGC TTTCA AGG-3'; HPSH tag is underlined). One-cell stage embryos were injected with 5 pg GFP-Cas9 protein (kind gift of Jean-Paul Concordet), 166 pg sgRNA, 5 pg ssDNA donor oligo and 0.45 pg morpholino against xrcc4 (morpholino: 5'-CACTA CTGCT G CGAC ACCTC ATTCC-3'; Gene Tools LLC). All adult fish (female and male) were individually genotyped to verify the integrity of both HPSH-tagged and wild type ctcf loci. The presence of the ctcf<sub>HPSH</sub> sequence was scored by PCR (genotyping primers: forward 5' GAGA CAGAA AGTGG TCGAG GC 3'; reverse 5' GGCTC CCCAT CTTTA GGCA 3'), the amplified DNA region was subjected to DNA sequencing to confirm tag integration (Figure S1A). Heterozygous ctcf<sub>HPSH</sub> animals were backcrossed to wild type AB fish for 3 generations before generating homozygous ctcf<sub>HPSH/HPSH</sub> animals; wild type siblings of ctcf<sub>HPSH</sub> fish with no HPSH-tag insertion were used as controls.

Zebrafish experimental procedures were approved by the ethics committee of the Institut Curie CEEA-IC #118 (project CEEA-IC 2017-017). Zebrafish were maintained according to standard protocols (Westerfield, 2000) that follow the current Directive 2010/63/EU.

Protein extraction and Western blot assays

Protein extracts were isolated from ~45 embryos or 3 male adult brains. Embryos were dechorionated (Pronase, Roche) and deyolked (55 mM NaCl, 1.8 mM KCl, 1.25 mM NaHCO<sub>3</sub>) prior to cell dissociation. Samples were homogenized in 400 μL of dissociation solution (1 cOmplete tablet, Roche REF 11873580001, in 5 ml of 1X PBS) and spun down at 2,000 rpm for 2 minutes. Protein extracts were obtained by sequential resuspension of pellets using cytoplasmic
(10 mM KCl, 10 mM Hepes (pH 7.9), 3 mM MgCl₂, 0.45% Triton X-100, 0.05% Tween-20) and nuclear (400 mM KCl, 10 mM Hepes (pH 7.9), 3 mM MgCl₂, 0.45% Triton X-100, 0.05% Tween-20) extract buffers. Proteins were separated on a NuPAGE 4-12% Bis-Tris gel (Life technologies, Lot 17022070). Western blot analysis was performed with antibodies detecting HA epitope (HA.11 Epitope tag antibody, BioLegend, Clone 16B12, Lot B224726) and γ-tubulin (Sigma, Clone GTU-88, Lot #026M4832V).

**ChIP-seq library preparation**

Two biological replicates were prepared. For each replicate, ~2000 24 hpf ctcf<sup>HP5H</sup> embryos were used. ChIP-seq was performed as previously described in (Pérez-Rico et al., 2017) using anti-HA.11 epitope tag antibody (BioLegend, Clone 16B12, Lot B224726). Purified chromatin was used for single-end library preparation following TruSeq – ChIP-seq Illumina protocol. Libraries were sequenced in a HiSeq 2500 system.

**Processing of CTCF ChIP-seq data**

Quality of mouse and zebrafish (published and newly generated) ChIP-seq libraries was assessed using FastQC (v0.9.3, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), 3’ adapters were removed using cutadapt v1.3 (Martin, 2011) with the following specifications –O 5 –match-read-wildcards –m 50. Reads were mapped to the danRer10 and mm10 genome versions with Bowtie 2 version 2.2.5 (Langmead and Salzberg, 2012) using –end-to-end –sensitive parameters. Alignment reports were converted to sorted BAM files using samtools version 1.1 (Li et al., 2009) to discard reads with mapping quality lower than 20. Duplicated reads were removed using MarkDuplicates from Picard Tools (version 1.45, https://broadinstitute.github.io/picard/) with parameters: REMOVE_DUPLICATES = true OPTICAL_DUPLICATE_PIXEL_DISTANCE = 100. Reads from technical replicates were combined with samtools merge. Peak calling was performed with MACS2 callpeak using the input libraries as control of the ChIP libraries and the following parameters: -g 9.9e8 for zebrafish and 2.3e9 for mouse –keep-dup all –bw [187|183] for zebrafish and 300 for mouse –call-summits (Zhang et al., 2008). Only peaks with q-values < 1x10⁻⁴ within chromosomes were used for further analyses. Pileup tracks were generated using MACS2 commands callpeak (--keep-dup all –B –model –extsize [187|183] –SPMR –g 9.9e8 or 2.3e9) and bdgcmp (-m subtract). Bedgraph files were filtered to keep only the signal in chromosomes and converted to bigwig format with bedGraphToBigWig version 4 (Kent et al., 2010). Pearson correlation between zebrafish biological duplicates was calculated with deepTools version 3.0.1.
(Ramírez et al., 2016) multBigwigSummary (--binSize 10000) and plotCorrelation (--corMethod pearson –log1p --removeOutliers). Zebrafish biological replicates were combined by first normalizing the signal for each replicate using MACS2 callpeak (-B –nomodel –extsize 185 –g 9.9e8 –keep-dup all) and bdgcmp (-m ppois). Replicates were combined using the generated bedgraph files as input of cmbreps (-m fisher) and peaks identified with bdgpeakcall using the bedgraph file with the combined signal (-l 200 –g 100 –c 8). The final bedgraph file was converted to bigwig format with bedGraphToBigWig.

**Sequence conservation**

Peak coordinates were converted from danRer10 to danRer7 using the liftOver tool (Karolchik, 2004). Conversion was performed for two peak sets: all peaks and peaks with no overlap to exons or with a fraction of overlap < 40% of the peak size. Ensembl 91 gene annotations of zebrafish (http://dec2017.archive.ensembl.org/Danio_rerio/Info/Index) were used as reference for exon coordinates. Control regions were obtained with bedtools shuffle (Quinlan and Hall, 2010) and different exclusion regions for all peaks (-excl danRer7_gaps.bed –chrom -noOverlapping) and the non-exonic peaks (-excl danRer7_gaps_exons.bed –chrom -noOverlapping). Profiles of conservation were generated using the 8 vertebrate NCBI PhastCons track of danRer7 as reference (Siepel et al., 2005), deepTools computeMatrix (-a 3000 –b 3000 –averageTypeBins mean –referencePoint center –missingDataAsZero) and plotProfile programs. Significance of the observed differences between the distributions of CTCF peaks and control regions were assessed by two-sided Kolmogorov-Smirnov tests performed with ks.test in R version 3.3.0 (R Development Core Team 2008).

**Processing of histone ChIP-seq and ATAC-seq data**

All zebrafish ChIP-seq and ATAC-seq data sets analyzed in this study correspond to 24 hpf of zebrafish development; all data sets analyzed in this study and their origin are listed in Table S1. The quality of these libraries was analyzed with FastQC. Only the ATAC-seq data showed biases in sequence content and therefore, fastx_trimmer (FASTX Toolkit 0.0.13, http://hannonlab.cshl.edu/fastx_toolkit/) was used to remove the first 15 bp of all reads (-f 15 –Q 33). For all libraries, reads were mapped to the danRer10 version of the genome with Bowtie 2, low quality (MQ < 20) alignments were filtered out and duplicates were removed with MarkDuplicates as described for CTCF ChIP-seq data. The only difference in mapping was for the ATAC-seq library that was mapped in paired-end mode with --very-sensitive parameters.
Mapping statistics for all ChIP-seq and ATAC-seq datasets are shown in Table S2. BAM files of ChIP-seq biological replicates were merged using samtools merge. Peak calling of ChIP-seq libraries was performed with SICER v1.1 (Zang et al., 2009) using the following parameters: redundancy threshold = 2, window size = 200, fragment size = [150|200], effective genome fraction = 0.7, gap size = 600. For ChIP-seq libraries with input controls, false discovery rate (FDR) was set to $1 \times 10^{-5}$, whereas for libraries without controls, an E-value equal to 10 was used. Pileup tracks of ChIP-seq libraries were generated with bamCompare (--binSize 10 –effectiveGenomeSize 938313030 –normalizeUsing RPKM –ignoreForNormalization chrM –skipNonCoveredRegions –extendReads [150|200]). For those libraries with input controls, input signal was subtracted from the ChIP signal (--operation subtract). Peak calling of the ATAC-seq data was done using MACS2 callpeak (-f BAMPE –keep-dup all –B –SPMR –g 9.3e8 –call-summits), and only peaks with q-values < 1$\times$10^{-4} were used for further analyses. The pileup track of the ATAC-seq library was obtained using the bedgraph files generated by callpeak, including the control signal generated, in combination with bdgcmp (-m subtract) and bedGraphToBigWig.

Heat maps over CTCF peaks

Heat maps were generated using normalized H3K4me1, H3K4me3, H3K27ac, H3K27me3 and ATAC-seq signal over CTCF peak regions centered in the summit position and ranked by decreasing CTCF ChIP-seq signal. Average signal was calculated with computeMatrix (-a 3000 –b 3000 –referencePoint center –sortRegions keep –averageTypeBins mean –misingDataAsZero) and plotted with plotHeatmap (--zMin 0 0 0 0 -1 –zMax 30 60 80 60 3.5 –sortRegions keep).

Motif analyses

The central core motif of the zebrafish CTCF matrix was first identified using all peak summit positions (summit ± 100 bp) as input of the HOMER v4.10.1 (Heinz et al., 2010) program findMotifsGenome.pl (danRer10 -size -100,100 -len 8,10,12 -S 25 -mis 2 -cpg). The motif was then optimized with the same program (danRer10r -opt motif -size given -len 20 -mis 4 -cpg) using as reference sequences identified as centrally enriched on CTCF and CTCFL (HOCOMOCO v11 Full) (Kulakovskiy et al., 2018) by the Centrimo program (MEME Suite 5.0.1 patch 1) (Bailey and MacHanick, 2012; Bailey et al., 2009). CTCF matrix models of zebrafish, human and Drosophila from JASPAR 2018 (Khan et al., 2018) were clustered using the RSAT program matrix-clustering (Castro-Mondragon et al., 2017; Nguyen et al., 2018) with the following clustering options: metric
for similarity = Ncor, agglomeration rule for hierarchical clustering = complete, merge matrices = sum. Binding sites within the CTCF ChIP-seq peaks were identified using matrix-scan (Turatsinze et al., 2008) with the zebrafish CTCF matrix (pseudo counts = 1) and an organism specific background model (GRCz10, upstream-noorf, pseudo-frequencies = 0.01). Sequences were scanned on both strands to report individual matches using the end of the sequence as origin. Only the first-rank matches and those with p-value ≤ 1x10^{-5} were used for further analyses. Spamo (Whittington et al., 2011) was used to identify significant spacings between the zebrafish CTCF motif and a previously identified upstream motif of CTCF in mammals (downloaded from CTCFBSDB (Ziebarth et al., 2013) and converted to meme format with transfac2meme). The sequences used as input for spamo were the sequences with binding sites identified with matrix-scan and extended in both directions to reach 350 bp long sequences (upstream - 168 bp, motif - 20 bp, downstream - 162 bp). Spamo was run with the following options: -numgen 1 -minscore 4 -dumpseqs -shared 0.8 -bgfile background_residues.txt (containing the nucleotide probabilities reported by matrix-scan; A 0.32028, C 0.17713, G 0.17903, T 0.32356). Clustering of sequences shown in Figure S2C was performed using extended binding site sequences identified by matrix-scan too, but restricted to a shorter central region of 50 bp. Distances between sequences were calculated using the Levenshtein method from the function stringdistmatrix of the stringdist package version 0.9.5.1 (https://cran.r-project.org/web/packages/stringdist) in R version (3.4.4). Sequences were clustered using calculated distances and complete linkage hierarchical clustering in R with the hclust function. All logos were generated using WebLogo version 3.6.0 (Crooks et al., 2004).

Enrichment of CTCF sites in repeat regions

The repeatMasker danRer10 track of NCBI was used (Casper et al., 2018) to assess enrichment of CTCF sites in repeats. Only repeat types with more than 50 repeats overlapped by extended CTCF binding site sequences used for clustering were tested for significance of the overlaps. Identification of overlapping regions between motif sequences and each repeat type was done with bedtools intersect (-f 0.76 to consider only those repeat annotations that overlapped at least 75% of the sequence) to calculate the fraction of total repeats with overlap. The same analysis was repeated 100 times using control regions (bedtools shuffle with the CTCF binding site extended coordinates and -chrom -noOverlapping options) to generate expected values of randomly distributed regions and assess the significance of the overlap with CTCF sites using Z-scores (Table S3).
Distribution of CTCF peaks in the genome

CTCF peaks were assigned to exonic, intronic, intergenic and promoter (2 kb upstream of TSSs) categories based on the location of their summit. Considering that genomic regions can be annotated as promoters, exons or introns of different transcriptional units, peak assignment was performed sequentially to first identify those overlapping promoters, then exons and finally introns. Genomic overlaps were identified with `bedtools intersect` using the Ensembl 90 gene annotations of zebrafish (http://aug2017.archive.ensembl.org/Danio_rerio/Info/Index). Those peaks that do not overlap with gene bodies or promoters were assigned as intergenic. The pie chart shown in Figure 3A was generated with `R pie` function.

Gene expression analyses

Expression levels of genes were obtained from RNA-seq data deposited in the Expression Atlas of zebrafish (White et al., 2017). Normalized counts per gene (TPM) for each developmental stage were downloaded from https://www.ebi.ac.uk/gxa/experiments/E-ERAD-475/Downloads. Only data from the “pharyngula prim-5” developmental stage were used in this study. Promoters overlapping summits of CTCF peaks were classified in three categories based on the `MACS2` scores of the CTCF peak. ‘Low’, ‘medium’ and ‘high’ categories corresponded to the following ppois values (Poison p-values): $x \leq 219$ (10th quantile, relative to all peaks), $219 > x < 948.2$, and $x \geq 948.2$ (90th quantile, relative to all peaks). Gene IDs of the promoters were used to filter the three categories and ensure that a given gene was assigned to only one category based on its CTCF peak with the highest score. TPM values for each category, including those genes with no CTCF peak at the promoter, were retrieved from the Expression Atlas table and used to generate the box plot shown in Figure 3C. Analyses of significance of detected differences were carried out using two-sided Wilcoxon rank-sum tests in `R (wilcox.test)`.

Histone and ATAC-seq profiles over gene bodies

Gene IDs of the three gene categories based on `MACS2` scores were used to extract genomic coordinates using Ensembl 90 annotations. Average CTCF and ATAC-seq signal over promoters was obtained with `computeMatrix reference-point` (-a 3000 –b 3000 –averageTypeBins mean – binSize 10), while `computeMatrix scale-regions` (-a 3000 –b 3000 –averageTypeBins mean – regionBodyLength 8000 --missingDataAsZero) was used for histone ChIP-seq signal. All profiles were plotted with `plotProfile`. Differences in the distribution were tested for significance with two-
sided Kolmogorov-Smirnov tests. ATAC-seq profile and heat map of genes without CTCF-bound promoters was performed with `computeMatrix reference-point` (-a 500 -b 3000 –averageTypeBins mean) and `plotHeatmap` (--kmeans 8).

**Generation of Hi-C maps**

Normalized Hi-C maps of 4 and 24 hpf zebrafish embryos were generated using the valid pairs reported by Kaaij et al. 2018 and deposited at the NCBI GEO data base. Files with valid pairs were used as input for Hi-C-Pro version 2.9.0 (Servant et al., 2015) to build contact maps at two different resolutions (1 Mb and 20 kb) and one iteration of ICE normalization was performed on those matrices. Contact maps were visualized with HiCPlotter version 0.6.6 (Akdemir and Chin, 2015) using the following parameters for whole genome maps: -tri 1 –wg 1 –r 1000000 –chr chr25 –hmc 1 –dpi 500. Hi-C maps showing examples of annotated TADs were also generated with HiCPlotter, but using the 20 kb normalized matrix of 24 hpf embryos. Mouse Hi-C analyses were done using biological duplicates of embryonic stem cells (Dixon et al., 2012). Raw data sets were processed with HiC-Pro to map reads, filter read pairs and generate normalized contact matrices by merging the valid pairs from both replicates at 20 kb resolution.

**Annotation of centromeres**

Centromeres were annotated based on enrichment of Type I transposable elements (TEs). The danRer10 nestedRepeats track from NCBI was downloaded and coordinates of Type I TEs (LINE|SINE|LTR) saved as a bed file. This bed file was used to generate a coverage track using TE counts per base and then calculating the average counts over 3 kb windows using `igvtools count` (version 2.3.57, http://www.broadinstitute.org/igv). The coverage track was converted to bigwig format with `wigToBigWig` and used to calculate the enrichment of Type I TEs over 3 Mb overlapping windows with a shift of 1 kb by `bigWigAverageOverBed`. Finally, the window with the highest enrichment for each chromosome was annotated as the centromere. Only chromosome 3 and chromosome 4 showed several prominent peaks and the selection of the centromere was performed using an average score including zero or discarding scores higher than 0.38 (values corresponding to the long arm that is known to be enriched on repeats) for chromosome 3 and 4, respectively.
Identification of TADs

Zebrafish and mouse TADs were annotated using perl scripts of the cworld::dekker module version 1.01 (https://github.com/dekkerlab/cworld-dekker) and the contact maps generated with HiC-Pro at 20 kb resolution. First, insulation was calculated with matrix2insulation.pl indicating the following parameters: --is 400000 --ss 80000 --ids 240000 --bmoe 0 --nt 0.1. Second, to obtain TAD coordinates, the calculated insulation scores and boundaries identified were used as input of insulation2tads.pl, indicating a value of 0 for the option –mts. Given that the danRer10 genome has evident misassembled regions, as indicated by the Hi-C maps, all TADs annotated within those regions that could have hampered the annotation (mainly telomeric, Table S4) were discarded. A total of 1,307 TADs were used for further analyses.

Identification of hierarchical domains

Hierarchical domains of zebrafish were annotated at 20 kb resolution using the reciprocal insulation score method (Zhan et al., 2017) implemented in CaTCH version 1.1 in R version 3.4.4. Boundaries of domains identified at 37 reciprocal insulation scores were analyzed to test for enrichment of CTCF. Boundaries for each set of domains were extended to obtain 60 kb regions and calculate the percentage of boundaries overlapping CTCF peak summits and the average number of CTCF summits within each boundary using bedtools intersect and bedtools coverage.

Enrichments at TAD boundaries

TAD boundaries and boundaries of domains identified at 72.5% reciprocal insulation were defined as the 20 kb regions located at both ends of domains. CTCF ChIP-seq peaks were categorized as associated or non-associated with boundaries according to the location of their summit relative to extended boundaries (± 20 kb) accounting for uncertainty in the exact definition of boundaries. CTCF and histone mark enrichments over boundaries were assessed by using the center of boundaries as reference point to extend regions on both sides (± 300 kb). These extended regions were divided in 10 kb or 30 kb sized bins for CTCF and histone mark analyses, respectively, and significant peaks identified by MACS and SICER were used to compute the number of overlapping peaks per bin for each boundary using bedtools coverage (-count option). These analysis generated matrices in which each row represents a boundary region and each column one of the bins. Total counts were obtained from each column to calculate the percentage of peaks overlapping each bin and plot their distribution. The same strategy was followed to assess
enrichment of predicted and *in vivo* identified CTCF motifs at boundaries using 10 kb sized bins. This strategy was also used to calculate the distribution of forward and reverse CTCF motifs along the ± 100 kb region around boundaries using 5 kb sized bins. Background controls were generated for each enrichment analysis by randomly distributing peaks in the same chromosome using *bedtools shuffle* (--chrom --noOverlapping options) and calculating percentage distributions as described above. This process was repeated 100 times to obtain the mean and standard deviation of expected distributions.

**Data and Code Availability**

CTCF ChIP-seq sequencing data generated in this study are available in the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). The accession number for the sequencing data reported in this paper is under accession NCBI GEO: GSE133437. No previously unreported algorithms were used to generate the results.
SUPPLEMENTAL REFERENCES


Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and


