Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo

Lars L. P. Hanssen1,2, Mira T. Kassoufa–d, A. Marieke Oudelara–d, Daniel Biggs3, Chris Preece3, Damien J. Downes1, Matthew Gosden1, Jacqueline A. Sharpe1, Jacqueline A. Sloane-Stanley1, Jim R. Hughes1, Benjamin Davies2,4 and Douglas R. Higgs1,4,5

The genome is organized via CTCF–cohesin-binding sites, which partition chromosomes into 1–5 megabase (Mb) topologically associated domains (TADs), and further into smaller sub-domains (sub-TADs). Here we examined in vivo an ~80 kb sub-TAD, containing the mouse α-globin gene cluster, lying within a ~1 Mb TAD. We find that the sub-TAD is flanked by predominantly convergent CTCF–cohesin sites that are ubiquitously bound by CTCF but only interact during erythropoiesis, defining a self-interacting erythroid compartment. Whereas the α-globin regulatory elements normally act solely on promoters downstream of the enhancers, removal of a conserved upstream CTCF–cohesin boundary extends the sub-TAD to adjacent upstream CTCF–cohesin-binding sites. The α-globin enhancers now interact with the flanking chromatin, upregulating expression of genes within this extended sub-TAD. Rather than acting solely as a barrier to chromatin modification, CTCF–cohesin boundaries in this sub-TAD delimit the region of chromatin to which enhancers have access and within which they interact with receptive promoters.

Whereas previous work has intensively studied the role of enhancers and promoters in regulating gene expression, it is becoming increasingly clear that their dynamic interactions in three dimensions within the nucleus provide a fundamentally important third component for switching genes on and off. We now know that this chromosomal topology is determined by a third class of regulatory elements defined by their binding of CCCTC-binding factor (CTCF) and components of the structural maintenance of chromosome (SMC) cohesin complex1–5. These elements appear to organize chromosomes into a series of increasingly complex topological structures (chromosome loops, sub-TADs, TADs and so on)3. However, not all CTCF–cohesin sites appear equivalent, and a variety of different functions have been attributed to such elements in different contexts, including: acting as boundaries to chromatin modifications4–6, facilitating interactions between regulatory elements2–6, and insulating genes from tissue-specific enhancers8–11. However, at present, how they interact with each other and with other regulatory elements in their natural chromosomal context is poorly understood. To address this, we have examined the interactions between CTCF–cohesin sites, enhancers and promoters, and determined their functional role(s) using chromosome engineering of an ~80 kb sub-TAD containing the well-characterized mouse α-globin cluster, in its natural chromosomal environment, in vivo. We find that CTCF–cohesin sites in this sub-TAD play a key role in regulating gene expression by delimiting the region of chromatin within which active enhancers can interact with receptive promoters.

Changes in chromatin and gene expression across the 1 Mb TAD containing the α-globin locus in erythroid cells

The mouse α-globin cluster is located in close proximity to a cluster of 10 widely expressed genes near the boundary of a ~1 Mb TAD as previously defined12 (Fig. 1a). We have previously characterized the chromatin in and around the α-globin cluster and noted that activation of α-globin expression in erythroid cells is associated with the appearance of a broad domain of histone acetylation and modification by H3K4me1 spanning ~80 kb including the α-globin genes and their regulatory elements (Fig. 1b)13,14. Using assays for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq), DNase I hypersensitive sites sequencing (DNase-seq) and chromatin immunoprecipitation with high-throughput DNA sequencing (ChIP-seq), we have identified all promoters and enhancers in this region via their characteristic chromatin signatures (Fig. 1b). We have shown that the α-globin genes are regulated by four

---

1MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, Oxford OX3 9DS, UK. The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, University of Oxford, Oxford OX3 7BN, UK. These authors contributed equally to this work. *These authors jointly supervised this work.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Figure 1 Regulation of the α-globin cluster in mouse ES and primary erythroid cells. (a) Heat map of Hi-C chromatin interactions surrounding the α-globin gene cluster in mouse ES cells. TADs are annotated by dashed lines as previously defined. Gene annotation is Refseq. (b) The α-globin locus with enhancer elements (R1, R2, R3, Rm, and R4), gene promoters and CTCF binding sites marked by peaks of open chromatin as indicated by ATAC-seq tracks (reads per kilobase per million mapped reads (RPKM)) in mouse ES and primary erythroid cells. 01 and 02 refer to CTCF binding sites at the promoters of the α-globin pseudogenes. All normalized ChIP-seq read densities (RPKM) represent an average of 2 independent experiments, in which 2 animals (erythroid cells) or 2 biologically independent samples (ES cells) were analysed in total, except for ES and erythroid H3K4me1, which represent single experiments. CTCF binding orientation is annotated with forward (red arrows) and reverse (blue arrows). Dashed boxes indicate clusters of convergent CTCF binding sites. Gene annotation is Refseq. The following data sets were previously published: ES SMC1 and SMC3; erythroid H3K4me3; ES CTCF, H3K27me3 and H3K4me1; ES H3K4me3; ES ATAC-seq. (c) Relative gene expression of mouse primary erythroid versus ES cells measured by real-time qPCR and representing n = 3 independent experiments, in which animals (erythroid cells) or n = 3 biologically independent samples (ES cells) were analysed in total. Bars represent the mean and the error bars represent the standard deviation (s.d.). Grey dots represent individual data points. P values are obtained via an unpaired, two-tailed Student’s t-test. Not significant (NS), P > 0.05; **P < 0.001; ***P < 0.0001. (d) DNaseI footprints of HS-38 and HS-39 CTCF binding sites. CTCF motifs are indicated by arrows; red arrow: forward core motif, blue arrow: reverse core motif, and green arrow: upstream motif. P values indicate the significance of the match of the HS-38 and HS-39 sequence to the core consensus motif (derived with the FIMO tool as described previously). The DNase data were previously published.
conserved erythroid-specific enhancers (R1–R4) and a mouse-specific element (Rm) located 14–38 kb upstream of the promoters\textsuperscript{15–17}. Four of these enhancers (R1, R2, R3 and Rm) lie within the introns of an adjacent widely expressed gene (\textit{Nprl3})\textsuperscript{15}. We show here that the remaining regions of open chromatin within and around the gene cluster, identified in all cell types analysed, correspond to binding sites for CTCF–cohesin (Fig. 1b).

The \(\alpha\)-globin genes and the closely linked, widely expressed gene (\textit{Nprl3}), which lie together within the H3K4me1-marked domain, are expressed at high levels in erythroid cells (Fig. 1c). Surrounding genes within the larger (1 Mb) TAD are unaffected by activation of the strong erythroid-specific enhancers and we show here that one of these genes (\textit{Rhbdf1}) is marked by high levels of the Polycomb-mediated repressive mark (H3K27me3) and completely silenced in erythroid cells (Fig. 1b,c). In this way, we have accounted for all regions of open chromatin, the corresponding regulatory elements, and the pattern of gene expression within and surrounding the \(\alpha\)-globin locus.

**An \(\alpha\)-globin sub-TAD is surrounded by convergent CTCF–cohesin-binding sites that interact with each other specifically in erythroid cells**

While the role of CTCF–cohesin-binding sites at boundaries between TADs are starting to be established\textsuperscript{21}, less is known about the CTCF–cohesin sites that are dispersed within TADs. These CTCF–cohesin sites are thought to contribute to the formation of smaller self-interacting domains that have been termed sub-TADs\textsuperscript{23}, contact domains\textsuperscript{19} and insulated neighbourhoods\textsuperscript{41,20}. In contrast to TADs, sub-TADs (40 kb to 3 Mb, median size of 185 kb (ref. 19)) often appear in a cell-specific manner and, as in the case of the 80 kb sub-TAD corresponding to the H3K4me1 and histone acetylation-marked \(\alpha\)-globin domain, may be identified via an increased density of chromatin interactions (as seen by Capture-C, Fig. 2a) and specific histone modifications in a specific cell type. Of interest, recent studies have shown a strong preference for interactions to occur between CTCF sites lying in a convergent orientation with respect to each other\textsuperscript{17–20}. We therefore established the orientation of CTCF binding sites surrounding and within the \(\alpha\)-globin sub-TAD (Fig. 1b). Motif orientations were predicted by inspecting each CTCF consensus core and flanking motifs and subsequently validated by analysing the directional orientation of associated DNaSe footprint in \textit{vivo} (Fig. 1d and Supplementary Fig. 1). This analysis revealed a striking pattern of CTCF orientations in which the regions flanking the \(\alpha\)-globin genes and their enhancers were shown to contain clusters of largely convergently oriented CTCF binding sites (Fig. 1b).

To investigate the mechanisms by which CTCF–cohesin-mediated domains may form, interact and influence the activity of strong tissue-specific enhancers and promoters, we performed next-generation Capture-C in mouse erythroid and non-erythroid, embryonic stem (ES) cells\textsuperscript{17,25}. In ES cells, \(\alpha\)-globin is not expressed whereas transcription of flanking genes (\textit{Snrnp25}, \textit{Mpg}, \textit{Rhbdf1} and \textit{Nprl3}) occurs in the absence of the erythroid-specific enhancer activity (Fig. 1b,c). Using viewpoints from the enhancer region (R1) and the \(\alpha\)-globin promoters (\(\alpha\)1 and \(\alpha\)2), in ES cells we observed broad, diffuse interactions extending across the entire gene cluster (Fig. 2a). By contrast, in erythroid cells, we observed much stronger interactions throughout the sub-TAD but especially between the enhancers and promoters (Fig. 2a).

Interaction profiles from nearby viewpoints located at the CTCF–cohesin-binding sites directly flanking the \(\alpha\)-globin cluster were very different. Despite their proximity to the \(\alpha\)-globin enhancer elements, these sites clearly do not interact with the enhancers within the sub-TAD: rather, they interact with the domains of chromatin containing convergent CTCF–cohesin sites extending from the \(\beta\)-globin promoters to the 3′ flanks of the \(\alpha\)-globin sub-TAD (Fig. 2b). Of particular interest, despite the near-identical CTCF–cohesin binding landscape across the \(\alpha\)-globin sub-TAD in erythroid and non-erythroid cells (Fig. 1b), we observed significantly increased interactions between flanking CTCF–cohesin clusters in erythroid cells, suggesting the development of a hairpin-like structure of the sub-TAD in erythroid cells. While this proposed structure would exclude interactions between flanking sequences and \(\alpha\)-globin enhancers, it would not prevent interactions between the two CTCF sites directly adjacent to the enhancers (HS-38 and HS-39) and the \(\alpha\)-globin promoters. Such interactions may occur with the CTCF binding sites at the \(\alpha\)-globin promoters. Profiles from viewpoints located at the promoters of flanking genes (\textit{Mpg} and \textit{Rhbdf1}) are consistent with this topological model (Fig. 2c), and suggest that CTCF-mediated chromatin interactions between domains flanking the \(\alpha\)-globin cluster may insulate promoters contained within these regions from the activity of the \(\alpha\)-globin enhancers by constraining their interactions with these strong enhancers. Consistent with this model, the \textit{Nprl3} gene, whose promoter is located within the \(\alpha\)-globin sub-TAD, shows a sixfold increase in expression in erythroid cells compared with non-erythroid cells, whereas the expression of \textit{Mpg} and \textit{Rhbdf1} lying outside the sub-TAD is unchanged and repressed in erythroid cells respectively (Fig. 1c).

**Deletion of CTCF–cohesin sites alters multiple chromatin interactions within the \(\alpha\)-globin sub-TAD**

Close inspection of the chromatin profile identified two prominent CTCF–cohesin-binding sites (HS-38 and HS-39) and a less prominent site (HS-29) lying close to and directly upstream of the \(\alpha\)-globin enhancers at the boundary of the erythroid-specific sub-TAD defined by histone acetylation and H3K4me1 enrichment. These CTCF–cohesin sites are positioned in between the erythroid enhancers and the widely expressed upstream genes (\textit{Mpg}, \textit{Rhbdf1} and \textit{Snrnp25}), suggesting that they may act individually or together as an insulator, shielding these upstream genes from enhancer activity. To test this hypothesis in \textit{vivo}, we used TALEN\textsuperscript{26} and CRISPR-mediated mutagenesis\textsuperscript{27} to generate mice with small deletions in the binding sequences of these three CTCF–cohesin sites, singly and in combination (Fig. 3 and Supplementary Fig. 2).

We first analysed erythroid cells of a mouse lacking both HS-38 and HS-39 (D3839). Mutation of the two CTCF core sequences resulted in a complete loss of CTCF binding at these sites (Figs 3, 4a and 5a), but in contrast to previous reports\textsuperscript{28,29} did not affect binding of CTCF to other, nearby sites. To investigate whether mutation of both HS-38 and HS-39 altered interactions between the regions of chromatin flanking the sub-TAD, we used the downstream CTCF binding site (HS48) as a Capture-C viewpoint. Although interactions between flanking domains remain intact in the D3839 mutant, the upstream boundary of the domain in erythroid cells shifts from the deleted HS-38/39 sites to the next adjacent upstream site (HS-59) site within the \textit{Rhbdf1} gene (Fig. 4a). Capture-C using the R1 enhancer...
Figure 2  Differential interactions of α-globin regulatory elements between mouse ES and primary erythroid cells. (a) Panels show overlaid, normalized Capture-C data for the α-globin promoter (α1, α2) and the R1 enhancer viewpoints in mouse ES and primary erythroid cells merged from 3 independent experiments, in which 3 animals (erythroid cells) or 3 biologically independent samples (ES cells) were analysed in total. Each of the α-globin promoters interacts with the enhancers independently, resulting in the expression of both genes (α1, α2)\(^{17,43}\). The mean, plus and minus one standard deviation (s.d.), of sliding 5 kb windows are visualized. Differential trends (ΔCapture-C) show a subtraction (erythroid—ES) of the mean number of meaningful interactions per restriction fragment. Red vertical bars indicate the position of the viewpoint. Also shown are normalized CTCF and cohesin (Rad21 or Smc3) ChIP-seq (RPKM) and ATAC-seq (RPKM) tracks for both ES and primary erythroid cells, all merged from 2 independent experiments, in which 2 animals were analysed in total. Gene annotation is Refseq. (b) Data presented as described in a for HS48 and HS-38 CTCF-site viewpoints. (c) As described in a, data for Mpg and Rhbdf1 promoter viewpoints.

as a viewpoint shows that, while interactions between R1 and the α-globin promoters appear to be unchanged, ablation of CTCF binding in the D3839 mutant results in increased interactions between R1 and a region of chromatin, directly upstream, containing the Mpg, Rhbdf1 and Snrnp25 genes (Fig. 4b). This is further confirmed by the interaction profiles obtained from the Rhbdf1 and Mpg promoters that show a strong increase of interactions with the R1 and R2 enhancers and the α-globin genes, while losing interactions with the downstream genomic region flanking the cluster (Fig. 4c). Thus, the elimination of CTCF binding in the D3839 mutant is associated with an entirely new set of contacts between the α-globin enhancers and the Rhbdf1 and Mpg genes. Importantly, these interactions occur with non-erythroid promoters and involve interactions with promoters located upstream, in the opposite direction to those normally seen from the α-globin enhancers. In the proposed hairpin analogy of the sub-TAD, contacts within the CTCF–cohesin stem of the hairpin have shifted to increase the region of chromatin within its loop that now includes the Mpg, Rhbdf1 and Snrnp25 genes.

**Mutation of CTCF–cohesin sites alters gene expression in the α-globin sub-TAD**

To examine whether the changes in local topology caused by the deletion of CTCF binding sites influence transcription in erythroid cells, we performed RNA sequencing (RNA-seq) on D3839 and wild-type primary erythroid cells. We found that expression of the three genes whose promoters are located in the genomic region that shows increased interactions with the R1 enhancer (Fig. 4b), Mpg, Rhbdf1 and Snrnp25, is strongly upregulated in D3839 mutant mice (Fig. 5a–c).

Housekeeping genes Mpg and Snrnp25 are normally expressed in wild-type erythroid cells but in the absence of HS-38 and HS-39 their expression increases by 12- and 6-fold, respectively. Interestingly, the Rhbdf1 gene, which is normally silenced by Polycomb group complexes in wild-type erythroid cells, increases its expression ~600-fold in the D3839 mutant. In ES cells, Rhbdf1 is transcribed at relatively high levels. Even when compared with this active gene regulatory state, expression was significantly increased under the influence of the α-globin enhancers in the absence of CTCF insulation in D3839 erythroid cells (Fig. 5c). By contrast, the Il9r gene, whose promoter is located within the chromatin region in which interactions between flanking CTCF–cohesin domains are retained in the D3839 mutant (Fig. 4a), remained inactive (Fig. 5b) and insulated from the influence of the α-globin enhancers. We also detected no significant changes in expression of the α- or β-globin genes (Fig. 5c), consistent with the identical interaction profile of the R1 enhancer with the α-like globin promoters in D3839 mutant mice (Fig. 4) and the lack of any detectable change in the haematological phenotype (Fig. 5d).

**Not all CTCF sites in the sub-TAD are equivalent**

Clearly, mutation of both HS-38 and HS-39 sites causes a change in the functional interactions between the α-globin enhancers and the...
promoters of the surrounding genes in the TAD. It has been suggested that effective chromatin boundaries are formed by two directly adjacent divergent CTCF binding sites\textsuperscript{39}. Therefore, we next made and analysed mice with single deletions of CTCF binding at either the HS-38 (D38) or HS-39 (D39) elements (Supplementary Fig. 2). Loss of HS-38 alone led to an upregulation of the upstream Mpg, Rhbdf1 and Snrnp25 genes, although to a lesser extent than that observed in the D3839 mice (Fig. 5e). Rhbdf1 was tenfold less upregulated in the D38 mutant than in the double D3839 mutant, suggesting that the presence of HS-39 in the D38 mutant prevents a complete loss of enhancer insulation. However, deletion of HS-39 alone did not have a strong effect on local gene expression (Fig. 5e). The observation that loss of HS-39 does not result in gene expression changes is consistent with the fact that only HS-38 is conserved across mammals including human and bound by higher levels of CTCF–cohesin, suggesting this site is sufficient for adequate enhancer insulation. These data do not exclude the opposite orientation of HS-38 and HS-39 or the difference in the composition of their CTCF binding motif as possible causes for this difference in insulator activity (Supplementary Fig. 1). Finally, we generated a mouse lacking the HS-29 CTCF binding site (D29), located between the R1 and R2 enhancers. While the loss of CTCF binding at HS-29 resulted in increased interactions between the enhancers and the a-globin promoters (Supplementary Fig. 3), these changes did not result in any detectable changes in local gene expression or histone modifications (Fig. 5f and Supplementary Fig. 4). However, minor changes in a-globin gene expression may not have been detected by quantitative PCR (qPCR), which cannot detect fractional changes in expression. Notably, the HS-29 CTCF site binds lower levels of CTCF and cohesin than HS-38, possibly providing an explanation for its lack of boundary activity.

**The CTCF–cohesin boundary constrains enhancer interactions rather than encroachment of chromatin modifications**

Our results demonstrate that the removal of a bona fide insulator lying within a TAD does not simply cause encroachment of one chromatin state into another, but rather extends the range of interactions from a strong enhancer to flanking promoters. This suggests that enhancer–promoter interactions may be promiscuous rather than specific and that such interactions are normally constrained, in some way, by CTCF–cohesin-binding sites. In this respect, it was of interest that the normally silent Rhbdf1 promoter acquired high levels of H3K4me3 in D3839 erythroid cells, consistent with its ectopic expression in these mutant cells (Fig. 6a,b). Similar increases in H3K4me3 were also noted at the Mpg promoter, and, to a lesser extent, at the Snrnp25 promoter, consistent with their transcriptional upregulation in D3839 erythroid cells. Following the smaller effects on gene transcription, the single disruption of HS-38 CTCF binding resulted in the recruitment of lower levels of H3K4me3 to the Rhbdf1 promoter, whereas loss of HS-39 had no detectable effect on local deposition of H3K4me3 (Supplementary Fig. 5). Changes in H3K4me3 in the D3839 mutant are accompanied by higher levels of RNA polymerase II (Pol II) recruitment and increased chromatin accessibility (ATAC-seq) at the gene promoters. Surprisingly, we could not detect a decrease in the levels of H3K27me3 at the Rhbdf1 gene promoter despite its strong transcriptional activation (Fig. 6a,c,d). To exclude the possibility that H3K27me3 was retained despite the loss of PRC2 recruitment.
Figure 4 Differential interactions of α-globin regulatory regions and flanking genes between WT and D3839 primary erythroid cells. Capture-C data for the indicated viewpoints (red vertical bars) in WT and CTCF mutant D3839 primary erythroid cells are shown as described in Fig. 2a. Data represent 3 independent experiments in which 3 animals were used in total. Differential tracks (ΔCapture-C) show a subtraction (WT–D3839) of the mean number of normalized meaningful interactions per restriction fragment. Mutated CTCF sites are indicated with a shaded grey vertical bar. Also shown are normalized CTCF ChIP-seq (RPKM) for both WT and D3839 primary erythroid cells and ATAC-seq (RPKM) for WT erythroid cells, all merged from 2 independent experiments in which 2 animals were used in total. Gene annotation is Refseq.

(a) HS48 CTCF viewpoint. The shaded blue box indicates a chromatin region with altered CTCF–CTCF interactions. (b) R1 enhancer viewpoint. The shaded blue box indicates altered R1 enhancer interactions. (c) Rhbdf1 and Mpg promoter viewpoints. The shaded blue box highlights the chromatin region with altered interactions.

Following α-globin enhancer activation, we verified that binding of the PRC2 complex component Ezh2 is retained in the D3839 mutant (Fig. 6a,d). Thus, it appears that insulation of the Rhbdf1 promoter from the α-globin enhancers by CTCF–cohesin is required for effective Polycomb-mediated transcriptional repression. This is not consistent with a model of simple chromatin encroachment.
Figure 5  Effects of individual and combined CTCF binding site deletions near the α-globin enhancers on local gene expression in primary erythroid cells. (a) The α-globin locus annotated with enhancer elements and genes (Refseq). Shown are normalized ATAC-seq (RPKM) data for WT and CTCF ChIP-seq (RPKM) for indicated WT and mutant mouse erythroid cells merged from 2 independent experiments in which 2 animals were analysed in total. (b) MA (log ratio (M) versus average (A)) plot of RNA-seq data derived from 2 independent experiments in which 3 animals were analysed in total. Mean RNA abundance is plotted on the x axis and enrichment is plotted on the y axis. Significant upregulation of local genes in the D3839 mutant is highlighted in blue and repressed in erythroid cells; Sh3pxd2b (red), repressed in erythroid cells; Mitoferrin 1 (green), a highly expressed housekeeping gene within the α-globin locus, unaffected by deletions. (c) Relative gene expression in WT and D3839 erythroid cells versus ES cells was measured by real-time qPCR. Data represent n = 3 independent experiments in which 3 animals (erythroid cells) or 3 biologically independent samples (ES cells) were analysed in total. Bars represent the mean and the error bars represent the standard deviation (s.d.). Grey dots represent individual data points. P values are obtained with an unpaired, two-tailed Student’s t-test. (d) Table summarizing the haematological parameters of erythroid cells in WT and D3839 mutant mice. RBC, red blood cell count; HGB, haemoglobin count; MCV, mean corpuscular volume (fL); MCH, mean corpuscular haemoglobin (g dL−1); spleen/body%, spleen weight as a percentage of body weight; WT+/−, wild-type; D3839+/−, heterozygous for D3839; D3839−/−, homozygous for D3839. (e) Relative gene expression in D38 and D39 versus WT erythroid cells (as in c, n = 3 independent experiments in which 3 animals were analysed in total). (f) Relative gene expression in WT versus D29 erythroid cells (as in c, n = 3 independent experiments in which 3 animals were analysed in total). No substantial difference in expression of local genes is detected (Rhbdf1; P = 0.03, unpaired, two-tailed Student’s t-test). NS, P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001.
Discussion

The above results indicate that following activation of the α-globin enhancers, selective convergent CTCF–cohesin-binding sites act as boundary elements and create an erythroid-specific chromatin structure, delimiting enhancer interactions and consequently ensuring an erythroid-specific transcriptional program (Fig. 6e). The ability of active enhancers to interact with an unexpectedly wide range of receptive promoters was revealed when critical CTCF–cohesin elements were removed. More widespread and bidirectional enhancer interactions appeared and were associated with the upregulation of three genes; in one case (Rhbd1) overcoming Polycomb-mediated repression. While genetic perturbation of CTCF binding has been shown to result in misregulation of gene expression in various cell lines and cancer,9-11,28, our more detailed investigation involving precise CTCF-site disruptions, and high-resolution chromatin conformation analysis (Capture-C) clearly link gene activation and acquisition of H3K4me3 to the establishment of aberrant enhancer contacts within a perturbed, tissue-specific sub-TAD. Importantly, we have shown that not all CTCF–cohesin sites subserve the same functions in the sub-TAD: HS-38 acts as a strong boundary element, HS-39 as a weaker element and HS-29 has no apparent insulator activity. The molecular basis of this is currently unknown.

Figure 6 Effects of combined deletion of HS-38 and HS-39 on the local chromatin state in primary erythroid cells. (a) Normalized ATAC-seq and ChIP-seq read densities (RPKM; 2 independent experiments in which 2 animals were analysed in total) for H3K4me3, H3K27me3, Ezh2, RNA Pol II and CTCF at the α-globin locus, both in WT and D3839 primary erythroid cells. The shaded grey bar indicates the position of HS-38 and HS-39. The dashed box highlights the region over the Rhbd1 and Mpg genes, magnified in the top panels for ease of data visualization. (b) MA plot of H3K4me3 ChIP-seq data derived from WT and D3839 erythroid cells (2 independent experiments in which 2 animals were analysed in total). Mean read abundance is plotted on the x axis and enrichment on the y axis. Changes in H3K4me3 detected as indicated on the plot by the genes highlighted in blue: Snrnp25: FDR < 0.1, Rhbd1: FDR < 0.05, Mpg: FDR < 0.05. Controls are shown in red (α-globin promoters) and yellow (Nprl3, unaffected by the combined disruption of HS-38 and HS-39). The FDR was calculated with the Diffbind package using DEseq2. (c) A model for α-globin cluster gene regulation. Interactions between clusters of flanking CTCF sites prevent contacts between the α-globin enhancers and upstream genes from forming. Following deletion of HS-38 and HS-39 CTCF binding sites (D3839), CTCF interactions shift towards more distally located upstream sites, allowing bidirectional α-globin enhancer interactions and upregulation of upstream genes.
In addition, we show that interactions between the two flanking clusters of CTCF sites are weaker or absent in ES cells despite the presence of CTCF and cohesin binding (Fig. 6e). This raises the question of what regulatory mechanisms drive the tissue specificity of these CTCF interactions. As cohesin is loaded at active enhancer–promoter junctions, one intriguing possibility is that additional cohesin recruitment in erythroid cells results in stabilization of flanking CTCF–CTCF interactions via a recently described loop extrusion mechanism. In addition, the enhancer–promoter interactions that occur in erythroid cells may further stabilize the interactions between flanking CTCF binding sites.

In conclusion, our findings suggest that rather than enhancers having inherent specificity for their cognate promoters, this communication is at least in part driven by the CTCF-mediated chromatin architecture that normally shields genes flanking a sub-TAD from the influence of enhancers in a tissue-specific manner. However, of interest, we have previously shown that the human α-globin enhancers may influence the expression of a gene (NM23) lying 400 kb away and outside of the orthologous region described here. Therefore, insulation may not be absolute. Nevertheless, given the dynamic genome partitioning through development and differentiation described here, it seems likely that in addition to variants in enhancers and promoters, intergenic variants within critical CTCF–cohesin-binding sites will underlie changes in gene expression associated with a wide variety of complex traits and diseases.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of this paper.

Note: Supplementary Information is available in the online version of this paper.

ACKNOWLEDGEMENTS

The authors would like to thank J. Telenius for her help with bioinformatic analysis of the data and D. Hoy, R. Gibbons and A. King for critically reading the manuscript. L.L.P.H. and A.M.O. would like to thank the Wellcome Trust for funding (Chromosome and Developmental Biology PhD Programme, grant code 099684/Z/12/Z; and Wellcome Trust Genomic Medicine and Statistics PhD Programme, grant code 083323/Z/07/Z respectively). The work was further supported by the Wellcome Trust core award 203412/Z/16/Z and the Medical Research Council (MRC Core Funding and Centenary Award reference 4050189188).

AUTHOR CONTRIBUTIONS

L.L.P.H. planned and carried out experiments, analysed the data, carried out bioinformatic analysis, and wrote the manuscript. M.T.K. coordinated and advised on the project and revised the manuscript. A.M.O. carried out Capture-C experiments and revised the manuscript. D.B.R. carried out in vitro culture and mouse experiments. C.P. carried out mouse microinjection experiments. D.J.D. and M.G. carried out ATAC-seq experiments. J.A.S. and J.A.S.-S. carried out mouse maintenance and hematology on the project and revised the manuscript. A.M.O. carried out Capture-C. The authors declare no competing financial interests.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://dx.doi.org/10.1038/nbt3573

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

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

ARTICLES

1. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. Genes Dev. 30, 881–891 (2016).
2. Nichols, M. H. & Corces, V. G. A CTCF code for 3D genome architecture. Nat. Rev. Genet. 17, 661–678 (2016).
3. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. Nat. Rev. Genet. 16, 713–727 (2015).
4. Weth, O. et al. CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. Nucleic Acids Res. 42, 11941–11951 (2014).
5. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. 20, 34–52 (2010).
6. Handoko, L. et al. CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. 43, 630–638 (2011).
7. Liu, Z.; Scannell, D. R., Eisen, M. B. & Tjian, R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell. 146, 720–731 (2011).
8. Hizanidis, K., Tarasawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. Cell Rep. 2, 345–357 (2012).
9. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 529, 110–114 (2016).
10. Hnisz, D. et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 351, 1454–1458 (2016).
11. Downen, J. M. et al. Control of cell identity genes occurs in isolated neighborhoods of mammalian chromosomes. Cell 159, 374–387 (2014).
12. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).
13. Anguita, E., Johnson, C. A., Wood, W. G., Turner, B. M. & Higgs, D. R. Identification of a conserved erythroid specific domain of histone acetylation across the alpha-globin gene cluster. Proc. Natl Acad. Sci. USA 98, 12114–12119 (2001).
14. Kowalczyk, M. et al. Intragenic enhancers act as alternative promoters. Mol. Cell 47, 447–458 (2012).
15. Hay, D. et al. Genetic dissection of the v-globin super-enhancer in vivo. Nat. Genet. 48, 895–903 (2016).
16. Anguita, E. Deletion of the mouse α-globin regulatory element (HS-26) has an unexpectedly mild phenotype. Blood 100, 3450–3456 (2002).
17. Davies, J. O. J. et al. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat. Meth. 13, 74–80 (2016).
18. Phillips-Cremins, J. E. et al. Architectural protein subclusters shape 3D organization of genomes during lineage commitment. Cell 153, 1281–1295 (2013).
19. Horikawa, M. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).
20. Hnisz, D., Day, D. S. & Young, R. A. Insulated neighborhoods: structural and functional units of mammalian gene control. Cell 167, 1188–1200 (2016).
21. Guo, Y. et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell 162, 900–910 (2015).
22. 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).
23. de Wit, E. et al. CTCF binding polarity determines chromatin looping. Mol. Cell 60, 676–684 (2015).
24. Keefe, M. V. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 10, 1297–1309 (2015).
25. Hughes, J. R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet. 46, 205–212 (2014).
26. Cao, W. et al. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res. 39, e82 (2011).
27. Ran, F. A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell 154, 1380–1389 (2013).
28. Narendra, V. et al. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science 347, 1017–1021 (2015).
29. Yang, R. et al. Differential contribution of cis-regulatory elements to higher order chromatin structure and expression of the CFTR locus. Nucleic Acids Res. 44, 3082–3094 (2016).
30. Gómez-Marín, C. et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc. Natl Acad. Sci. USA 112, 7542–7547 (2015).
31. Kappio, M. H. et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature 467, 430–435 (2010).
32. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. Cell Rep. 15, 2038–2049 (2016).
33. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. Cell 164, 1110–1121 (2016).
34. Zabidi, M. A. & Stark, A. Regulatory enhancer–core-promoter communication via transcription factors and cofactors. Trends Genet. 32, 801–814 (2016).
35. Oh, M. A., Faic, J., Huynen, M., Juretic, M. & Yalcin, T. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. BMC Genom. 17, 25 (2016).
36. Narendra, V.; Bulajic, M.; Dekker, J., Mazzoni, E. D. & Reinenberg, D. CTCF-mediated topological boundaries during development foster appropriate gene regulation. Genes Dev. 30, 2657–2662 (2016).
37. Lower, K. M. et al. Adventitious changes in long-range gene expression caused by polymeric structural variation and promoter competition. Proc. Natl Acad. Sci. USA 106, 21771–21776 (2009).
38. Stadler, P. B. et al. CTCF shapes the mouse methylome at distal regulatory regions. Nature 480, 490–495 (2011).
39. Consortium, T. E. P. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **488**, 57–74 (2012).

40. Simon, C. S. et al. Functional characterisation of cis-regulatory elements governing dynamic Eomes expression in the early mouse embryo. *Development* **144**, 1249–1260 (2017).

41. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

42. Hosseini, M. et al. Causes and consequences of chromatin variation between inbred mice. *PLoS Genet.* **9**, e1003570 (2013).

43. Leder, A., Daugherty, C., Whitney, B. & Leder, P. Mouse zeta- and alpha-globin genes: embryonic survival, alpha-thalassemia, and genetic background effects. *Blood* **90**, 1275–1282 (1997).

44. Wallace, H. A. C. et al. Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**, 197–209 (2007).
Methods
Animal procedure. C57BL/6j mice were sourced from MRC Harwell/Charles River Laboratories. The mutant mouse strains reported in this study were maintained on a C57BL/6j background and were generated and phenotyped in accordance with Animal [Scientific Procedures] Act 1986, with procedures reviewed by the clinical medicine Animal Welfare and Ethical Review Body (AWERB), and conducted under project licences PPL 30/2966 and PPL 30/3339. Animals were housed in specific pathogen-free conditions, with the only reported positives on health screening over the entire time course of these studies being Entamoeba spp. All animals were singly housed, provided with food and water ad libitum and maintained on a 12 h light:12 h dark cycle (150–200 lux cool white LED light, measured at the cage floor). No statistical method was used to predetermine sample size. Experiments to determine haematological parameters were blinded. Mice were given neutral identifiers and analysed by research technicians unaware of mouse genotypes during outcome assessment. Experiments for Capture-C, gene expression and ChiP-seq analysis were not randomized and the investigators were not blinded to allocation during these experiments and outcome assessments. No statistical method was used to predetermine sample size. No samples or animals were excluded from the analysis.

Isolation and selection of ter119+ cells from mice. Mature primary erythrocyt cells were obtained from young adult mice of both genders between 2 and 6 months of age that were pre-treated with acetylatedphydrazine (APh) as described previously46. Spleens of APh-treated mice were mechanically disrupted to single-cell suspension in RPMI media (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS, Gibco). To isolate late-stage erythrocytes, cells from a single spleen were resuspended in 5 ml of cold PBS/2% BSA and stained with 120 µl PE anti-ter119 antibody (Ly-76, BD Biosciences) at 4°C for 15 min (ref. 46). After washing stained cells in PBS/0.5% BSA, cells were resuspended in 1.6 ml of PBS/0.5% BSA and 400 µl of anti-PE magnetic beads (Miltenyi Biotec) and incubated for 20 min at 4°C. Ter119-positive cells were isolated via auto-magnetic-activated cell sorting (autoMACS, Miltenyi Biotec) and processed for downstream applications. Purity of the isolated erythrocyt cells was routinely verified by FACS.

Cell lines. The embryonic stem cell line ES-E14TG2a was used for gene expression and Capture-C analysis and cultured according to standard conditions (ATCC CRL-1821). The E14TG2a line was a gift from A. Smith and was tested negative for mycoplasma and has been extensively authenticated by blastocyst injection. This cell line is not found in the database of commonly misidentified cell lines that is maintained by ICLAC and NCRi Biosample.

Preparation of TALEN expression constructs. For TALEN construction, a 500 bp sequence centred around the HS-38 CTCT consensus sequence was submitted to the TALE-NT Targeter using NN for G recognition (Golden Gate TALEN and TAL Effectors, Addgene)34, 35. Two TALEN pairs with a differential spacer region that targeted the HS-38 CTCT binding sequence were selected and constructed via the Golden Gate assembly method34. TALEN-FF targeted the sequence 5′-TCT GGGTAGGCTCTT-3′ with the RVD array HD-HD-NG-NN-NN-NG-NI-NN-NN-HD-HD-HD-NG and TALEN-AR targeted the sequence 5′-GAGTC CCACGTATCGT-3′ on the reverse strand with the RVD array HD-HD-NG-NI-NN-NN-HD-HD-HD-NG-NI-NN. The vector RCIscript-Goldytalen (38412, Addgene) was used as the scaffold vector in the final step of the Golden Gate cloning protocol.

Preparation of CRISPR-Cas9 expression construct. To generate the single guide-RNAs (sgRNAs) used to target TCF binding sequences, oligonucleotides corresponding to the target protospacers were cloned into plx330-U6-Chimeric_BR-CRBN-hspCas9 (Addgene plasmid no. 42230, px330) or px335-U6-Chimeric_BR-CRBN-hspCas9 (Addgene plasmid no. 42335, px335) or px335 and px330 were modified to contain a purmoycin and neomycin selection cassette respectively. DNA oligonucleotides containing the 20 nt protospacer sequences are shown in Supplementary Table 1.

Preparation and injection of TALEN mRNA and CRISPR sgRNA. TALEN microinjections were performed as previously described44. DNA templates for use in in vitro transcription reactions were generated from CRISPR–Cas9 expression constructs by PCR. The forward, sgRNA-specific primer was modified with a 5′ extension that contained a T7 polymerase binding site, and used to amplify the gRNA with a reverse primer binding downstream of the mature gRNA sequence (gRNA-R) (see Supplementary Table 1). The MEGAscriptT7 Transcription Kit (Thermo Scientific) was used for in vitro transcription of the gRNAs. Diluted in 10 mM Tris-HCl pH 7.5, 0.1 mM EDTA pH 8.0 before microinjection. Manipulations using wild-type Cas9 were performed using a Cas9-expressing mouse line to provide a maternal supply of Cas9 to zygotes as previously described45. Briefly, female mice homozygous or heterozygous for the CAG-Cas9 transgene insertion were superovulated and mated with C57BL/6j or, for the production of double (D3839) mutants, with DEL38 studs. Oocytes were prepared for microinjection from naive females and 20 ng/µl of gRNA was injected into the pronucleus. Depending on the experiment, ssODN templates for HDR (Eurogentec) were added at a final concentration of 20 ng/µl (see Supplementary Table 1). For the single mutation of HS-39, D10A Cas9 protein (PNA Bio) was injected with two sgRNAs at a concentration of 40 ng/µl into C57BL/6j oocytes. The microinjected zygotes were immediately transferred to pseudopregnant CD1 foster mothers.

Next-generation Capture-C. Next-generation Capture-C was performed as previously described48. A total of 2 × 10⁴ mouse ES cells or isolated ter119+ mouse spleen cells were used per biological replicate experiment and processed in parallel. To visualize differences in Capture-C profiles, normalized interactions in ES cells or erythroid cells of CTCT binding site mutants were subtracted from wild-type erythroid interactions to generate a differential Capture-C track. For plotting of multiple interaction profiles simultaneously, Capture-C interactions were binned in 250 bp bins and a sliding 5 kb window was used. The mean of three biological replicates and standard deviation were plotted in R.

ATAC-sequencing. ATAC-seq was performed on 65,000 ter119+ cells isolated from APh-treated mouse spleens and mouse ES cells as previously described44. Data was mapped to the genome using an in-house analysis pipeline as previously described44.

RNA expression analysis. Isolation of total RNA was performed by lysing 1 × 10⁴ mouse ES or purified ter119+ cells in TRI reagent (Sigma) according to the manufacturer’s instructions. To remove genomic DNA from RNA samples, samples were treated with TURBO DNase with the DNA-free DNA removal kit (Ambion). DNase-treated RNA samples were stored at −80°C. To assess relative changes in gene expression by qPCR, 1 µg of total RNA was used for cDNA synthesis using the Superscript III first-strand synthesis SuperMix (Invitrogen). The ΔΔCt method was used for relative quantitation of RNA abundance using the primers in Supplementary Table 1. For RNA-seq libraries, RNA and globin mRNA species were removed using the Globin-Zero Gold kit (Illumina) with 5 µg of total RNA according to the manufacturer’s instructions. To further enrich for mRNA, poly(A)+ mRNA was isolated using the NEBNext Poly(A) mRNA magnetic isolation module (New England Biolabs) followed by a NEBNext Ultra directional RNA library preparation (New England Biolabs) according to the manufacturer’s instructions. Fragmentation of mRNA was achieved by incubating samples at 95°C for 12 min. To achieve strand specificity, actinomycin D was added (5 µM) to the first-strand cDNA synthesis reaction. Poly(A)+ libraries (4 nM) were sequenced on the Illumina NextSeq platform. All RNA-seq data sets were aligned to the mm9 mouse genome build using STAR49. Deeptools bamCoverage was used to calculate normalized (RPKM) and strand-specific read coverage, which was visualized in the UCSC genome browser. Mapped RNA-seq reads were assigned to genes using Subread featureCounts using ReSeq gene annotation. Normalized differential gene expression, between biological triplicate data from littermate wild-type and CTCT binding site mutant mice extracted in parallel, was calculated with the DESeq2 R package.

De novo CTCT motif analysis in ter119+ cells. Motif analysis was performed as previously described45. Briefly, 2,000 CTCT peak sequences from ter119+ cells were retrieved and used for de novo motif discovery using the MEME suite. The motif with the highest score matched the previously published consensus CTCT core binding motif. Significant matches (P < 10⁻⁵) for the CTCT core motif within all CTCT peak regions were identified using FIMO. When multiple core motifs were detected within the same peak region, only the best match was retained. Motifs up- and downstream of the core motif were identified from 6,000 randomly selected 20 bp sequences of up- and downstream flanks and were similar to those previously identified45. Analysis of spacing between the core and flanking motifs revealed a preferential spacing for both up- and downstream motifs. Significant upstream or downstream motif (P-value threshold of 10⁻⁵) were added to CTCT peak annotation only if the motifs were present at the preferred spacing (5–6 bp for upstream motif, 4–6 for downstream motif).

DNasel footprint analysis. DNasel footprints and meta-plots at CTCT binding sites were generated using a custom perl script based on Samtools using previously published C57BL/6 DNA seq data45. DNasel-seq cuts were counted as the 5′ end of the first peak (P-value threshold of 10⁻⁵) were added to CTCT peak annotation only if the motifs were present at the preferred spacing (5–6 bp for upstream motif, 4–6 for downstream motif).

Nature cell biology
Chromatin immunoprecipitation (ChIP) was performed on purified ter119-positive primary erythroid cells (1 × 10^7 cells per ChIP) using the ChIP Assay Kit (Cat. No. 17-295, Millipore). For ChIP of cohesin component Rad21, cells were subjected to dual crosslinking with 2 mM disuccinimidyl glutarate (DSG, Thermo Fischer Scientific) for 50 min and 1% (v/v) formaldehyde for 10 min, whereas a single 10 min 1% formaldehyde fixation was used for all other antibodies (see details in the Reporting Summary). Chromatin fragmentation was performed with the Bioruptor sonicator (Diagenode) for 15 min at 4°C to obtain an average fragment size between 200 and 500 bp. Immunoprecipitated DNA fragments were analysed by qPCR or sequencing. Primers used for qPCR are listed in Supplementary Table 1. DNA libraries for sequencing were prepared with the NEBNext Ultra II DNA library prep kit (New England Biolabs) and sequenced on the Illumina platform.

**Analysis of ChIP-seq data.** ChIP-seq data was mapped to the genome using an in-house analysis pipeline as previously described. The MACS (Model-based analysis of ChIP-seq) peak finding algorithm was used to identify regions of ChIP-seq enrichment over background in an unbiased manner. The MACS2 callpeak function was used on biological duplicate ChIP-seq data of CTCF (-q 10^-5) and H3K4me3 (-q 10^-6). For H3K27me3, the MACS2 callpeak function was used with the broad-cuto option (-broad-cuto 0.05) on biological duplicate ChIP-seq data. To identify regions that were differentially enriched between wild-type and CTCF binding site mutant mice, the R package DiffBind was used. Two biological duplicate data sets and independent peak calls of CTCF, H3K4me3 and H3K27me3 were used to identify differentially enriched regions with a false discovery rate (FDR) of 0.05. Differential analysis within the DiffBind package was performed with DESeq2.

**Statistics and reproducibility.** Statistical analysis was carried out with Graphpad Prism (version 7.0c) unless otherwise indicated. All gene expression experiments (both RT-qPCR and RNA-seq) were performed on three biological replicates with similar results (standard deviation (s.d.) is shown for all measurements). Statistical analysis was performed using multiple unpaired, two-tailed t-tests and corrected for multiple comparisons using the Holm-Sidak method where appropriate. The statistical analysis of RNA-seq data was performed in R using the Bioconductor DESeq2 package. All Capture-C experiments were performed on three biological replicates with similar results. The standard deviation of 250 bp bins was calculated in R and visualized to illustrate the reproducibility of this chromatin interaction analysis. All ChIP experiments newly generated for this study were performed at least in biological duplicate with similar results. The analysis of ChIP-seq data was performed with Bioconductor package DiffBind, using DEseq2 to determine false discovery rate (FDR). P values are represented as *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001.

**Data availability.** RNA-sequencing, ChIP-sequencing, ATAC-sequencing and Capture-C data generated for this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE97871. Previously published ChIP-seq data that were reanalysed here are available under the following accession codes: GSE27921, GSE30189, GSE30203. Previously published Capture-C data that were reanalysed here are available under the accession code GSE27959. Previously published DNaseI- and ATAC-seq data that were reanalysed here are available under the accession codes GSE49460 and GSE94249. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

45. Spivak, J. L., Toretti, D. & Dickerman, H. W. Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. *Blood* **42**, 257–266 (1973).
46. Kina, T. et al. The monoclonal antibody TER-119 recognizes a molecule associated with glycophorin A and specifically marks the late stages of murine erythroid lineage. *Br. J. Haematol.* **109**, 280–287 (2000).
47. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
48. Davies, B. et al. Site specific mutation of the Zic2 locus by microinjection of TALEN mRNA in mouse CD1, C3H and C57BL/6J oocytes. *PLoS ONE* **8**, e60216–e60217 (2013).
49. Cebrian-Serrano, A. et al. Maternal supply of Cas9 to zygotes facilitates the efficient generation of site-specific mutant mouse models. *PLoS ONE* **12**, e0169887–e01698820 (2017).
50. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
51. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
52. Nakahashi, H. et al. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3**, 1678–1689 (2013).
Supplementary Figure 1 De novo analysis of the CTCF binding motifs and footprints in mouse erythroid cells. A, CTCF consensus binding motif in forward and reverse orientation resulting from MEME motif analysis of CTCF ChIP-seq (2 independent experiments in which 2 animals were analysed in total) in erythroid cells. Core (C), upstream (U), and downstream (D) sequence elements are identified with preferential spacing to the core motif (spacer, red bar histograms). B, Plots of the average DNaseI footprints of C, UC, and UCD motif containing CTCF binding sites in forward orientation (top panel) and reverse orientation (lower panel). Upper (red, +) and lower (blue, -) strand specific DNaseI cleavage signals are shown. Footprints are averaged over the total number of sites in each category (indicated between parentheses). C, Normalised CTCF ChIP-seq (RPKM, 2 independent experiments in which 2 animals were analysed in total) annotated with the CTCF site names and orientation at the α-globin locus in erythroid cells. Gene annotation is Refseq. DNaseI footprints and top CTCF binding motif hit for each of the CTCF binding sites. Motif P-values are shown (as explained in Fig 1C). Orientation is indicated by the colour of the arrow over the core motif (C); forward (red) or reverse (blue). Upstream motif (U) is shown in green and downstream motif (D) in yellow.
**SUPPLEMENTARY INFORMATION**

Supplementary Figure 2 Genome editing strategies for disruption of CTCF binding at the γ-globin locus. Schematic overview of the mutagenesis method (CRISPR-Cas9 and/or TALEN) and the resulting mutations induced at HS-38 (D38), HS-39 (D39), HS-38 and HS-39 (D3839), and HS-29 (D29). DNaseI footprints are annotated with matches to the C, U, and D CTCF motif (coloured bar under C, red = forward, blue = reverse). Where homology directed repair was used to replace part of the CTCF core motif with a restriction site, the single-stranded oligodeoxynucleotides (ssODNs) used as repair templates are shown.
Supplementary Figure 3 Differential interactions of α-globin regulatory regions and flanking genes between wild-type and D29 primary erythroid cells. Capture-C interaction data for the indicated viewpoints (red vertical bars) in wild-type (WT) and CTCF mutant D29 (deletion of HS-29) primary erythroid cells is shown as described in Fig 2A. Data represent 3 independent experiments in which 3 animals were analysed in total. Shaded grey bar indicates the mutated CTCF site (HS-29). Also shown are normalised CTCF ChIP-seq (RPKM) for both WT and D29 erythroid cells and ATAC-seq (RPKM) for WT erythroid cells, all merged from 2 independent experiments in which 2 animals were analysed. Gene annotation is Refseq.
Supplementary Figure 4 Effects of the deletion of HS-29 on local gene expression and chromatin state in primary erythroid cells. A. MA plot for RNA-seq data derived from WT and D29 erythroid cells. Data represent n=3 independent experiments in which 3 animals were analysed in total. Mean RNA abundance is plotted on the x-axis and enrichment is plotted on the y-axis. No significant changes in local gene expression were detected as shown on the plot by the genes highlighted in blue: Snrnp25, Rhbdf1, Mpg. Controls are shown in different colours: Mitoferrin-1 (Slc25a37, green), a highly expressed erythroid gene; Sh3pxd2b and Il9r (red), repressed in erythroid; Nprl3 (yellow), a housekeeping gene within the α-globin locus that is unaffected by HS-29 deletion. The five most significant outliers were investigated and only one gene was located in cis to α-globin (Irgm2, >25Mb removed), making it unlikely that these expression changes represent direct effects from HS-29 deletion. Significance was tested with a Wald test (DEseq2, n=3 independent experiments in which 3 animals were analysed). B. Normalised ATAC-Seq and ChIP-seq read-densities (RPKM; 2 independent experiments in which 2 animals were analysed in total) for H3K27me3, H3K4me3, and CTCF at the α-globin locus both in WT and D29 primary erythroid cells. Shaded grey bar indicates HS-29. The dashed box highlights the region over the Rhbdf1 and Mpg genes, magnified in top panels for ease of data visualisation.
Supplementary Figure 5  Effects of the individual deletion of HS-38 or HS-39 on local chromatin state in primary erythroid cells. Normalised ATAC-seq and ChIP-seq read-densities (RPKM; 2 independent experiments in which 2 animals were analysed in total) for H3K27me3, H3K4me3, and CTCF at the α-globin locus in WT, D38 and D39 primary erythroid cells. Shaded grey bar indicates HS-38 and HS-39 (mutated CTCF sites in D38 and D39). The dashed box highlights the region over the Rhbd1 and Mpg genes, magnified in top panels for ease of data visualisation.
Supplementary Table Legends

Supplementary Table 1 DNA sequences of sgRNAs, primers and ssODNs. Contains the DNA sequences of sgRNAs and ssODNs used for the generation of CTCF mutant mice and primers used for RT-qPCR and ChIP-qPCR analysis of genes in close proximity to the α-globin gene cluster.
Experimental design

1. Sample size
   
   Describe how sample size was determined. No statistical method was used to predetermine sample size. Experiments were performed in triplicate (Capture-C, expression analysis, ChIP-qPCR) or duplicate (ChIP-seq, ATAC-seq) and this is mentioned where relevant in figure legends.

2. Data exclusions
   
   Describe any data exclusions. No samples or animals were excluded from the analysis.

3. Replication
   
   Describe whether the experimental findings were reliably reproduced. Gene expression experiments (RNA-seq and RT-qPCR) and Capture-C were performed in biological triplicate with similar results. ChIP experiments were performed at least in duplicate with similar results.

4. Randomization
   
   Describe how samples/organisms/participants were allocated into experimental groups. The experiments were not randomised.

5. Blinding
   
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. Experiments to determine haematological parameters were blinded. Mice were given neutral identifiers and analysed by research technicians unaware of mouse genotype during outcome assessment. Experiments for Capture-C, gene expression and ChIP-seq analysis were not randomised and the investigators were not blinded to allocation during these experiments and outcome assessments.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

- The exact sample size ($n$) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. $p$ values) given as exact values whenever possible and with confidence intervals noted
- A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Policy information about availability of computer code

Describe the software used to analyze the data in this study.

Several previously published software packages were used for the analysis of genome-wide data. These include FASTQC 0.11.4, Bowtie 1.1.2, Samtools 0.1.19, Bedtools 2.25.0, Deeptools 2.2.2, the MEME suite 4.9.1_1 and FIMO, MACS 2.0.10, Bioconductor packages DEseq2 1.8.2 and DiffBind 1.14.6, and the Capture-C analysis pipeline (DpnII2E and CCanalyser, available at https://github.com/telenius/captureC/releases) as published by Davies et al. 2016. ChIP-seq and ATAC-seq data was mapped using an in-house analysis pipeline (http://userweb.molbiol.ox.ac.uk/public/ telenius/PipeSite.html). The R (3.2.1) plot function was used for visualisation.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The Nature Methods guidance for providing algorithms and software for publication may be useful for any submission.

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

The following antibodies were used in this study for ChIP. Prior to use in ChIP-seq in C57BL/6 mouse erythroid cells, these antibodies were validated for enrichment at known positive sites by ChIP-qPCR. All of the antibodies have previously been used and published in ChIP-seq studies.

Target, Manufacturer, Product code, Amount used per ChIP
CTCF, Millipore, 07-729, 10 μL
Rad21, Abcam, ab992, 15 μL
H3K4me3, Abcam, ab8580, 4 μL
H3K27me3, Cell signalling, #9733, 4 μL
Ezh2, Cell signalling, #5246, 4 μL
Pol II, Santa Cruz, sc-56767, 10 μL
10. Eukaryotic cell lines
   a. State the source of each eukaryotic cell line used.
      The E14TG2a line was a kind gift from Andrew Smith
   b. Describe the method of cell line authentication used.
      The cell line has been extensively authenticated by blastocyst injection
   c. Report whether the cell lines were tested for mycoplasma contamination.
      The E14TG2a line was regularly tested negative for mycoplasma
   d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.
      No commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals
   Provide details on animals and/or animal-derived materials used in the study.
   C57BL/6J mice were sourced from MRC Harwell/Charles River Laboratories. Young adult mice of both genders between 2 and 6 months of age were used in experiments. Mice were pre-treated with acetylphenylhydrazine (APH) after which mice were sacrificed and erythroid cells (ter119+) were isolated from the removed spleen.

Policy information about studies involving human research participants

12. Description of human research participants
   Describe the covariate-relevant population characteristics of the human research participants.
   The study did not involve human research participants
ChIP-seq Reporting Summary

Data deposition

1. For all ChIP-seq data:
   a. Confirm that both raw and final processed data have been deposited in a public database such as GEO.
   b. Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

2. Provide all necessary reviewer access links. (The entry may remain private before publication.)

   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97871

3. Provide a list of all files available in the database submission.

   In the ChIP-seq subset of the database submission, the following files are included (fasta and BigWig files). In the filenames below, C57 indicates WT C57BL/6 erythroid cells, DXX indicate erythroid cells of the indicated CTCF mutant:

   BigWig:
   C57_CTCF_1_RPKM.bw BigWig
   C57_CTCF_2_RPKM.bw BigWig
   C57_Ezh2_1_RPKM.bw BigWig
   C57_Ezh2_2_RPKM.bw BigWig
   C57_H3K27me3_1_RPKM.bw BigWig
   C57_H3K27me3_2_RPKM.bw BigWig
   C57_H3K4me3_1_RPKM.bw BigWig
   C57_H3K4me3_2_RPKM.bw BigWig
   C57_PolII_1_RPKM.bw BigWig
   C57_PolII_2_RPKM.bw BigWig
   C57_Rad21_1_RPKM.bw BigWig
   C57_Rad21_2_RPKM.bw BigWig
   D29_CTCF_1_RPKM.bw BigWig
   D29_CTCF_2_RPKM.bw BigWig
   D29_H3K4me3_1_RPKM.bw BigWig
   D29_H3K4me3_2_RPKM.bw BigWig
   D3839_CTCF_1_RPKM.bw BigWig
   D3839_CTCF_2_RPKM.bw BigWig
   D3839_Ezh2_1_RPKM.bw BigWig
   D3839_Ezh2_2_RPKM.bw BigWig
   D3839_H3K27me3_1_RPKM.bw BigWig
   D3839_H3K27me3_2_RPKM.bw BigWig
   D3839_H3K4me3_1_RPKM.bw BigWig
   D3839_H3K4me3_2_RPKM.bw BigWig
   D3839_PolII_1_RPKM.bw BigWig
   D3839_PolII_2_RPKM.bw BigWig
   D38_CTCF_1_RPKM.bw BigWig
   D38_CTCF_2_RPKM.bw BigWig
   D38_H3K4me3_1_RPKM.bw BigWig
   D38_H3K4me3_2_RPKM.bw BigWig
   D39_CTCF_1_RPKM.bw BigWig
   D39_CTCF_2_RPKM.bw BigWig
4. If available, provide a link to an anonymized genome browser session (e.g. UCSC).

Methodological details

5. Describe the experimental replicates.

All ChIP data generated for this study was performed at least in duplicate (on chromatin isolated from 2 animals (erythroid cells) or independent flasks (ES cells))

6. Describe the sequencing depth for each experiment.

Libraries were submitted to between 9 and 11 cycles of PCR (NEBNext kit). Insert size was verified to be 250bp by Agilent Tapestation and libraries were subsequently sequenced paired-end on the following illumina platforms and red lengths listed below. All experiments contained >50m reads of which at least 40% of reads mapped to the reference genome in all experiments.

In the filenames below, C57 indicates WT C57BL/6 erythroid cells, DXX indicate erythroid cells of the indicated CTCF mutant. PE
indicates paired-end sequencing:

C57_CTCF_1_R1.fastq  C57_CTCF_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_CTCF_2_R1.fastq  C57_CTCF_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_Ezh2_1_R1.fastq  C57_Ezh2_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_Ezh2_2_R1.fastq  C57_Ezh2_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_H3K27me3_1_R1.fastq  C57_H3K27me3_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_H3K27me3_2_R1.fastq  C57_H3K27me3_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_H3K4me3_1_R1.fastq  C57_H3K4me3_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_H3K4me3_2_R1.fastq  C57_H3K4me3_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_PolII_1_R1.fastq  C57_PolII_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_PolII_2_R1.fastq  C57_PolII_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
C57_Rad21_1_R1.fastq  C57_Rad21_1_R2.fastq  Illumina HiSeq 2500 100bp (50bp PE)
C57_Rad21_2_R1.fastq  C57_Rad21_2_R2.fastq  Illumina HiSeq 2500 100bp (50bp PE)
D29_CTCF_1_R1.fastq  D29_CTCF_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D29_CTCF_2_R1.fastq  D29_CTCF_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D29_H3K4me3_1_R1.fastq  D29_H3K4me3_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D29_H3K4me3_2_R1.fastq  D29_H3K4me3_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_CTCF_1_R1.fastq  D3839_CTCF_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_CTCF_2_R1.fastq  D3839_CTCF_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_Ezh2_1_R1.fastq  D3839_Ezh2_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_Ezh2_2_R1.fastq  D3839_Ezh2_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_H3K27me3_1_R1.fastq  D3839_H3K27me3_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_H3K27me3_2_R1.fastq  D3839_H3K27me3_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_H3K4me3_1_R1.fastq  D3839_H3K4me3_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_H3K4me3_2_R1.fastq  D3839_H3K4me3_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_PolII_1_R1.fastq  D3839_PolII_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D3839_PolII_2_R1.fastq  D3839_PolII_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D38_CTCF_1_R1.fastq  D38_CTCF_1_R2.fastq  Illumina NextSeq 80bp (40bp PE)
D38_CTCF_2_R1.fastq  D38_CTCF_2_R2.fastq  Illumina NextSeq 80bp (40bp PE)
7. Describe the antibodies used for the ChIP-seq experiments. The following antibodies were used in this study for ChIP. Prior to use in ChIP-seq, these antibodies were validated for enrichment at known positive sites by ChIP-qPCR. All of the antibodies have previously been used and published in ChIP-seq studies.

- CTCF Millipore 07-729
- Rad21 Abcam ab992
- H3K4me3 Abcam ab8580
- H3K27me3 Cell signaling #9733
- Ezh2 Cell signalling #5246
- Pol II Santa Cruz sc-56767

8. Describe the peak calling parameters. Peak calling was only performed for CTCF, H3K4me3, and H3K27me3. The MACS (Model based analysis of ChIP-seq) peak finding algorithm was used to identify regions of ChIP-seq enrichment over background in an unbiased manner. The MACS2 callpeak function was used on biological duplicate ChIP-seq data of CTCF (-q 10^-5) and H3K4me3 (-q 10^-3). For H3K27me3, the MACS2 callpeak function was used with the --broad-cutoff option (--broad-cutoff 0.05).

9. Describe the methods used to ensure data quality. ChIP enrichment was verified by qPCR prior to sequencing (>5 fold). Quality of sequencing was verified by FastQC. For CTCF, H3K27me3, and H3K4me3 MACS peakcall revealed >10000 peaks with the cutoff q-values (=FDR) as specified above. Other ChIP-seq tracks (Ezh2, RNA PolII, Rad21) were verified by visual-inspection of enrichment only. However, these data were not used for genome-wide analyses in this study, but only to show local enrichment at the alpha-globin locus.

10. Describe the software used to collect and analyze the ChIP-seq data. Several previously published software packages were used for the analysis of genome-wide ChIP-seq data. ChIP-seq data was mapped using an in-house analysis pipeline, details of which can be found online (http://userweb.molbiol.ox.ac.uk/public/telenius/ PipeSite.html). Other software used includes UCSC genome browser for visualisation, Bowtie 1.12 for alignment, Samtools 0.1.19 for sorting and duplicate removal, Bedtools 2.25.0, Deeptools 2.2.2 for normalisation and generation of BigWigs, the MEME suite 4.9.1_1 and FIMO, MACS 2.0.10 for peak calling, Bioconductor package DiffBind 1.14.6 to investigate differential peak occupancy in H3K4me3 and H3K27me3 ChIP.