Defining genome architecture at base pair resolution

Peng Hua[1], Mohsin Badat[1*], Lars L. P. Hanssen[1*], Lance Henteges[1,2], Nicholas Crump[1], Damien J. Downes[1], Danuta M. Jeziorska[1], Marieke A. Oudelaar[1,2], Ron Schwessinger[1,2], Stephen Taylor[2], Thomas A. Milne[1], Jim R. Hughes[1,2], Doug R. Higgs[1] and James O. J. Davies[1]

Affiliations:

1. MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK
2. MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK

* These authors contributed equally

Abstract

Complex gene regulation is one of the key requirements for the evolution of higher eukaryotes.[1] In these organisms, many genes are regulated by enhancers that are 10^4-10^6 base pairs (bp) distant from the promoter. Enhancer sequences usually contain multiple small transcription factor binding sites (typically ~10bp), and physical contact between the promoter and enhancer is thought to be required to modulate gene expression.[2] Current methods have extensively defined chromatin architecture at scales above 1 kb but until now it has not been possible to define physical contacts at the scale of the key proteins determining gene expression. Here we define the interactions between different classes of regulatory elements (enhancers, promoters and boundary elements) in unprecedented detail, using a novel chromosome conformation capture method (Micro Capture-C (MCC)), which allows physical contacts to be determined at base-pair resolution. We find that highly punctate contacts occur between enhancers, promoters and CCCTC-binding factor (CTCF) sites and we show, using base pair resolution plots of ligation junctions, that transcription factors generate a key component of the contacts between enhancers and promoters. Our data show that contacts from CTCF sites highly correlate with co-occupancy of cohesin and that interactions between CTCF sites are increased when active promoters and enhancers are located within the intervening chromatin. We also find that promoters make the strongest contacts with both enhancers and CTCF sites and that while CTCF sites contact promoters strongly they only make weak contacts with enhancers. The highly punctate nature of the contacts is an unexpected finding because the current view is that physical contacts are constrained by much larger domains such as topological associated domains (TADs).[3] Our results support a model in which chromatin loop extrusion[4-6] is dependent on cohesin loading at active promoters and enhancers, explaining the formation of tissue-specific chromatin domains without changes in CTCF binding. The data suggest that a separate mechanism to loop extrusion underlies enhancer-/promoter contacts, which likely involves DNA binding proteins at enhancers and promoters. The unprecedented

resolution obtainable with MCC is likely to transform the way in which we understand the 3D structure of chromatin.

Main

The macroscopic structure of the genome has been defined by a variety of methods including microscopy, chromosome conformation capture (3C) and ligation free approaches.[7] This has shown that the nucleus is highly structured with a number of layers of organisation at different scales, ranging from chromosomal territories and TADs[8,9] down to smaller scale contact domains between regulatory sequences. These approaches have successfully defined large and medium scale chromatin architecture, but to date it has not been possible to define physical contacts at the scale of the key proteins determining gene expression. This is because proteins such as transcription factors bind short DNA sequences (~7-22 bp),[10] which is well below the kilobase resolution attainable with conventional 3C techniques.[11] In order to fully understand gene expression, it is important to define interactions at this scale, which is impossible with the resolution of currently available imaging techniques and 3C methods. We therefore developed a new high resolution 3C method (Micro Capture-C (MCC)), which allows us to determine the physical contacts between regulatory elements at the resolution of individual transcription factors, and thereby provides unique insights into gene regulation.

MCC is able to map physical interactions at base pair resolution. This is a marked increase in resolution compared to all other available 3C methods including Hi-C,[3] Micro-C,[12-16] Promoter Capture Hi-C,[17] 4C[18] and NG Capture-C[19] (Fig. 1). This was achieved by combining five advances with the NG Capture-C[19] method (Extended Data Fig. 1), which currently provides the greatest sensitivity and resolution 3C data from individual viewpoints.[20] First, micrococcal nuclease (MNase) was used in place of restriction enzymes. This fragments the genome largely independently of DNA sequence[12-14,21] and we found that sub-nucleosome detail could be resolved by titrating MNase to maintain inter-nucleosomal linkers (Extended Data Fig. 2a). Second, we found that the resolution could be substantially improved through minimising the disruption of nuclear architecture by using intact cells permeabilised with digitonin,[22] whereas previous 3C methods have largely used chromatin in solution or purified nuclei. Third, we generated extremely deep data from individual viewpoints (up to 500,000 unique contacts (mean 140,000) per 120 bp viewpoint, Extended Data Table 1). This equates to over 1000-fold the depth of data obtained with “all vs all” approaches such as Hi-C and Micro-C (over 3 trillion ligation junctions would be required for this depth of coverage genome wide). Fourth, we generated contact maps with base pair accuracy by directly sequencing the ligation junctions between reads at different sites. This was achieved by reconstructing single reads from paired end sequencing by sonicating the MNase 3C library to 200 bp fragments and sequencing with 300 bp reads (Extended Data Fig. 1). Fifth, a new analysis pipeline was developed to allow ligation junctions to be precisely located and protein-protein interactions to be reconstructed (Extended Data Fig. 1). As with other Capture-C approaches, MCC

allows the simultaneous, parallel study of a large number of viewpoints in a single experiment. Genome-wide profiles were generated from 330 viewpoints, including promoters, enhancers and CTCF sites in primary murine erythroid and embryonic stem (ES) cells (6 replicates of each cell type (3 biological x 2 technical)). The data are highly reproducible and sequencing of the MNase digest controls showed cutting across the genome and no bias in favour of hypersensitive sites (Extended Data Fig. 2b&c). Sequencing of the uncaptured Hi-C like MNase 3C libraries showed no clear sequence biases but minor evidence of biases towards hypersensitive sites was seen, which is likely due to variability in ligation rather than MNase cutting. Importantly, correction for this effect does not appreciably alter the contact profile or peaks detected (Extended Data Fig. 2b-f, Extended Data Table 2).

The increased resolution afforded by MCC delineated a number of novel, biologically important findings. First, it showed that extremely localised contacts occur between active gene promoters and enhancers at all genes studied, including very well characterised loci such as the globin genes (*Hba-a1/Hba-a2* and *Hbb-b1*/*Hbb-bs*) in mouse erythroid cells and *Nanog, Pou5f1 (Oct4), Klf4, Myc and Sox2* in mouse ES cells, (Fig.2 and Extended Data Figs. 3&4). The promoter contacts identified the location of enhancers with near base pair precision. By performing base pair mapping of ligation junctions we were able to detect the DNA sequences bound by proteins such as transcription factors in a similar fashion to DNase I hypersensitivity footprinting data (Fig. 1f, Extended Data Fig. 1).[23] These results show that it is possible to identify the precise location of the sequences bound by regulatory proteins in the enhancers controlling a specific promoter using a single technique.

To further characterise the potential importance of these interactions, at 25 promoters we included viewpoints 1 kb up or downstream from the promoter, which showed a more generalised signal, with less prominent contacts with the enhancers (Extended Data Fig. 3a-d). This was confirmed genome wide through analysis of the unenriched Hi-C like data (Extended Data Fig. 3e). The localised contacts were also highly tissue-specific; being absent in ES cells at genes only expressed in erythroid cells, and conversely, absent in erythroid cells at ES cell specific genes (Extended Data Figs. 3&4). In addition, inactive genes such as the fetal *Hbg* gene at the beta globin locus did not make contacts with the active enhancers despite being interposed between the active promoter and enhancer. In addition, reciprocal contacts with promoters were seen from enhancers (Extended Data Fig. 4c). At the *Myc* and *Sox2* loci, which are well characterised loci with long-range (~1 Mb) contacts, we were able to define highly punctate contacts with enhancers and CTCF binding sites which aligned precisely with the known location of these elements (Fig. 2, Extended Data Fig. 4).

With the increased resolution, we were able to define contacts with enhancers located very close to the promoter (<2 kb), which have previously been impossible to visualise (at the *Pou5f1 locus,* for example). We were also able to separate signals from contacts between regulatory elements in very close proximity such as at the erythroid

gene *Pnpo*, at which we have previously described a strain specific enhancer that is located 1 kb upstream of the promoter of the adjacent gene *Cdk5rap3* (Fig. 2).

We were also able to study contacts from the promoter to the gene body. It has previously been postulated that active genes form a loop between the 3' end of the gene and the promoter to facilitate unidirectional transcription and re-initiation.[24] However, we found no evidence of contacts between the promoter and the 3'-UTR at any of the 30 promoters we studied unless other elements, such as enhancers or CTCF sites were in proximity to the 3'-UTR (Extended Data Fig. 4e&f).

The increased resolution allowed us to interrogate gene dense loci that have previously been difficult to characterise with 3C methods. The promoter of the erythroid transcription factor *Klf1* was found to contact at least 15 other promoters and enhancers in its vicinity (Fig. 2) and other promoters were seen to contact more than 20 surrounding promoters (Extended Data Fig. 5). This shows that promoters colocalise very specifically in gene dense regions.

CTCF has an important role in determining genome structure[7] and we therefore went on to study the interaction profiles from CTCF binding sites. Surprisingly, we found highly specific interactions between CTCF binding sites, which were dependent on the cell type (Fig. 3a, Extended Data Figs. 6&7). For example, at the alpha globin locus the well-studied HS-38 CTCF site[25] interacts strongly with CTCF binding sites between the promoters of the *Hba-a1&2* genes in erythroid cells, when both the enhancers and promoters are highly active. These contacts are absent in ES cells when the locus is inactive despite similar levels of CTCF binding in the locus (Fig. 3a). In erythroid cells, very few contacts were seen between the HS-38 CTCF site and the enhancers of the gene despite the proximity of these hypersensitive sites; demonstrating that the method does not simply report contacts with the nearest open chromatin regions. Similar tissue-specific increases in CTCF to CTCF contacts were seen at many other loci in both erythroid and ES cells upon activation of promoters or enhancers in the intervening chromatin (Extended Data Figs. 6&7) and there was a correlation between contacts between CTCF sites and interspaced active enhancers and promoters as measured by H3K27ac (Extended Data Fig. 9a).

To determine whether specific CTCF-CTCF interactions correlate with cohesin we generated data from 20 CTCF sites classified as having high or low co-occupancy of RAD21 (a component of cohesin) by ChIP-seq. We find that when cohesin is not present significant peaks of interaction virtually do not occur between CTCF bound sites (Fig. 3b, Extended Data Fig. 8). In keeping with previous studies,[4,26,27] the interaction frequency was heavily determined by the relative orientation of CTCF binding sites, with greater numbers of contacts between convergent sites than divergent sites or sites in the same orientation (Fig. 3c). Metaplots of junctions between CTCF sites confirm this orientation dependence and show that there are

consistently strong contacts with the neighbouring nucleosomes when the CTCF sites are in a convergent orientation (Extended Data Fig. 9b).

The data support a model of loop extrusion whereby activation of promoters and enhancers results in increased extrusion activity through increased loading of cohesin (Extended Data Fig. 9f). To support of this, we demonstrated that the cohesin loader NIPBL is enriched at active promoters and enhancers in comparison to CTCF sites (Fig. 3 d&e, Extended data Fig. 9c-e). This provides an explanation for the generation of tissue-specific chromatin interactions by a general, tissue-invariant extrusion mechanism. This model explains how it is possible to have tissue specific changes in chromatin structure that, are defined by CTCF, without alterations in CTCF binding.

It has previously been challenging to identify peaks of interaction in 3C data sets systematically without using complex modelling algorithms, which makes it difficult to perform unsupervised analysis of these data sets. The punctate nature of MCC data meant that it was possible to use conventional peak callers such as MACS2.[28] We also analysed data with a customised Poisson model of the background data as well as a deep neural network based approach (LanceOtron[29]). Irrespective of the method used, highly significant peaks of interaction were demonstrated (Extended Data Fig. 10a, Extended data table 2). At the majority of genes all of the significant peaks were contained within the TAD and overall 87% of all ligation junctions in *cis* were located within the local TAD (Extended Data Fig. 10b&c). Using these approaches combined with annotation using GenoSTAN[30] we were able to show that promoters preferentially contact enhancers and other promoters (Fig. 3f), whilst CTCF sites preferentially contact other CTCF sites and promoters (Fig. 3g). This suggests that enhancer-promoter contacts are at least in part mediated or maintained by a different mechanism to CTCF / cohesin mediated interactions.

Based on our data we hypothesise that contacts from CTCF sites to other CTCF sites and promoters are mediated predominantly via cohesin. By contrast, additional mechanisms that involve regulatory DNA-binding proteins may stabilise and maintain contacts between enhancers and promoters.

Using MCC it is possible to increase resolution by mapping ligation junctions between fragments rather than whole reads. Using this approach, we were able to identify transcription factor and CTCF binding footprints in a similar fashion to DNase I hypersensitivity footprinting (Fig. 1f, Fig. 4, Extended Data Fig. 1c). Furthermore, we could identify the binding sites of proteins contributing to chromatin interactions in greater detail by generating separate profiles depending on whether the read containing the transcription factor binding site was up or downstream of the junction, analogous to ChIP nexus[31] (Fig. 4, Extended Data Fig. 1c).

We went on to derive single base pair maps of ligation junctions between different CTCF sites and between promoters and enhancers (Fig. 4b, Extended Data Fig. 10a-

b). Superimposition of the directionality of the reads onto these maps allowed the position of the protein leading to the contact to be determined (Fig. 4c) and it was possible to generate maps of the protein-protein contacts at high resolution between viewpoints and interacting elements (Fig. 4d, Extended Data Fig. 1c). At CTCF binding sites the data show that protein binding at the core motif is a key determinant of the physical interactions and that strong contacts occur with the surrounding nucleosomes (Fig. 4d&i). The contacts between enhancers and promoters generated complex patterns of interactions (Fig. 4e-h&j, Extended Data Fig. 10d&e) compared to CTCF binding sites, which correlated well with transcription factor binding motifs; providing evidence that there may be dynamic contacts or multiple interaction states.

We found that the patterns of footprints altered in a cell type specific manner by capturing from promoters which are active in both erythroid and ES cells, which use the same enhancer sequences in both cell types (Fig. 4k-m, Extended Data Fig. 10e-g). In these reused enhancers we found clearly different footprinting patterns, suggesting that contacts are dependent on different cohorts of transcription factors in the two cell types (Fig. 4l&m, Extended Data Fig. 10f-h). To examine the specificity and functional importance of these protein/protein interactions we went on to show that the strength of interactions is significantly reduced when a 2-4bp sequence was deleted to disrupt an NF-E2 binding site at the main enhancer at the alpha globin locus (Extended Data Fig. 11). Deletion of this sequence results in reduced NF-E2 binding at the enhancer and transcription of the alpha globin genes.

In summary, we have defined 3C interactions in mammalian cells at unprecedented, base pair resolution. We show that extremely well defined, highly punctate physical contacts occur precisely between active promoters and enhancers in a tissue-specific manner. Single base pair resolution plots of ligation junctions delineate transcription factor binding sites, which shows that the proteins directly binding DNA at these sites are important for mediating physical contacts between enhancers and promoters. Extremely well-defined contacts also occur between CTCF sites which have co-occupancy of cohesin, but sites with little cohesin show very few interactions with anything other than the immediate surrounding chromatin. Our data clearly show that the relative orientation of CTCF sites is a key determinant of physical contacts between CTCF sites. Interestingly there is discordance in the interactions between different classes of element. Promoters contact enhancers and CTCF strongly and CTCF sites contact other CTCF sites and promoters but do not appear to contact enhancers strongly. This suggests that enhancer-promoter contacts are maintained by a separate mechanism to loop extrusion. Overall these data support the loop extrusion model but they also show that additional mechanisms involving the DNA binding proteins, such as transcription factors, play an important role in enhancer-promoter contacts. The high resolution and signal to noise ratio achievable with MCC is a very significant advance in contact mapping technologies and it is likely to redefine the way in which we view chromatin structure. It is likely to facilitate our understanding of how genes are controlled by regulatory sequences in the non-coding genome.

Main Figures

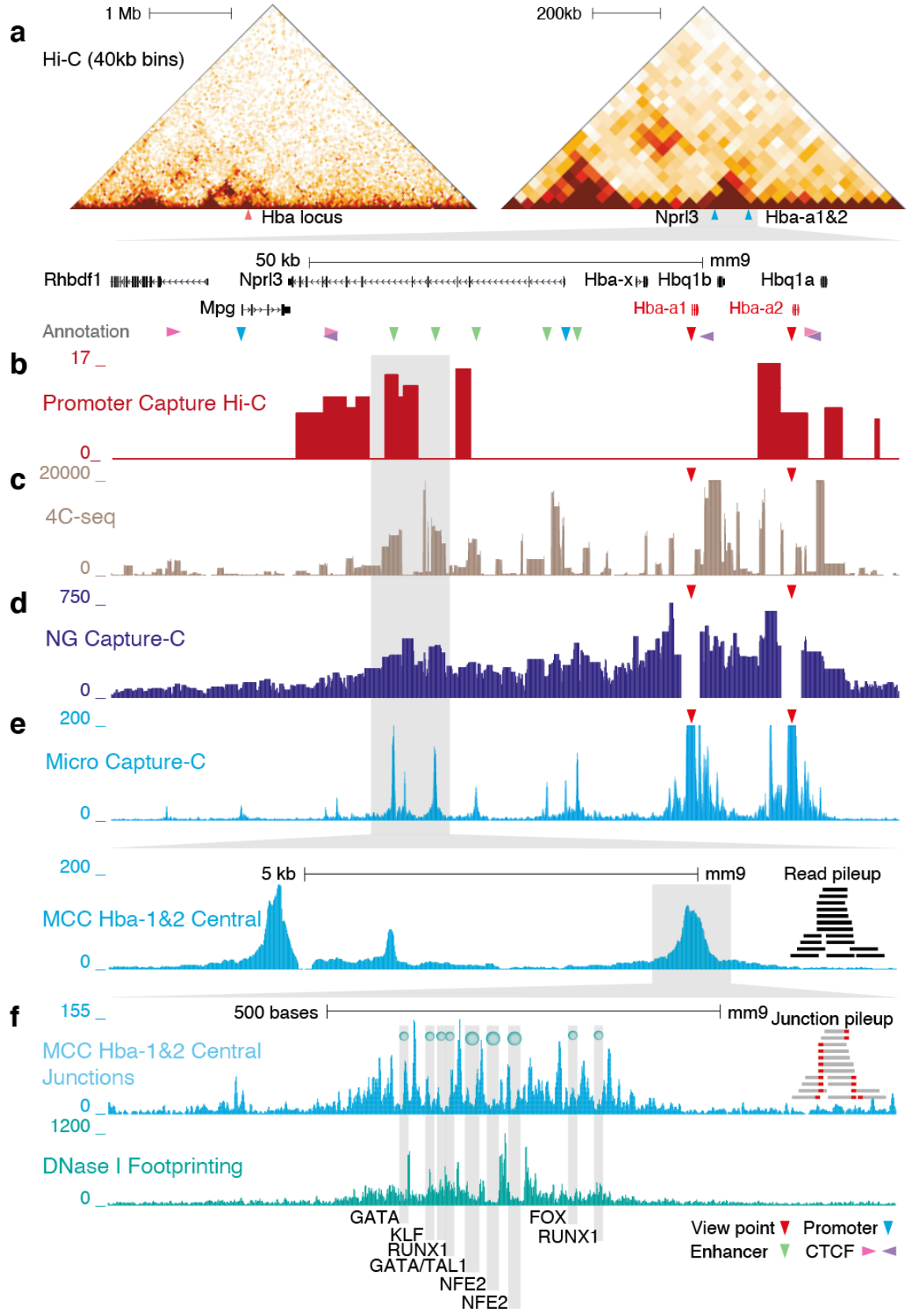


Fig. 1 | Comparison of Micro Capture-C (MCC) with other 3C techniques in erythroid cells from the promoters of the alpha globin genes (*Hba-a1&2*) showing the profoundly increased resolution afforded by the new method. a, Hi-C data of the region surrounding the locus (40 kb resolution).[32] **b,** Promoter Capture Hi-C[17] (output from GOTHiC) **c,** 4C-seq[18] and **d,** NG Capture-C[19] plotted to individual DpnII restriction fragments without windowing **e,** MCC read pileup without windowing. Data was normalised to total number of unique reads across the genome. **f,** Plots of the precise location of ligation junctions (number of unique junctions +/-1 bp) allow footprinting of DNA binding proteins to be derived, similar to DNase I digestion; the sites of protein binding can be determined because they protect the DNA from cleavage by MNase. The Y-axis shows number of unique reads; profiles for 4C / NG Capture-C and MCC are a composite from *Hba-a1&2* promoters, which are identical.

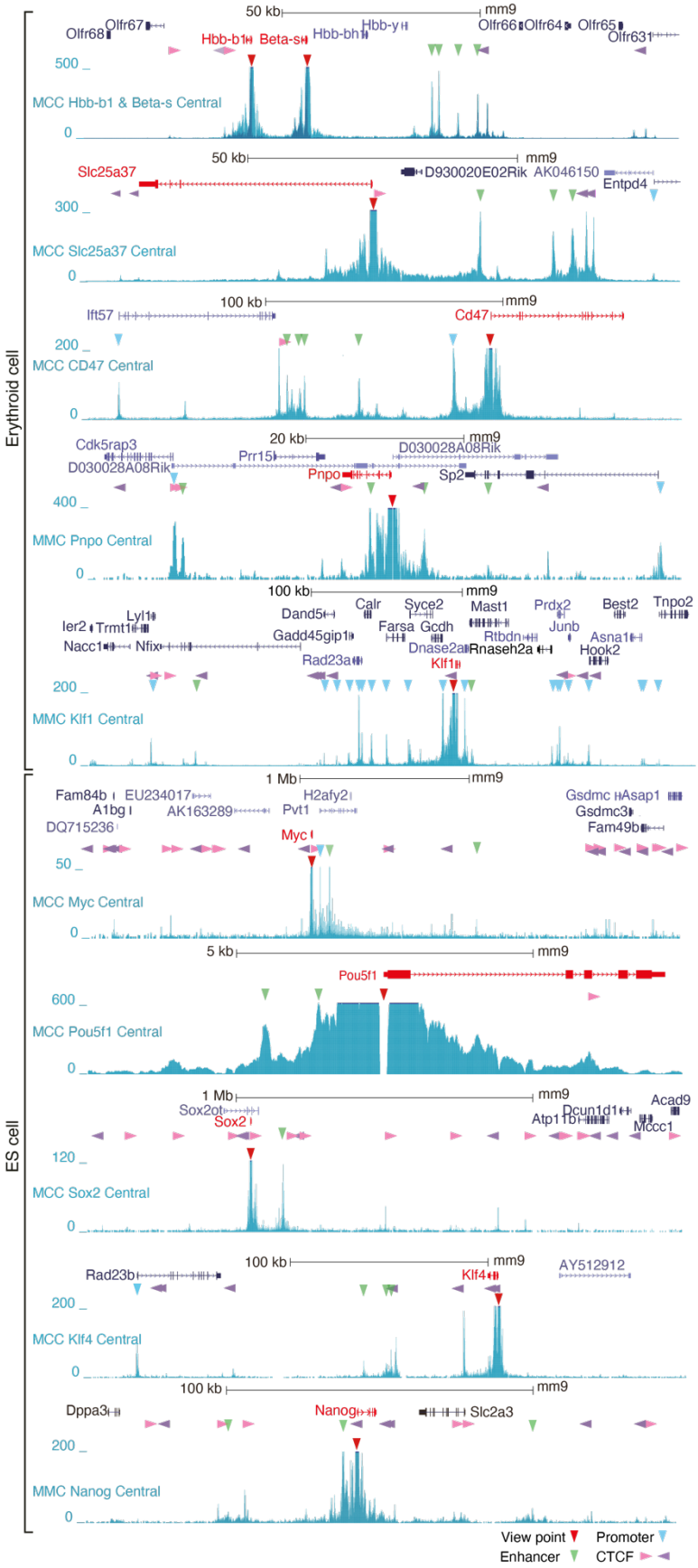

Erythroid cell
ES cell
50 kb
mm9
Olfr68
Olfr67
Hbb-b1
Beta-s
Hbb-bh1
Hbb-y
Olfr66
Olfr64
Olfr65
Olfr631
500
MCC Hbb-b1 & Beta-s Central
0
50 kb
mm9
Slc25a37
D930020E02Rik
AK046150
Entpd4
300
MCC Slc25a37 Central
0
100 kb
mm9
Ift57
Cd47
200
MCC CD47 Central
0
Cdk5rap3
20 kb
mm9
D030028A08Rik
Prr15l
D030028A08Rik
Pnpo
Sp2
400
MMC Pnpo Central
0
100 kb
mm9
Lyl1
Dand5
Calr
Syce2
Mast1
Prdx2
Best2
Tnpo2
Ier2
Trmt1
Gadd45gip1
Farsa
Gcdh
Junb
Nacc1
Nfix
Rtbdn
Asna1
Dnase2a
Rnaseh2a
Hook2
Rad23a
Klf1
200
MMC Klf1 Central
0
1 Mb
mm9
Fam84b
EU234017
H2afy2
Gsdmc
Asap1
A1bg
AK163289
Pvt1
Gsdmc3
DQ715236
Fam49b
Myc
50
MCC Myc Central
0
5 kb
mm9
Pou5f1
600
MCC Pou5f1 Central
0
1 Mb
mm9
Acad9
Sox2ot
Dcun1d1
Atp1b1
Mccc1
Sox2
120
MCC Sox2 Central
0
100 kb
mm9
Rad23b
Klf4
AY512912
200
MCC Klf4 Central
0
100 kb
mm9
Dppa3
Nanog
Slc2a3
200
MMC Nanog Central
0
View point
Promoter
Enhancer
CTCF

Fig. 2 | MCC defines highly specific contacts between promoters and enhancers at many well characterised loci. These include the beta globin locus (*Hbb-b1* and *Beta-s*) in erythroid cells and *Myc*, *Sox2* in ES cells, where highly specific long range contacts can be seen with enhancers up to 1 Mb from the promoter. The profile at *Pnpo* shows that the contact with a strain-specific enhancer we have previously described,[33] can be separated from the adjacent promoter of *Cdk5rap* 1 kb upstream. At *Pou5f1* a super-enhancer is located 2 kb upstream of the promoter, which is too close to distinguish by conventional 3C methods. The track for *Klf1* shows that the promoter at this gene dense locus forms distinct contacts with at least 15 surrounding promoters.

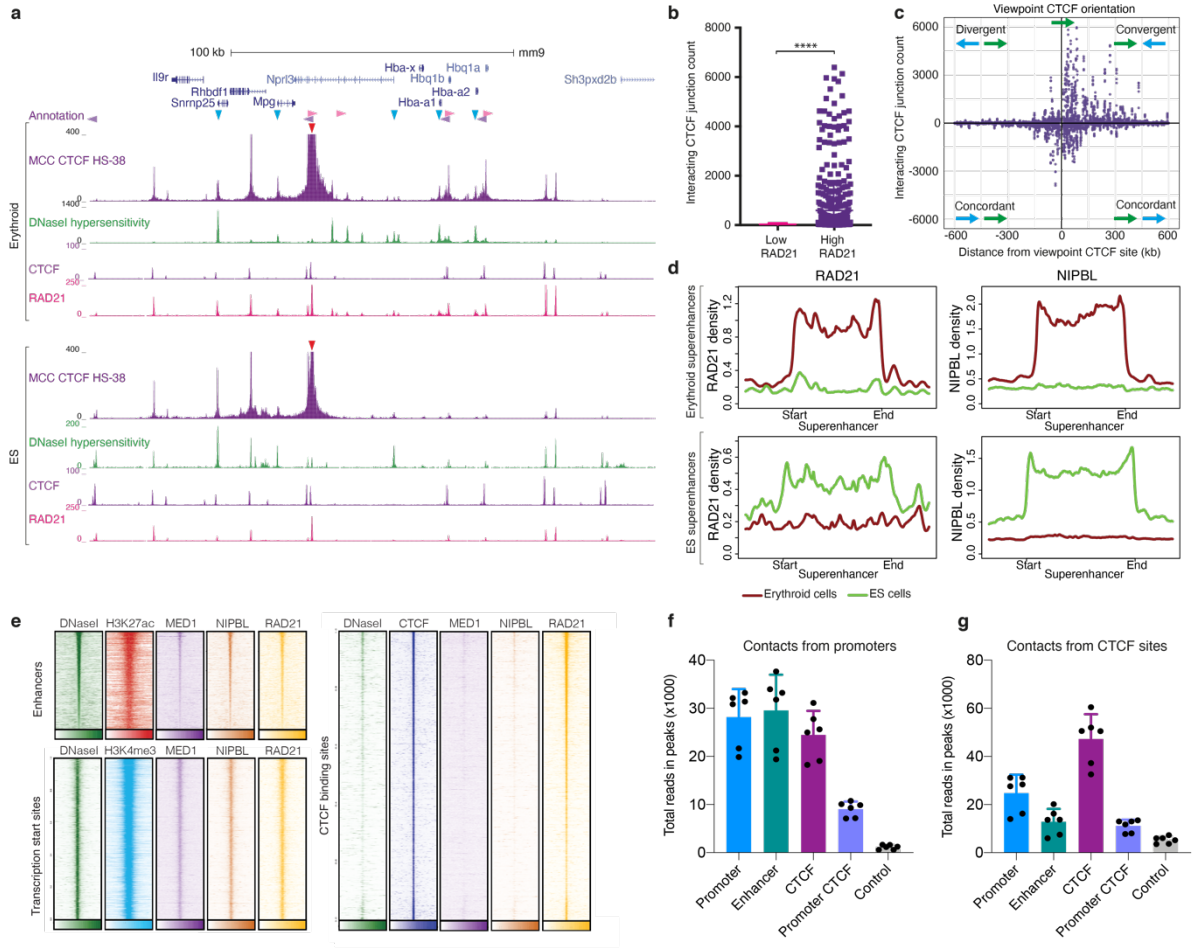


Fig. 3 | Contact profiles from the main CTCF boundary element (HS-38) at the alpha globin locus alter with changes in gene expression a, MCC data, DNase I hypersensitivity and ChIP-seq for CTCF and the cohesin component RAD21 in erythroid cells in which the gene is active and ES cells in which the gene is not expressed. **b,** Plot of the number of ligation junctions between CTCF sites called from ChIP-seq data within the 1.2 Mb surrounding the viewpoint CTCF site, with high or low co-occupancy of cohesin quantified by RAD21 ChIP-seq (erythroid cells, 6 replicates, 10 viewpoints, CTCF sites analysed low RAD21 n=468; high RAD21 n=498; **** p<0.0001). **c,** Effect of CTCF orientation on ligation junction counts. The x-axis displays distance relative to the viewpoint corrected for orientation of the CTCF at the viewpoint; the y-axis displays junction count corrected for the orientation of the interacting CTCF site; 6 replicates, 82 viewpoints, n=14010. **d,** Metaplots of RAD21 and NIPBL binding density (RPKM) at erythroid and ES cell specific superenhancers in erythroid (red) and ES (green) cells, showing high levels of enrichment of both proteins at active superenhancers compared to the same sequence in inactive cell types[34,35]. **e,** Heatmaps of 10kb genomic regions surrounding promoters, enhancers, or CTCF binding sites showing DNase I hypersensitivity and ChIP-seq data for H3K27ac, H3K4me3, mediator (MED1), NIPBL, RAD21 and CTCF. The chromatin loader NIPBL is highly enriched at enhancers and promoters compared to CTCF sites.

f&g, Analysis of the strength of contacts from promoters with different classes of elements as categorised by GenoSTAN showing that promoters preferentially contact enhancers whereas CTCF sites preferentially contact other CTCF sites. Total numbers of reads in the 1kb region surrounding different classes elements within 400 kb of the viewpoint (mean, SD).

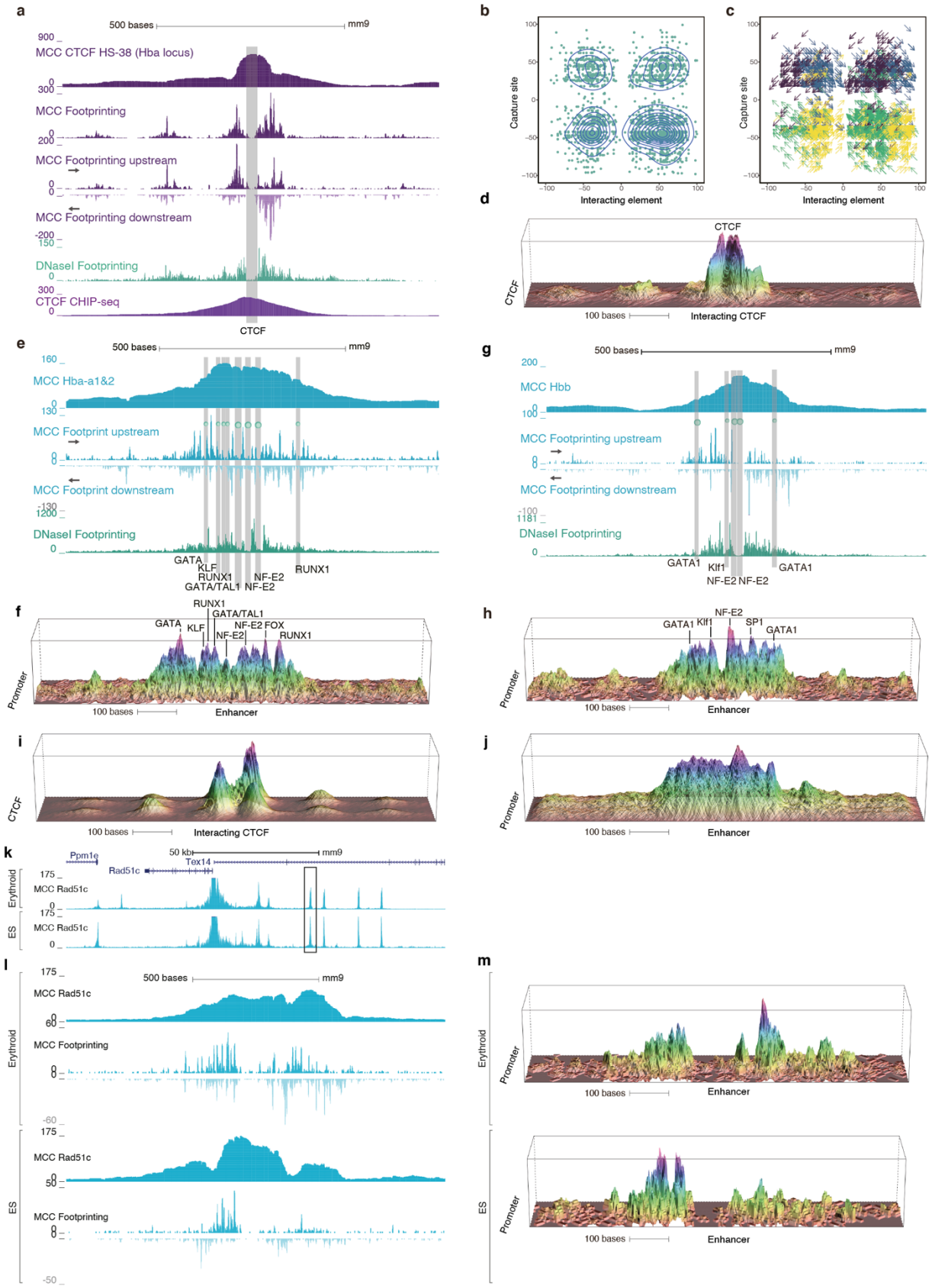
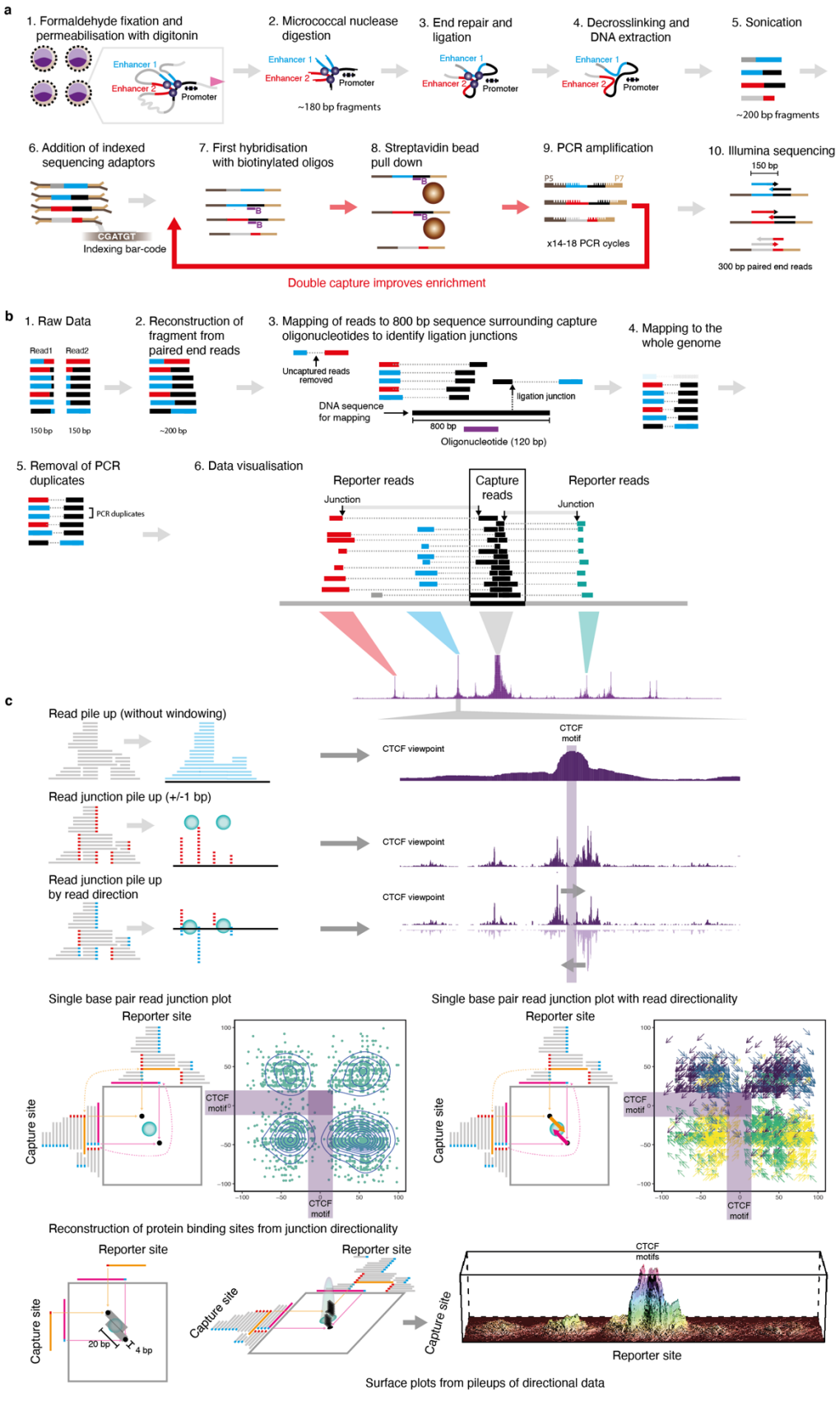


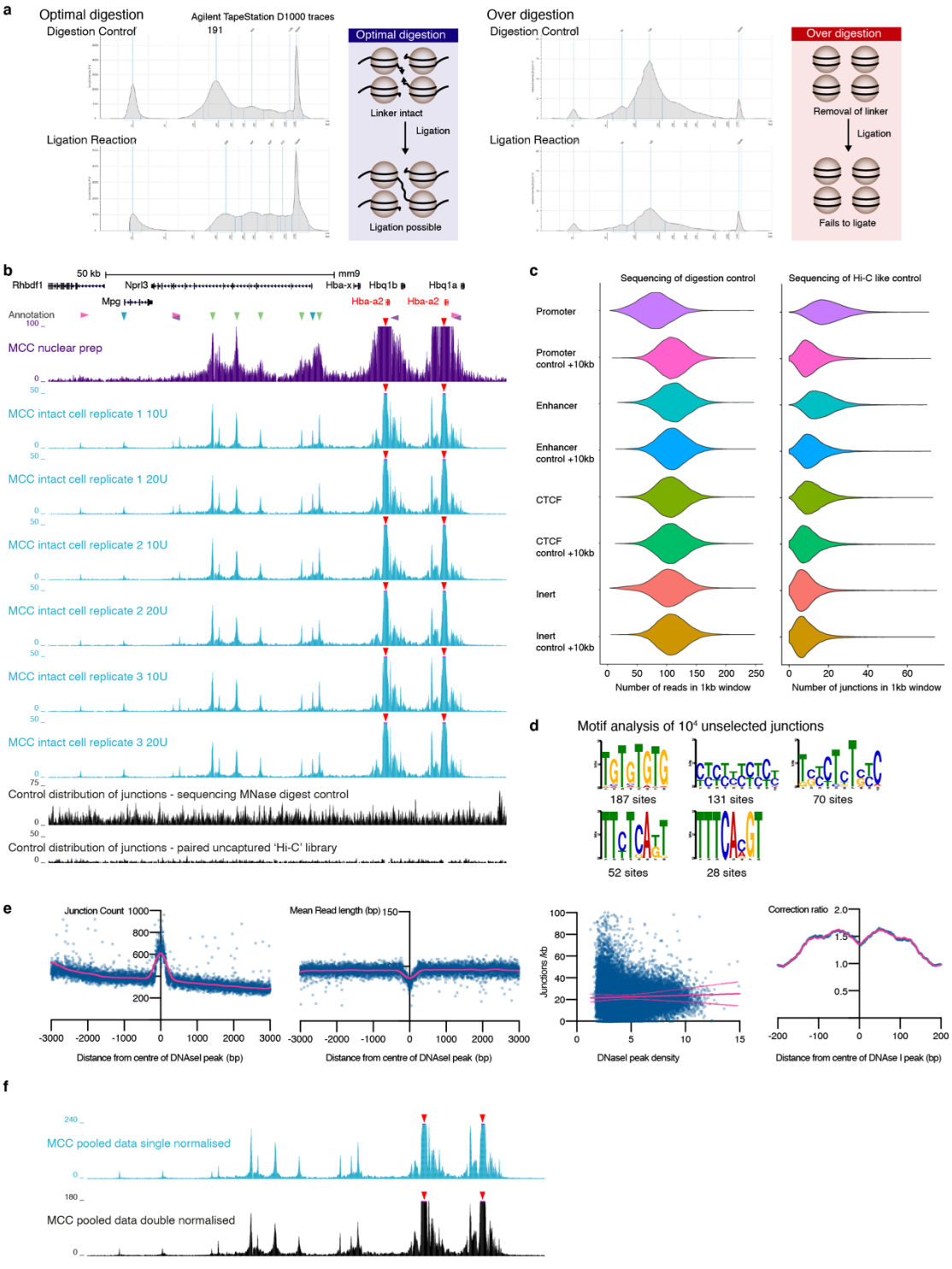
Fig. 4 | Single base pair resolution analysis of MCC ligation junctions. a, Fine details of the contacts between two CTCF sites at the alpha globin locus (HS-38 and HS-56). Protein binding sites can be identified as footprints from single base pair plots

of ligation junctions because the protein binding prevents MNase digestion. Localisation of protein binding can be improved by separating profiles depending on whether the ligation junction is upstream or downstream of the read and therefore likely protein bound DNA sequence giving rise to the contact (Extended Data Fig. 1c). At CTCF binding sites this highlights the central consensus sequence as the site leading to inter CTCF contacts. **b,** Single base pair resolution plots of ligation junctions between two CTCF sites showing that the junctions surround the central binding motif, which is protected from MNase digestion. **c,** Shows the direction of the reads resulting in the junctions (Extended Data Fig. 1) **d,** Reconstruction of position of DNA binding proteins at the capture site and interacting site showing the central position of the two CTCF sites is critically important for binding. Data from the main enhancer at alpha globin (**e&f**) and beta globin (**g&h**) showing that much more complex patterns of ligation junctions occur compared CTCF sites. The reconstructions of the protein binding sites correlates well with key transcription factor binding sites identified by JASPAR and ChIP-seq (Extended Data Fig. 9c&d).[10] **i&j**, Show reconstruction metaplots of data from the strongest 150 CTCF-CTCF interactions and 65 enhancer-promoter interactions. **k,** To show that MCC could detect subtle changes in chromatin architecture, we captured from gene promoters that are active in erythroid and ES cells, which use the same enhancers. **l&m,** MCC at the shared enhancers was able to delineate different patterns of footprinting in erythroid and ES cells.

Extended Data

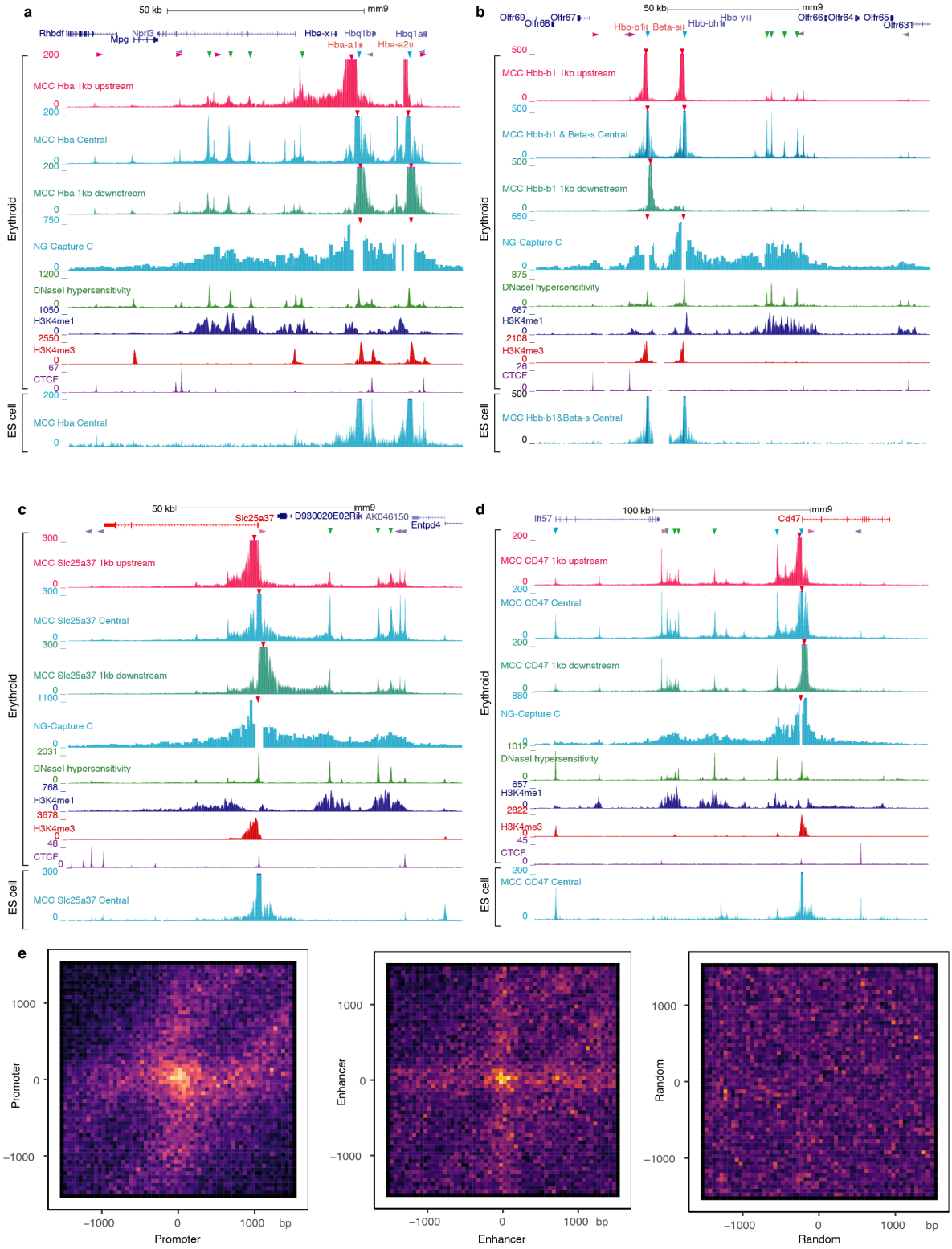


Extended Data Fig. 1 | Overview of experimental and computational workflow. a, Cells are initially fixed with formaldehyde and then permeabilised with digitonin. They are subsequently treated with MNase at different concentrations. An end repair and ligation reaction are then performed. This results in ligation of sequences in close proximity in the nucleus. DNA is then extracted to make an MNase 3C library. This is sonicated to a fragment size of ~200 bp. Illumina sequencing adaptors are added to the library. This library manufacture process is scaled up to maximise the amount of DNA available and complexity of the libraries. Multiple samples with different sequencing indices are then mixed. The DNA is then denatured and mixed with a pool of biotinylated oligonucleotides. These 120mer oligonucleotides were designed to capture the central portion of the hypersensitive site at the promoter or the central sequence of CTCF sites guided by a combination of motif analysis and DNase I footprinting. Following a hybridisation reaction, a streptavidin bead pull down is performed and the uncaptured material is washed away. The material is PCR amplified and the oligonucleotide capture repeated to improve the purity. The reads are then sequenced with 300 bp paired end reads, which allows the entire sequence of each read to be determined since the DNA is fragmented to 200 bp by sonication. **b,** Shows the overview of the data analysis. The raw fastq file is processed to reconstruct a single read from the paired end sequencing. This is then mapped to the 800bp sequence surrounding the capture oligonucleotide using the non-stringent aligner BLAT. This allows the reads to be cut into 'slices' depending on whether they align to the sequence around the capture site. This strategy allows ligation junctions in the read to be determined with base pair accuracy. The resulting fastq file is aligned to the genome using bowtie. This file is processed to remove PCR duplicates and the junctions of the 'slices' within the reads are identified. **c,** Shows the different methods used for data visualisation. Simple read pileups are generally used. However, the resolution can be further increased by reporting the precise base pair position of the ligation junctions. Since protein binding protects against DNA digestion by MNase, the regions of protein binding can be inferred from footprints in the junction plots similar to DNase I footprinting. More detailed localisation of the protein binding site that results in the interaction can be achieved by separating the junction profiles based on whether the read and therefore protein binding site is upstream of the junction or downstream of the junction. Finally, single base pair resolution maps of junctions between the capture site and the peaks at regulatory elements can be generated. In the example above the central binding site of the two interacting CTCF sites is protected and the ligation junctions surround this. The direction of the reads at the capture and reporter sites can be used to identify the site of the proteins giving rise to the ligation junctions. This can be plotted with arrow plots. Here this shows that the central CTCF motif at both the capture and reporter site are the origin of the contacts between the two sites. This is more easily visualised using 3D surface plots. These were constructed by converting each data point into a rectangle 20 bp long and 4 bp wide in the direction of the reads giving rise to the interactions. This shows the central binding site of the two CTCF sites giving rise to the interactions and contacts between these central CTCF motif and the neighbouring nucleosomes.



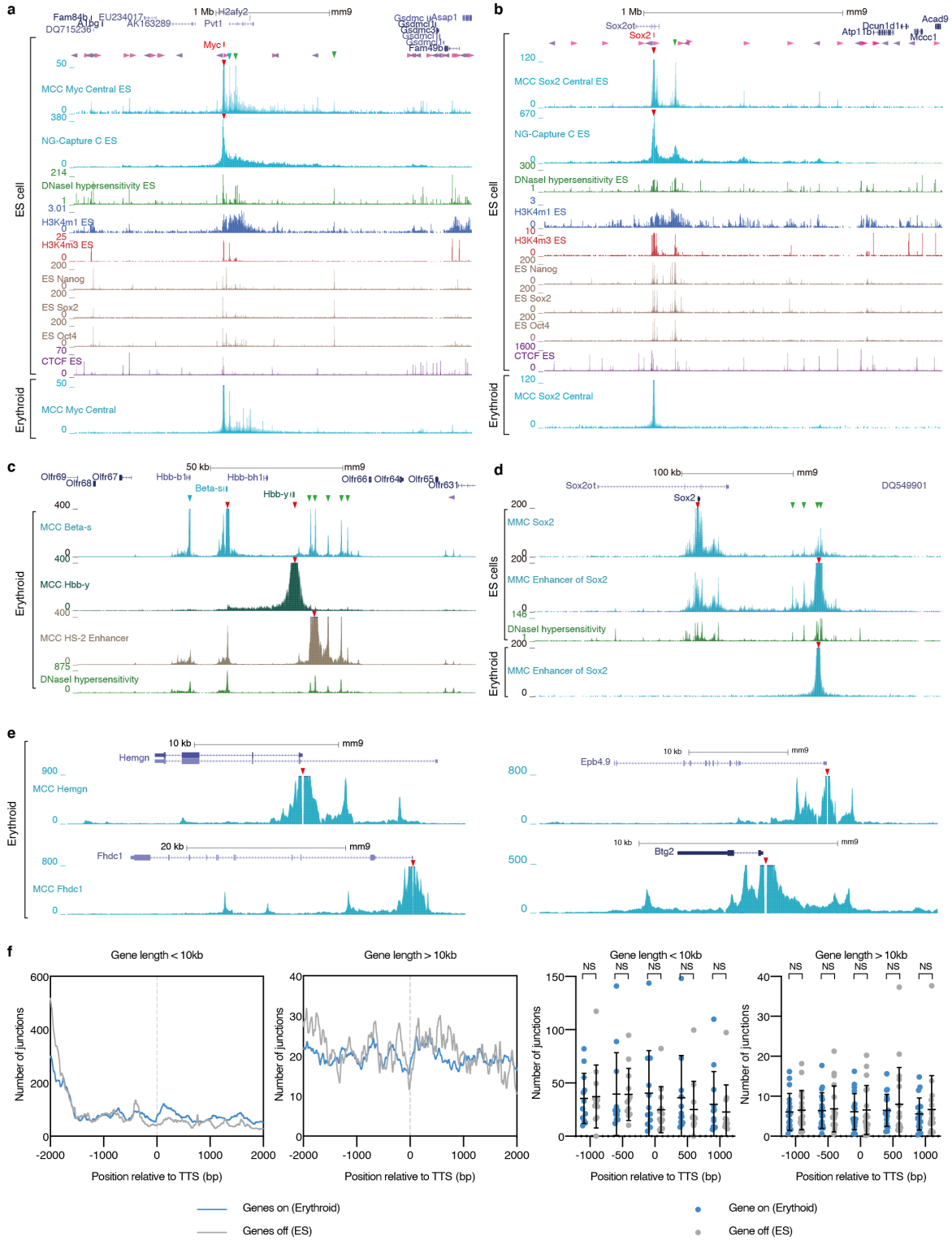
Extended Data Fig. 2 | Technical details of library preparation, reproducibility and biases. a, Optimal MNase 3C library digestion keeps the nucleosome tails intact whereas over digestion leads to loss of nucleosome linkers. This results in a failure of fragments to religate. Note that the digestion controls show that MNase cuts chromatin into mononucleosomes and that there are very few fragments over 1000bp. The fragment size considerably smaller than DpnII digested chromatin. **b,** MCC profiles

from the *Hba-a1&2* promoters in erythroid cells showing interaction profiles derived from MNase-based 3C library preparation using a conventional NP-40 based nuclear extract, compared with data generated from intact cells permeabilised with digitonin. Data from 3 biological replicates from two different concentrations of MNase are shown. Counts are normalised to the total number of reads across the genome. The assay is highly reproducible between replicates. In order to look for biases caused by MNase digestion we sequenced the digestion controls. In addition, we sequenced the Hi-C like 3C library without oligonucleotide capture to look for biases resulting from the ligation reaction. The global distribution of reads from the MNase digestion and the ligation junctions from the uncaptured library shows a very similar distribution to background (bottom two panels), without obvious biases towards the hypersensitive sites **c,** violin plots from genome wide analysis at different classes of element show the number of reads from sequencing the MNase digestion controls and ligation junctions in the uncaptured Hi-C like MNase 3C library in a 1 kb window around different classes of element compared to control regions 10 kb downstream of the element. Sequencing of the digestion controls shows no evidence of biases in MNase digestion at enhancers or CTCF sites. There is a small reduction in the number of reads at promoters, possibly due to loss of smaller fragments from histone depeleted regions in the DNA extraction and sequencing library preparation. Conversely sequencing of the ligation junctions reveals a slightly higher numbers of junctions at promoters and regulatory elements including CTCF sites, which is likely due to the ligation process. **d,** Analysis of the DNA sequence at ligation junctions detected no biases towards ligation junctions AT rich sequences (which MNase is reported to cut preferentially). **e,** Metaplots of the junction count from the uncaptured Hi-C library at DNase I hypersensitive sites show a small bias to the central 200 bp where there are more junctions, but this is partially offset because the fragment size is reduced in the hypersensitive sites. There was no correlation between the strength of the hypersensitive site and the number of junctions per kb within the site. A model of the background distribution of reads was generated to correct for this effect was generated using a 20 bp moving window across the metaplot of the hypersensitive sites. **f,** Plots of single normalised MCC data (to the total number of reads across the genome from the viewpoint) compared to double normalisation, which corrects for the small bias at hypersensitive sites. This shows that double normalisation for the hypersensitive site effect does not significantly change the interaction profile compared to single normalisation. Peak calling with the machine learning based peak caller LanceOtron of both single and double normalised data showed that 94% of significant peaks remain unchanged by this correction (Extended Data Table 2).



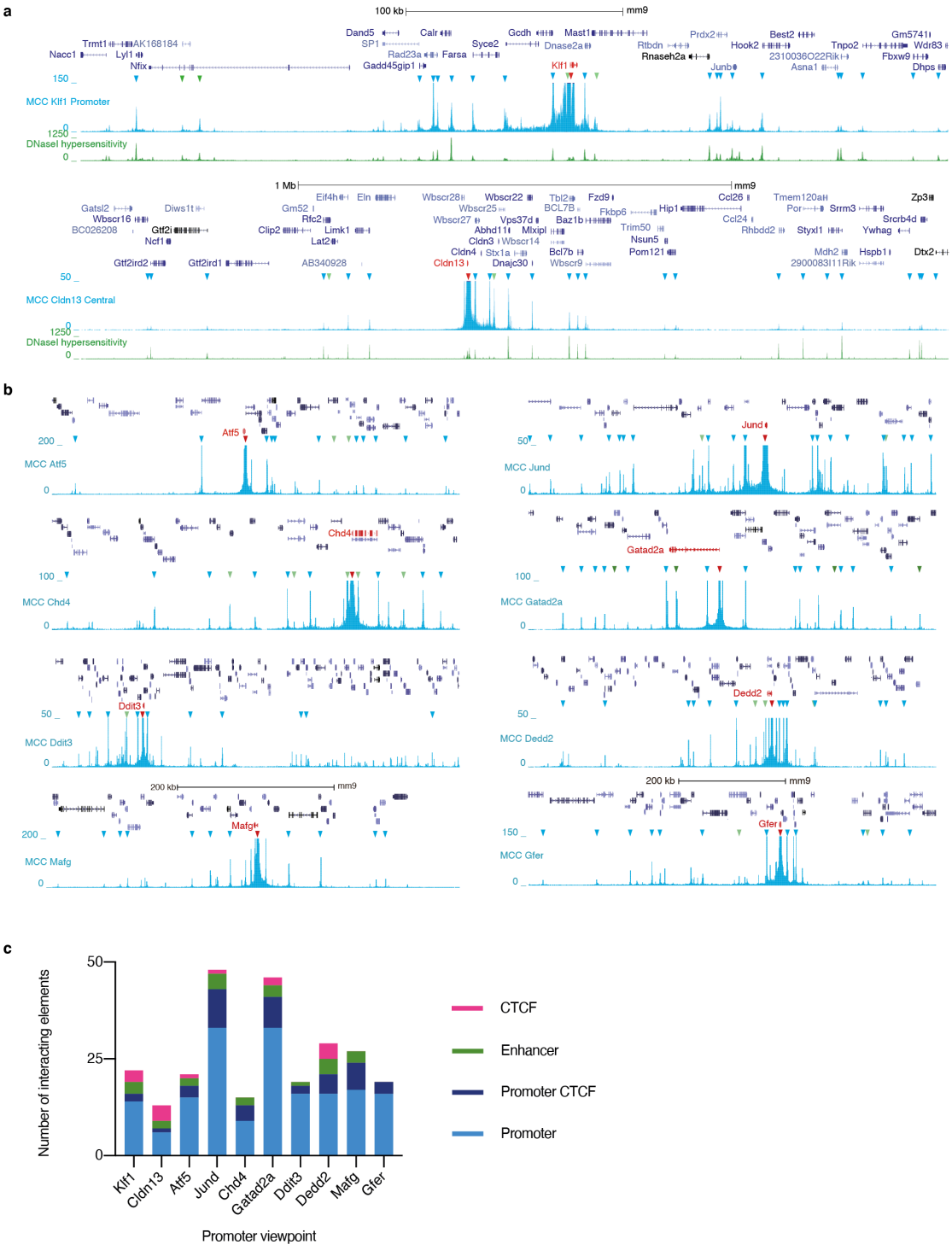
Extended Data Fig. 3 | Comparison of MCC from different capture sites from active genes in erythroid cells. MCC profiles from the promoters of **a,** *Hba-a1&2*, **b,** *Hbb-b1 & Beta-s*, **c,** *Slc25a37* and **d,** *Cd47* in erythroid and ES cells (anchor points denoted by red arrows). The interaction profile was markedly different depending on

whether the viewpoint was directly over the hypersensitive site at the promoter (central; blue) or shifted 1kb upstream (red) or 1kb downstream (green). Data are reported as read pileups without windowing and the number of reads is normalised to the total number of reads across the genome. Profiles of NG-Capture C data from the same viewpoint, DNase-seq and ChIP-seq data of H3K4me1, H3K4me3 and CTCF. The bottom panel shows the MCC profile from the central viewpoint in ES cells as control. At all genes there are significantly stronger interactions between the central promoter and the known enhancers of the genes (denoted by green arrows) and CTCF sites. **e,** Heatmap shows the punctate nature of promoter-promoter and enhancer-enhancer contacts (compared to randomly chosen sites) using data from the unenriched Hi-C library. The data presented show a 3kb region around the centre of the hypersensitive site with a 50bp bin size.



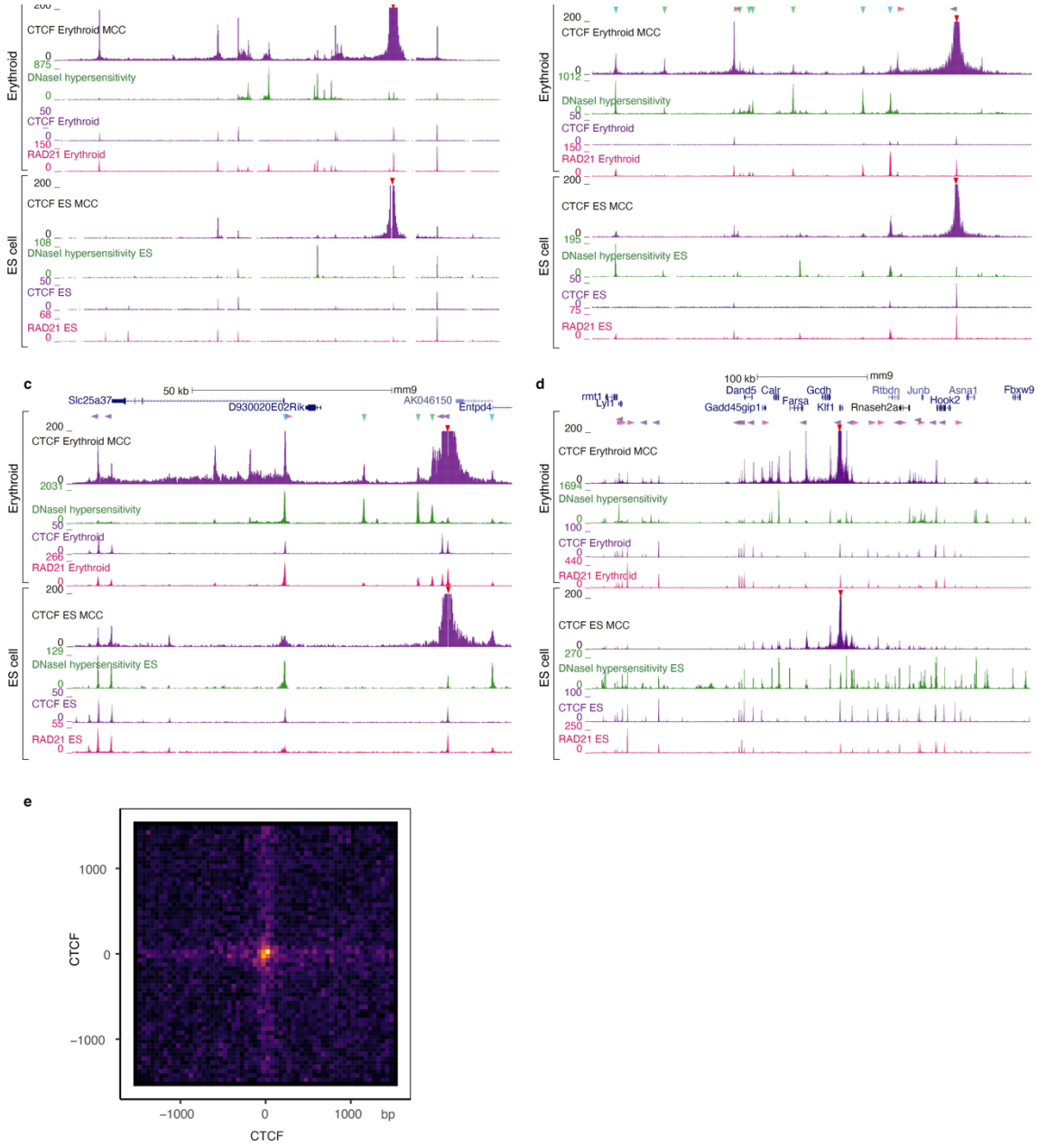
Extended Data Fig. 4 | Comparison of MCC from the promoters of active genes in ES cells. a, MCC profiles from the promoter of *Myc* showing long range interactions with enhancers of the gene over 1 Mb from the promoter. **b,** Analysis of the *Sox2* locus showing long range interactions 750 kb from the promoter of the gene, which precisely localise with CTCF binding sites, DNase-seq (green, ENCODE UW) and ChIP-seq

data of H3K4me3, H3K4me1, NANOG, SOX2, OCT4 and CTCF. **c**, MCC profiles from the promoter of *Hbb* (*Beta-s*), the inactive *Hbb-y* gene and from enhancer of second enhancer, DNase-seq (green). Note that the profile from the inactive gene does not show contacts with the hypersensitive enhancers but that it has contacts with the surrounding chromatin compartment. **d,** MCC data from the promoter or enhancer of *Sox2* in ES cells, DNase-seq (green) and data from enhancer in erythroid cells. **e**, Shows contacts from the promoter with the gene body at 4 genes that are transcribed in erythroid cells (*Hemgn*, *Fhdc1*, *Epb4.9* and *Btg2*). These data show no evidence of gene looping between the promoter and the 3' end of the gene. **f**, Metaplot of the number of junctions detected between the promoter and the 4 kb region surrounding the 3'UTR. To account for distance effects, short genes (<10 kb) (left) and long genes (>10kb) (right) are plotted separately. The comparison between erythroid cells, in which the genes are active and ES cells, in which the genes are not transcribed controls for distance and mapping effects. The dot plots show that there are no significant differences between the number of ligation junctions in 500bp bins surrounding the 3'UTR when genes are in active and inactive states (mean +/- SD). This shows that there is no change in the number of contacts when the genes change from an active to an inactive state.



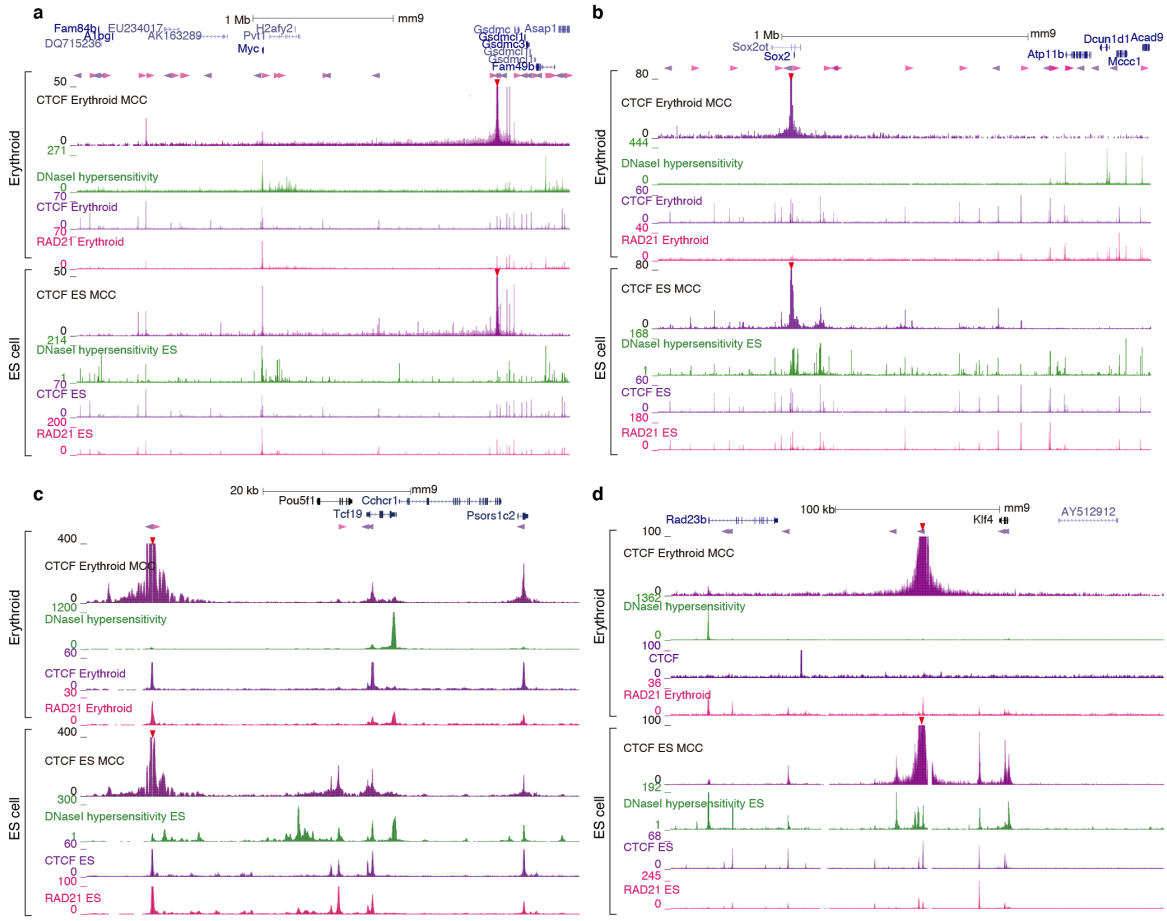
Extended Data Fig. 5 | Promoters in gene dense regions make contacts with multiple promoters. MCC profiles from the promoters of **a,** *Klf1* showing long range interactions with multiple promoters in the 400 kb surrounding the gene. However, it is likely that this gene is at least in part controlled locally as the region surrounding the promoter is monomethylated and bound by the erythroid transcription factor GATA1 (data not shown)**;** MCC data from *Cldn13* showing contacts with 25 promoters and

enhancers in the surrounding 1.5 Mb. Again, this gene is likely to be controlled at least in part by local regulatory elements. **c**, *Atf5*, *Jund*, *Chd4*, *Gatad2a*, *Ddit3*, *Dedd2*, *Mafg* and *Gfer* showing the interactions in gene dense regions. **d**, Plot showing the number of interacting elements within the 2 Mb region of each viewpoint (capture from promotors) and promotor-promotor interactions dominated in these gene dense regions.

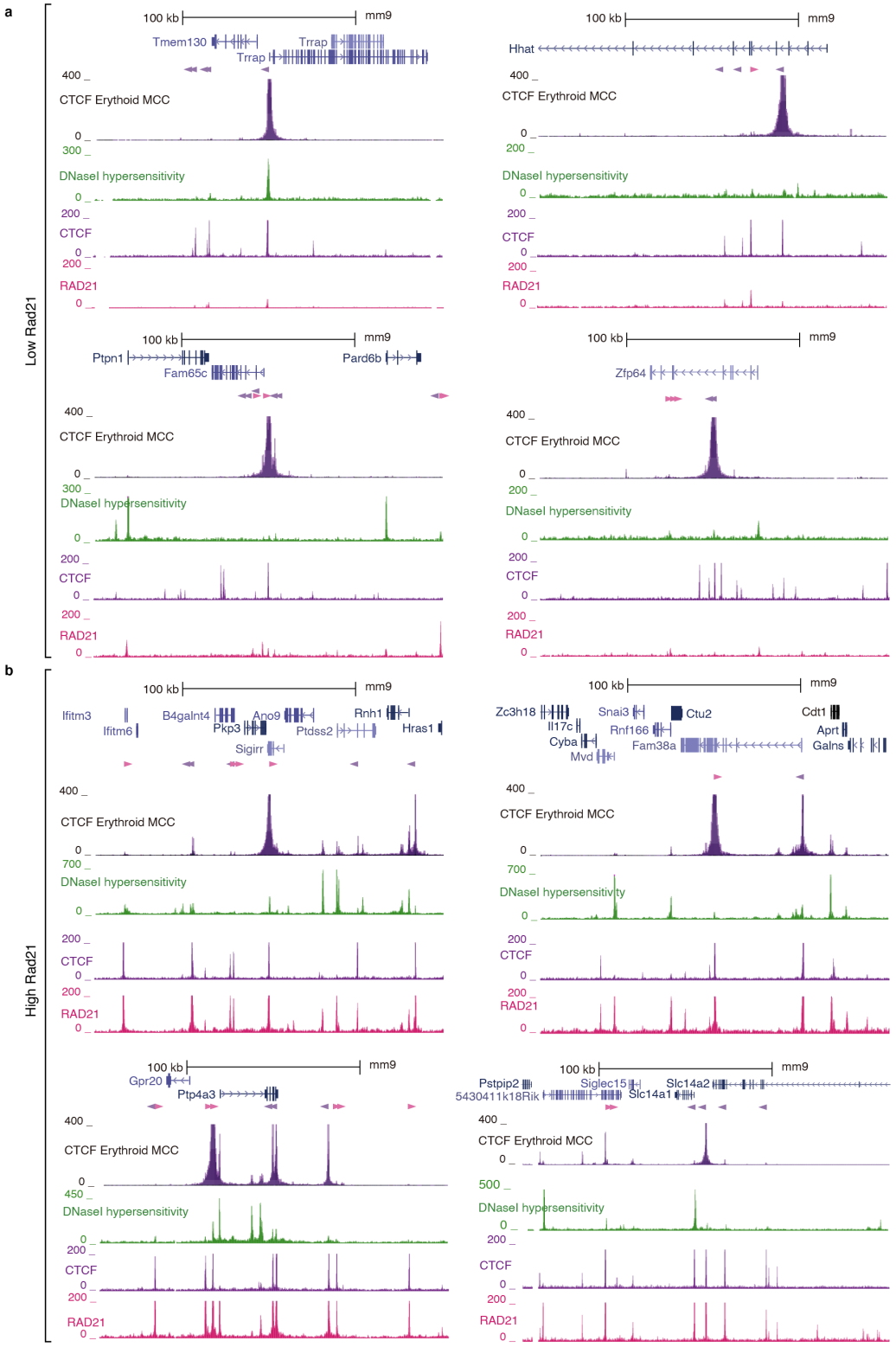


Extended Data Fig. 6 | MCC from CTCF sites at loci, which are active in erythroid cells shows that highly punctate interactions occur between CTCF sites and that these correlate with the activity of the intervening chromatin. Note that contacts from CTCF sites do not correlate strongly with DNaseI hypersensitivity. **a,** Capture from the CTCF site downstream of the enhancers at the beta globin locus. This site interacts strongly and precisely with the CTCF site upstream of the genes. These sites are in a convergent orientation. These interactions are not present in ES cells when the gene is inactive despite there being similar levels of CTCF occupancy at these sites in both tissues. **b,** Capture from an intergenic CTCF site at the *Cd47* locus in erythroid cells shows strong interactions tissue specific interactions between convergent sites upstream, that are not present in ES cells when the gene is inactive.

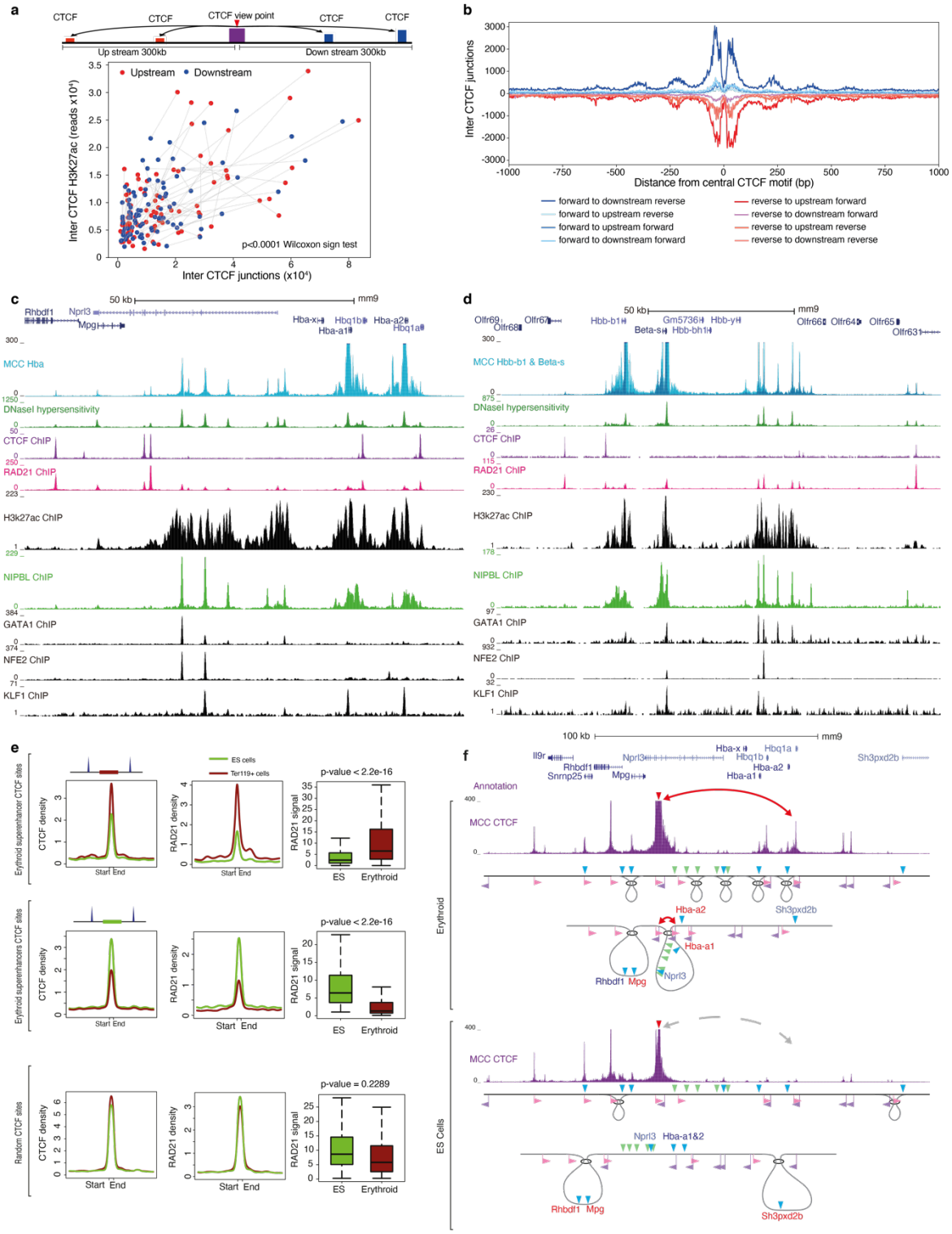
c, At the *Slc25a37* (mitoferrin) locus the CTCF site downstream of the enhancers interacts strongly with the convergent CTCF site at the promoter of the gene in a tissue specific manner. **d,** Strong and highly punctate contacts are seen between multiple CTCF sites at the gene dense *Klf1* locus. DNase-seq (green), ChIP-seq of CTCF and RAD21 are shown for both erythroid cells and ES cells. The RAD21 and CTCF ChIP-seq are normalised using a spike in of human cells. **e,** Analysis from sequencing of the unenriched Hi-C like library confirms highly punctate contacts between CTCF sites (heatmap with 20bp bins +/- 1000 bp from the centre of interacting CTCF sites).



Extended Data Fig. 7 | MCC from CTCF sites in ES cells shows that highly punctate interactions occur between CTCF sites over very long ranges. a, At the *Myc* locus highly specific, punctate contacts occur between the CTCF sites on either side of the gene and its regulatory elements. These contacts correlate with transcription (the gene is transcribed more in ES cells than erythroid cells). **b,** Similarly at the *Sox2* locus highly specific long range contacts occur between CTCF sites and these are highly tissue specific. **c,** at *Pou5f1* tissue specific contacts occur with a tissue specific CTCF binding site that is found in ES cells but not erythroid cells. **d,** At *Klf4* we captured from a tissue specific CTCF site, which contacts several CTCF sites in the same orientation in the vicinity. In erythroid cells this sequence forms no specific contacts with surrounding chromatin. DNase-seq, ChIP-seq of CTCF and RAD21 are shown for both erythroid and ES cells.

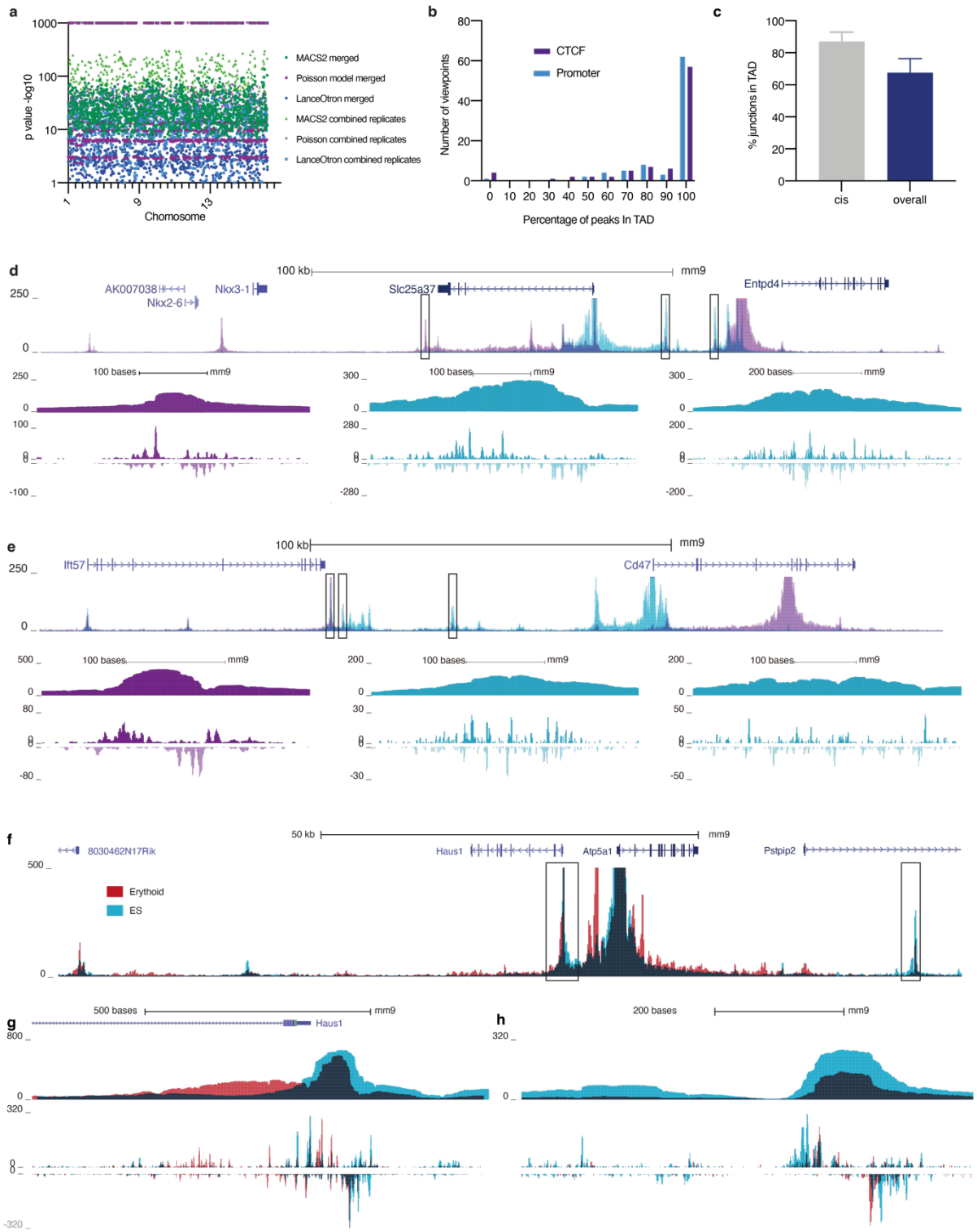


Extended Data Fig. 8 | Contacts from CTCF sites are highly correlated with cohesin. a, MCC profiles from 4 CTCF sites with low co-occupancy of cohesin as measured by RAD21 ChIP-seq. These sites form no peaks with surrounding CTCF sites. **b,** When high levels of RAD21 coincide with CTCF the sites form multiple contacts with surrounding CTCF sites. Profiles are included for DNase-seq, ChIP-seq of CTCF and RAD21 in erythroid cells.



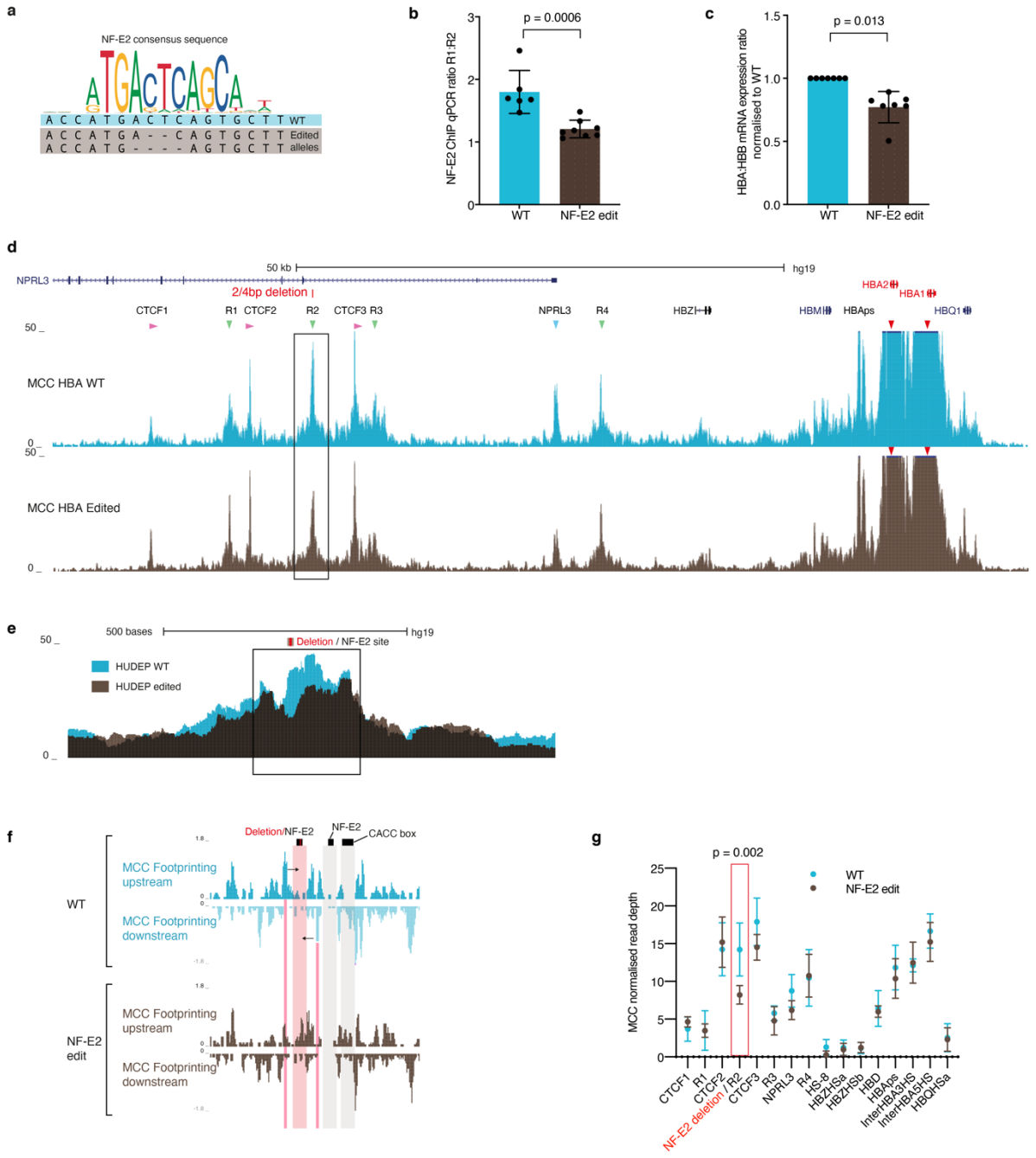
Extended Data Fig. 9 | Genome-wide analysis of inter CTCF contacts. a, Correlation between the number of contacts between the CTCF site used as a viewpoint and all of the CTCF sites within a 300 kb window either upstream (red) or downstream (blue) of the viewpoint with the activity of enhancers and promoters within the same 300 kb window as measured by H3K27ac ChIP-seq. The linked data points from individual loci show that virtually all CTCF sites make more frequent contacts with

other CTCF sites in the direction of the region with the highest intervening levels of H3K27ac. This shows that contacts between CTCF sites correlate with the activity of the intervening chromatin. **b,** Metaplot of ligation junctions between CTCF sites. These are separated by the relative orientation and position of the viewpoint and interacting CTCF sites. This clearly shows that the orientation and relative position of the CTCF sites stongly determines contact frequency. **c&d,** MCC profile of promotor of *Hba* and *Hbb*, DNase-seq (green, ENCODE UW) and ChIP seq data of CTCF, RAD21, H3K27ac, NIPBL, GATA1, NF-E2 and KLF1. **e**. Metaplots of RAD21 and CTCF binding density (RPKM) at the two nearest CTCF binding sites flanking erythroid and ES cell-specific superenhancers in erythroid (red) and ES (green) cells, showing higher levels of enrichment of both proteins at CTCF sites flanking active superenhancers compared to the same sequence near inactive enhancer sequences. CTCF binding density of RAD21 and CTCF at 500 random sites in both cell types are shown as a control. Boxplots of RAD21 binding (RPKM) at the two nearest CTCF binding sites (1 kb region around the centre of the CTCF site) flanking erythroid and ES cell-specific superenhancers in erythroid (red) and ES (green) cells, showing higher levels of enrichment of both proteins at CTCF sites flanking active superenhancers compared to the same sequence near inactive enhancer sequences (p-values calculated using a two-sided Student T-test; box plots show mean+/-SD). **f**, Proposed model for the way in which cohesin loading at active enhancer and promoter elements alters CTCF-CTCF contacts using the alpha globin locus as a model. In erythroid cells activity of the alpha globin genes and enhancers leads to increased cohesion loading at these sites. This results in loop extrusion and subsequently increased contacts between the CTCF sites upstream of the enhancers and distal to the *Hba1&2* promoters. These contacts do not occur in ES cells, despite very similar levels of CTCF binding, because the enhancers and promoters are inactive and do not load cohesin.



Extended Data Fig. 10 | Analysis of peak calling and single base pair resolution analysis of MCC ligation junctions. a, Manhatten plot showing highly significant peaks of interaction irrespective of the method of peak calling (-$\log_{10}$ of the p values are plotted on the Y axis). The data have been peak called with 3 different orthogonal methods. MACS2, a custom poisson based model and a machine learning based

model. All of these methods calculate the enrichment over the background data, which has undergone targeted capture. **b,** Histogram of the percentage of peaks falling within the TAD in erythroid cells from promoters and CTCF sites. **c,** Percentage of ligation junctions falling within the TAD in cis from erythroid promoters **d&e**, MCC footprinting shows more complex patterns of ligation junctions at enhancers than at CTCF sites at the *Slc25a37* and *Cd47* loci**. f,** The gene *Atp5a1* is active in both erythroid (red) and ES cells (blue) and there are contacts in both tissues with the promoter of *Haus1* (**g**) and a local enhancer (**h**). At both of these sites the footprinting is clearly different in the two tissues, at sites with the same DNA sequence, showing fine scale changes in the contact pattern resulting from different patterns of DNA binding proteins in the two cell types.



Extended Data Fig. 11 | MCC profiles at the main alpha globin enhancer (R2) of showing specific loss of contacts when an NF-E2 site is deleted a, Genome editng was used to make a small 2-4 bp deletion in an NF-E2 consensus motif in the main R2 enhancer at the alpha globin locus (determined by Illumina sequencing). **b,** ChIP qPCR showing loss of NF-E2 binding at the R2 enhancer compared to enrichment at the adjacent R1 enhancer showing. Note that complete abrogation of NF-E2 binding would not be expected because there are two NF-E2 binding sites in the enhancer, which are separated by 26 bp. **c,** Deletion of the NF-E2 binding site at the R2 enhancer results in a significant reduction in expression of the alpha globin

genes in erythroid cells from normal donors that have undergone genome editing at the R2 NF-E2 site (editing efficiencies in excess of 90% with Cas9 ribonuclear protein, data not shown). **d,** MCC profiles from the promotors of *HBA1&2* in human HUDEP2 cells, showing that the interactions with the main enhancer (termed R2) reduce very specifically (e) at the site of an engineered 2-4 bp deletion at the binding site of NF-E2. In addition, the MCC footprint alters in specifically at the NF-E2 binding site (**f**). **g**, Quantification of read depth at all other hypersensitive sites at the alpha globin locus showing that the only statistically significant change is at the site of the deletion (Mann-Whitney U test) (data from the edited clone aligned to a modified hg19 genome with the deletion, mean +/- SD).

Acknowledgements
JD and PH are also funded by an MRC Clinician Scientist Award (MRC Clinician Scientist Fellowship ref. MR/R008108) to JD. This work was supported by a Medical Research Council Discovery Award led by Prof. Doug Higgs (MC_PC_15069). LH, JH and ST developed LanceOtron with support from the National Institutes of Health (USA) grant number R24DK106766. JH is supported by the MRC Molecular Haematology Unit (MC_UU_00016/14). Dr Kurita and Dr Nakamura from the RIKEN Tsukuba Branch kindly provided HUDEP2 cells.

Authorship Contributions
J.D. conceived the project, designed, performed, analysed experiments, performed the majority of bioinformatic analyses and wrote the first draft of the manuscript. P.H. analysed data and performed experiments. L.L.P.H assisted with the design of experiments, performed experiments and assisted with data analysis. M.B. performed experiments and analysis. L.D.H, M.O. and R.S. contributed to analysis. D.J., J.B. and N.C. assisted with experiments. T.M. and J.H. assisted with experimental design and data analysis. D.H. provided funding and assisted with experimental design. All authors contributed to writing the manuscript.

Competing Interests
J.D., D.J. and J.H. are co-founders of Nucleome Therapeutics Ltd. J.D. and J.H. provide consultancy to the company and D.J. is an employee.

Code availability
The code required for analysis will be available for academic use through the Oxford University Innovation software store.

Data availability
Sequencing data has been submitted to the NCBI Gene Expression Omnibus (GSE144336), which can be accessed by token code: atsjwwiuljgjjmz.

References

1 Sebe-Pedros, A. *et al.* The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. *Cell* **165**, 1224-1237, doi:10.1016/j.cell.2016.03.034 (2016).

2 Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339, doi:10.1016/j.cell.2011.01.024 (2011).

3 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

4 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 U S A* **112**, E6456-6465, doi:10.1073/pnas.1518552112 (2015).

5 Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049, doi:10.1016/j.celrep.2016.04.085 (2016).

6 Nasmyth, K. Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet* **35**, 673-745, doi:10.1146/annurev.genet.35.102401.091334 (2001).

7 Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet*, doi:10.1038/s41576-019-0195-2 (2019).

8 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).

9 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

10 Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D260-D266, doi:10.1093/nar/gkx1126 (2018).

11 Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219-226, doi:10.1038/nature23884 (2017).

12 Hsieh, T. H. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108-119, doi:10.1016/j.cell.2015.05.048 (2015).

13 Hsieh, T. S., Fudenberg, G., Goloborodko, A. & Rando, O. J. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat Methods* **13**, 1009-1011, doi:10.1038/nmeth.4025 (2016).

14 Krietenstein, N. *et al.* Ultrastructural details of mammalian chromosome architecture. *bioRxiv*, 639922, doi:10.1101/639922 (2019).

15 Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol Cell* **78**, 554-565 e557, doi:10.1016/j.molcel.2020.03.003 (2020).

16 Hsieh, T. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Mol Cell* **78**, 539-553 e538, doi:10.1016/j.molcel.2020.03.002 (2020).

17 Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res* **25**, 582-597, doi:10.1101/gr.185272.114 (2015).

18 van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* **9**, 969-972, doi:10.1038/nmeth.2173 (2012).

19 Davies, J. O. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* **13**, 74-80, doi:10.1038/nmeth.3664 (2016).

20 Davies, J. O., Oudelaar, A. M., Higgs, D. R. & Hughes, J. R. How best to identify chromosomal interactions: a comparison of approaches. *Nat Methods* **14**, 125-134, doi:10.1038/nmeth.4146 (2017).
21 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871, doi:10.1126/science.184.4139.868 (1974).
22 Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, doi:10.7554/eLife.21856 (2017).
23 Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90, doi:10.1038/nature11212 (2012).
24 Tan-Wong, S. M. *et al.* Gene loops enhance transcriptional directionality. *Science* **338**, 671-675, doi:10.1126/science.1224350 (2012).
25 Hanssen, L. L. P. *et al.* Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat Cell Biol* **19**, 952-961, doi:10.1038/ncb3573 (2017).
26 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038 (2015).
27 de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell* **60**, 676-684, doi:10.1016/j.molcel.2015.09.023 (2015).
28 Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**, 1728-1740, doi:10.1038/nprot.2012.101 (2012).
29 Hentges, L. D., Sergeant, M. J., Downes, D. J., Hughes, J. R. & Taylor, S. LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq. *bioRxiv*, 2021.2001.2025.428108, doi:10.1101/2021.01.25.428108 (2021).
30 Zacher, B. *et al.* Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. *PLoS One* **12**, e0169249, doi:10.1371/journal.pone.0169249 (2017).
31 He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **33**, 395-401, doi:10.1038/nbt.3121 (2015).
32 Oudelaar, A. M. *et al.* Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat Genet* **50**, 1744-1751, doi:10.1038/s41588-018-0253-2 (2018).
33 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, doi:10.1038/ng.2871 (2014).
34 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
35 Hay, D. *et al.* Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet* **48**, 895-903, doi:10.1038/ng.3605 (2016).