Article

The structural basis for cohesin–CTCF-anchored loops

https://doi.org/10.1038/s41586-019-1910-z

Received: 21 February 2019

Accepted: 5 December 2019

Published online: 6 January 2020

Yan Li[1,6], Judith H. I. Haarhuis[2,6], Ángela Sedeño Cacciatore[2,6], Roel Oldenkamp[2], Marjon S. van Ruiten[2], Laureen Willems[2], Hans Teunissen[3], Kyle W. Muir[1,4*], Elzo de Wit[3*], Benjamin D. Rowland[2*] & Daniel Panne[1,5*]

Cohesin catalyses the folding of the genome into loops that are anchored by CTCF[1]. The molecular mechanism of how cohesin and CTCF structure the 3D genome has remained unclear. Here we show that a segment within the CTCF N terminus interacts with the SA2–SCC1 subunits of human cohesin. We report a crystal structure of SA2–SCC1 in complex with CTCF at a resolution of 2.7 Å, which reveals the molecular basis of the interaction. We demonstrate that this interaction is specifically required for CTCF-anchored loops and contributes to the positioning of cohesin at CTCF binding sites. A similar motif is present in a number of established and newly identified cohesin ligands, including the cohesin release factor WAPL[2,3]. Our data suggest that CTCF enables the formation of chromatin loops by protecting cohesin against loop release. These results provide fundamental insights into the molecular mechanism that allows the dynamic regulation of chromatin folding by cohesin and CTCF.

The interphase genome is folded in 3D through the concerted action of cohesin and CTCF. These architectural factors regulate the interactions between regulatory elements along chromosomes to control gene expression[1,4,5]. Cohesin is thought to catalyse genome folding through a process known as 'loop extrusion', which involves the formation of chromosome loops that are progressively enlarged[6–10]. Genomic regions within which cohesin forms loops are also known as topologically associating domains (TADs), or loop domains. TADs are flanked by CTCF sites that are thought to act as barriers to the loop extrusion process[11,12]. CTCF acts as such a boundary only when the 3′ ends of CTCF binding motifs are oriented towards the inside of the TAD[9,13,14]. Consequently, only convergently oriented pairs of CTCF sites form CTCF-anchored loops[15,16].

This model is supported by genetic manipulation of cohesin and CTCF. Depletion of the core cohesin subunit SCC1 leads to loss of TADs[12,17]. By contrast, depletion of the cohesin release factor WAPL increases the size of chromatin loops[10,12,18]. CTCF depletion leads to a marked loss of CTCF-anchored loops[11,12]. However, how CTCF can act as a directional boundary that controls cohesin loop extrusion remains unknown.

Here we have investigated the mechanism of cohesin interaction with CTCF, and how this interaction contributes to genome organization. We have identified an N-terminal segment of CTCF that directly engages the SA2–SCC1 subcomplex of cohesin. Our crystal structure of the SA2–SCC1–CTCF complex elucidates the molecular basis of the interaction. CTCF-anchored loops are abolished in mutants of key amino acids in the interface, but the accumulation of cohesin at CTCF binding sites across the genome is only partially impaired. In addition to its function as a translocation barrier, CTCF thus possesses a distinct loop-stabilizing activity, which is realized through a direct interaction with cohesin. Furthermore, we observe intermolecular competition between CTCF and the cohesin release factor WAPL for this interface, which suggests a mechanism by which chromatin loop formation may be dynamically regulated.

Structure of the SA2–SCC1–CTCF complex

Previous data indicate that CTCF directly interacts with the SA2 subunit of the cohesin complex[19,20]. To map this interaction, we produced a series of CTCF truncations as proteins fused to glutathione *S*-transferase (GST), and performed pulldown assays against a complex of SA2 and SCC1[2]. CTCF fragments that contained amino acids 227–235 generally retained SA2–SCC1 on GST beads (Extended Data Fig. 1a, b). Isothermal calorimetry experiments further showed that the interaction is largely driven by amino acids 222–231 of CTCF, as the interaction involving this truncated CTCF retained an equilibrium dissociation constant ($K_d = 1.04 \pm 0.20$ μM) comparable to that of an extended CTCF construct ($K_d = 0.62 \pm 0.07$ μM) (Extended Data Fig. 1c, Extended Data Table 1a). To understand the molecular details, we produced crystals of the SA2–SCC1 complex in the presence of a peptide comprising the CTCF binding motif, and determined the structure by molecular replacement at a resolution of 2.7 Å (Extended Data Table 1b). An $F_o - F_c$ omit electron density Fourier map exhibited clear features that correspond to the CTCF peptide (Extended Data Fig. 1d).

The CTCF peptide is bound to the convex surface of SA2 (Fig. 1a, b). The CTCF binding surface is predominantly hydrophobic and composed of amino acids that are contributed by both SA2 and SCC1. The lead 'anchoring' amino acids of CTCF, which bury the largest solvent-accessible surface area upon binding, are Y226 and F228 (Fig. 1b). F228 inserts into a pocket comprising amino acids from SCC1 (S334, I337 and

[1]European Molecular Biology Laboratory, Grenoble, France. [2]Division of Gene Regulation, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [3]Division of Gene Regulation, Oncode Institute, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [4]MRC Laboratory of Molecular Biology, Cambridge, UK. [5]Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Leicester, UK. [6]These authors contributed equally: Yan Li, Judith H. I. Haarhuis, Ángela Sedeño Cacciatore.
*e-mail: kmuir@mrc-lmb.cam.ac.uk; e.d.wit@nki.nl; b.rowland@nki.nl; daniel.panne@le.ac.uk

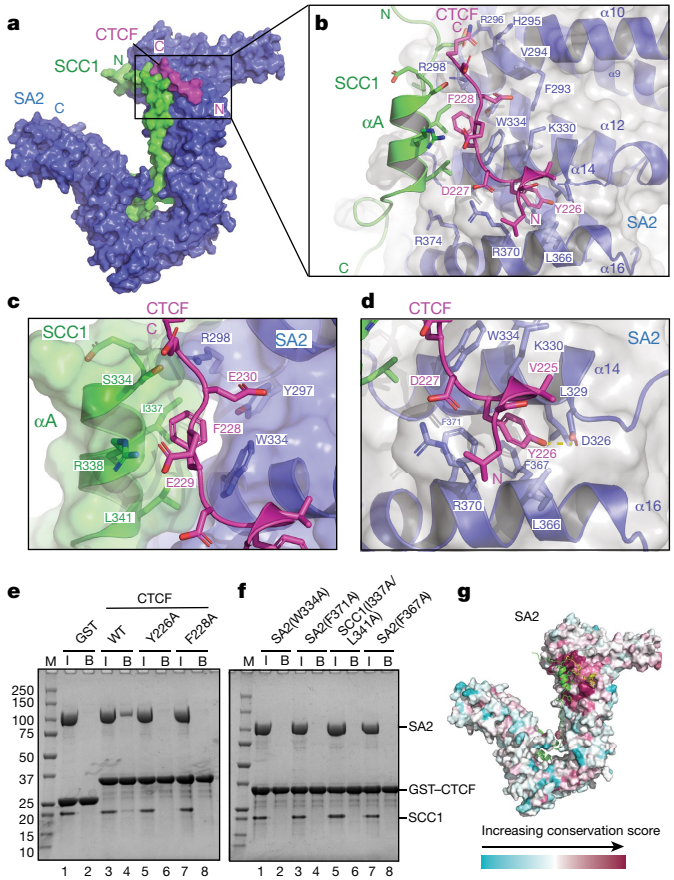


Fig. 1 | Structure of the SA2–SCC1–CTCF complex. a, Surface-rendered cartoon of the SA2–SCC1–CTCF complex, with components coloured in blue, green and magenta, respectively. C, C terminus; N, N terminus. **b**, Detailed view of the binding interface with SA2 residues in blue, SCC1 in green and CTCF in magenta. **c**, **d**, Details of the composite binding pocket around CTCF F228 (**c**) and CTCF Y226 (**d**). **e**, **f**, GST pulldown analysis of CTCF (**e**) and SA2 or SCC1 variants (**f**). B, bound fraction; I, input; M, molecular weight marker. Controls are shown in **e** (lanes 1 and 2). Experiments were done once. **g**, SA2 is surface-rendered and coloured according to sequence conservation.

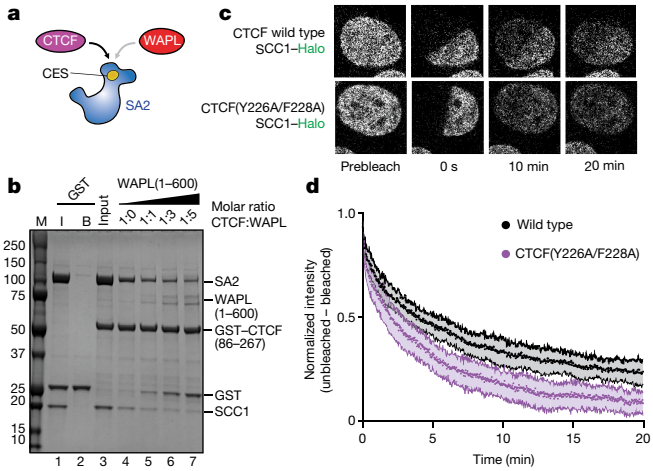


Fig. 2 | CTCF interaction stabilizes cohesin on DNA. a, Schematic of competition between CTCF and WAPL. **b**, Increasing amounts of WAPL residues 1–600 (WAPL(1–600)) (lane 4–7; molar ratios are indicated) were incubated with GST–CTCF and SA2–SCC1 and the bound fraction was analysed. Three independent experiments were done, with consistent results. A representative example is shown. **c**, Example images of cells used in **d** at the indicated time points after photobleaching. FRAP was performed in G1 cells (Extended Data Fig. 3d). **d**, Quantification of the FRAP experiments. Averages and standard deviations for 21 wild-type cells and 17 *CTCF*$^{Y226A/F228A}$ cells, measured over 3 independent experiments.

L341) and SA2 (Y297 and W334) (Extended Data Fig. 1e). The hydroxyl group of Y226 hydrogen-bonds with D326 of SA2 in a deep hydrophobic pocket lined by L329, L366 and F367 (Fig. 1d). E229 and E230 of CTCF constitute secondary anchoring residues, which presumably contribute to binding specificity by forming salt bridges with R298 of SA2 and R338 of SCC1 (Fig. 1c). As CTCF engages a composite binding surface containing amino acids from SCC1 and SA2, previous mapping studies that used isolated SA2 may have been misleading[20].

Analysis of the CTCF binding interface

Mutagenesis of Y226A or F228A in CTCF abolished SA2–SCC1 binding in a GST pulldown assay (Fig. 1e). Likewise, the substitution of critical amino acid residues—including W334A, F371A or F367A in SA2 or I337A/L341A in SCC1—abolished CTCF binding (Fig. 1f). SA2 contains an 86-amino-acid motif termed the 'stromalin conservative domain'[21,22] or 'conserved essential surface' (CES)[2,23], which is conserved from fungi to mammals and coincides with the CTCF binding pocket. For simplicity, we refer to the composite SA2–SCC1 binding pocket as the CES. Mapping of sequence conservation onto the structure confirms that the CES is highly conserved (Fig. 1g, Extended Data Fig. 2a). A series of missense mutations are found in *SA2* (also known as *STAG2*), *SCC1* (also known as *RAD21*) and *CTCF* in various types of cancer[24]. The mapping of mutation frequencies onto the structure shows that amino acids that are largely buried in the interface are hotspots in cancer (Extended Data Fig. 2b).

Previous data indicate that the SA2–SCC1 complex interacts with multiple cohesin regulators[2,23,25]. This includes two factors with opposing functions: WAPL, the general cohesin release factor, and shugoshin (SGO1), a factor that is crucial for the protection of centromeric cohesion during mitosis[2,26–28]. This antagonism arises as a result of direct competition for binding to the CES of SA2–SCC1[2]. As mutants reported to interfere with both SGO1 and WAPL binding cluster in the CES, we investigated whether these proteins bind to SA2–SCC1 by a mechanism comparable to that of CTCF. In SGO1, the reported CES-binding domain (amino acids 313–353) contains a conserved FGF-like motif that strongly resembles that of the CTCF peptide. Vertebrate WAPL also contains several FGF motifs in its N-terminal region that are potentially involved in cohesin regulation[3,29]. A minimal fragment of WAPL capable of competing with SGO1 for access to the CES (amino acids 410–590) contains two such FGF motifs[2]. We observed that a peptide that spans the second and third FGF motif of WAPL (amino acids 423–463) bound to SA2–SCC1 with a K_d of about 32.8 μM (Extended Data Fig. 2c), whereas a peptide that comprises only the third motif bound more weakly (Extended Data Table 1a). The peptide containing the CES motif of CTCF therefore binds with higher affinity than do peptides that contain the WAPL motif(s).

CTCF stabilizes cohesin on chromatin

The observation that CTCF and WAPL can bind to the same surface on SA2–SCC1 raises the possibility that their interaction with the CES is mutually exclusive (Fig. 2a). To determine whether WAPL competes with CTCF for binding to the CES of SA2–SCC1, we performed GST-pulldown competition assays. Titration of WAPL residues 1–600 against a preformed complex of GST–CTCF and SA2–SCC1 depleted the latter from the beads (Fig. 2b). Similarly, titration of a peptide of SGO1 phosphorylated at T346—which has previously been reported to preclude WAPL binding[2]—also displaced SA2–SCC1 from GST–CTCF (Extended Data Fig. 2d). Hence, the CES of SA2–SCC1 is a general interaction hub for multiple regulators of cohesin (Extended Data Fig. 2e). Whereas SGO1 precludes WAPL binding (thus stabilizing centromeric

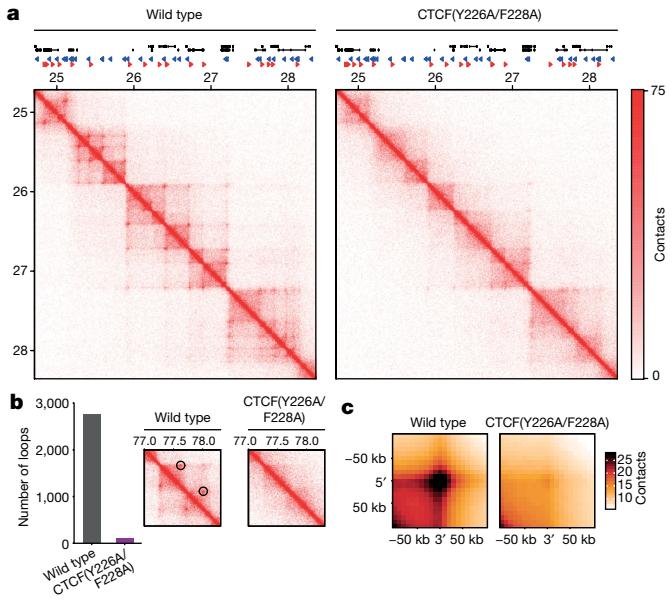


Fig. 3 | CTCF–CES interaction is required for CTCF-anchored loops. a, Hi-C contact matrices of the HOXA locus at 10-kb resolution, normalized to 100 million contacts per sample. Genes and CTCF sites are depicted above the contact matrices. **b**, Genome-wide quantification of loops using HICCUPS[15]. The inset shows an example of called loops for a region of chromosome 16. **c**, Aggregate peak analysis for the loops defined genome-wide in wild-type cells. The Hi-C signal is averaged across these locations for both cell lines.

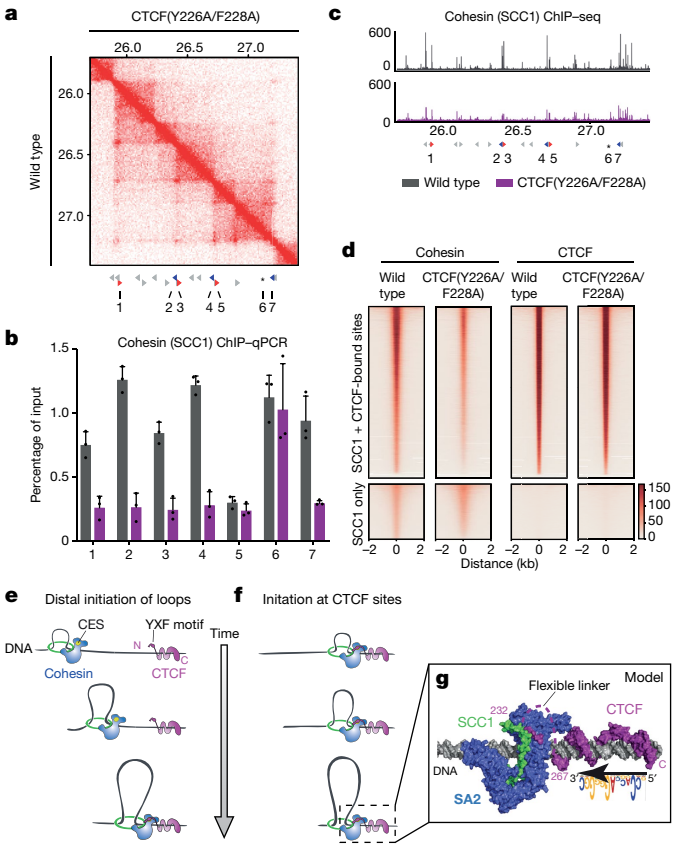


Fig. 4 | CTCF–CES interaction promotes localization of cohesin to CTCF sites. a, Hi-C contact matrix of a region of chromosome 7 at 10-kb resolution. CTCF sites are depicted below; those selected for qPCR are shown in colour (forward motifs in red, reverse motifs in blue). The numbers underneath indicate the loci used for qPCRs in **b**. Locus 6 (indicated with *) is the *HOTAIRM1* transcription start site. **b**, ChIP–qPCR analysis of SCC1 (cohesin) at the loci depicted in **a**. The mean of three independent ChIP experiments is shown with standard deviations. **c**, ChIP–seq tracks for SCC1 of the same region of chromosome 7 as is depicted by Hi-C in **a**. The ChIP–qPCR loci of **b** are depicted below. **d**, ChIP–seq heat map of the cohesin subunit SCC1 (left) and CTCF (right). The depicted sites are selected for being bound in wild-type cells by both SCC1 and CTCF (top), or only by SCC1 (bottom). **e**, Cohesin-mediated looping initiates at distal sites until encounter of the N-terminal end of CTCF. **f**, Cohesin-mediated looping starts at CTCF sites. **g**, Molecular model of CTCF and SA2–SCC1 bound to DNA (grey). The YXF motif is separated by a flexible linker spanning residues 232–267 (magenta dotted line) to the C-terminal DNA-binding domain of CTCF.

cohesin in mitosis), CTCF could exert a similar function at CTCF sites in interphase.

To test whether the CTCF–CES interaction stabilizes cohesin on chromatin, we mutated the endogenous allele of *CTCF* in the human haploid HAP1 cell line using CRISPR–Cas9 technology. We thereby obtained HAP1 cells that contained the $CTCF^{Y226A/F228A}$ mutation as their sole copy of *CTCF* (Extended Data Fig. 3a, b). These cells displayed no obvious proliferation defects. To study the consequences of the *CTCF* mutations on cohesin turnover on chromatin, we endogenously tagged the core cohesin subunit SCC1 with a Halo tag in both wild-type and $CTCF^{Y226A/F228A}$ cells (Extended Data Fig. 3c), and performed fluorescence recovery after photobleaching (FRAP) experiments. In wild-type cells, we found that—over a period of 20 min—a fraction of the fluorescent cohesin population did not recover. However, in $CTCF^{Y226A/F228A}$ cells we observed a near-complete recovery by FRAP, which demonstrates that cohesin is more mobile in these cells (Fig. 2c, d). The CTCF–CES interaction therefore stabilizes a subpopulation of cohesin on chromatin.

Loops require CTCF–CES binding

To investigate the role of the cohesin–CTCF interaction in chromosome organization, we generated chromosome conformation capture (Hi-C) profiles of wild-type and $CTCF^{Y226A/F228A}$ cells. Wild-type HAP1 cells displayed clear loops connecting CTCF sites (Fig. 3a, Extended Data Fig. 4), however, Hi-C matrices of $CTCF^{Y226A/F228A}$ cells revealed a robust ablation of CTCF-anchored loops (Fig. 3a). By systematically scoring the number of loops, we found that in the $CTCF^{Y226A/F228A}$ mutant the vast majority of detectable loops across the genome were lost (from 2,756 in the wild-type cells, to 98 in the mutant cells) (Fig. 3b). An aggregate peak analysis, which quantifies the contact frequency of all the loops identified in wild-type cells, likewise showed a marked loss of these contacts (Fig. 3c, Extended Data Fig. 4d).

CTCF sites not only lie at the bases of CTCF-anchored loops, but also form the boundaries of TADs (Extended Data Fig. 4a). We then assessed the effect of the $CTCF^{Y226A/F228A}$ mutation on TADs, and found that these structures were—to a considerable degree—still present in $CTCF^{Y226A/F228A}$ cells but that they have less-clear edges (Fig. 3a). Aggregate TAD analysis further confirmed that TAD-like structures do exist in $CTCF^{Y226A/F228A}$ cells, but that these structures have less-clear boundaries (Extended Data Fig. 4b, c, e, f) and completely lack CTCF loops at their corners (Extended Data Fig. 4b, e). Our results therefore support the notion that in $CTCF^{Y226A/F228A}$ cells cohesin can form the loops along DNA that make up the contacts within TADs, but that cohesin is not stabilized at CTCF sites to allow for the formation or maintenance of CTCF-anchored loops.

Cohesin localization to CTCF sites

To assess whether the CTCF–CES interaction affects cohesin abundance at loop anchors, we performed chromatin immunoprecipitation with quantitative PCR (ChIP–qPCR) experiments. We selected CTCF sites at the base of loops (Fig. 4a, Extended Data Fig. 5a) and found that in the $CTCF^{Y226A/F228A}$ mutant the abundance of cohesin was reduced at the majority of these loci. By contrast, cohesin levels at a nearby locus

that did not contain a CTCF site were not affected (Fig. 4b, Extended Data Fig. 5b). CTCF binding to the corresponding CTCF sites was also largely unaffected (Extended Data Fig. 5c–e). We then assessed cohesin distribution genome-wide by chromatin immunoprecipitation with sequencing (ChIP–seq) and found that the *CTCF*$^{Y226A/F228A}$ mutation decreased cohesin localization to CTCF sites, but had little-to-no effect on cohesin localization at unrelated sites (Fig. 4c, d). Although cohesin levels at CTCF sites were reduced in *CTCF*$^{Y226A/F228A}$ cells, cohesin was—to a considerable degree—still present at CTCF sites. Our data therefore support a model in which CTCF influences cohesin in two ways: (i) it halts cohesin at CTCF sites and (ii) it stabilizes cohesin at the base of CTCF-anchored loops. The former function could be important for defining TAD boundaries. The binding of CTCF to the CES of cohesin could affect the latter function and may thereby prevent the disruption of CTCF-anchored loops.

To evaluate the consequences of the loss of CTCF-anchored loops on gene expression, we performed RNA-sequencing analyses. The *CTCF*$^{Y226A/F228A}$ mutation affected the expression of more than 2,000 genes. Although the number of genes that were upregulated was comparable to the number of genes that were downregulated, the most strongly affected genes were more frequently downregulated (Extended Data Fig. 7a). Thus, the interface of CTCF formed by Y226 and F228 and (by extension) cohesin–CTCF anchored loops are apparently key to correct expression of these genes. Despite this effect on gene expression and the loss of virtually all CTCF-anchored loops, cells that contain only this mutant form of CTCF are viable. CTCF has previously been shown to be essential for viability of mouse embryonic stem cells[11]. We therefore tested whether CTCF is essential for the viability of HAP1 cells, and found that CTCF depletion mediated by short interfering RNA was lethal to both control HAP1 cells and *CTCF*$^{Y226A/F228A}$ mutant cells (Extended Data Fig. 7b, c). Thus, CTCF has essential roles that are apparently independent of CES engagement and the formation of CTCF-anchored loops in these cells.

Identification of CES ligands

To investigate the prevalence of the CES-binding factors, we compiled an alignment of known cohesin partners and derived a regular expression motif (Extended Data Fig. 2e, f). We used this motif to query the human and budding yeast proteomes for proteins that contain similar binding motifs[30]. From the set of nuclear proteins that arose from this search, we were able to identify known cohesin regulators as well as several additional potential binding factors. We generated peptide arrays that bear these sequences and assayed the binding of SA2–SCC1, using an SA2(F371A)–SCC1 mutant complex as a negative control. We observed clear signal for the CTCF peptide that spans amino acids 222–231, which was abolished in the SA2(F371A)–SCC1 mutant (Extended Data Fig. 7d, e). A CTCF(Y226F) mutant showed approximately 1.5-fold reduced binding, apparently due to loss of the hydrogen bond between the hydroxyl group of CTCF Y226 and D326 of SA2 (Extended Data Table 2). Consistent with our pulldowns, the CTCF(Y226A), CTCF(F228A) and CTCF(Y226A/F228A) peptide variants did not retain SA2–SCC1. The WAPL peptides showed considerably weaker binding as compared to CTCF, and we could not detect binding for ligands such as SGO1 (Extended Data Table 1a, Methods). Robust binding was observed for MCM3 (a subunit of the replicative helicase), SYCP3 (a component of the synaptonemal complex), ZGPAT (a transcriptional repressor) and CENPU (a subunit of the inner kinetochore). Thus, the CES of SA2–SCC1 potentially facilitates cohesin regulation for a number of functionally divergent chromosomal processes.

Discussion

Our study reveals that CTCF binds to a CES on the SA2–SCC1 subcomplex of cohesin. The ablation of this interaction results in a near-complete loss of CTCF–cohesin anchored loops. Thus, CTCF does not simply present a passive barrier to cohesin-mediated loop extrusion, but specifically interacts with the CES to stabilize cohesin at these loci and to prevent loop disruption. Accordingly, impairment of the CTCF–CES interaction renders cohesin more dynamic (Fig. 2c, d).

SA2 and *SCC1*, as well as *CTCF*, are frequently mutated in a number of tumour types[31] and the mutations cluster in the CES (Extended Data Fig. 2b). Therefore, the dysregulation of chromatin looping may be causally related to carcinogenesis[32,33].

We envisage two possible scenarios for the formation of CTCF-anchored chromatin loops. In the first model (Fig. 4e), cohesin initiates loop enlargement at distal chromatin loci. These cohesin complexes remain dynamic because the cohesin release factor WAPL directly binds to cohesin by engaging the CES[2,3,29] and PDS5[12,27,34], and promotes the opening of cohesin rings at the SMC3–SCC1 interface[35–37]. Alternatively, loop enlargement commences at CTCF sites[38]. Cohesin then catalyses DNA looping at these sites because CTCF counteracts DNA release (Fig. 4f). These models are not necessarily mutually exclusive, as a cohesin complex that initiates looping in the former mode may well be converted into the latter upon encountering CTCF. As CTCF directly competes with WAPL for binding to the CES (Fig. 2a, b), we suggest that this interaction stabilizes chromatin loops.

We propose a model for how cohesin and CTCF co-associate on DNA (Fig. 4g). Our model indicates that cohesin engages CTCF only when approaching the N terminus of CTCF. Specifically, the 34-amino-acid flexible linker that connects the YXF motif to the first DNA-binding zinc finger of human CTCF is sufficiently long to allow SA2–SCC1 DNA binding towards the N, but not the C, terminus of CTCF (Fig. 4g), thus confirming previous mapping studies[39]. Stabilization of cohesin by engagement of the CTCF N terminus may explain why TAD boundaries arise preferentially when CTCF binding sites are convergently oriented[9,13–16,39,40]. If an individual cohesin complex anchors itself at the N terminus of CTCF, and then reels in DNA until it encounters a cohesin that is likewise reeling from the opposite CTCF site, this would bring together CTCF sites[38]. Loop formation by the related *Saccharomyces cerevisiae* condensin complex appears to involve a DNA anchoring function of its HEAT-repeat subunit Ycg1, a paralogue of human SA2 and *Saccharomyces cerevisiae* Scc3[41–43]. These different complexes may therefore use a similar anchoring principle to build loops and provide structure to genomes. As the CES interface is conserved between isoforms of SA, we anticipate that ligand binding will affect all cohesin variants in a similar manner. Similarly, this interface is also conserved through Scc3 in fungi eukaryotes, despite the absence of CTCF in these organisms. The CES therefore is likely to represent an ancient interaction hub on cohesin.

The observation that CTCF–CES interaction controls DNA looping indicates that this aspect of cohesin function can be regulated by an F/YXF motif containing cohesin ligands. A number of other genome regulatory factors contain F/YXF motifs—including SGO1 (Extended Data Fig. 7d, e), which protects centromeric chromatid cohesion by antagonizing WAPL binding to the CES of SA2–SCC1[2]. We therefore predict that a number of proteins that contain F/YXF motifs engage the CES and thereby modulate the ability of cohesin to catalyse genome folding in functionally divergent chromosomal processes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1910-z.

1. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164**, 1110–1121 (2016).

2. Hara, K. et al. Structure of cohesin subcomplex pinpoints direct shugoshin–Wapl antagonism in centromeric cohesion. *Nat. Struct. Mol. Biol.* **21**, 864–870 (2014).
3. Shintomi, K. & Hirano, T. Releasing cohesin from chromosome arms in early mitosis: opposing actions of Wapl–Pds5 and Sgo1. *Genes Dev.* **23**, 2224–2236 (2009).
4. Merkenschlager, M. & Nora, E. P. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genomics Hum. Genet.* **17**, 17–43 (2016).
5. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).
6. Yatskevich, S., Rhodes, J. & Nasmyth, K. Organization of chromosomal DNA by SMC complexes. *Annu. Rev. Genet.* **53**, 445–482 (2019).
7. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* **40**, 11202–11212 (2012).
8. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
9. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
10. Haarhuis, J. H. I. et al. The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693–707 (2017).
11. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 (2017).
12. Wutz, G. et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36**, 3573–3599 (2017).
13. Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
14. de Wit, E. et al. CTCF binding polarity determines chromatin looping. *Mol. Cell* **60**, 676–684 (2015).
15. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
16. Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
17. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 (2017).
18. Gassler, J. et al. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600–3618 (2017).
19. Rubio, E. D. et al. CTCF physically links cohesin to chromatin. *Proc. Natl Acad. Sci. USA* **105**, 8309–8314 (2008).
20. Xiao, T., Wallace, J. & Felsenfeld, G. Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol. Cell. Biol.* **31**, 2174–2183 (2011).
21. Pezzi, N. et al. *STAG3*, a novel gene encoding a protein involved in meiotic chromosome pairing and location of *STAG3*-related genes flanking the Williams-Beuren syndrome deletion. *FASEB J.* **14**, 581–592 (2000).
22. Orgil, O. et al. A conserved domain in the Scc3 subunit of cohesin mediates the interaction with both Mcd1 and the cohesin loader complex. *PLoS Genet.* **11**, e1005036 (2015).
23. Roig, M. B. et al. Structure and function of cohesin's Scc3/SA regulatory subunit. *FEBS Lett.* **588**, 3692–3702 (2014).
24. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
25. Beckouët, F. et al. Releasing activity disengages cohesin's Smc3/Scc1 interface in a process blocked by acetylation. *Mol. Cell* **61**, 563–574 (2016).
26. Gandhi, R., Gillespie, P. J. & Hirano, T. Human Wapl is a cohesin-binding protein that promotes sister-chromatid resolution in mitotic prophase. *Curr. Biol.* **16**, 2406–2417 (2006).
27. Kueng, S. et al. Wapl controls the dynamic association of cohesin with chromatin. *Cell* **127**, 955–967 (2006).
28. Liu, H., Rankin, S. & Yu, H. Phosphorylation-enabled binding of SGO1–PP2A to cohesin protects sororin and centromeric cohesion during mitosis. *Nat. Cell Biol.* **15**, 40–49 (2013).
29. Ouyang, Z. et al. Structure of the human cohesin inhibitor Wapl. *Proc. Natl Acad. Sci. USA* **110**, 11355–11360 (2013).
30. Krystkowiak, I. & Davey, N. E. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Res.* **45**, W464–W469 (2017).
31. Lawrence, M. S et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
32. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in *IDH* mutant gliomas. *Nature* **529**, 110–114 (2016).
33. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).
34. Ouyang, Z., Zheng, G., Tomchick, D. R., Luo, X. & Yu, H. Structural basis and IP6 requirement for Pds5-dependent cohesin dynamics. *Mol. Cell* **62**, 248–259 (2016).
35. Chan, K. L. et al. Cohesin's DNA exit gate is distinct from its entrance gate and is regulated by acetylation. *Cell* **150**, 961–974 (2012).
36. Buheitel, J. & Stemmann, O. Prophase pathway-dependent removal of cohesin from human chromosomes requires opening of the Smc3–Scc1 gate. *EMBO J.* **32**, 666–676 (2013).
37. Eichinger, C. S., Kurze, A., Oliveira, R. A. & Nasmyth, K. Disengaging the Smc3/kleisin interface releases cohesin from *Drosophila* chromosomes during interphase and mitosis. *EMBO J.* **32**, 656–665 (2013).
38. Sedeño Cacciatore, Á. & Rowland, B. D. Loop formation by SMC complexes: turning heads, bending elbows, and fixed anchors. *Curr. Opin. Genet. Dev.* **55**, 11–18 (2019).
39. Tang, Z. et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
40. Nagy, G. et al. Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics* **17**, 637 (2016).
41. Kschonsak, M. et al. Structural basis for a safety-belt mechanism that anchors condensin to chromosomes. *Cell* **171**, 588–600 (2017).
42. Ganji, M. et al. Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018).
43. Li, Y. et al. Structural basis for Scc3-dependent cohesin recruitment to chromatin. *eLife* **7**, e38356 (2018).

Methods

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

Constructs, protein expression and purification

The human SA2 fragment amino acids 80–1060, cloned into pGEX-6P and codon-optimized for expression in *Escherichia coli*, was obtained from H. Yu. The construct encodes an N-terminal GST tag and C-terminal SA2 separated by a PreScission protease cleavage site. A plasmid encoding SCC1 was obtained from J.-M. Peters. SA2 was co-expressed with an N-terminally 6×His-tagged fragment of SCC1 spanning residues 281–420 cloned into the NcoI–NotI sites of a pACYCDuet-1 vector (Merck Millipore). CTCF constructs were cloned into the BamHI and NotI sites of pGEX-6P1. Mutagenesis was performed using a QuikChange Lightning site-directed mutagenesis kit (Agilent). All the proteins were expressed in *E. coli* BL21(*DE3*) by autoinduction[44]. Cells were grown at 37 °C until an optical density at 600 nm ($OD_{600\,nm}$) = 0.6 and then shifted to 18 °C for 16 h. Cells were collected with a JLA-8.1 rotor (Beckman) and washed once with ice-cold PBS buffer. Pellets were resuspended in buffer 1 (40 mM TRIS, pH 7.5, 500 mM NaCl and 0.5 mM TCEP), lysed using a microfluidizer (Microfluidics) and centrifuged at 4 °C for 1 h at 15,000 rpm using JA-20/14 rotors (Beckman).

The GST- and His-tagged SA2–SCC1 complex was applied to Co^{2+} conjugated IMAC sepharose resin (GE Healthcare) using a Minipuls3 peristaltic pump (Gilson), washed with buffer 1 supplemented with 20 mM imidazole and eluted using buffer 1 supplemented with 300 mM imidazole. Co^{2+} eluate was then bound to Glutathione Sepharose 4 Fast Flow resin (GE Healthcare) using a Minipuls3 peristaltic pump (Gilson), washed with buffer 1 and eluted by adding 10 mM reduced L-Glutathione (Sigma-Aldrich) into buffer 1. The GST tag was cleaved by PreScission protease (EMBL core facilities) during overnight incubation at 4 °C. Cleaved protein was concentrated using an Amicon Ultra -15 concentrator (Millipore) and applied to a MonoQ 5/50 GL column (GE Healthcare) in buffer 2 (40 mM TRIS, pH7.5, 150 mM NaCl and 0.5mM TCEP) and eluted via a linear gradient of buffer 2 containing 1 M NaCl and further purification using a HiLoad 16/60 Superdex 200 prep-grade column (GE Healthcare) in buffer 3 (20 mM TRIS, pH7.7, 300 mM NaCl and 5 mM TCEP). The final purified proteins were concentrated using an Amicon Ultra -15 concentrator (Millipore) and flash-frozen in liquid N_2 for storage at −80 °C.

Crystallization and structure determination

Crystals of SA2(80–1060) in complex with SCC1 amino acids 281–420 (otherwise denoted the SA2–SCC1 complex) were grown by hanging-drop vapour diffusion at 20 °C by mixing equal volumes of protein at 8 mg ml^{-1} and crystallization solution containing 0.06 M Morpheus Divalents mix, 0.1 M Morpheus buffer system 2, 48% (v/v) Morpheus EOD_P8K (Molecular Dimensions). Crystals were soaked for 24–48 h with a peptide (obtained from peptid.de) including amino acid residues 222–231 of CTCF (Uniprot ID Q8NI51; DVSVYDFEEE). Crystals were cryo-protected by adding 15% glycerol to the well solution and flash-frozen in liquid nitrogen.

Diffraction data for all crystals were collected at 100 K at an X-ray wavelength of 0.966 Å at beamline ID30A-1/MASSIF-1[45] of the European Synchrotron Radiation Facility, with a Pilatus3 2M detector, using automatic protocols for the location and optimal centring of crystals[46]. The beam diameter was selected automatically to match the crystal volume of highest homogeneous quality[47]. Data were processed with XDS[48] and imported into CCP4 format using AIMLESS[49].

The structure was determined by molecular replacement using Phaser[50]. A final model was produced by iterative rounds of manual model-building in Coot[51] and refinement using PHENIX[52]. The CTCF-containing model was refined to a resolution of 2.7 Å with an R_{work} and an R_{free} of 25% and 27%, respectively (Extended Data Table 1b). Analysis by MolProbity[53] showed that there are no residues in disallowed regions of the Ramachandran plot and the all-atom clash score was 7.2. The model shown in Fig. 4f was generated by superposition on DNA of SA2–SCC1–CTCF (RCSB Protein Data Bank code (PDB) 6QNX) with DNA-bound *Saccharomyces cerevisiae* SCC3–SCC1 (PDB 6H8Q)[43] and a composite model of DNA-bound CTCF zinc fingers assembled from PDB 5YEF and PDB 5YEL[54].

GST pulldowns and peptide arrays

For GST pulldowns, 10 μM GST-tagged CTCF constructs and 2.5 μM SA2–SCC1 were mixed in 50 μl buffer 4 (20 mM TRIS, pH7.7, 300 mM NaCl and 0.5 mM TCEP) + 0.1% Tween-20 containing 25 μl of a 50% slurry of GST sepharose beads per reaction. For WAPL and SGO1 competition assays, 2.5 μM GST-tagged CTCF(86–267) was incubated with 1 μM of SA2–SCC1 and increasing concentrations of WAPL(1–600) or a SGO1 phosphorylated at T346 peptide spanning amino acids 331–349 (molar ratios are indicated in each figure), under reaction conditions that were otherwise identical to GST pulldowns. Reactions were incubated at for 1 h at 4 °C. Twenty-five microlitres of the reaction were withdrawn as the reaction input and the remainder was washed 5 times with 500 μl of buffer 4 + 0.1% Tween-20. Samples were boiled in 1× SDS sample loading buffer (NEB) for 5 min to obtain the bound fraction, followed by SDS–PAGE analysis.

Isothermal calorimetry (ITC) was performed using a MicroCal iTC 200 (Malvern Panalytical) at 25 °C. SA2–SCC1 and the CTCF, SGO1 and WAPL peptide ligands were dialysed overnight at 4 °C against 20 mM TRIS, pH7.7, 150 mM NaCl, 0.5 mM TCEP. For each titration, 300 μl of 50 μM SA2–SCC1 was added to the calorimeter cell. The concentration of peptides was adjusted to 500 μM and injected into the sample cell as 16× 2.5-μl syringe fractions. Results were analysed and displayed using the Origin 7.0 software package supplied with the instrument. Data were analysed using the one-site binding model.

Peptide arrays, with an area of 3 cm^2, were obtained from R. Volkmer (http://immunologie.charite.de). Arrays were washed with 100% ethanol for 5 min on a shaker at 21 °C, followed by 3 washes, for a total of 10 min in TBS-T buffer (50 mM Tris pH7.5, 150 mM NaCl and 0.05% Tween-20). For the blocking step, arrays were incubated in 1× blocking buffer (Sigma B6429) for 3 h at 21 °C, followed by 3 washes in TBS-T for a total of 10 min. SA2–SCC1 and SA2(F371A)–SCC1 were added to 1× blocking buffer at a final concentration of 1.2 μM and incubated with the array overnight at 4 °C under gentle agitation. The membrane was washed 3 times (1× 30 s, and then 2× 5 min) at 21 °C. The anti 6× anti-poly His–HRP antibody (Sigma A-7058) was diluted 1:2,000 in 1× blocking buffer and incubated with the arrays for 1 h at 21 °C. The array was washed 3 times (1× 30 s, and then 2× 5 min) and developed by addition of 3,3′-diaminobenzidine (Sigma D4293) for 1 min followed by quenching in deionized H_2O. To measure non-specific binding of the anti-6×His antibody, all steps were identical except that no SA2–SCC1 protein solution was added to 1× blocking buffer during the overnight-incubation step. Arrays were imaged with a BioRad Gel Doc XR+ documentation system. Spot intensities were measured using ImageJ 1.52k. Three independent experiments were done and the apparent dissociation constants determined by normalization with ITC data from CTCF(222–231) (Extended Data Table 1a).

Genome editing and cell culture

Cells were cultured in Iscove's modified Dulbecco's medium supplemented with 10% FCS (Clontech), 1% penicillin–streptomycin (Invitrogen) and 1% UltraGlutamin (Lonza). The guide (g)RNA targeting exon 1 of *CTCF* was designed and annealed into pX330 (primer, 5′- CGATTTTGAGGAAGAACAGC-3′). To modify the targeted locus, we cotransfected a 120-base-pair repair oligonucleotide containing the desired mutation and a silent mutation (repair oligonucleotide: 5′-CCAAAAAGAGCAAACTGCGTTATACAGAGGAGGGCAAAGATGTAGAT

GTGTCTGTCGCCGATGCTGAAGAAGAACAGCAGGAGGGTCTGCTATCAGAGGTTAATGCAGAGAAAGTGGTTG-3′). pBabePuro was cotransfected in a 10:1 ratio to the pX330. Transfected clones were selected using 2 µg/µl puromycin for 2 days. Colonies were picked when they were clearly visible, gDNA of clones was isolated and mutations were validated by Sanger sequencing.

To target the C terminus of SCC1, a gRNA (primer: 5′-CCAAGGTTCCATATTATATA-3′) was cloned into px459 V2.0 (Addgene plasmid no 62988). The SCC1–Halo tag HR template was a gift from J. Rhodes[55]. SCC1–Halo cell lines were generated by cotransfection of pX459 and the SCC1–Halo tag HR vector using FuGENE HD Transfection Reagent. Cells were selected with puromycin (2 µg/ml) for 2 days. Colonies were picked when they were clearly visible and validated using western blot analysis and immunofluorescence.

Antibodies

The following antibodies were used for western blots: SMC1 (A300-055A, Bethyl), CTCF (07-729, Millipore and ab128873, Abcam), HSP90 (F-8, Santa Cruz), SCC1 (05-908, Millipore), tubulin (T5168, Sigma) and H4 (05-858, Millipore). All primary antibodies were used at a 1:1,000 dilution, with the exception of HSP90 and tubulin (1:10,000). Secondary antibodies for western blot analysis were used in a 1:2,000 dilution: goat anti-rabbit-PO and goat anti-mouse-PO (DAKO). For ChIP–seq, we used the following antibodies: SCC1 (ab992, Abcam), CTCF (3418S, Cell Signaling) and IgG (I5006, Sigma-Aldrich).

FRAP

Cells were grown on LabTekII-chambered cover glass (Thermo Scientific Nunc). Two days before imaging, cells were transfected with DNA helicase B fragment fused with near-infrared fluorescent protein (DHB–iRFP) using FuGENE HD Transfection Reagent. Before imaging, cells were incubated with 300 nM fluorescent Halo tag ligand JF585 for 30 min. Cells were washed 3 times with normal medium and incubated for 1 h to allow exit of excess of ligand. Medium was replaced twice more with prewarmed Leibovitz L-15 medium (Invitrogen). Live-cell imaging was performed on a Leica SP5 confocal microscope with a 63× 1.2 NA water objective using the LAS-AF FRAP-Wizard. Before bleaching, five images were taken. Half of the nucleus of G1 cells was photobleached using 6 pulses of 100% transmission of a 561-nm laser. Subsequently, 600 frames were taken every 2 s. Fluorescence intensity was measured in the bleached and unbleached area by user-defined regions using ImageJ v.1.52q, and adjusted by hand for nucleus movement. Measurements were corrected for photobleaching by monitoring a nonbleached cell. Recovery was quantified by calculating the difference in intensity in the bleached and unbleached regions after background correction. Nondiffusive SCC1–Halo (Extended Data Fig. 3f) was quantified by the relative loss in fluorescence intensity in the unbleached region between the first frame postbleaching and five frames prebleaching.

Colony-formation assay

Cells were seeded at equal density and transfected with short interfering (si)RNAs targeting either no oligonucleotide, luciferase, *CTCF* or *SMC1A*. All siRNAs were ON-TARGETplus SMARTpools manufactured by Dharmacon. Transfection was repeated after 3 days, and after an additional 4 days samples were fixed for 10 min with 96% methanol and stained with 0.25% crystal violet. Cells treated by the same protocol were collected for western blot analysis; samples were collected two days before fixation to have enough cells for western blot analysis.

Chromatin fractionation

For the chromatin fractionation experiment shown in Extended Data Fig. 3e, 50 million cells per cell line were collected and fractionation was performed using Subcellular Protein Fractionation Kit for Cultured Cells (78840, Thermo Fisher Scientific) according to the manufacturer's protocol, with minor changes. The pellet was washed twice after centrifugation. Western blots were performed as previously described[10].

Hi-C

Samples for Hi-C were prepared as previously described[10]. Raw sequence data were mapped and processed using HiC-Pro v.2.9[56] with hg19 as reference. Statistics on the number of valid pairs and percentage of *cis* contacts are summarized in Extended Data Table 3b, c. Replicates 1 and 2 are highly similar, with a reproducibility >0.98 as assessed by HiCRep v.1.8.0[57], and were subsequently combined into one Hi-C dataset. The valid pair files generated by HiC-Pro were used to create juicebox ready files using juicebox-pre (juicer tools v.0.7.5)[58]. For visualization, contact matrices were ICE-normalized[59] and counts were normalized for 100 million contacts per sample.

Loops were then called with HICCUPS v.1.11.09[15] at 5-kb, 10-kb and 25-kb resolution. To visualize the genome-wide effect of the introduced *CTCF* mutations in loops, we performed aggregate peak analysis[15] as implemented in GENOVA v.0.9.8 (https://github.com/robinweide/GENOVA), using loops that had previously been defined in wild-type HAP1 cells[10]. In brief, for a set of loop coordinates a square submatrix is selected such that it is centred on the corresponding coordinates, with a 100-kb flanking region upstream and downstream. These submatrices are then averaged to obtain a mean contact map for these locations.

Similar to the aggregated peak analysis, aggregate TAD analysis was done to visualize how TAD structures are affected by the *CTCF* mutations. For this analysis, we used TADs that had previously been defined for wild-type HAP1 cells[10]. In brief, these TADs were called using HiCseg[60] on 10-kb matrices as input, Poisson distribution, the extended diagonal model and a maximum number of change points of 50. To compensate for TADs of different sizes, the selected regions are resized before averaging the contact maps. These regions are comprised of the TAD itself and a flanking region of half its size. We calculated the insulation score as previously described[61]. The insulation score was computed using the implementation of GENOVA, with a rolling window size of 25 kb. The insulation score was then aligned to TAD borders to create heat maps.

For Extended Data Fig. 6, the compartment-score was calculated as previously described[62]. In brief, the compartment score is computed per chromosome arm by obtaining the first eigenvector of the observed over expected matrix, minus 1. Then, this eigenvector is multiplied by the square root of its eigenvalue to obtain the compartment score. To correctly orient the scores so that positive values correspond to compartment-A regions, we used the correlation of the compartment score to H3K4me1 peaks in wild-type cells (J.H.I.H. et al., manuscript in preparation).

For Extended Data Fig. 6e, we compared the effect of *CTCF*$^{Y226A/F228A}$ mutation on genome organization to that of CTCF and cohesin degradation[12]. Raw Hi-C data from Gene Expression Omnibus (GEO) accession GSE102884 were converted to HiC-Pro format and ICE-normalized. Relative contact probability profiles were generated using GENOVA.

ChIP–seq

Samples for ChIP–seq were prepared and sequenced as previously described[10], with minor changes. The DNA was sheared using Biorupter Pico (Diagenode), 5 cycles of 15-s on and 90-s off. Reads were first trimmed using TrimGalore v.0.6.0[63], then mapped to hg19 using Bowtie2 v.2.3.4[64] with default settings. Bigwig files were generated with DeepTools v.3.1.3[65] with the following settings: minimum mapping quality of 15, bin length of 10 bp, extending reads to 200 bp and reads per kilobase per million reads normalization.

Peaks were called for all samples using MACS2 v.2.1.1[66] with default options. Overlaps between the sets of identified peaks across samples were obtained using BEDtools v.2.25.0[67]. Heat maps were generated using DeepTools[65] for the different sets of peaks identified in the wild-type cell line, excluding those overlapping blacklisted regions of the genome[68].

CTCF sites shown in Hi-C contact matrices were obtained from a previous publication[10]. In brief, these sites were generated by intersecting CTCF peaks with CTCF motifs from JASPAR CORE 2014[69], using FIMO[70] to annotate their motif orientation.

ChIP–qPCR

ChIP–qPCR analysis was performed to assess SCC1 and CTCF abundance at specific genomic loci. SCC1 ChIP was performed three times, and on each ChIP three qPCRs were performed in duplicate. A representative qPCR analysis of each ChIP was used for quantification. For CTCF and IgG, two ChIPs were performed in duplicate. Reactions were performed using SYBR No-Rox Mix 2× (BIOLINE) and run on a LightCycler 480 II (Roche). C_t values were determined for input and ChIP samples, and subsequently the ΔC_t value was converted into a percentage of input. The primers are listed in Extended Data Table 3a.

RNA sequencing

Samples for RNA sequencing were prepared and sequenced as previously described[10]. Reads were aligned to hg19 using TopHat v.2.1.1[71] and later counted with HTSeq v.0.11.1[72] using Gencode v.19 gene-build as reference. Differentially expressed genes were identified with DESeq2 v.1.18.135[73], with an adjusted *P* value threshold of 0.05 and considering only protein-coding genes.

Reporting summary

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

Data availability

Coordinates are available from the PDB under accession number 6QNX for the SA2–SCC1–CTCF complex. The generated Hi-C, RNA sequencing and ChIP–seq data have been deposited in GEO, accession number GSE126637. Any other relevant data are available from the corresponding authors upon reasonable request.

44. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
45. Bowler, M. W. et al. MASSIF-1: a beamline dedicated to the fully automatic characterization and data collection from crystals of biological macromolecules. *J. Synchrotron Radiat.* **22**, 1540–1547 (2015).
46. Svensson, O., Malbet-Monaco, S., Popov, A., Nurizzo, D. & Bowler, M. W. Fully automatic characterization and data collection from crystals of biological macromolecules. *Acta Crystallogr. D* **71**, 1757–1767 (2015).
47. Svensson, O., Gilski, M., Nurizzo, D. & Bowler, M. W. Multi-position data collection and dynamic beam sizing: recent improvements to the automatic data-collection algorithms on MASSIF-1. *Acta Crystallogr. D* **74**, 433–440 (2018).
48. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66**, 133–144 (2010).
49. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
50. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
51. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
52. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
53. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
54. Yin, M. et al. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res.* **27**, 1365–1377 (2017).
55. Rhodes, J. D. P. et al. Cohesin can remain associated with chromosomes during DNA replication. *Cell Rep.* **20**, 2749–2755 (2017).
56. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
57. Yang, T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).
58. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
59. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
60. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386–i392 (2014).
61. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
62. Flyamer, I. M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).
63. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
65. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
66. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP–seq enrichment using MACS. *Nat. Protocols* **7**, 1728–1740 (2012).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
68. Amemiya, H.M., Kundaje, A., & Boyle, A.P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep.* **9**, 9354 (2019).
69. Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
70. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
71. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
72. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
73. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
74. Landgraf, C. et al. Protein interaction networks by proteome peptide scanning. *PLoS Biol.* **2**, e14 (2004).

Acknowledgements This work was funded by EMBL. J.H.I.H., Á.S.C. and B.D.R. were supported by an ERC CoG (772471 'CohesinLooping'), M.S.v.R. by the Boehringer Ingelheim Fonds and H.T. and E.d.W. by an ERC StG (637587 'HAP-PHEN'). H.T. and E.d.W. are part of the Oncode Institute, which is partly financed by the Dutch Cancer Society. We thank the staff at the ESRF beamline Massif-1; T. Gibson for advice concerning short linear motifs; J. Rhodes and K. Nasmyth for reagents and advice on Halo tagging; R. van der Weide for advice and bioinformatic analyses; and R. Kerkhoven and the NKI Genomics Core Facility for sequencing.

Author contributions Y.L. and K.W.M. initiated the project and proposed the CES motif. Y.L. performed biochemical studies and structural analyses with support from K.W.M. J.H.I.H., R.O., M.S.v.R., L.W. and H.T. performed wet-laboratory cell-based experiments and Á.S.C. performed bioinformatic analyses. K.W.M., E.d.W., B.D.R. and D.P. provided supervision. Y.L., K.W.M., B.D.R. and D.P. were involved in conceptualization, project administration and wrote the original and revised draft with input from all authors.

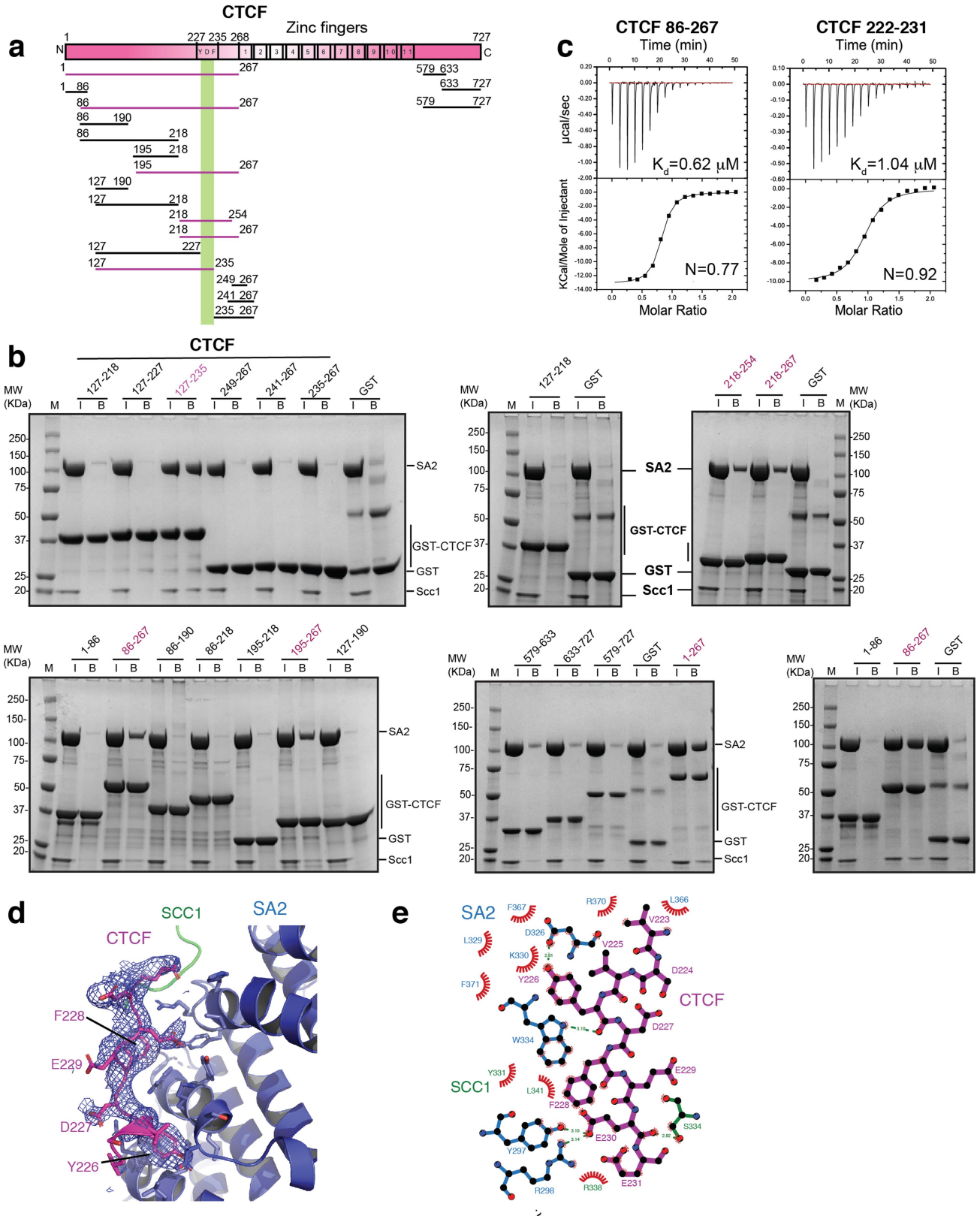
Competing interests The authors declare no competing interests.

Additional information

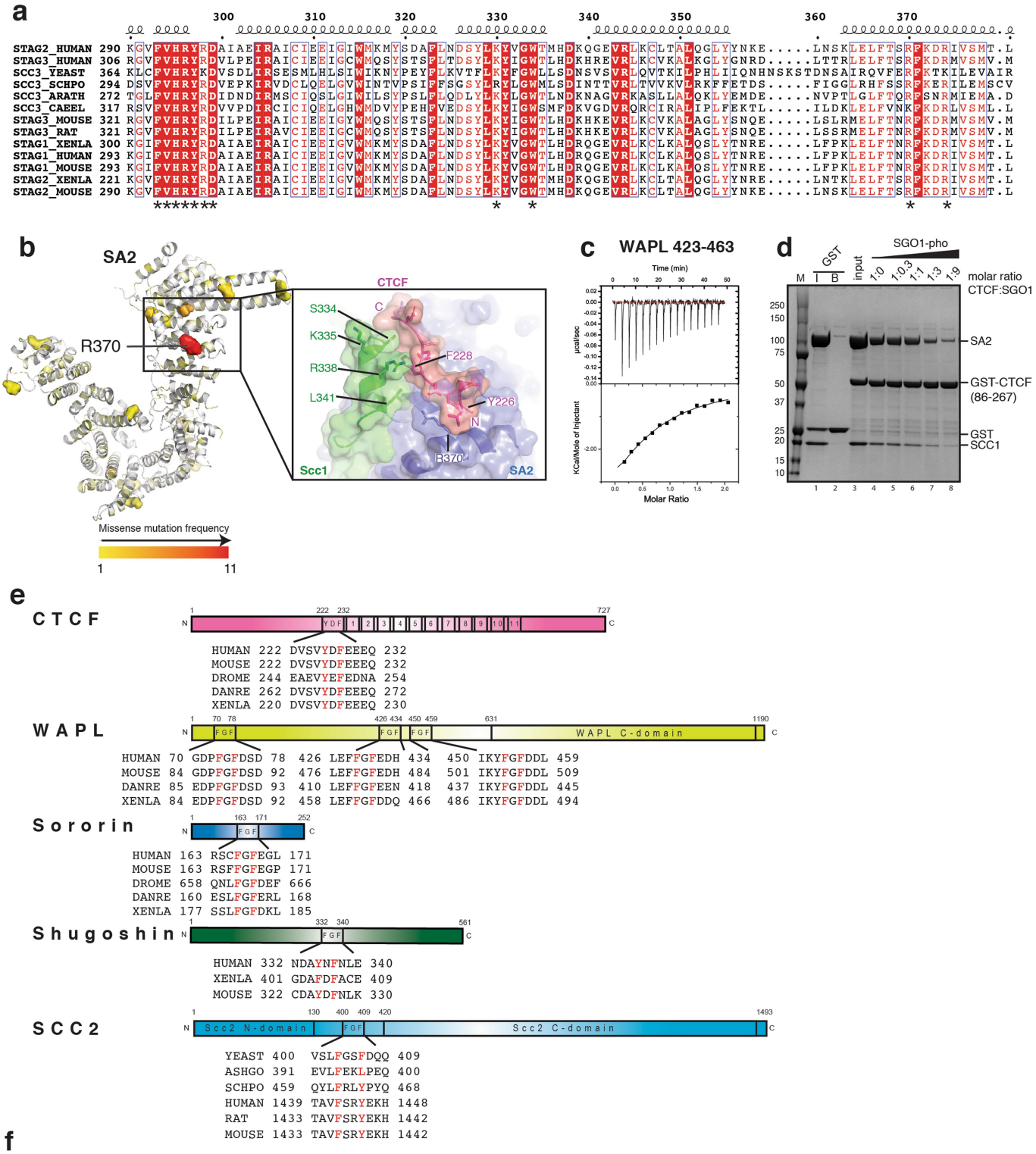
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1910-z.

Correspondence and requests for materials should be addressed to K.W.M., E.d.W., B.D.R. or D.P.

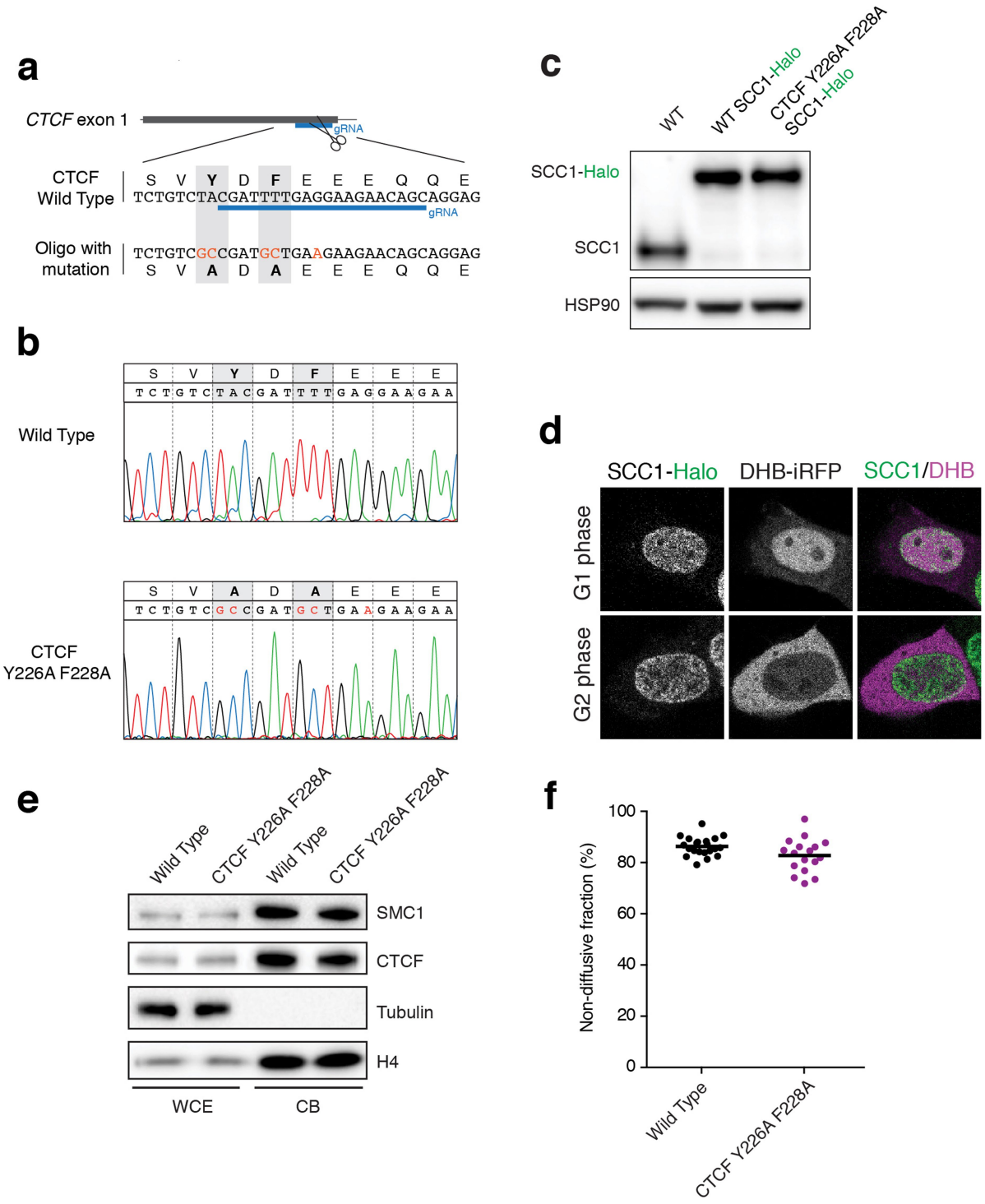
Peer review information *Nature* thanks Victor Corces, Karl-Peter Hopfner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



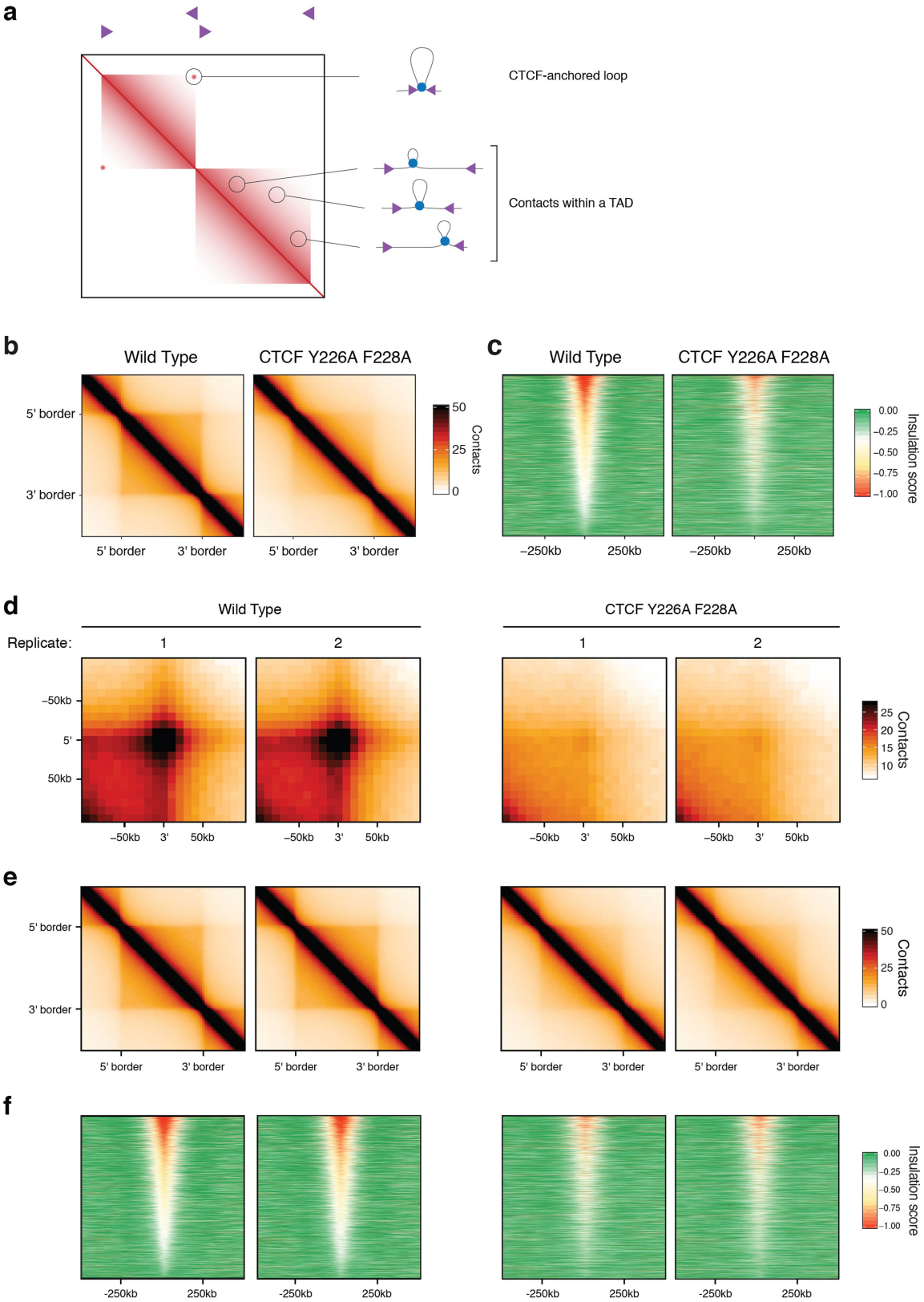
Extended Data Fig. 1 | Biochemical analysis of CTCF binding to SA2–SCC1. **a**, Domain architecture of CTCF. CTCF fragments tested for SA2–SCC1 binding by GST pulldown analysis are indicated. The region that retains SA2–SCC1 is highlighted in magenta. **b**, Summary data showing results of GST pulldowns. The input and the bound fractions were analysed by SDS–PAGE. CTCF fragments that bind SA2–SCC1 are shown in magenta. The experiment was repeated once. **c**, ITC curves. The binding stoichiometry (*N*) and dissociation constants (K_d) are indicated. The experiment was repeated three times, with consistent results. **d**, $F_o - F_c$ omit electron-density Fourier map contoured at 3σ. **e**, LIGPLOT representation of the interaction between the CTCF peptide and SA2–SCC1. The CTCF peptide is shown in magenta, SA2 in blue and SCC1 in green bonds.



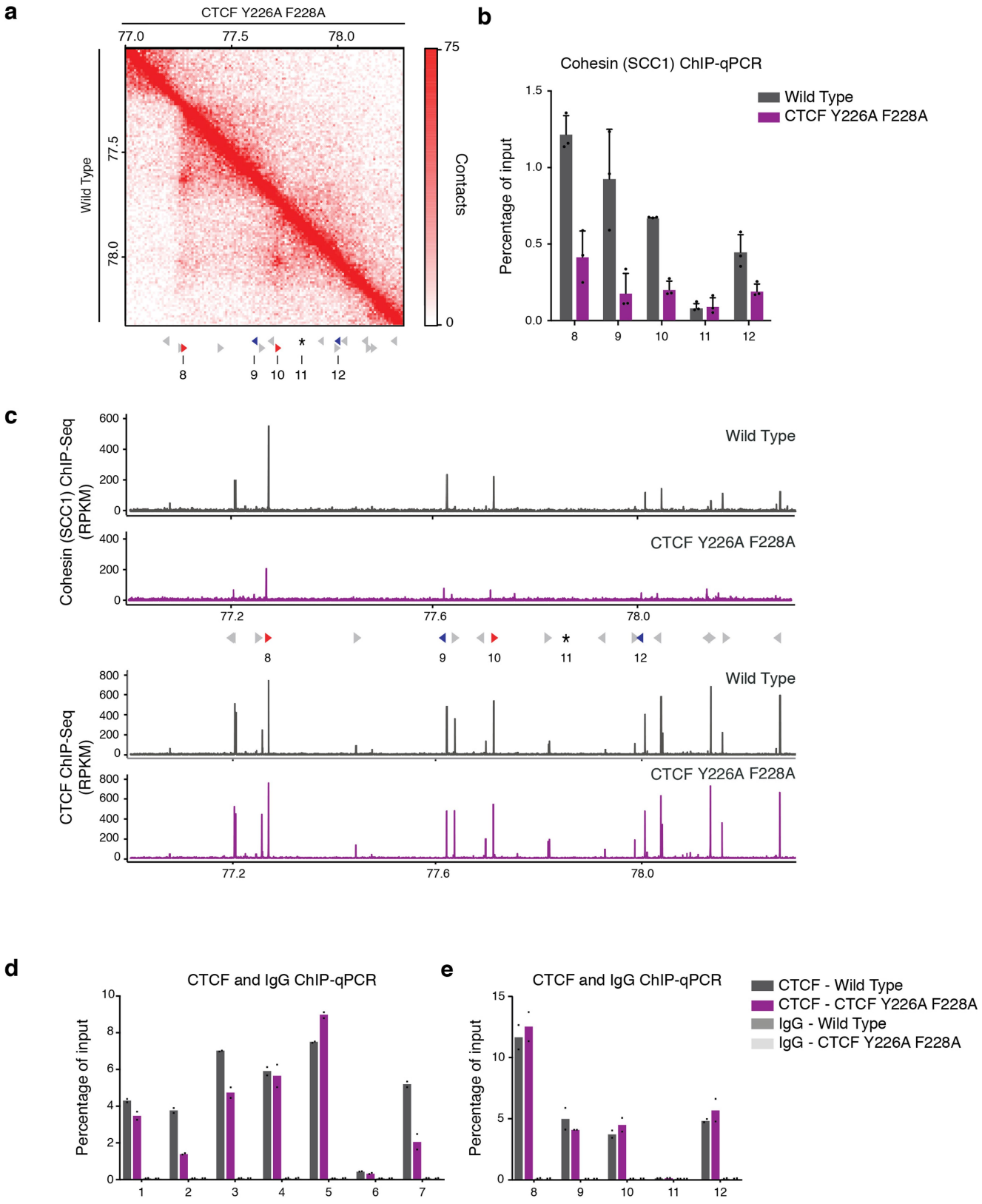
Extended Data Fig. 2 | Analysis of the SA2–SCC1–CTCF structure. a, Multiple sequence alignment of SA2 (here denoted STAG2) orthologues and paralogues. *Key amino-acid residues that engage CTCF. **b**, Missense mutation frequencies plotted onto the SA2 structure. R370 (a hotspot in SA2) is indicated. The inset shows an overview of the mutation hotspots R370 of SA2), Y226 and F228 of CTCF, and S334, K335, R338 and L341 of SCC). **c**, ITC progress curves of binding between WAPL(423–463) and SA2–SCC1. **d**, Competition between SGO1 and CTCF for SA2–SCC1 binding. SA2–SCC1 was incubated with GST–CTCF(86–267). Increasing amounts (lanes 4–8) (molar ratios are indicated) of the SGO1 phosphorylated at T346 peptide (spanning residues 331–349) were added and the input and the bound fraction analysed by SDS–PAGE. The experiment was repeated twice. One representative example is shown. **e**, Domain architecture and sequence alignments of cohesin regulators that contain F/YXF motifs. Putative CES-interacting residues are highlighted in red. **f**, Regular expression motif used to query the human and yeast proteomes for factors containing F/YXF motifs. Regular expression syntax: letters denote a specific amino acid; square brackets denote a subset of allowed amino acids; curly brackets denote length variability.



Extended Data Fig. 3 | Generation of *CTCF*$^{Y226A/F228A}$ cells. a, Schematic of CRISPR–Cas9-based generation of *CTCF*$^{Y226A/F228A}$ cells. The guide targets cleavage of exon 1 of the *CTCF* gene. The repair oligonucleotide renders the gene noncleavable by Cas9, and simultaneously introduces mutations in the codons that encode Y226 and F228. **b**, The *CTCF*$^{Y226A/F228A}$ mutation was confirmed by Sanger sequencing, including a silent mutation at position 229. **c**, Western blot depicting Halo-tagged SCC1 in wild-type and *CTCF*$^{Y226A/F228A}$ cells. The parental wild-type cells are included as a control. This experiment was performed once. **d**, Representative images of cells in G1 and G2, as indicated by their nuclear and cytoplasmic localization of DHB–iRFP, respectively. **e**, Chromatin-bound levels of CTCF and SMC1 analysed by western blot. Histone H4 is used as a control for the chromatin fraction. The *CTCF*$^{Y226A/F228A}$ mutation does not evidently affect overall CTCF and cohesin levels on chromatin. WCE, whole-cell extract; CB, chromatin-bound fraction. This experiment was performed twice with similar results. **f**, Relative SCC1–Halo fluorescence intensity quantified in the unbleached area directly after photobleaching, as a proxy for the chromatin-bound fraction of SCC1. This nondiffusive fraction is not evidently affected by the *CTCF*$^{Y226A/F228A}$ mutation. Individual cells of three independent experiments are plotted as dots and their mean is indicated (21 wild-type cells and 17 *CTCF*$^{Y226A/F228A}$ cells were scored).

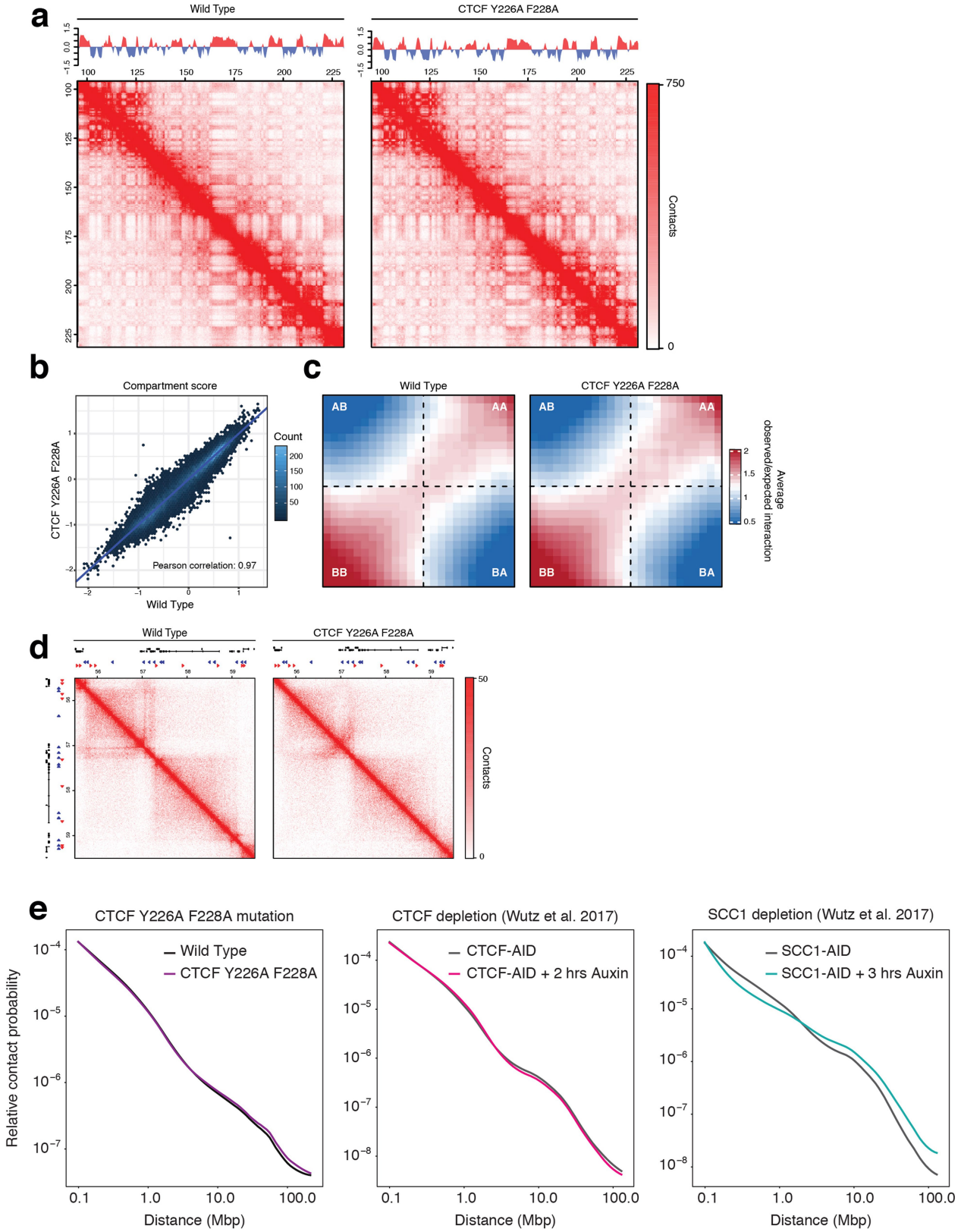


Extended Data Fig. 4 | TAD analyses and Hi-C replicates. a, Schematic of a Hi-C matrix displaying DNA–DNA contacts across a genomic region that includes two TADs. TADs in general are flanked by inwards-pointing CTCF sites (magenta arrows). Signal close to the diagonal line reflects short-range contacts, and contacts that span longer distances are found further away from the diagonal. The contacts within a TAD are formed by cohesin complexes (blue circles). Cohesin builds loops that it can enlarge until it encounters CTCF. Some TADs are enriched for contacts between the two CTCF sites that lie at their boundaries. These contacts are referred to as CTCF-anchored loops. **b**, Aggregate TAD analysis depicting the average contact frequency across TADs defined in wild-type cells. **c**, Heat map of the insulation score[61] at TAD borders, as defined for wild-type cells. **d**, Aggregate peak analysis as in Fig. 3c, using two independent library preparations per genotype. **e**, Aggregate TAD analysis for wild-type and *CTCF*$^{Y226A/F228A}$ cells as in **b**. **f**, Heat map of insulation scores at TAD borders for wild-type and *CTCF*$^{Y226A/F228A}$ cells as in **c**.

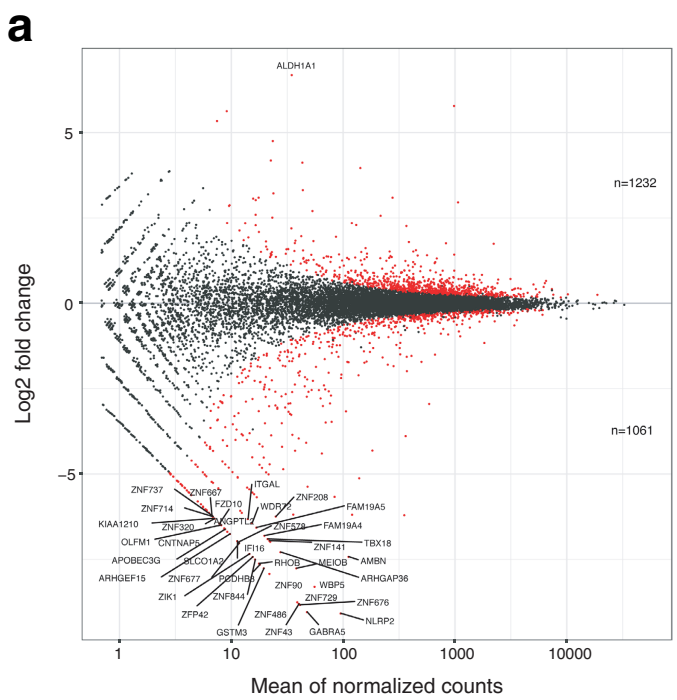


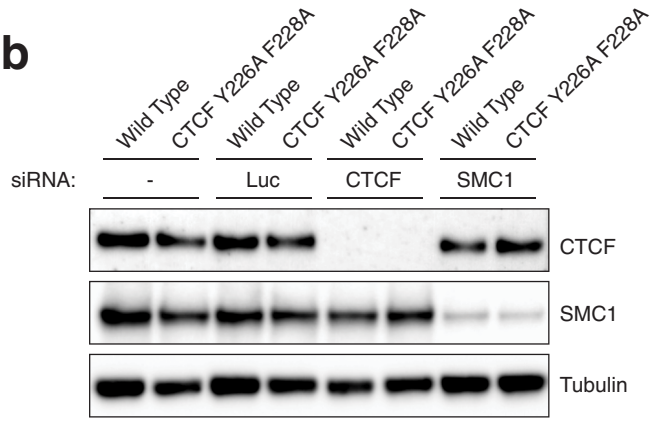
Extended Data Fig. 5 | See next page for caption.

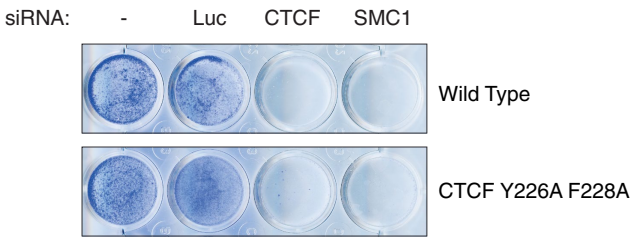
Extended Data Fig. 5 | *CTCF*$^{Y226A/F228A}$ mutation has little effect on CTCF levels at CTCF sites. a, Hi-C contact matrix of region chromosome 16: 77000000–78300000 at 10-kb resolution for the wild-type cell line (bottom triangles) and the *CTCF*$^{Y226A/F228A}$ cell line (top triangles). CTCF sites are depicted below; those selected for qPCR are shown in colour. Red triangles indicate sites with a forward motif and blue triangles indicate sites with a reverse motif. The numbers underneath indicate the qPCR primer pairs shown in **b**. Primer pair 11 (indicated with *) is at a locus devoid of SCC1 and CTCF. **b**, ChIP–qPCR analysis of SCC1 (cohesin) enrichment at the aforementioned CTCF sites and control locus (*) in wild-type and *CTCF*$^{Y226A/F228A}$ cells. The mean of three independent ChIP experiments is shown with the s.d. **c**, ChIP–seq tracks for SCC1 and CTCF at region chromosome 16: 77000000–78300000 in wild-type and *CTCF*$^{Y226A/F228A}$ cells. The loci used for ChIP–qPCR analysis are indicated below the SCC1 ChIP–seq tracks. RPKM, reads per kilobase per million reads. **d**, ChIP–qPCR analysis of CTCF abundance at loci 1–7, as described in Fig. 3d. Analysis includes IgG as a control. The mean of two independent ChIP experiments is shown. Details of replicates are given in the Methods. **e**, ChIP–qPCR analysis of CTCF abundance at loci 8–12, as described in Extended Data Fig. 4a. Analysis includes IgG as a control. The mean of two independent ChIP experiments is shown. Details of replicates are given in the extended methods.

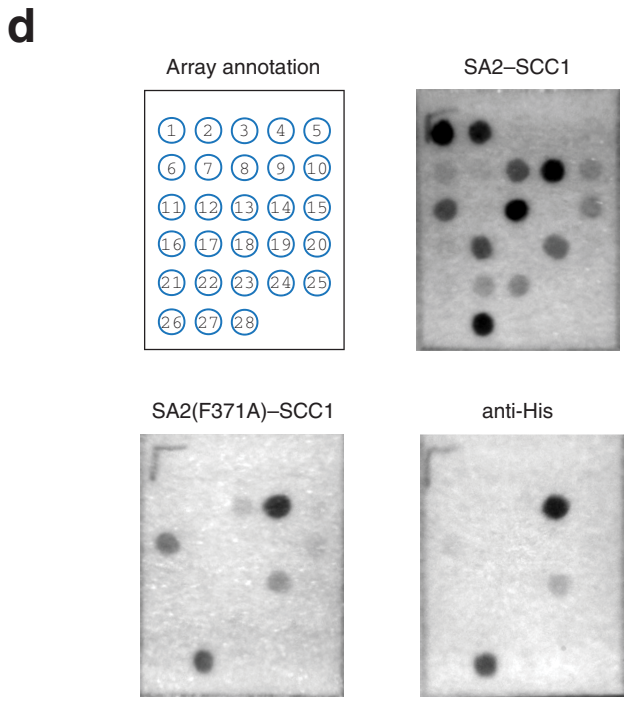


Extended Data Fig. 6 | Compartmentalization is largely maintained in cells that contain the *CTCF*$^{Y226A/F228A}$ mutation. a, Hi-C contact matrices of the q-arm of chromosome 2 at 500-kb resolution. The corresponding compartment scores are plotted above. **b**, Genome-wide comparison of compartment scores for wild-type and *CTCF*$^{Y226A/F228A}$ cells. Pearson correlation = 0.97. **c**, Saddle plots representing the interaction between A and B compartments. **d**, A region of chromosome 1 (55500000–59500000) at 10-kb resolution that contains no obvious CTCF-anchored loops. **e**, Relative contact probability profiles for wild-type and *CTCF*$^{Y226A/F228A}$ mutant cells (left), compared to previously published[12] contact profiles upon degradation of CTCF (middle) or SCC1 (right). The contact probability profile is affected only slightly in the *CTCF*$^{Y226A/F228A}$ mutants, similar to the effects of CTCF depletion.









e

1.	hCTCF	222	DVSVYDFEEEQ
2.	hCTCF_m1	222	DVSVFDFEEEQ
3.	hCTCF_m2	222	DVSVADFEEEQ
4.	hCTCF_m3	222	DVSVYDAEEEQ
5.	hCTCF_m4	222	DVSVADAEEEQ
6.	hWapl_1	70	TGDPFGFDSDD
7.	hWapl_2	425	KLEFFGFEDHE
8.	hWapl_3	450	KIKYFGFDDLS
9.	hWapl_4	557	ARHWNHPDSEE
10.	yWapl	108	KLSAFNFLDGS
11.	hSororin	162	RRSCFGFEGLL
12.	hShugoshin	332	SNDAYNFNLEE
13.	hMCM3	704	SYDPYDFSDTE
14.	hPOGZ	1398	TESFYGFEEAD
15.	hSYCP1	882	RKMAFEFDINS
16.	hSYCP2	867	VNDVYNFNLNG
17.	hSYCP3	21	FTRAYDFETED
18.	yScc1	278	EPTDFGFDLDI
19.	hScc2	1439	TAVFSRYEKHR
20.	yScc2	400	VSLFGSFDQQR
21.	hBTBD7	942	YPDFYDFSNAA
22.	hCENPU	40	PIDVFDFPDNS
23.	hZGPAT	411	PRNVFDFLNEK
24.	hZFHX4	2144	KDSPYNFSNPP
25.	hZFHX3	2205	KDSPYNFSNPP
26.	hMDC1	320	RAQPFGFIDSD
27.	hCHD6	2177	HRRPYEFEVER
28.	hTFIIH	245	PMDLFDFYEQM

Extended Data Fig. 7 | Identification of CES ligands. a, Plot depicting the $\log_2$-transformed fold change in gene expression in relation to the mean of the normalized counts for each gene. Differentially expressed genes (adjusted *P* value < 0.05, two-tailed Wald test adjusted for multiple testing using the Benjamini–Hochberg procedure) are shown in red. Gene names are included for the 40 genes with the highest fold change. **b**, Western blot assessing knockdown of CTCF and the cohesin subunit SMC1 upon transfection with a control siRNA targeting luciferase (luc) or siRNAs targeting *CTCF* or *SMC1A*. This experiment was performed twice with similar results. **c**, Colony-formation assay of wild-type and $CTCF^{Y226A/F228A}$ cells upon transfection with a control siRNA targeting luciferase or siRNAs targeting *CTCF* or *SMC1A*. CTCF remains essential for viability in $CTCF^{Y226A/F228A}$ cells. This experiment was performed four times with similar results. **d**, Peptide array annotation (top left), binding of SA2–SCC1 (top right) or SA2(F371A)–SCC1 mutant (bottom left) and antibody control (bottom right). Three independent experiments were done, with consistent results. One representative example is shown. **e**, Amino acid sequences of the peptides. Predicted lead-anchoring residues are coloured red.

Extended Data Table 1 | Summary of ITC data, and X-ray data collection and refinement statistics

Protein	Residues	K_d (μM)	ΔH (kcal/mol)	TΔS (kcal/mol)	ΔG (kcal/mol)	N‡
CTCF#	222-231	1.04 ± 0.20	-11.08 ± 0.70	-2.92 ± 0.82	-8.16 ± 0.09	0.93 ± 0.04
CTCF†	86-267	0.62	-13.16	-4.61	-8.54	0.78
Wapl†	423-463	32.8	-6.66	-0.54	-6.11	0.62
Wapl†	447-462	78.7	-6.81	-1.20	-5.60	1.00
Shugoshin†	331-341	13.5	-10.67	-4.02	-6.64	0.83
Shugoshin†	331-349 (pT346)§	2.32	-20.00	-12.30	-7.69	0.89

	SA2-Scc1-CTCF PDB 6QNX
Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, *b*, *c* (Å)	79.02, 107.25, 176.49
Resolution (Å)	45.81–2.70^
R_{sym} or R_{merge}	6.9 (175)*
$I / \sigma I$	12.0 (0.8)*
CC 1/2	0.99 (0.33)*
Completeness (%)	99.6 (99.7)*
Redundancy	4.4 (4.3)*
Refinement	
Resolution (Å)	45.81–2.70
R_{work} / R_{free}	0.25 / 0.27
No. reflections	46759
No. atoms	16487
SA2	15088
Scc1	1235
Ligand	140_{CTCF}
B-factors (mean; Å^2)	
SA2	133.4
Scc1	111.3
Ligand	143.6_{CTCF}
R.m.s deviations	
Bond lengths (Å)	0.004
Bond angles (°)	0.53

a, ITC data summary. **b**, X-ray data collection and refinement statistics.

#Three independent experiments were performed. The mean values ± s.d. are shown.

†Experiment was performed once.

‡Binding stoichiometry.

§pT346, phosphothreonine.

^Data derived from one crystal.

*Values in parentheses are for the highest-resolution shell.

Extended Data Table 2 | Quantification of peptide arrays

Position	Protein	species	Mutation	Uniprot	Sequence	K_d [μM]
1	**CTCF**	Human	Wild-type	P49711	222 DVSVYDFEEEQ	1.04 ± 0.16*
2	**CTCF**	Human	Y226F	P49711	222 DVSVFDFEEEQ	1.60 ± 0.03†
3	**CTCF**	Human	Y226A	P49711	222 DVSVADFEEEQ	n.b.
4	**CTCF**	Human	F228A	P49711	222 DVSVYDAEEEQ	n.b.
5	**CTCF**	Human	Y226/F228A	P49711	222 DVSVADAEEEQ	n.b.
6	**WAPL**	Human	Wild-type	Q7Z5K2	70 TGDPFGFDSDD	5.90 ± 0.95†
7	**WAPL**	Human	Wild-type	Q7Z5K2	425 KLEFFGFEDHE	12.43 ± 7.72†
8	**WAPL**	Human	Wild-type	Q7Z5K2	450 KIKYFGFDDLS	2.17 ± 0.16†
9	**WAPL**	Human	Wild-type	Q7Z5K2	557 ARHWNHPDSEE	ND
10	**WAPL**	Yeast	Wild-type	Q99359	108 KLSAFNFLDGS	10.65 ± 3.83†
11	**CDCA5**	Human	Wild-type	Q96FF9	162 RRSCFGFEGLL	ND
12	**SGO1**	Human	Wild-type	Q5FBB7	332 SNDAYNFNLEE	n.b.
13	**MCM3**	Human	Wild-type	P25205	704 SYDPYDFSDTE	1.02 ± 0.05†
14	**POGZ**	Human	Wild-type	Q7Z3K3	1398 TESFYGFEEAD	n.b.
15	**SYCP1**	Human	Wild-type	Q15431	882 RKMAFEFDINS	6.71 ± 3.18†
16	**SYCP2**	Human	Wild-type	Q9BX26	867 VNDVYNFNLNG	n.b.
17	**SYCP3**	Human	Wild-type	Q8IZU3	21 FTRAYDFETED	1.96 ± 0.26†
18	**SCC1**	Yeast	Wild-type	Q12158	278 EPTDFGFDLDI	n.b.
19	**SCC2**	Human	Wild-type	Q6KC79	1439 TAVFSRYEKHR	ND
20	**SCC2**	Yeast	Wild-type	Q04002	400 VSLFGSFDQQR	n.b.
21	**BTBD7**	Human	Wild-type	Q9P203	942 YPDFYDFSNAA	n.b.
22	**CENPU**	Human	Wild-type	Q71F23	40 PIDVFDFPDNS	4.03 ± 0.22†
23	**ZGPAT**	Human	Wild-type	Q8N5A5	411 PRNVFDFLNEK	4.03 ± 1.19†
24	**ZFHX4**	Human	Wild-type	Q86UP3	2144 KDSPYNFSNPP	n.b.
25	**ZFHX3**	Human	Wild-type	Q15911	2205 KDSPYNFSNPP	n.b.
26	**MDC1**	Human	Wild-type	Q14676	320 RAQPFGFIDSD	n.b.
27	**CHD6**	Human	Wild-type	Q8TD26	2177 HRRPYEFEVER	ND
28	**TFIIH**	Human	Wild-type	Q13888	245 PMDLFDFYEQM	n.b.

Peptide spot signal intensities were correlated to the K_d of CTCF wild type, thus yielding a semiquantitative binding assay[74]. Data points are indicated as mean ± s.d. n.b., no apparent binding. ND, not determined owing to nonspecific binding of the anti-6×histidine antibody.

*Value for CTCF wild type, based on ITC measurement shown in Extended Data Table 1a. The mean values ± s.d. are shown.

†Apparent K_d determined on the basis of three independent peptide array experiments.

Extended Data Table 3 | Primers and Hi-C statistics

Primerset	Primer orientation	Sequence
Primerpair 1	**Forward**	GGCACTACAGGACCACGTTT
	Reverse	CCCAATTGTGTCTGCCTTTT
Primerpair 2	**Forward**	GTGGTGTGGGGAAGAGTGTT
	Reverse	GTCAGCTAAACGCCCAGGTA
Primerpair 3	**Forward**	CAAGTTTTCCACCCGCTTTA
	Reverse	GAGCCCTAACACCACTCCAC
Primerpair 4	**Forward**	GGCTTGGAACTGTTGGTCAT
	Reverse	AGATGGCAGCAGCTTTTCAT
Primerpair 5	**Forward**	TGATTGTGTACAACAGCTGCAA
	Reverse	ATTTTTAGGTGCCTCGCAGT
Primerpair 6	**Forward**	CTGAGCCTCCTGCAAAAGTT
	Reverse	CTCTTCTTCGCTCCAGCACT
Primerpair 7	**Forward**	ACTGCAGCCTCAGCTACCTC
	Reverse	TTTATTGGCATTGCCTCCTC
Primerpair 8	**Forward**	CAGTCCTTGTGGCTCCTAGC
	Reverse	TCTGGTGTGCCCTGAACATA
Primerpair 9	**Forward**	CACCTTGTGGACAGTGGTTG
	Reverse	AGCCTGTGAAACAGGGTGAG
Primerpair 10	**Forward**	TACACGGGTGGCTAAAGGAG
	Reverse	AGCCAGCCAGATGTCAAACT
Primerpair 11	**Forward**	CATGCCCAGCCAATTATTTT
	Reverse	CTCTCCTCCACTTCCCCATT
Primerpair 12	**Forward**	CACTTTTCCGACCCAGAAGA
	Reverse	GGCCTGGAGAACTCAAACTG

Genotype	Replicate	Total Pairs	Valid Pairs	Cis	Cis%	Cis < 20kb	Cis > 20kb	Cis ratio
Wild type	1	61118122	60166198	47100811	78,28	7085049	40015762	5,65
Wild type	2.1	62631817	61755440	48127333	77,93	7114243	41043090	5,76
Wild type	2.2	190892790	152708260	122528381	80,24	18087008	104441373	5,77
CTCF Y226A F228A	1	63339779	62197640	47164621	75,83	72900092	39874529	5,47
CTCF Y226A F228A	2.1	62326840	61227593	46569997	76,06	7419962	39150035	5,28
CTCF Y226A F228A	2.2	148586127	118816672	93071165	78,33	14814014	78257151	5,28

Genotype	Replicate	Total Pairs	Valid Pairs	Cis	Cis%	Cis < 20kb	Cis > 20kb	Cis ratio
Wild type	1+2.1+2.2	312814428	270823907	217755919	80,40	32286097	185469852	5,74
CTCF Y226A F228A	1+2.1+2.2	272011360	238342505	186805272	78,38	29523837	157281435	5,33

a, Primers. **b**, Hi-C statistics for replicate library preparations. Libraries 1 and 2 are independent preparations; 2.2 is a deeper resequencing of sample 2.1. The independent libraries 1 and 2.1 were used for Extended Data Fig. 4. A merge of replicates 1, 2.1 and 2.2 of the wild-type cells, and a merge of replicates 1, 2.1 and 2.2 of *CTCF*$^{Y226A/F228A}$ mutant cells, was used for Figs. 3, 4a, Extended Data Figs. 5a, 6. **c**, Hi-C statistics after merging replicates of wild-type and *CTCF*$^{Y226A/F228A}$ libraries.

Author Queries

Journal: **Nature**
Paper: **s41586-019-1910-z**
Title: **The structural basis for cohesin–CTCF-anchored loops**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q1	This proof has been produced on the basis of your corrections to the preproof. For this later stage of production we use an online 'eproof' tool, where you can make corrections directly to the text within the tool and also mark up corrections to the copyedited figures. Please check that the display items are as follows: Figs 0 (black & white); 4 (colour); Tables: None; Boxes: None; Extended Data display items: 7 figures, 3 tables; SI: yes. The eproof contains the main-text figures edited by us and (if present) the Extended Data items (unedited except for the legends) and the Supplementary Information (unedited). Please note that the eproof should be amended in only one browser window at any one time, otherwise changes will be overwritten. Please check the edits to all main-text figures (and tables, if any) very carefully, and ensure that any error bars in the figures are defined in the figure legends. Extended Data items may be revised only if there are errors in the original submissions. If you need to revise any Extended Data items please upload these files when you submit your corrections to this preproof.
Q2	reference/citation amendments were taken over from the preproof comments; please can you double-check all citations carefully to ensure they correspond to the correct references?
Q3	Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.
Q4	Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.
Q5	Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly. Note that changes here will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding email address(es).

Author Queries

Journal: **Nature**
Paper: **s41586-019-1910-z**
Title: **The structural basis for cohesin–CTCF-anchored loops**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q6	You cannot alter accepted Supplementary Information files except for critical changes to scientific content. If you do resupply any files, please also provide a brief (but complete) list of changes. If these are not considered scientific changes, any altered Supplementary files will not be used, only the originally accepted version will be published.
Q7	please note that, per our house style, I have retained (e.g.) SA2(F371A) (on the line, with substitution in parentheses) where the protein is intended, versus SA2F371A (italic, with superscript) for the gene/genotype; please can you check that this accords with your intention throughout?
Q8	I updated SMC1 to SMC1A in legend of Extended Data Fig. 7 (to match the change that was made in the Methods); is this correct? Please amend if necessary

nature research

Corresponding author(s): D.Panne, B. Rowland, K. Muir, E. De Wit

Last updated by author(s): Nov 1, 2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	MXCuBE2, XDS (20180808), UNICORN 6, LAS-AF FRAP-Wizard 2.7.4.10100
Data analysis	Molecular replacement was done with Phaser (Phenix 1.14-3260). Hi-C sequencing data was processed with HiC-Pro 2.9. Replicates similarity was assessed with HiCRep 1.8.0. Hi-C data analysis was performed with GENOVA 0.9.98 (github.com/deWitLab/GENOVA). HiC-Pro output was converted to juicer files using juicebox-pre (juicer tools 0.7.5). We performed loop calling with HICCUPs 1.11 on juicer files. ChIPseq reads were trimmed with TrimGalore 0.6.0. Mapping of ChIPseq data was performed with bowtie 2.3.4.130 to hg19. We performed peak calling with MACS2 2.1.131. Overlaps between sets of identified peaks across samples were obtained using BEDtools 2.25.0. ChIPseq alignment plots were created with deeptools 3.1.3. RNAseq data was mapped with TopHat 2.1.133 and count-tables were generated with HTSeq 0.11.1 with the stranded=reverse setting using the Gencode v19 gene-build. Differential expression analysis was performed with DESeq2 1.18.135. Fluorescence intensity in FRAP experiments was measured using ImageJ 1.52q. Structure refinement was done with Phenix (1.14-3260), Structure building was done with COOT 0.8.0-3, Structure renderings were done with Pymol (2.2.3), Structure analysis was done with MolProbity (4.3), Gel band quantification was done with imageJ (1.8.0_112), ITC data were analyzed with Origin 7.0

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

Data

Policy information about availability of data

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

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

Coordinates are available from the Protein Data Bank under accession number 6QNX for the SA2-SCC1-CTCF complex. The generated Hi-C, RNA-Seq and ChIP-Seq data has been deposited in GEO (accession number GSE126637). The current status of these entries is HPUB ('Hold for Publication') which indicates that they are released when the article is published. We will instruct the data depositories to make these entries publicly accessible prior to publication.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample size. We have added comments on the sample sizes in the legends. For Figures showing GST-pulldown analyses (Fig.1, Extended Data Fig. 1), appropriate controls are used to compare binding side-by-side, as is customary. Single replicates are sufficient in this case. For assays measuring dissociation constants (Extended Data Fig.1, Extended Data Table 1 and 3), three independent experiments were performed as required for determination of measurement errors. Hi-C library preps was performed in duplicate. These replicates showed high similarity and were subsequently merged. RNAseq libraries were generated in triplicate. SCC1 ChIP experiments were performed in triplicate, and CTCF ChIPs in duplicate. These were all analysed by ChIPqPCRs, and a representative of each ChIP was analysed by ChIP-Seq. FRAP was performed on 21 wild type cells, and on 17 CTCF Y226A F228A cells.
Data exclusions	No data was excluded in our analyses.
Replication	We have indicated the number of repeat measurements made and consistency of the results obtained in the figure legends. RNAseq experiments were performed in triplicate. Hi-C was performed in duplicate and data was later pooled together. SCC1 ChIPs were performed in triplicate, and CTCF ChIPs in duplicate. FRAP was performed in three in dependent experiments. All attempts at replication were successful.
Randomization	Randomization is not relevant to this study, as protein and crystal samples are not required to be allocated into experimental groups. No animals or human research participants are involved in this study.
Blinding	Blinding is not relevant to this study, as protein and crystal samples are not required to be allocated into experimental groups in protein structural studies. No animals or human research participants are involved in this study.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	His-HRP antibody (Sigma A-7058 lot number 088M4865V) was used for analysis of peptide arrays at a dilution of 1:2000.

The following primary antibodies were used for ChIP:
SCC1: Abcam ab992 lot.GR3253930-2 and lot.GR3253930-3 5 ug per ChIP
CTCF: Cell Signaling 3418S 5 ug per ChIP
IgG: Sigma-Aldrich I5006 5 ug per ChIP
For Western blot, the following antibodies were used:
SMC1: Bethyl A300-055A 6 dilution 1:1000
SCC1: Millipore 05-908 lot.3055582 dilution 1:1000
CTCF: Millipore 07-729 lot.3059608 and Abcam ab128873 dilution 1:1000
HSP90: Santa Cruz F-8 #I0518 dilution 1:10.000
H4: Millipore 05-858 dilution 1:1000
Tubulin: Sigma T5168 lot.047M4760V dilution 1:10.000
Goat anti-Rabbit: DAKO P0447 lot.20046248 dilution 1:2000
Goat anti-mouse: DAKO P0448 lot.20053537 dilution 1:2000

Validation

https://www.abcam.com/rad21-antibody-chip-grade-ab992.html
https://www.cellsignal.com/products/primary-antibodies/ctcf-d31h2-xp-rabbit-mab/3418
https://www.sigmaaldrich.com/catalog/product/sigma/i5006
https://www.bethyl.com/product/A300-055A/SMC1+Antibody
https://www.emdmillipore.com/INTL/en/product/Anti-RAD21-Antibody,MM_NF-05-908
http://www.merckmillipore.com/INTL/en/product/Anti-CTCF-Antibody,MM_NF-07-729
https://www.abcam.com/ctcf-antibody-epr7314b-ab128873.html
https://www.scbt.com/scbt/product/rapgef6-antibody-f-8
https://www.merckmillipore.com/INTL/en/product/Anti-Histone-H4-Antibody-pan-clone-62-141-13-rabbit-monoclonal,MM_NF-05-858
https://www.sigmaaldrich.com/catalog/product/sigma/t5168
https://www.agilent.com/store/en_US/Prod-P044701-2/P044701-2
https://www.agilent.com/store/en_US/Prod-P044801-2/P044801-2

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HAP1 wild type cells from Carette et al., Nature 2011 a gift from the authors.
HAP1 CTCF Y226A F228A generated in this study in HAP1 wild type background cells using CRISPR/Cas gene editing.

Authentication

Karyotyping. Mutants were confirmed by Sanger sequencing.

Mycoplasma contamination

All cell lines were negative for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified line was used.

ChIP-seq

Data deposition

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

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

Data access links
May remain private before publication.

Go to https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126637
Enter token qzshmsyavvcndmj into the box

Files in database submission

GSM4052950_SCC1_ChIPseq_WT_L001.fastq.gz
GSM4052950_SCC1_ChIPseq_WT_L002.fastq.gz
GSM4052950_SCC1_ChIPseq_WT_peaks.narrowPeak.gz
GSM4052951_SCC1_ChIPseq_CTCF.Y226A.F228A_L001.fastq.gz
GSM4052951_SCC1_ChIPseq_CTCF.Y226A.F228A_L002.fastq.gz
GSM4052951_SCC1_ChIPseq_CTCF.Y226A.F228A_peaks.narrowPeak.gz
GSM4052952_CTCF_ChIPseq_WT_L001.fastq.gz
GSM4052952_CTCF_ChIPseq_WT_peaks.narrowPeak.gz
GSM4052953_CTCF_ChIPseq_CTCF.Y226A.F228A_L001.fastq.gz
GSM4052953_CTCF_ChIPseq_CTCF.Y226A.F228A_peaks.narrowPeak.gz

Genome browser session
(e.g. UCSC)

https://genome.ucsc.edu/s/asedeno/CTCF_Y226A_F228A_HAP1

Methodology

Replicates

SCC1 ChIP experiments were performed in triplicate, and CTCF ChIPs in duplicate. These were all analysed by ChIP-qPCRs, and a representative of each ChIP was analysed by ChIP-Seq

Sequencing depth

sample total_reads uniquely_mapped length type
5512_11_SCC1_WT_CCGTCC_S25_R1_001 30249392 29761787 65 single

	5512_12_SCC1_CTCFmut_GTGAAA_S26_R1_001 30623586 30175451 65 single 5588_5_CTCF_WT_2_GCCAAT_S146_R1_001 54584946 53634127 65 single 5588_6_CTCF_CTCF103_2_CAGATC_S147_R1_001 54586898 53657421 65 single
Antibodies	SCC1: Abcam ab992 lot.GR3253930-2 and lot.GR3253930-3 CTCF: Cell Signaling 3418S
Peak calling parameters	We performed peak calling with MACS2 2.1.131 for SMC1 and CTCF with standard settings.
Data quality	sample >-log10(0.05) >5FC 5512_11_SCC1_WT_CCGTCC_S25_R1_001_peaks.narrowPeak 52350 29930 5512_12_SCC1_CTCFmut_GTGAAA_S26_R1_001_peaks.narrowPeak 47710 19299 5588_5_CTCF_WT_2_GCCAAT_S146_R1_001_peaks.narrowPeak 71900 57656 5588_6_CTCF_CTCF103_2_CAGATC_S147_R1_001_peaks.narrowPeak 84009 67305
Software	We performed peak calling with MACS2 2.1.131 for SMC1 and CTCF with standard settings.