CTCF loss has limited effects on global genome architecture in *Drosophila* despite critical regulatory functions

Anjali Kaushal, Giriram Mohana, Julien Dorier, Isa Özdemir, Arina Omer, Pascal Cousin, Anastasiia Semenova, Michael Taschner, Oleksandr Dergai, Flavia Marzetta, Christian Iseli, Yossi Eliaz, David Weisz, Muhammad Saad Shamim, Nicolas Guex, Erez Lieberman Aiden & Maria Cristina Gambetta

Vertebrate genomes are partitioned into contact domains defined by enhanced internal contact frequency and formed by two principal mechanisms: compartmentalization of transcriptionally active and inactive domains, and stalling of chromosomal loop-extruding cohesin by CTCF bound at domain boundaries. While *Drosophila* has widespread contact domains and CTCF, it is currently unclear whether CTCF-dependent domains exist in flies. We genetically ablate CTCF in *Drosophila* and examine impacts on genome folding and transcriptional regulation in the central nervous system. We find that CTCF is required to form a small fraction of all domain boundaries, while critically controlling expression patterns of certain genes and supporting nervous system function. We also find that CTCF recruits the pervasive boundary-associated factor Cp190 to CTCF-occupied boundaries and co-regulates a subset of genes near boundaries together with Cp190. These results highlight a profound difference in CTCF requirement for genome folding in flies and vertebrates, in which a large fraction of boundaries are CTCF-dependent and suggest that CTCF has played mutable roles in genome architecture and direct gene expression control during metazoan evolution.
A wide range of animal genomes are partitioned into a series of contact domains (CDs) that exhibit increased physical proximity among loci within them. An evolutionarily conserved mechanism of such genome folding is thought to be compartmentalization, reflecting the segregation of chromosomal domains based on their transcriptional and epigenetic states. In vertebrates, chromosomal loops are additionally extruded on underlying compartmental domains through a process involving DNA-bound CTCF molecules which stall loop-extruding cohesin complexes at domain boundaries. CTCF-dependent extrusion-based boundaries either reinforce or counteract compartmental domain boundaries, depending on the locus. Overall, a large fraction of boundaries in the vertebrate genome are CTCF-dependent.

Intriguingly, although Drosophila has widespread CDs and CTCF, it is currently unclear whether CTCF-dependent domains exist in Drosophila. High-resolution genome-wide Hi-C maps of formaldehyde-crosslinking frequencies between Drosophila CTCF, it is currently unclear whether CTCF-dependent developmental genes in cultured cells and mice identified regulatory elements and their target promoters. Types of CDs may have different functions.

Intriguingly, although Drosophila has widespread CDs and CTCF, it is currently unclear whether CTCF-dependent domains exist in Drosophila. High-resolution genome-wide Hi-C maps of formaldehyde-crosslinking frequencies between pairs of DNA fragments (as a measurement of their proximity in 3D-space) were recently generated in Drosophila tissue culture cells. These studies highlighted the lack of hallmarks of CTCF-mediated domains observed in vertebrate cells. Rather, evidence suggests that CDs in flies are formed by CTCF-independent compartmentalization and other transcription-related processes, as most boundaries lie between domains with different histone modifications or at promoters of highly transcribed genes.

Crucially, the functional importance of genome folding into CTCF-dependent domains is not fully understood in any organism. CTCF is essential for the viability of mammalian cells, whereas it is dispensable for early development in Drosophila. Assessing whether or not CTCF-mediated domains exist in Drosophila is important for understanding their relevance for genome function. Recent studies have perturbed specific CDs in flies to address their biological roles without knowing whether they are CTCF-mediated or compartmental, yet different types of CDs may have different functions.

CTCF-dependent domains in mammals generally comprise regulatory elements and their target promoters. This suggested that CTCF somehow limits regulatory crosstalk between CDs, and fosters regulatory interactions within them. This model is, however, difficult to test in mammals because global perturbation of CTCF leads to cell death. Acute depletion of CTCF protein in mouse embryonic stem cells followed by transcriptional profiling did not reveal widespread transcriptional changes. Alternatively, deletion of CTCF binding sites near developmental genes in cultured cells and mice identified some sites where CTCF appears to critically prevent developmental defects and disease, and many CTCF sites that did not appear functional. These diverse results paint an opaque picture of how CTCF impacts gene expression. Previous studies that partially knocked-down CTCF in Drosophila cell lines also did not reveal clear effects on transcription. Analysis of the homoeotic phenotype of CTCF mutants completely lacking both maternal and zygotic CTCF suggested that CTCF blocks regulatory crosstalk between elements on either side of some CTCF binding sites. A fundamental question arising from comparative studies in flies and humans is how CTCF impacts transcription, and how this relates to its uncertain architectural function in flies. Whether CTCF stably associates with partner proteins to effect its functions also remains unclear.

Here, we show using CTCF mutant Drosophila that CTCF is critically required in neurons for fly viability. We examine the effects of CTCF loss on genome folding and transcriptional regulation in the central nervous system (CNS) and investigate the molecular basis of CTCF function.

Results
CTCF expression in neural stem cells (NSCs) or neurons is essential for fly viability. To identify a biologically relevant tissue in which to study CTCF function in Drosophila, we used previously described CTCF knock-out (CTCFKO) mutants and CTCF mutants that additionally lack maternally inherited CTCF. Some CTCFKO mutants (60%) hatch into adults with spasmatic movements suggesting a neurological phenotype that might be the cause of their short lifespan. We tested the relevance of CTCF expression in the nervous system by performing tissue-specific knock-out and rescue experiments. Specifically, we used Gal4 drivers active in NSCs, mature neurons or muscles to drive conditional excision of a CTCF rescue transgene (knock-out) or UAS-CTCF expression (rescue) in CTCF mutant genetic backgrounds. Loss of CTCF expression in NSCs or neurons compromised the ability of flies to hatch to a comparable extent as loss of all zygotic CTCF expression (Fig. 1a) and severely shortened the life span of flies that did hatch (Fig. 1b, Supplementary Movie 1). On the other hand, loss of CTCF in muscle only slightly impaired adult hatching and life span (Fig. 1a, b).

In contrast to CTCFKO, CTCF0 mutants never hatch from the pupal case (Fig. 1c). Conditional expression of CTCF in NSCs or neurons of CTCF0 mutants strongly rescued hatching (Fig. 1c) and adults were capable of coordinated movements and survived for several days (Fig. 1d, Supplementary Movie 3). On the other hand, expressing CTCF in muscles of CTCF0 mutants barely rescued hatching (Fig. 1c, d).

Together, these results show that CTCF expression is critically required in neurons for pupal hatching and adult viability. Consistently, CTCF is more highly expressed in the nervous system than in other tissues. Analyses of molecular phenotypes of CTCF0 mutants described hereafter were therefore performed in dissected CNSs of third instar larvae, a developmental stage at which CTCF0 mutants are fully viable.

Physical insulation defects in CTCF0 mutants. To address whether CTCF is required to form CD boundaries in flies, Hi-C was performed on CNSs dissected from wildtype (WT) and CTCF0 larvae in biological triplicate using two 4-cutter restriction enzymes for enhanced resolution. Hi-C maps consisting of 200 million reads per genotype were obtained by combining the correlated biological replicates (see Methods, Supplementary Table 1). Hi-C maps from whole bodies as control. Only 740 CTCF peaks were defined in WT relative to CTCF0 CNSs, of which 77% overlapped a CTCF consensus motif (Supplementary Fig. 2b, Supplementary Data 1).

To assess the relation between CTCF peaks and CD boundaries genome-wide in WT CNS Hi-C maps, boundaries were identified at 2 kb resolution with TopDom (see “Methods”, Supplementary Table 2, Supplementary Data 2 and 3). Very few (<1%) boundaries defined in this study potentially correspond to small CDs defined in even higher resolution Hi-C studies (see “Methods”). Domain boundaries were enriched within ±1 kb of several (36%) CTCF peaks (Fig. 2a). Conversely, a CTCF peak was located within ±1 kb of only 8% of all boundaries (Fig. 2b). This indicates that while CTCF peaks are frequently at domain boundaries, CTCF is only present at a small fraction of all boundaries in flies.
WT and CTCF^0 Hi-C maps were globally similar, and most (84%) domain boundaries were detected in both WT and CTCF^0 mutants. Nevertheless, specific CD s were visibly less physically insulated from the neighboring domain in CTCF^0 mutants (Fig. 2c, Supplementary Fig. 2c, Supplementary Table 3). Clearly disrupted domain boundaries in CTCF^0 mutants frequently occurred at former CTCF peaks (Fig. 2d). Of 135 strongly affected domain boundaries that were lost in CTCF^0 mutants, 89 (66%) were at former CTCF peaks (Supplementary Table 2). To determine how generally physical insulation defects are observed at former CTCF peaks in the absence of CTCF (irrespective of their localization at CD boundaries identified by TopDom), physical insulation score differences between WT and CTCF^0 mutants were measured across all 740 CD boundaries bound by CTCF (Fig. 2f). These observations are not due to the presence of contaminating CTCF, as CTCF RNA and protein are undetectable by RNA-seq and ChIP-seq (Fig. 2c and next section). As CTCF^0 mutants lack CTCF from the beginning of development, residual boundaries can also not be explained by a role of CTCF in the establishment but not maintenance of boundaries. Rather, this observation suggests that at some sites, CTCF reinforces boundaries redundantly established by other mechanisms, a scenario also observed in mammalian cells. We define CTCF-occupied CD boundaries present only in WT as strictly CTCF-dependent, and those that are present in CTCF^0 (generally weaker than in WT) as partially CTCF-dependent. These two types of CTCF-dependent boundaries are contrasted later in the “Results” section.

A region in the N-terminus of human CTCF directly interacts with cohesin and stabilizes cohesin on DNA, partially explaining how human CTCF forms CD boundaries. Vertebrate and fly CTCF N-termini are highly diverged, yet a 10 amino acid residue stretch in CTCF’s N-terminus that binds to cohesin in human cells is present at a similar distance from the zinc finger domain in fly CTCF (boxed in Supplementary Fig. 2f). We therefore tested whether two residues critical for cohesin interaction in human CTCF (Y226 F228, homologous to Y248 F250 in fly CTCF) mediate direct interaction of fly CTCF with the SA-Vtd (homologous to human SA2-SCC1) complex. For this, GFP-tagged recombinant WT and Y248A F250A point mutant peaks 5 and 6). Of 343 WT CD boundaries bound by CTCF, only 125 (36%) were fully lost in CTCF^0 mutants (Supplementary Table 2). This resulted in a lower average physical insulation score at former CTCF peaks in CTCF^0 mutant CNS Hi-C maps (Fig. 2f). These observations are not due to the presence of contaminating CTCF, as CTCF RNA and protein are undetectable by RNA-seq and ChIP-seq (Fig. 2c and next section). As CTCF^0 mutants lack CTCF from the beginning of development, residual boundaries can also not be explained by a role of CTCF in the establishment but not maintenance of boundaries. Rather, this observation suggests that at some sites, CTCF reinforces boundaries redundantly established by other mechanisms, a scenario also observed in mammalian cells. We define CTCF-occupied CD boundaries present only in WT as strictly CTCF-dependent, and those that are present in CTCF^0 (generally weaker than in WT) as partially CTCF-dependent. These two types of CTCF-dependent boundaries are contrasted later in the “Results” section.

A region in the N-terminus of human CTCF directly interacts with cohesin and stabilizes cohesin on DNA, partially explaining how human CTCF forms CD boundaries. Vertebrate and fly CTCF N-termini are highly diverged, yet a 10 amino acid residue stretch in CTCF’s N-terminus that binds to cohesin in human cells is present at a similar distance from the zinc finger domain in fly CTCF (boxed in Supplementary Fig. 2f). We therefore tested whether two residues critical for cohesin interaction in human CTCF (Y226 F228, homologous to Y248 F250 in fly CTCF) mediate direct interaction of fly CTCF with the SA-Vtd (homologous to human SA2-SCC1) complex. For this, GFP-tagged recombinant WT and Y248A F250A point mutant peaks 5 and 6). Of 343 WT CD boundaries bound by CTCF, only 125 (36%) were fully lost in CTCF^0 mutants (Supplementary Table 2). This resulted in a lower average physical insulation score at former CTCF peaks in CTCF^0 mutant CNS Hi-C maps (Fig. 2f). These observations are not due to the presence of contaminating CTCF, as CTCF RNA and protein are undetectable by RNA-seq and ChIP-seq (Fig. 2c and next section). As CTCF^0 mutants lack CTCF from the beginning of development, residual boundaries can also not be explained by a role of CTCF in the establishment but not maintenance of boundaries. Rather, this observation suggests that at some sites, CTCF reinforces boundaries redundantly established by other mechanisms, a scenario also observed in mammalian cells. We define CTCF-occupied CD boundaries present only in WT as strictly CTCF-dependent, and those that are present in CTCF^0 (generally weaker than in WT) as partially CTCF-dependent. These two types of CTCF-dependent boundaries are contrasted later in the “Results” section.

A region in the N-terminus of human CTCF directly interacts with cohesin and stabilizes cohesin on DNA, partially explaining how human CTCF forms CD boundaries. Vertebrate and fly CTCF N-termini are highly diverged, yet a 10 amino acid residue stretch in CTCF’s N-terminus that binds to cohesin in human cells is present at a similar distance from the zinc finger domain in fly CTCF (boxed in Supplementary Fig. 2f). We therefore tested whether two residues critical for cohesin interaction in human CTCF (Y226 F228, homologous to Y248 F250 in fly CTCF) mediate direct interaction of fly CTCF with the SA-Vtd (homologous to human SA2-SCC1) complex. For this, GFP-tagged recombinant WT and Y248A F250A point mutant peaks 5 and 6). Of 343 WT CD boundaries bound by CTCF, only 125 (36%) were fully lost in CTCF^0 mutants (Supplementary Table 2). This resulted in a lower average physical insulation score at former CTCF peaks in CTCF^0 mutant CNS Hi-C maps (Fig. 2f). These observations are not due to the presence of contaminating CTCF, as CTCF RNA and protein are undetectable by RNA-seq and ChIP-seq (Fig. 2c and next section). As CTCF^0 mutants lack CTCF from the beginning of development, residual boundaries can also not be explained by a role of CTCF in the establishment but not maintenance of boundaries. Rather, this observation suggests that at some sites, CTCF reinforces boundaries redundantly established by other mechanisms, a scenario also observed in mammalian cells. We define CTCF-occupied CD boundaries present only in WT as strictly CTCF-dependent, and those that are present in CTCF^0 (generally weaker than in WT) as partially CTCF-dependent. These two types of CTCF-dependent boundaries are contrasted later in the “Results” section.

A region in the N-terminus of human CTCF directly interacts with cohesin and stabilizes cohesin on DNA, partially explaining how human CTCF forms CD boundaries. Vertebrate and fly CTCF N-termini are highly diverged, yet a 10 amino acid residue stretch in CTCF’s N-terminus that binds to cohesin in human cells is present at a similar distance from the zinc finger domain in fly CTCF (boxed in Supplementary Fig. 2f). We therefore tested whether two residues critical for cohesin interaction in human CTCF (Y226 F228, homologous to Y248 F250 in fly CTCF) mediate direct interaction of fly CTCF with the SA-Vtd (homologous to human SA2-SCC1) complex. For this, GFP-tagged recombinant WT and Y248A F250A point mutant peaks 5 and 6). Of 343 WT CD boundaries bound by CTCF, only 125 (36%) were fully lost in CTCF^0 mutants (Supplementary Table 2). This resulted in a lower average physical insulation score at former CTCF peaks in CTCF^0 mutant CNS Hi-C maps (Fig. 2f). These observations are not due to the presence of contaminating CTCF, as CTCF RNA and protein are undetectable by RNA-seq and ChIP-seq (Fig. 2c and next section). As CTCF^0 mutants lack CTCF from the beginning of development, residual boundaries can also not be explained by a role of CTCF in the establishment but not maintenance of boundaries. Rather, this observation suggests that at some sites, CTCF reinforces boundaries redundantly established by other mechanisms, a scenario also observed in mammalian cells. We define CTCF-occupied CD boundaries present only in WT as strictly CTCF-dependent, and those that are present in CTCF^0 (generally weaker than in WT) as partially CTCF-dependent. These two types of CTCF-dependent boundaries are contrasted later in the “Results” section.
CTCF N-termini were mixed with an untagged SA-Vtd subcomplex and purified on GFP binder beads. WT but not mutant CTCF versions retained SA-Vtd (Fig. 2g). Therefore, despite profound divergence, the fly CTCF N-terminus interacts directly with cohesin in vitro. This interaction was suggested to impart directionality to CTCF-dependent boundaries in mammalian cells\textsuperscript{10,39}, but we find that CTCF has at best a very weak preference to establish directional boundaries (Supplementary Fig. 2g) consistent with a previous study\textsuperscript{2}.

We conclude that Drosophila CTCF is required to form physical boundaries with strengths generally proportional to its occupancy on DNA. Other mechanisms reinforce CTCF-dependent boundaries at some sites and explain the formation of most boundaries in flies.
CTCF impacts expression patterns of genes near CTCF peaks. To understand how CTCF impacts transcription, we performed RNA sequencing (RNA-seq) on cDNA libraries from mRNA purified from WT and CTCF0 larval CNSs in triplicate. This confirmed the absence of CTCF mRNA in CTCF0 samples (Supplementary Fig. 3a). 392 (~3% of all) genes were significantly differentially expressed (DE) in CTCF0 mutants with adjusted p-value<0.05 and |fold-change| > 1.5 (Fig. 3a, Supplementary Data 4). CTCF0 mutants therefore do not show widespread transcriptional defects, though changes occurring in subsets of cells in the CNS such as CTCF’s previously validated target gene Abdominal-B elude our analysis21.

Some DE genes had decreased expression in CTCF0 mutant CNSs compared to WT (Fig. 3b). Several DE genes with increased expression in CTCF0 CNSs are normally not expressed in the CNS but rather restricted to other specialized tissues like testes (Intraflagellar transport 52), tendons (Thrombospondin), and the peripheral nervous system (Odonate receptor 67d) (Figs. 3c, 3d, Supplementary Fig. 3b). Some ectopic transcripts lacked annotated start and termination sites suggesting that they are cryptic (Supplementary Fig. 3b). RNA fluorescent in situ hybridization (RNA-FISH) analysis showed that genes with increased expression in CTCF0 CNSs were misexpressed in various patterns, possibly driven by locus-specific enhancers (Fig. 3e).

Indirect transcriptional changes are expected in CTCF0 mutants, which lack CTCF since the beginning of development, and we asked whether CTCF regulates genes in the vicinity of its binding sites. 10% of DE genes had a CTCF peak within ±1 kb of their transcriptional start site (TSS) (ninefold enrichment over randomly sampled matched non-DE genes) (Fig. 3f), a result that was not very different for genes with increased versus decreased expression in CTCF0 mutants (Supplementary Fig. 3c). Conversely, 5% of CTCF peaks were located within ±1 kb of a DE gene TSS (9-fold enrichment over randomly sampled matched non-DE genes) (Fig. 3g). These results suggest that, depending on the locus, CTCF may directly repress or activate the transcription of nearby genes, or alternatively CTCF may shield promoters from inappropriate enhancers or silencers as observed at Hox gene loci21,40.

Could the structural defects observed in CTCF0 Hi-C maps be secondary consequences of gene misregulation in the vicinity of former CTCF peaks? Some CTCF-dependent domain boundaries were located far from genes (Fig. 2c, CTCF peak 6 is 9 kb away from the closest gene) and are thus unlikely to be impacted by transcription. Others were located near genes whose expression increased (Supplementary Fig. 2c, peak 3), decreased (Supplementary Fig. 2c, peak 6) and in most cases remained unchanged (Supplementary Fig. 2c, peak 7). Few (8%) DE genes were located in different A/B compartments in CTCF0 mutants relative to WT, indicating that differential gene expression mostly occurred without large changes in higher-order spatial chromatin configuration (Supplementary Fig. 3a, Supplementary Data 4). Together, these results indicate that the pervasive weakening of physical boundaries observed at former CTCF peaks in CTCF0 mutants (Fig. 2c-f) is not a mere consequence of altered transcription.

CTCF occupancy scales with enhancer-blocker activity in a reporter assay. Previous studies of the functionality of CTCF binding sites stably integrated into the fly genome suggested that most of them lack insulator activity (i.e., the ability to block regulatory crosstalk)36, at least in single copies40. Here, we tested CTCF peaks in a quantitative reporter assay. The reporter comprises an enhancer positioned between two fluorescent reporter genes (EGFP and mCherry) driven by minimal Heat-shock-protein-70 (Hsp70) promoters (Fig. 4a). Test fragments were cloned in between EGFP and the enhancer, maintaining the enhancer at a similar distance from both reporter genes. Reporter plasmids were then transiently transfected into Drosophila S2 cells, and relative EGFP and mCherry intensities were measured in thousands of single cells with a cell analyzer (Supplementary Fig. 4a). An insulator should reduce EGFP expression while mCherry expression should remain high. Control experiments with a neutral spacer or the well-characterized gypsy insulator41 validated the assay (Fig. 4b, lanes 1 and 2). Two CTCF peaks near genes whose expression decreased (peak G from Fig. 3b) or increased (peak N from Fig. 3e) in CTCF0 mutants had similar effects as gypsy (Fig. 4b, lanes 3 and 4). EGFP levels in the presence of gypsy or CTCF peaks were not strongly reduced below basal levels measured in enhancer-less control reporters (Supplementary Fig. 4b), suggesting that these tested sequences mostly impaired enhancer-mediated EGFP expression. Additional CTCF peak regions (Supplementary Fig. 4c, average size 360 bp) were tested and their relative insulator strengths were estimated from the median ratio of mCherry-over-EGFP fluorescence measured in single cells. Eleven out of 14 tested CTCF peaks selectively reduced EGFP intensities to various degrees that globally scaled with CTCF ChIP-seq occupancy measured in S2 cells42 (Fig. 4c) and that appeared independent of the endogenous locations of CTCF peaks relative to their nearest genes (Supplementary Fig. 4c) and of combinatorial co-binding with other fly insulator-binding proteins on the cloned fragments (Supplementary Fig. 4d). Mutating two base pairs of a CTCF motif in one of these fragments abolished its activity (Fig. 4c, fragment N mut); thus, the reporter specifically reveals the activity of a single CTCF
Fig. 3 CTCF impacts expression patterns of genes near CTCF peaks. a RNA-seq MA plot of CTCF<sup>0</sup> versus WT CNSs with mean abundance (in x) plotted as a function of enrichment (in y). Differentially expressed (DE) genes (pad < 0.05 and |fold change| > 1.5) are red. b-d RNA-seq signals in WT (black) and CTCF<sup>0</sup> (red) larval CNSs, and CTCF ChIP-seq signals in WT (green) and CTCF<sup>0</sup> (red) larval CNSs at CG1354 (b), IFT52 (c) and Tsp (d) loci. Differentially transcribed regions are shaded in red. Scales in tracks of all figures indicate reads per million. In all figures, CTCF peaks labeled by capital letters were tested in Fig. 4c. e RNA-FISH with antisense probes (red) against indicated transcripts in CNSs of wildtype and CTCF<sup>0</sup> larvae stained by DAPI (blue) (scale bars 100 μm). mRNAs of SP1029, IFT52 and an antisense transcript overlapping can (shown in Supplementary Fig. 3b) are normally not expressed in wildtype CNSs (background signal is sometimes visible in trachea) and are misexpressed in different patterns in CTCF<sup>0</sup> mutants. All animals showed similar misexpression patterns for a given transcript. f Percentage (in y) of n = 386 DE genes in CTCF<sup>0</sup> larval CNSs (black) or n = 386 randomly sampled expression-level-matched non-DE genes (gray) with at least one of 740 CTCF peaks at a given distance (per 2 kb bins) around the gene TSS, measured in the direction of transcription (in x). Ten percent of DE genes have at least one CTCF peak within ±1 kb of their TSS, which is ninefold higher than the average enrichment at the sampled non-DE genes. g Percentage (in y) of CTCF peaks with at least one of n = 386 DE gene TSSs (black) or n = 386 randomly sampled expression-level-matched non-DE gene TSSs (gray) at a given distance (per 2 kb bins) around CTCF peaks, measured in the direction of transcription (in x). Five percent of CTCF peaks have at least one DE gene TSS within ±1 kb, which is 9-fold higher than at the sampled non-DE TSSs.

CTCF recruits Cp190 to a subset of Cp190-bound domain boundaries. To further understand how CTCF functions, we asked whether it stably associates with partner proteins that contribute to its activity. Unbiased identification of CTCF partners from Drosophila embryonic nuclear extracts in biological duplicates by mass spectrometry reproducibly identified known insulator-binding proteins Centrosomal protein 190 kDa (Cp190) and Insulator binding factors 1 and 2 (Ibf1 and Ibf2) as enriched CTCF interactors relative to negative control (Supplementary Fig. 5a). Reciprocal Cp190 purifications published by others also identified Ibf1, Ibf2 and CTCF among other proteins.<ref>386</ref> Traces of the cohesin complex also co-purified with CTCF (Supplementary Fig. 5a) reminiscent of transient interactions between cohesin and CTCF seen in mammalian cells.<ref>39</ref>

CTCF was previously shown to directly interact with Cp190<sup>43</sup> yet the relevance of this interaction remained unclear. No common target genes are known<ref>46</ref> and a mutant version of CTCF reported to no longer interact with Cp190 was largely functional in vivo<ref>45</ref>. We performed pull-downs of GFP-tagged CTCF fragments co-expressed in bacteria with Cp190’s BTB (Broad-Complex, Tramtrack and Bric-a-brac) domain and found that amino acids 698-771 in CTCF C-terminus directly interact with Cp190 BTB (Supplementary Fig. 5b). Importantly, this stretch in binding site. Taken together, these observations indicate that CTCF sites in the reporter do not strongly directly repress or activate transcription but rather insulate a promoter from an enhancer.
CTCF does not overlap the previously deleted region (amino acid residues 774–818) that was used to conclude that CTCF’s interaction with Cp190 was unimportant in vivo.

To assess the genome-wide overlap between CTCF and Cp190 binding sites in larval CNSs, specific Cp190 peaks were identified by ChIP-seq with a polyclonal anti-Cp190 antibody in WT and in Cp190KO animals with a CRISPR-Cas9 mediated deletion of the Cp190 open reading frame as control (Supplementary Fig. 5c). 6,473 Cp190 peaks were enriched in WT relative to Cp190KO CNSs (Fig. 5a, Supplementary Data 6). Cp190 colocalized with CTCF at most (79%) CTCF peaks and was additionally present at many other sites (Fig. 5a), consistent with other studies.35,36,47. We profiled Cp190 binding sites in WT and CTCF0 larval CNSs and found that Cp190 was normally recruited to most Cp190 peaks in CTCF0 mutants with the exception of former CTCF peaks, at which Cp190 was globally reduced (Figs. 5a, 5b, Supplementary Data 7 and 8). In CTCF0 mutants, Cp190 was lost from former higher-occupancy CTCF peaks but only reduced at former lower-occupancy CTCF peaks (Fig. 5b). We therefore distinguish between strictly CTCF-dependent Cp190 peaks (lacking a detectable Cp190 peak when comparing CTCF0 and Cp190KO mutants) and partially CTCF-dependent Cp190 peaks (with a detectable Cp190 peak in CTCF0 relative to Cp190KO mutants, generally weaker in CTCF0 than in WT).

Unlike CTCF, Cp190 binding was enriched at CD boundaries genome-wide (Fig. 5c lane 3, Supplementary Figs. 5d, e).2,15,17. Outside of CTCF peaks, Cp190-occupied domain boundaries were often proximal to transcribed TSSs (Fig. 5c, lane 6). In CTCF0 mutants, residual Cp190 binding at former CTCF-occupied boundaries was significantly associated with boundary retention (Figs. 5d–f). Seventy-five percent of strictly CTCF-dependent boundaries lacked a residual Cp190 peak, and 80% of residual Cp190 peaks were associated with a residual boundary in CTCF0 mutants (Fig. 5e). CD boundary defects in CTCF0 mutants were also less severe at former TSS-proximal CTCF peaks (within 200 bp of a gene TSS) than at former TSS-distal CTCF peaks (Fig. 5f). This suggests that either Cp190 itself, its associated factors, or transcription at Cp190-bound TSSs may redundantly contribute to the formation of physical boundaries independently of CTCF and may synergize with CTCF at partially CTCF-dependent Cp190 peaks (see examples in Fig. 5g).

CTCF and Cp190 co-regulate a subset of target genes. To assess whether loss of Cp190 results in transcriptional changes shared with CTCF0 mutants, RNA-seq was performed on Cp190KO larval CNSs in biological triplicate. Overall, 440 DE genes were observed in Cp190KO mutant CNSs compared to WT, of which 192 went up and 248 went down relative to WT (with adjusted p-value < 0.05 and [fold-change] > 1.5) (Supplementary Fig. 6a, Supplementary Data 9). Since Cp190 is bound to more many sites than CTCF (Fig. 5a), we did not expect that many transcriptional changes in Cp190KO mutants would be shared in CTCF0 mutants. Surprisingly, however, a considerable fraction of DE genes in CTCF0 and Cp190KO mutants were common (31% of all DE genes in CTCF0 and 26% of all DE genes in Cp190KO) and concordantly changed in similar directions and to similar degrees relative to WT (Fig. 6a). This is exemplified at the SP1029 (Fig. 6b–c) and CG15478 (Fig. 6d–e) genes that are proximal to a CTCF and Cp190 co-bound peak (peak 1/N in Fig. 6b, peak 2 in Fig. 6d). In the absence of CTCF, Cp190 is additionally lost from these peaks (Figs. 6b and d, middle), a CD boundary is disrupted (Supplementary Figs. 6b and c), and the gene is expressed at increased (SP1029 in Fig. 6b, middle) or decreased (CG15478 in Fig. 6d, middle) levels relative to

Fig. 4 CTCF occupancy scales with enhancer-blocker activity in a reporter assay. a In the reporter plasmid, a test insulator I is cloned in between an enhancer E and EGFP, and mCherry serves as a reference (elements are drawn to scale, arrowheads represent Hsp70 minimal promoters). A gypsy insulator G is present downstream of EGFP to block EGFP activity by the enhancer (which in a circular plasmid molecule is both upstream and downstream of EGFP) from the left. b Split violin plots (thick lines mark medians, boxes mark interquartile ranges) show distributions of mCherry (left) and EGFP (right) fluorescence intensities (log10 values in y) measured in thousands of single S2 cells transiently transfected with reporters with indicated I fragments (in x). mCherry-to-EGFP ratios (log2 values in y) in single cells are shown below. For each reporter, merged biological triplicates are plotted. c Median mCherry-to-EGFP ratios in single transfected S2 cells (log2 values in y) relative to CTCF ChIP-seq counts in S2 cells42 (log10 values in x) on selected CTCF peaks in CNSs (Fig. 5a, Supplementary Data 6). Cp190 colocalized with CTCF at most (79%) CTCF peaks and was additionally present at many other sites (Fig. 5a), consistent with other studies.35,36,47. We profiled Cp190 binding sites in WT and CTCF0 larval CNSs and found that Cp190 was normally recruited to most Cp190 peaks in CTCF0 mutants with the exception of former CTCF peaks, at which Cp190 was globally reduced (Figs. 5a, 5b, Supplementary Data 7 and 8). In CTCF0 mutants, Cp190 was lost from former higher-occupancy CTCF peaks but only reduced at former lower-occupancy CTCF peaks (Fig. 5b). We therefore distinguish between strictly CTCF-dependent Cp190 peaks (lacking a detectable Cp190 peak when comparing CTCF0 and Cp190KO mutants) and partially CTCF-dependent Cp190 peaks (with a detectable Cp190 peak in CTCF0 relative to Cp190KO mutants, generally weaker in CTCF0 than in WT).

Unlike CTCF, Cp190 binding was enriched at CD boundaries genome-wide (Fig. 5c lane 3, Supplementary Figs. 5d, e).2,15,17. Outside of CTCF peaks, Cp190-occupied domain boundaries were often proximal to transcribed TSSs (Fig. 5c, lane 6). In CTCF0 mutants, residual Cp190 binding at former CTCF-occupied boundaries was significantly associated with boundary retention (Figs. 5d–f). Seventy-five percent of strictly CTCF-dependent boundaries lacked a residual Cp190 peak, and 80% of residual Cp190 peaks were associated with a residual boundary in CTCF0 mutants (Fig. 5e). CD boundary defects in CTCF0 mutants were also less severe at former TSS-proximal CTCF peaks (within 200 bp of a gene TSS) than at former TSS-distal CTCF peaks (Fig. 5f). This suggests that either Cp190 itself, its associated factors, or transcription at Cp190-bound TSSs may redundantly contribute to the formation of physical boundaries independently of CTCF and may synergize with CTCF at partially CTCF-dependent Cp190 peaks (see examples in Fig. 5g).

CTCF and Cp190 co-regulate a subset of target genes. To assess whether loss of Cp190 results in transcriptional changes shared with CTCF0 mutants, RNA-seq was performed on Cp190KO larval CNSs in biological triplicate. Overall, 440 DE genes were observed in Cp190KO mutant CNSs compared to WT, of which 192 went up and 248 went down relative to WT (with adjusted p-value < 0.05 and [fold-change] > 1.5) (Supplementary Fig. 6a, Supplementary Data 9). Since Cp190 is bound to more many sites than CTCF (Fig. 5a), we did not expect that many transcriptional changes in Cp190KO mutants would be shared in CTCF0 mutants. Surprisingly, however, a considerable fraction of DE genes in CTCF0 and Cp190KO mutants were common (31% of all DE genes in CTCF0 and 26% of all DE genes in Cp190KO) and concordantly changed in similar directions and to similar degrees relative to WT (Fig. 6a). This is exemplified at the SP1029 (Fig. 6b–c) and CG15478 (Fig. 6d–e) genes that are proximal to a CTCF and Cp190 co-bound peak (peak 1/N in Fig. 6b, peak 2 in Fig. 6d). In the absence of CTCF, Cp190 is additionally lost from these peaks (Figs. 6b and d, middle), a CD boundary is disrupted (Supplementary Figs. 6b and c), and the gene is expressed at increased (SP1029 in Fig. 6b, middle) or decreased (CG15478 in Fig. 6d, middle) levels relative to
WT. In the absence of Cp190, CTCF remains bound at SP1029 (Fig. 6b, bottom) and CG15478 (Fig. 6d, bottom) which are nevertheless also similarly misexpressed relative to WT (Figs. 6b and d, bottom). This suggests that Cp190 is required for CTCF function independently of CTCF binding to DNA. To more stringently compare SP1029 and CG15478 misexpression in the absence of CTCF or Cp190, we visualized their mRNAs in embryos completely lacking maternal and zygotic CTCF (Figs. 6b, bottom). This suggests that Cp190 is required for CTCF

**Discussion**

CTCF-dependent CDs have been proposed to regulate the communication between genes and their regulatory elements. Here, we analyzed Drosophila that developed in the complete absence of CTCF and reached the following conclusions: (1) CTCF is most critically required in neuronal cells for adult viability (Fig. 1). (2) Domain boundary defects in CTCF mutants are overwhelmingly associated with CTCF-bound sites, consistent with a mechanism in which CTCF can form boundaries (Fig. 2). At the same time, the vast majority of boundaries are CTCF-independent. (3) CTCF prevents ectopic activation and silencing of certain genes in its vicinity (Fig. 3). (4) Sites bound by CTCF do not directly repress or activate transcription, but rather functionally insulate promoters and enhancers in a reporter assay in S2 cells (Fig. 4). (5)

![](image.png)

**Fig. 2**. CTCF and Cp190 at CD boundaries. In (a) and (b), examples of Cp190 and CTCF ChIP-seq signals in S2 cells. In (c-f), examples of Hi-C contacts between genomic loci in wild-type embryos. In (g), examples of CTCF occupancy across the genome.
CTCF0 directly binds to the C-terminus of CTCF and is recruited to CTCF peaks in a strictly or partially CTCF-dependent manner (Fig. 5). Residual CTCF binding at former CTCF peaks coincides with residual boundary retention in CTCF0 mutants (Fig. 5), (6) CTCF binding to DNA alone is not sufficient for correct expression patterns of a subset of genes that also rely on CTCF. Below we discuss how this work further our understanding of genome folding in Drosophila, CTCF’s role in transcriptional regulation and the molecular basis thereof.

CTCF0 binding interfaces that are not all conserved in mammals but lack clear directionality in Drosophila, but which CTCF interacts with extruding cohesin polymerase II recruitment which may account for the observed boundary defects16.

Whether Drosophila CTCF, like its mammalian counterpart, forms CD boundaries in concert with loop-extruding cohesin remains unclear because of discrepancies between flies and mammals. (1) In mammalian Hi-C maps, CTCF sites at both anchors of an extruded loop often engage in high-frequency contacts4 not seen in Drosophila2, Supplementary Fig. 2c). (2) CTCF and cohesin colocalize genome-wide in mammals49,50,51, but cohesin does not colocalize specifically with CTCF in Drosophila13,17. Fly CTCF may therefore not have a robust or unique ability to stall or stabilize loop-extruding cohesin complexes, despite their ability to interact in vitro (Fig. 2g). (3) CTCF-dependent boundaries are directional in mammals4,5,51 but lack clear directionality in flies (Supplementary Fig. 2g).52 All these discrepancies could nevertheless be expected given the probable differences in how fly CTCF interacts with extruding cohesin (Supplementary Fig. 2f). Indeed, previous in silico simulations5 and experiments affecting loop-extrusion processivity across CTCF-dependent boundaries in human cells7,9,10 described CD boundaries with weaker corner interactions more similar to domains observed in Drosophila (Fig. 2). This strongly contrasts with the mammalian genome where extrusion-based mechanisms are responsible for the formation of a large fraction of boundaries. This demonstrates that although domain formation is ubiquitous in different species, the contributions of different mechanisms can vary widely. The limited role that CTCF plays in global genome architecture in flies is nevertheless consistent with our finding that CTCF binding sites are an order of magnitude less frequent in flies (~800 peaks in 130 Mb genome) than in humans (~80,000 peaks in 3 billion bp genome)49, and the fact that alternative boundary-forming mechanisms exist in flies.

CTCF0 directly binds to the C-terminus of CTCF and is recruited to CTCF peaks in a strictly or partially CTCF-dependent manner (Fig. 5). Residual CTCF binding at former CTCF peaks coincides with residual boundary retention in CTCF0 mutants (Fig. 5), (6) CTCF binding to DNA alone is not sufficient for correct expression patterns of a subset of genes that also rely on CTCF. Below we discuss how this work further our understanding of genome folding in Drosophila, CTCF’s role in transcriptional regulation and the molecular basis thereof.

Relaxed requirement of CTCF for Drosophila genome architecture. In comparison to vertebrates, the principles of genome folding into CDs in Drosophila are less clear. On the one hand, the majority of fly CDs were proposed to form by compartmentalization of domains with different transcriptional states or because actively transcribed genes cluster, with little contribution from architectural proteins acting independently of transcription3,48. On the other hand, analyses of enriched transcription factor motifs at domain boundaries defined at high-resolution revealed that 77% were enriched in core promoter motifs (and called promoter boundaries) and the remaining 23% were enriched in motifs of insulator-binding proteins like CTCF, su(Hw) and Ibf1 (and called non-promoter boundaries)17. This suggested that architectural proteins may form some domain boundaries. By completely abating CTCF in vivo, we definitively show that CTCF contributes to the formation of a small fraction (below 10%) of domain boundaries in Drosophila (Fig. 2). This strongly contrasts with the mammalian genome where extrusion-based mechanisms are responsible for the formation of a large fraction of boundaries. This demonstrates that although domain formation is ubiquitous in different species, the contributions of different mechanisms can vary widely. The limited role that CTCF plays in global genome architecture in flies is nevertheless consistent with our finding that CTCF binding sites are an order of magnitude less frequent in flies (~800 peaks in 130 Mb genome) than in humans (~80,000 peaks in 3 billion bp genome)49, and the fact that alternative boundary-forming mechanisms exist in flies.

At strictly CTCF-dependent boundaries, CTCF can form boundaries independently of the presence/absence of a nearby TSS and of detectable transcriptional changes in nearby genes (Figs. 2c and 5d). At partially CTCF-dependent boundaries, defects in CTCF0 mutants are limited by redundant boundary-forming mechanisms often associated with CTCF-independent recruitment of Csp190, Csp190-associated factors or the presence of Csp190-bound transcribed gene TSSs (Figs. 5c–g and 6f). Csp190 marks both promoter and non-promoter boundaries (Fig. 5c)15,17, and it remains to be clarified whether Csp190 or its associated factors directly contribute to domain boundary formation (through similar or unrelated mechanisms as CTCF) or whether boundary formation is governed by transcription of Csp190-bound TSSs. Pervasive transcriptional perturbation globally affects Hi-C contact maps2,16,48,49, indicating that transcription itself or the transcription machinery at least reinforces CDs. Finally, we note that apart from CTCF, the transcription factor Zelda has also been shown to affect CD boundaries in flies: Zelda depletion in early Drosophila embryos led to partial disruption of former Zelda-occupied domain boundaries, and to concurrent loss of RNA polymerase II recruitment which may account for the observed boundary defects16.

Impact of CTCF on transcriptional regulation. Functional studies of how CTCF impacts expression are challenging in mammalian cells. Recent studies that manipulated CTCF binding sites at specific loci have moderated our view of how critical CTCF is for patterned gene expression, but a limitation is that effects can be masked by unperturbed CTCF sites nearby that function redundantly31–33.
Our transcriptional analyses of *Drosophila* CTCF<sup>−</sup> CNSs showed that CTCF is required for patterned expression of selected genes in the CNS while at the same time being dispensable for orchestrating other complex gene expression programs. Gene misexpression may result from defective gene insulation from local regulatory elements, as supported by the binding of CTCF between certain neuronal and non-neuronal genes in vivo (Figs. 3c, d), the increased expression of these genes in CTCF<sup>−</sup> larval CNSs (Figs. 3c–e) and the enhancer-blocking activity of CTCF peaks in S2 cells (Fig. 4b–c). Our reporter assay is independent of chromatin environment, allowing quantitative measurements of insulator activity that reveal a direct relation to
the efficiency of CTCF recruitment. These findings are consistent with our previous characterization of Hox gene misexpression in CTCF0 mutants, which phenocopies deletions of insulator boundaries that maintain the independence of some Hox regulatory domains21. Our ability to detect gene misregulation in CTCF0 larval CNSs likely depends on genomic context, notably the presence of regulatory elements active in this organ in a sufficiently large number of cells to detectably alter transcription.

Why aren’t gene misexpression defects in CTCF0 mutants more widespread? Recent studies have emphasized that specific communication between regulatory elements and gene promoters is controlled at many levels, of which CTCF provides one. In particular, enhancer-promoter compatibility54 and regulation of the chromatin properties of regulatory elements themselves55 also determine whether or not regulatory elements and promoters functionally communicate. CTCF may also function redundantly with other insulator-binding proteins in Drosophila to limit regulatory crosstalk in this compact genome. Unlike what is known in mammals, flies have a family of insulator-binding proteins, many of which have DNA binding domains with which they target specific loci56.

**Molecular basis of how CTCF impacts gene regulation.** Whether CTCF’s ability to form boundary explanations its conserved genetic insulator activity remains an open question15,57. An ideal scenario to address this would be to separate boundary formation from gene insulator function. Human CTCF with mutated critical cohesin-interacting residues was largely functional, but CD boundaries were only partially disrupted10. We observed that some DE genes in CTCF0 mutants are close to partially CTCF-dependent boundaries (Fig. 5d, lane 6). Gene misregulation in the absence of CTCF may therefore occur despite significant retention of a physical boundary, but we did not definitively confirm that these DE genes are direct CTCF targets.

We found that CTCF functionally cooperates with a stably bound regulatory cofactor, expanding the view of how CTCF may impact gene regulation. The relevance of the CTCF-Cp190 interaction has been debated. On the one hand, Cp190 was assumed to be required for CTCF’s insulator function based on the observations (1) that the enhancer-blocking activity of a Hox gene insulator in transgenic reporter assays depended on both CTCF and Cp190, and (2) that CTCF failed to be recruited to many sites on polytene chromosomes in Cp190 mutants58,59. The latter observation was, however, not reproduced in genome-wide ChIP experiments in Cp190 knock-down cells36. On the other hand, no common CTCF and Cp190 target genes were known46, and the interaction between CTCF and Cp190 was recently concluded to be dispensable in vivo45. The latter conclusion was based on deleting residues in CTCF that did not interact with Cp190 in our pull-down experiments (Supplementary Fig. 5b). We identified genes with concordant transcriptional changes upon loss of either CTCF and Cp190 that are potentially directly regulated by both proteins.

**Is this interaction conserved in vertebrates?** Around 40 Cp190-like proteins comprising an N-terminal BTB domain and zinc fingers exist in humans56, but Cp190 does not have a direct ortholog. The C-terminus of human CTCF is capable of interacting with the BTB domain of a Cp190-like protein called KAISO in yeast two-hybrid experiments61, reminiscent of the interaction between fly CTCF C-terminus and the BTB domain of Cp190 (Supplementary Fig. 5b). Whether CTCF transiently interacts with a BTB domain-containing protein in human cells or whether this interaction has not been maintained in vivo remains to be clarified.

**How do Cp190 and CTCF collaborate?** Incomplete overlap of DE genes in CTCF0 and Cp190KO mutants suggests that CTCF requires Cp190 at some loci but not others (Fig. 6a). Alternatively, additional common targets may be masked by other transcriptional changes in Cp190KO mutants or by maternal Cp190 rescuing early defects in these mutants. How Cp190 functions is not known, but it may contribute to CTCF’s insulator activity similarly to how Cp190 contributes to the activities of gypsy and some Hox gene boundary insulators46,62. Cp190 may help CTCF form CD boundaries, or Cp190 may function independently of boundary formation through unknown mechanisms that could uncover paradigms for controlling the communication between genes and regulatory elements.

**Methods**

**Tissue-specific CTCF loss-of-function.** CTCFKO, UAS-FLP/TM6B heterozygotes were crossed to CTCF0/TM6B heterozygotes for an independently isolated CTCF0 allele that also carried an FRT-flanked genomic CTCF rescue transgene and one of various Gal4 drivers: expressed in neuroblasts [w{oriP-Gal} (Bloomington stock 65553)], mature neurons [eln-Gal4 (Bloomington stock 25750)], or muscles [Mef2-Gal4 (Bloomington stock 25756)]. Resulting non-TM6B animals were transheterozygous for CTCF0 alleles, derived from a WT maternal germline, and expressed UAS-FLP under the control of a Gal4 driver leading to tissue-specific excision of the CTCF rescue transgene. D. melanogaster (wildtype) and CTCF0 transheterozygous animals were used as controls.

**Tissue-specific rescue of CTCF0 mutants.** Females trans-heterozygous for two independently isolated CTCF0 alleles were rescued with an FRT-flanked genomic CTCF rescue transgene that was excised in their germline by expressing FLP
recombine under the control of nano regulatory sequences. These females were crossed to CTCTfo-TM6B males carrying a UAS-CTCF-3xHA transgene (HydroFLyte, Bethesda, MA) and a Gal4-VP16 transgene. Resulting non-TM6B animals were transheterozygous for CTCTfo alleles, derived from a maternal germine devoid of CTCT (CTCTo mutant background) and expressed UAS-CTCF under the control of a Gal4 driver. w118 stocks were used as WT control.

Drosophila viability tests. Three sets of 30–40 third instar larvae of desired genotypes were transferred into separate vials and the number of pupae and fully hatched adults was recorded. The average percentage and standard deviation of animals alive at each developmental stage and over a 30-day period after hatching were scored and plotted in Kaplan-Meier survival plots with 5% confidence intervals from the triplicate experiments.

Antibodies. For this study, polyclonal rabbit antibodies were raised against CTCT–293 and Cpf1–1096. Proteins were recombinantly purified in E. coli by tandem affinity purification using N-terminal GFP- and C-terminal His-tags. Tags were cleaved off by 3C protease and used for immunization.

Western blotting. Forty third-instar larval CNSs per biological replicate were dissected in ice-cold PBS. Samples were sonicated in 100 µl of 20 mM Tris pH 7.5, 250 mM LiCl, 1 mM EDTA, 0.5% Igepal CA-630, 0.5% sodium deoxycholate) and pre-mixed Protein A and G Dynabeads (Thermo Fisher 100-01D and 100-03D) elution from a QIAGEN Minelute PCR purification kit. ChIP-seq analysis

Hi-C library preparation. 60 third-instar larval CNSs (~600,000 cells) per biological replicate were dissected in ice-cold PBS. CNSs or a single whole-bodied female fly were crushed in RPMI supplemented with 10% fetal bovine serum using a micro-pestle. Cells were fixed in 1% (v/v) paraformaldehyde for 10 min at room temperature. The Hi-C libraries were prepared using MboII and MseI as restriction enzymes. Restricted ends were marked with biotin, then ligated. Fragmented DNA was enriched for pairwise DNA junctions by biotin pull-down using Dynabeads MyOne Streptavidin T1 beads following the manufacturer’s instructions. Illumina sequencing libraries were prepared with standard protocols. 4×106 equilibrum pools of multiplexed Hi-C libraries were subjected to paired-end sequencing on Illumina HiSeqX Ten and HiSeq4000 instruments.

Hi-C data processing. We pre-computed a table containing the positions of all restriction sites used for Hi-C present in the dm6 genome. The FASTQ read pairs were analyzed with a Perl script available for download in the Mcmapc package (see Code Availability) to locate and separate fusion sites using the patterns /CTCF/ (CTCF/TATAAA) or /CTCF/ (CTCF/TATAAA) to estimate local copy number status of the underpinning genomic region. Reads were filtered out before use to estimate local copy number status in the reference genome. Genomes were indexed with the Bismark software, v0.5.267. Low coverage regions (bins with no contacts and those with the 2% smallest total number of contacts among bins) were filtered out. Pearson correlation coefficients were determined for every pair of normalized matrices by flattening each matrix and evaluating the Pearson correlation coefficient for the resulting vector, using only pairs of bins at a genomic distance below 1 Mb. The limitation on the distance was introduced to compare contacts at a scale relevant to the analyses performed in this manuscript which were based on sets of CNVs. Resulting Pearson correlation matrices were thus split into 50449 for all replicates, showing that they were well correlated and that WT and CTCT Hi-C matrices were globally similar. For the analyses presented in the main figures, pooled replicates of the same genotype were downsampled to 200 million contacts per genotype. Raw Hi-C contact matrices were obtained by binning Hi-C pairs at 1 kb resolution. These matrices were then normalized with the ICE normalization implemented in ice v0.5.267. Low coverage regions (bins with no contacts and those with the 2% smallest total number of contacts among bins) were filtered out before normalization (these regions are marked by gray lines in Hi-C maps shown in the figures).

For each normalized Hi-C contact matrix, CD boundaries were called using TopDom68. Given a window size w, a physical insulation score was defined for each bin i as:

\[
\log_{2} \left( \frac{\text{binSignal}_i}{\frac{1}{w} \sum_{j \in \text{neighbors of } i} \text{binSignal}_j} \right)
\]

where binSignal is the average normalized Hi-C contact frequency between w bins upstream of bin i and w bins downstream of bin i determined by TopDom. The strength of a boundary at bin i was thus estimated as the log2 of the binSignal value at bin i normalized by its local average on a window of size w. With this definition, lower insulation scores indicate stronger boundaries. We extracted CD boundaries and physical insulation scores for Hi-C matrices at 2 kb resolution using window sizes 2, 80, and 160 kb. CD boundaries found in the 2 kb window sizes were merged, and the average insulation score obtained with all window sizes was retained. To facilitate comparisons of CD boundaries found in WT and CTCTO genotypes and avoid mismatches due to small fluctuations of CD boundary positions obtained with different window sizes or genotypes, groups of consecutive boundaries (i.e., missing bins between two other) were merged. Groups of consecutive boundaries were replaced by the boundary with the lowest insulation score (average of both genotypes for boundaries common to WT and CTCTo).
Hi-C maps were visualized in R and Juicebox69 (see Supplementary Table 3 for links to interactive maps for browsing).

A/B compartment calling. A/B compartment calling was performed following the method proposed in Lieberman Aiden et al.70. Each individual chromosome arm (chr2L, chr2R, chr3L, chr3R, chr4, chrX) was analyzed separately. Normalized Hi-C contact matrices at 2 kb resolution were considered after discarding invalid bins (low coverage regions) and bins around centromeres (chosen for exclusion as dm6 coordinates >22,170,000 for chr2L, <5,650,000 for chr2R, >22,900,000 for chr3L, <4,200,000 for chr3R). Observed-over-expected matrices were generated by dividing the normalized Hi-C contact matrices by the average number of normalized Hi-C contacts at the corresponding genomic distance. For each chromosome arm, the first eigenvector of the correlation matrix was obtained by principal component analysis of the observed-over-expected matrix. Each eigenvector was then centered around zero by subtracting its mean value, then multiplied by the sign of the Pearson correlation between the eigenvector and the number of expressed gene TSSs per 2 kb bin. 2 kb bins with positive eigenvector values were assigned to compartment A, those with negative eigenvector values were assigned to compartment B. chr4 eigenvectors appeared to reflect a large-scale structure that separated the chromosome into two halves, and were thus excluded from Supplementary Fig. 3d.

Comparison with CD boundaries from other Hi-C studies. To assess whether CD boundaries called in our study could correspond to small CDs resolved in higher resolution Hi-C contact maps (analyzed at 500 bp resolution instead of 2 kb used here), we aligned our Hi-C CD boundaries (PCR to end) with the relevant boundaries from published Hi-C contact maps. The general result is that Eagen et al. did not report CDs smaller than 6 kb. Only 31% of our domain boundaries were within 2 kb of a ≤4 kb CD identified by Ramirez et al. Thus, very few (31/3970, or <1%) of our domain boundaries may correspond to a small domain defined by Ramirez et al. We next asked: How many domain boundaries that disappear in CTCP mutants could correspond to small domains? At least 70% of our domain boundaries identified only in WT were within 2 kb of a ≤4 kb CD identified by Ramirez et al. Domain boundaries identified by Ramirez et al. are displayed together with domain boundaries identified in this study in all Hi-C screenshots throughout the manuscript for comparison.

Differential Hi-C RNA-seq analysis. RNA-seq reads were mapped both to the dm5 Drosophila melanogaster reference genome and to Flybase gene models and transcripts (dmel-all-r6.26.gtf.gz) using Micmap80. The results of both mappings were combined into spliced alignments in BAM format file. Then, htseq-count (v0.9.1) was used to produce read counts per gene.1 Statistical analysis was performed in R (v3.5.1). Genes with <1 count per million in at least three replicate samples were filtered out using EdgeR (v3.22.5)66. Normalization and differential expression analysis were performed in DESeq2 (v1.21.2)72 individually for both WT versus CTCP and WT versus ChrIp0073 and WT control.

Differential RNA-seq analysis. RNA-seq reads were mapped both to the dm5 Drosophila melanogaster reference genome and to Flybase gene models and transcripts (dmel-all-r6.26.gtf.gz) using Micmap80. The results of both mappings were combined into spliced alignments in BAM format file. Then, htseq-count (v0.9.1) was used to produce read counts per gene.1 Statistical analysis was performed in R (v3.5.1). Genes with <1 count per million in at least three replicate samples were filtered out using EdgeR (v3.22.5)72. Normalization and differential expression analysis were performed in DESeq2 (v1.21.2)72 individually for both WT versus CTCP and WT versus ChrIp0073 and WT control. Statistical significance was tested by Wald test and the Benjamini-Hochberg method was used for multiple testing adjustment. A significance threshold of $|\text{fold change}| > 1.5$ and $p$-adjusted < 0.05 was used to identify DE genes. The R package ggplot2 (v2.2.1) was used for data visualization.

RNA-FISH. Labeled RNA probes were generated by in vitro transcription with DIG-UTP labeling mix (Roche 11277073910) and T7 RNA polymerase (Roche 10881876001) antisense to full-length complementary DNA clones of Spin029 (F22034) and IFT52 (MIP14440), genomic DNA amplified from dm5 coordinates chr3L:16666244, or cDNAs amplified using gene-specific primers from a cDNA library prepared from Drosophila embryo (see Supplementary Data 10 for primer sequences). After DNase I digestion for 20 min at 37 °C, RNA probes were fragmented by incubating 20 min at 65 °C in 60 mM NaClO4, 40 mM NaHCO3, pH 10.2, precipitated in 300 mM sodium acetate pH 5.2, 1.25 M LiCl, 50 mg/ml RNAse A and 80 µg/ml RNAse T1, resuspended in 50% formamide, 75 mM sodium citrate pH 5, 750 mM NaCl, 100 µg/ml salmon sperm DNA, 50 µg/ml heparin and 0.1% Tween20, and stored at −20 °C. Embryos or third instar larval cuticles were fixed in 4% paraformaldehyde for 30 min at room temperature, washed, and then stored in 100% MeOH at −20 °C for at least overnight. Samples were rehydrated in PBS with 0.1% Tween 20, post-fixed in 4% paraformaldehyde for 20 min at room temperature, progressively equilibrated to hybridization buffer (50% formamide, 75 mM sodium citrate pH 5, 750 mM NaCl) and heated to 65 °C. RNA probes were diluted 1:50 in hybridization buffer, denatured at 80 °C for 10 min then placed on ice, and added to the samples overnight shaking at 65 °C. Samples were washed 6 times 10 min each in 5× SSC followed by 5× SSC, then probed 1:150 with 0.1% Triton X-100. Samples were incubated overnight at 4 °C in anti-dig peroxidase (Roche 11207733910) diluted 1:2000 in PBS, 0.1% Triton X-100, 1× Western blocking reagent (Sigma 21921783). Samples were washed six times 10 min in PBS with 0.1% Tween20, labeled with Cy3 (Avidin in the TSA Plus Kit (Perkin Elmer NEL73001KT) for 3 min at room temperature, washed 6 times 10 min in PBS with 0.1% Tween20, and finally mounted with DAPI to stain DNA. Images were acquired on a Zeiss LSM 880 microscope with a ×20 objective and visualized with Fiji software v2.1.0.133.
NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-21366-2
ARTICLE

potassium acetate, 10% glycerol) and allowed to bind to each other at 4°C for 2 h. Twenty microilters of this solution was removed as ‘input’ sample and boiled in SDS-PAGE loading buffer. GFP-binder beads (Agarose beads covalently bound to GFP-nanobody; 20 µl per reaction) were washed in binding buffer and added to the binding reactions for 30 min at 4°C on a rotating wheel to bind to the GFP-tagged CTCF construct. Beads were harvested by centrifugation (1 min, 700 g) and washed twice with 1 ml of binding buffer. The final immobilized material was eluted by boiling in 50 µl of SDS-PAGE loading buffer. Input and pulldowns were loaded onto a 12% SDS-PAGE gel, and the proteins were visualized by staining with Coomassie.

Pull-downs between C-terminal CTCF constructs and Cp190 BTB domain. Expression plasmids encoding GFP-His-tagged constructs of the C-terminal domain of CTCF (all with Ampicillin resistance) were co-transformed with an expression plasmid carrying a His-tagged Cp190 BTB-domain (with Kanamycin resistance) into the E. coli Rosetta strain. Colonies were inoculated in 10 ml TB cultures and grown at 37°C to an OD600 of 1. The culture temperature was then reduced to 18°C, and 0.5 mM IPTG was added to induce protein expression. Cells were harvested after overnight incubation at 18°C, and the pellets were resuspended in 2 volumes of lysis buffer (50 mM Tris pH 7.5, 200 mM NaCl, 5% glycerol, 25 mM Imidazole). Cells were lysed by sonication and the lysate was clarified by centrifugation at 16000 g for 10 min at 4°C. The lysate was split into two halves, which were incubated for 1 h at 4°C with either 20 µl of GFP-binder resin, pHH-DsRed-attP vector78. Guide RNAs close to the start and stop codons of the Cp190 open reading frame were cloned into pCFD3 vector25. Plasmids were co-injected into nanos-Cas9 embryos26. Experiments were performed in animals transheterozygous for two independent knockodds.

Generation of Cp190KO animals. Cp190KO mutants were rescued into viable and fertile adults with an FRT-flanked 7 kb Cp190 genomic rescue transgene (dnm6 coordinates chr13:15276425-15276409) amplified by PCR. The Cp190 rescue cassette was excised from male and female germlines through nanos-Gal4/VP16 driven expression of nanos-Cas9. GFP-3C-CTCF14 animals were collected from crosses between such males and females.

Statistics and reproducibility. All described replicate experiments are biological (not technical) replicates. For all box plots: center line, median; box limits, upper and lower quartiles; upper whisker extends to the largest value no further than 1.5x interquartile range from the upper hinge; lower whisker extends to the smallest value no further than 1.5x interquartile range from the lower hinge; points, outliers. Figure 2g: This experiment was repeated twice from independently grown bacterial cultures, with similar results. Figure 3e and Supplementary Fig. 1a–b: n = 10 independent third instar larvae per genotype were examined over two independent experiments each. All animals showed similar expression patterns for a given gene, that was characteristic of each genotype. RNA-FISH probes for additional genes were tested on larval nervous systems but discarded because they showed an inconsistent pattern (variable, asymmetric signal in the optic lobes in all genotypes) that we concluded was non-specific background. Figure 8c, e, r, n = 50 independent embryos per genotype were examined over two independent RNA-FISH experiments each. All animals showed similar expression patterns for a given gene, that was characteristic of each genotype. Supplementary Fig. 2a: The experiment was repeated twice with independently prepared extracts, with similar results. Supplementary Fig. 5b: The pull-down experiments were repeated twice from independently grown bacterial cultures, with similar results.

Data availability All sequencing data (Hi-C, ChIP-seq, RNA-seq) that support the findings of this study were deposited in Gene Expression Omnibus with accession code GSE146752. Hi-C maps are browsable on Juicerbox (links in Supplementary Table 3). Mass spectrometry proteomics data were deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD019487. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. Additional information is provided in Supplementary Data files 1–10 and a reporting summary for this Article is available as a Supplementary Information file. Source data are provided with this paper.

Code availability All software used as described in the Methods to map, visualize and analyze data is available open source and freely available for download in the following links: “Micmap v2.20.200223 [https://github.com/sib-swiss/micmap]”; “DESeq v1.22.2 [https://bioconductor.org/packages/release/bioc/html/DESeq2.html]”; “HTSeq v0.9.1 [https://github.com/simon-anders/htseq]”; “csaw v0.5.2 [https://github.com/hichib/csaw]”; “TopDom v0.0.2 [https://github.com/jasminezhoulab/TopDom]”; “R v3.5.1 [https://www.R-project.org/]” with packages “caw v1.16.1 [https://bioconductor.org/packages/release/bioc/html/caw.html]”, “edgeR v3.22.5 [https://bioconductor.org/packages/release/bioc/html/edgeR.html]”, “Enviser v6.0 [https://cran.r-project.org/package=enviser]” and “ggplot2 v3.1.0 [https://ggplot2.tidyverse.org/]”; “bedtools multivc v2.29.2 [https://bedtools.readthedocs.io/en/latest/]”; “Juicerbox v1.5 [https://alienlab.org/juicerbox]. Custom scripts are provided in “link [https://github.com/gambettalab/kauhla252020].”

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

The following databases used for searching were Drosophila melanogaster and Escherichia coli reference proteomes based on the UniProt database (www.uniprot.org; versions of 31 January 2019, containing 21,939 and 1915 sequences respectively), and a D. melanogaster database containing the most usual environmental contaminants and the enzymes used for digestion (keratins, trypsin, etc). Mass tolerance was 4.5 ppm on precursors (after recalibration) and 0.5 Da on CID fragments. Both peptide and protein identifications were filtered at 1% FDR relative to hits against a decoy database built by reversing protein sequences. The MaxQuant24 search engine was used to search against the contaminant database with Perseus software27 to remove proteins matched to the contaminants database as well as proteins identified only by modified peptides or reverse database hits next. The table was filtered to retain only proteins identified by a minimum of two peptides, the IBAQ quantitative values were log-2 transformed and missing values imputed with a constant value of 9.

Generation of Cp190KO animals. Cp190KO mutants were rescued into viable and fertile adults with an FRT-flanked 7 kb Cp190 genomic rescue transgene (dnm6 coordinates chr13:15276425-15276409) amplified by PCR. The Cp190 rescue cassette was excised from male and female germlines through nanos-Gal4/VP16 driven expression of nanos-Cas9. GFP-3C-CTCF14 animals were collected from crosses between such males and females.

---

The text describes experiments involving the isolation and analysis of proteins using GFP and Coomassie staining. It mentions the use of E.coli expression plasmids for generating C-terminal CTCF constructs and Cp190 BTB domain. Pull-down experiments were performed using GFP-His-tagged constructs. The proteins were eluted by boiling in SDS-PAGE loading buffer. The resulting protein extracts were analyzed by SDS-PAGE and stained with Coomassie. The study also involved the generation of Cp190KO animals by using FRT-flanked genomic rescue transgenes. Data availability and code availability sections are provided, along with links to repositories for the datasets and the software used in the study.
1. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* 171, 305–309. e24 (2017).

2. Rowley, M. J. et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* 67, 837–852.e7 (2017).

3. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* 19, 1–800 (2018).

4. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution chromosome domains from genomic compartmentalization. *Nucleic Acids Res.* 45, 1–1730 (2017).

5.coordinates of CTCF protein in Drosophila melanogaster. *Mol. Cell* 708, 1–9 (2020).

6. Moore, J. M. et al. Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. *PLoS ONE* 7, e34915 (2012).

7. Shoshnikova, N., Montavon, T., Leleu, M., Galjart, N. & Duboule, D. Functional analysis of CTCF during mammalian limb development. *Dev. Cell* 19, 819–830 (2010).

8. Gambetta, M. C. & Furlong, E. E. M. The insulator protein CTCF is required for correct Hox gene expression, but not for embryonic development in *Drosophila*. *Genetics* 210, 129–136 (2018).

9. Arzate-Mejia, R. G., Cerdeiro-Castillo, A. J., Guerrero, G., Furlan-Magaril, M. & Recillas-Targa, F. In situ dissection of domain boundaries affect genome topology and gene transcription in *Drosophila*. *Nat. Commun.* 11, 894 (2020).

10. Ghavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* 51, 1272–1282 (2019).

11. Yokoshi, M., Segawa, K. & Fukaya, T. Visualizing the role of boundary elements in enhancer-promoter communication. *Mol. Cell* 78, 224–235.e5 (2020).

12. Andrey, G. et al. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* 340, 1234167–1234167 (2013).

13. Nara, V. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385 (2012).

14. Symmons, O. et al. Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* 24, 390–400 (2014).

15. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 529, 110–114 (2016).
60. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 43, D257–D260 (2015).

61. Defossez, P.-A. et al. The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso. J. Biol. Chem. 280, 43017–43023 (2005).

62. Pai, C.-Y., Lei, E. P., Ghosh, D. & Corces, V. G. The centrosomal protein CP190 is a component of the gypsy chromatin insulator. Mol. Cell 16, 737–748 (2004).

63. Iseli, C., Ambrosini, G., Bucher, P. & Jungenneel, C. V. Indexing strategies for rapid searches of short words in genome sequences. PLoS ONE 2, e579 (2007).

64. Lun, A. T. L. & Smyth, G. K. cscav: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. 44, e45 (2015).

65. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25 (2010).

66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. Oxf. Engl. 26, 139–140 (2009).

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

68. Shin, H. et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 44, e70 (2015).

69. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3, 99–101 (2016).

70. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).

71. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).

72. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550–521 (2014).

73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

74. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372 (2008).

75. Cox, J. et al. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J. Proteome Res 10, 1794–1805 (2011).

76. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. Nature 473, 337–342 (2011).

77. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (pro)teomics data. Nat. Methods 13, 731–740 (2016).

78. Gratz, S. J. et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics 196, 961–971 (2014).

79. Port, F., Chen, H.-M., Lee, T. & Bullock, S. L. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. Proc. Natl Acad. Sci. USA 111, E2967–E2976 (2014).

Acknowledgements

We thank Winship Herr, Richard Benton, Jean-Yves Rognant and Naoko Mizuno for critical comments on the manuscript. We thank Patrice Waridel and Manfredo Quadrani for mass spectrometry analyses. We thank René Drees for advice on statistical analyses. MCG thanks Eileen Furlong for support during early phases preceding this work. Deep sequencing was performed at the Genomic Technologies Facility (GTF); mass spectrometry was performed at the Protein Analysis Facility (PAF) and imaging was performed at the Cellular Imaging Facility (CIF) at the Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, Switzerland. This work was supported by the Swiss National Science Foundation (SNSF #184715 to M.C.G.) and the University of Lausanne.

Author contributions

M.C.G. conceived the project and designed experiments. E.L.A. and A.O. conceived and designed Hi-C experiments. A.K., G.M., I.O., A.O., M.T., P.C., A.S., and M.C.G. performed the experiments. J.D., P.C., A.S., O.D., F.M., C.I., Y.E., D.W., M.S.S., and N.G. analyzed data. Y.E. created links interactive browsing of Hi-C and ChIP-seq data in Juicebox. M.C.G. prepared the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21366-2.

Correspondence and requests for materials should be addressed to E.L.A. or M.C.G.

Peer review information Nature Communications thanks Elphège Nora, Sergey Razin, Félix Recillas-Targa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

© The Author(s) 2021