nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01116-w

DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell Hi-C data at kilobase resolution

Shanshan Zhang[1,2,6], Dylan Plummer[3,6], Leina Lu[1,6], Jian Cui[1,6], Wanying Xu[1,2], Miao Wang[4], Xiaoxiao Liu[1], Nachiketh Prabhakar[3], Jatin Shrinet[4], Divyaa Srinivasan[4], Peter Fraser[4], Yan Li[1,7] ✉, Jing Li[3,5,7] ✉ and Fulai Jin[1,3,5,7] ✉

Mapping chromatin loops from noisy Hi-C heatmaps remains a major challenge. Here we present *DeepLoop*, which performs rigorous bias correction followed by deep-learning-based signal enhancement for robust chromatin interaction mapping from low-depth Hi-C data. *DeepLoop* enables loop-resolution, single-cell Hi-C analysis. It also achieves a cross-platform convergence between different Hi-C protocols and micrococcal nuclease (micro-C). *DeepLoop* allowed us to map the genetic and epigenetic determinants of allele-specific chromatin interactions in the human genome. We nominate new loci with allele-specific interactions governed by imprinting or allelic DNA methylation. We also discovered that, in the inactivated X chromosome (X_i), local loops at the *DXZ4* 'megadomain' boundary escape X-inactivation but the *FIRRE* 'superloop' locus does not. Importantly, *DeepLoop* can pinpoint heterozygous single-nucleotide polymorphisms and large structure variants that cause allelic chromatin loops, many of which rewire enhancers with transcription consequences. Taken together, *DeepLoop* expands the use of Hi-C to provide loop-resolution insights into the genetics of the three-dimensional genome.

Hi-C has transformed our understanding of mammalian genome organization and can reliably identify high-order three-dimensional genome features such as compartments and topological-associated domains (TADs)[1–4]. However, when resolution is at the kilobase scale, Hi-C contact heatmaps quickly become noisy due to the increasingly complex bias structure and severe data sparsity[5–9]. To date, genome-wide mapping of chromatin loops, especially enhancer–promoter interactions within TADs (sub-TADs), remains a major challenge in Hi-C analyses. Consequently, scientists often turn to focused technologies, such as chromatin interaction analysis by paired-end tag sequencing (ChIA–PET), promoter capture Hi-C (pcHi-C), HiChIP/proximity ligation-assisted ChIP–seq (PLAC–seq) and so on, hoping for better signal-to-noise ratios at selected loci[10–15], even though these approaches identify only a subset of all interactions

Bias and noise are two distinct types of error in Hi-C data. Here we define biases as 'unwanted pattern in a Hi-C heatmap'. This definition is goal oriented. For example, to distinguish relatively weak loop signals, the strong nonspecific diagonal Hi-C signal needs to be corrected as bias. Hi-C protocols using different digestion enzymes have different bias structures determined by fragment size, distance, genomic copy (GC) content and the interactions between these factors[5,6], and the bias structure becomes more complicated when the resolution increases, especially at the sub-TAD mid-range (that is, within 1–2 Mb). While several methods have been developed to model and correct known sources of Hi-C biases explicitly with joint functions, the most commonly used strategy is to 'normalize' the Hi-C matrices and correct Hi-C biases implicitly with matrix-balancing algorithms[5–7,16–18]. However, both explicit and implicit strategies have drawbacks[5,6,8,9,16,17]. To improve the rigor of Hi-C bias correction, we recently developed a *HiCorr* pipeline that performs both explicit and implicit correction[9]. Unlike normalization methods[7,17], which preserve a strong diagonal signal in the contact heatmaps, *HiCorr* corrects distance effects in a joint function with other biases and outputs the observed/expected ratio heatmaps for chromatin interaction profiling. When read depth is high, *HiCorr* generates sharper contact heatmaps and is more robust in identification of sub-TAD chromatin loops[9].

Theoretically, when all biases are corrected, only data sparsity contributes to Hi-C noises. Therefore, reduction of Hi-C noises is mathematically equivalent to signal enhancement. Several recent studies have pioneered the application of deep learning techniques to enhance Hi-C signal at the compartment, TAD and loop levels[19–23]. These pipelines share a similar framework to impute high-depth contact matrices from low-depth raw or normalized Hi-C data. It is, however, important to point out that this strategy 'learns' Hi-C biases in the input matrices, which may no longer be properly corrected after enhancement. This flaw is critical for loop analysis because distance effect is a major bias for loop analysis, and other Hi-C biases are also much worse at high resolution. To address this issue, here we developed a strategy to enhance *HiCorr*-corrected ratio heatmaps. The resulting *DeepLoop* pipeline achieved striking robustness in calling loops from low-depth Hi-C data. This study highlights the application of *DeepLoop* to single-cell

[1]Department of Genetics and Genome Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, USA. [2]The Biomedical Sciences Training Program, School of Medicine, Case Western Reserve University, Cleveland, OH, USA. [3]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH, USA. [4]Department of Biological Science, Florida State University, Tallahassee, FL, USA. [5]Department of Population and Quantitative Health Sciences, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA. [6]These authors contributed equally: Shanshan Zhang, Dylan Plummer, Leina Lu, Jian Cui. [7]These authors jointly supervised this work: Yan Li, Jing Li, Fulai Jin. ✉e-mail: yxl1379@case.edu; jingli@cwru.edu; fxj45@case.edu

and allele-resolved Hi-C data analyses, both impacted by the challenge of severe data sparsity.

Results

***LoopDenoise* improves the robustness of Hi-C loop mapping.** We begin with denoising of high-depth Hi-C heatmaps with a five-layer autoencoder (Fig. 1a and Extended Data Fig. 1a). We picked three replicates of Hi-C data in human fetal cerebral cortex[24] for model training; each replicate has ~140–150 million mid-range (<2 Mb) *cis* contacts. (In this paper we use the number of mid-range *cis* contacts to measure 'read depth', or the total amount of data for a Hi-C experiment.) We applied *HiCorr* to each replicate and extracted ~18,000 submatrices at fragment resolution (~5 kb) as training sets (Fig. 1a,b). As previously reported, *HiCorr* generates sharper distance-corrected ratio heatmaps than ICE/KR[9] (compare rows 3 and 4 in Fig. 1c), but noise pixels are still present. When pooling the reads from all three replicates, *HiCorr* heatmaps are only slightly cleaner (Fig. 1b). Since true loop pixels are more reproducible than noise pixels between biological replicates, we set up 'training target' heatmaps by retaining only reproducible loop pixels (Fig. 1b, Extended Data Fig. 1b and Methods).

LoopDenoise removes all visible noise pixels from the *HiCorr*-corrected ratio heatmaps (Fig. 1c, compare rows 4 and 6). Denoised heatmaps are cleaner than the training targets (Fig. 1c, compare rows 5 and 6). When applied to biological replicates, *LoopDenoise* improves pairwise reproducibility to ~70–80% at the pixel level (Extended Data Fig. 1c,d). When applied to independent Hi-C datasets in human embryonic stem cells (hESCs), IMR90, GM12878 and mouse embryonic stem cells (mESCs)[1,6,7,9,25–27] (Supplementary Table 1), the benefits of *LoopDenoise* are also obvious (Fig. 1d–g and Extended Data Fig. 2a). Loop pixels are better concentrated near CTCF, H3K4me3 and H3K27ac peaks after denoising[1,28–31] (Extended Data Fig. 2b–d). *LoopDenoise* successfully reveals loop interactions at loci with well-established long-range gene regulation, such as *Sox2*, *Wnt6* and *Malt1* in mESCs, and at *HOXAs*, *FTO* and *SHH* in hESCs (Extended Data Fig. 2e)[32–35].

To test whether this improved reproducibility would facilitate the identification of dynamic chromatin loops, we compared human cortex Hi-C data from the germinal zone (GZ) and cortical plate (CP), which are two layers of developing cortex enriched with neuron progenitors and postmitotic neurons[24,36]. Indeed, R^2 between GZ and CP improved (from 0.31 to 0.65) after denoising (Extended Data Fig. 3a). When picking those genes associated with the top 3,000 GZ- or CP-specific loop pixels, we found that GZ loop genes are enriched with terms related to neural development while CP loop genes are enriched with neuronal function terms (Extended Data Fig. 3b). After denoising, the dynamic loop pixels are clearly recognizable at GZ-specific (such as *SOX2*, *FOXP2* and *EOMES*) and CP-specific (such as *TGFB2* and *NELL2*) genes, in agreement with GZ- or CP-specific assay for transposase-accessible chromatin using sequencing (ATAC–seq) peaks[37] (Extended Data Fig. 3c).

***LoopEnhance* reliably maps Hi-C loops from low-depth data.** We then developed a method to analyze low-depth Hi-C data. We trained a series of U-Net[38] *LoopEnhance* models using downsampled cortex Hi-C data with ~10–250 million mid-range *cis* contacts. Notably, we used *LoopDenoise* outputs from high-depth data as training targets, which should be better representations of the 'ground truth' (Fig. 2a and Supplementary Fig. 1a). Strikingly, although loop signals are hardly recognizable when read depth is <50 Mb, the enhanced heatmaps from low-depth Hi-C data are nearly identical (Fig. 2b). *LoopEnhance* models created with cortex data also performed very well in the independent GM12878 datasets (Fig. 2c). When comparing enhanced heatmaps to full data (~380 million mid-range *cis* contacts), we found no compromise in performance (pixel-level reproducibility >70%) when read depth was reduced to 100 million; pixel-level reproducibility remained >50% even when sequencing depth was reduced to 12.5 million (Fig. 2d). We also trained new *DeepLoop* models (*LoopDenoise* and *LoopEnhance*) with Hi-C data from H9 hESCs and confirmed that the choice of training set did not affect results (Supplementary Fig. 1b,c). Because pixel intensity in the *DeepLoop* heatmaps represents Hi-C signal enrichment, we can directly term top loop pixels as interactions. Note that *DeepLoop* does not output an explicit list of discrete 'loops'; conversion of 'loop pixels' to 'loops' requires new algorithms and parameters, which will inevitably introduce new biases. Therefore, we retain *DeepLoop* as a 'what-you-see-is-what-you-get' method.

We next compared the *DeepLoop* pixels in GM12878 cells with the ~83,000 loops called by pcHi-C in the same cell line[39]. We classified pcHi-C loops into promoter–promoter interactions (PP, the fragments of both ends were captured with promoter probes) and promoter–other interactions (PO, only one end of the interaction was captured), and further divided these loops into long-range (>100 kb) and short-range (<100 kb) categories. *DeepLoop* improved receiver operating characteristic (ROC) curves in all categories, especially the long-range examples (Fig. 2e); this is consistent with *DeepLoop*'s noise reduction function, because Hi-C matrices are noisier at long range due to more severe data sparsity.

We also collected five sets of ChIA–PET or HiChIP data in GM12878 cells performed with CTCF, PolII, RAD21, SMC1A and H3K27ac antibodies[14,40–43]. The numbers of loops from these datasets were highly variable (3,600–48,000), with a grand total of ~64,000 (Fig. 2f). Clearly, each experiment captures only a subset of all interactions. We classified all loops based on their recurrence among these experiments and examined the recovery efficiency of Hi-C for each category. With *DeepLoop*, a downsampled 50-million-depth Hi-C map recovered 7,051 (62%) and 8,260 (72%) of the 11,401 'recurrent' (in least two experiments) loops when calling 500,000 and 1,000,000 top loop pixels, respectively, in contrast to only 23 and 29% before enhancement. Recovery of the ~53,000 'non-recurrent' loops improved even more. In fact, the enhanced 50-million map outperformed the unenhanced 380-million full-data map in all loop categories (Fig. 2f). Notably, the cost involved in generation of 50-million-depth Hi-C data was already lower than that for one ChIA–PET or HiChIP experiment.

***DeepLoop* Hi-C maps converge with micro-C maps.** Although *DeepLoop* is trained with six-cutter Hi-C data, because its bias correction is independent from the noise-reduction module we need only to adjust *HiCorr* for *DeepLoop* to work for other Hi-C-like data. Indeed, both *LoopDenoise* and *LoopEnhance* work very well on *MboI*-based GM12878 in situ Hi-C data[7] (examples in Extended Data Fig. 4). Interestingly, although with a conventional pipeline, four-cutter Hi-C heatmaps are sharper than six-cutter heatmaps and *DeepLoop* heatmaps are very similar, indicating that *HiCorr* is more efficient at removal of platform-specific biases and supports cross-platform comparison. For the same reason, *DeepLoop* substantially outperformed other Hi-C enhancing pipelines including *HiCPlus*[21], *HiCNN2* (ref. [44]) and *SRHiC*[23] (Extended Data Fig. 4).

To further explore cross-platform consistency we compared the published ultradeep Hi-C data in H1 hESCs prepared with *HindIII*, *DpnII* and micro-C[9,27,45,46]. As expected[47], for raw, KR and KR-ratio heatmaps, micro-C was sharper than both *HindIII* and *DpnII* Hi-C while *DpnII*-Hi-C was sharper than *HindIII*-Hi-C (examples in Fig. 3a,b and Extended Data Fig. 5a). However, *DeepLoop* heatmaps from *HindIII* and *DpnII* Hi-C were much more similar at the pixel level, regardless of read depth (Fig. 3c). Importantly, when digestion resolution was increased (from Hi-C to micro-C), KR-ratio heatmaps become sharper and more similar to *DeepLoop* outputs (Fig. 3a,b and Extended Data Fig. 5a). When we compared other signal enhancement methods using micro-C KR-ratio heatmaps as reference, *DeepLoop* showed the highest correlation coefficient

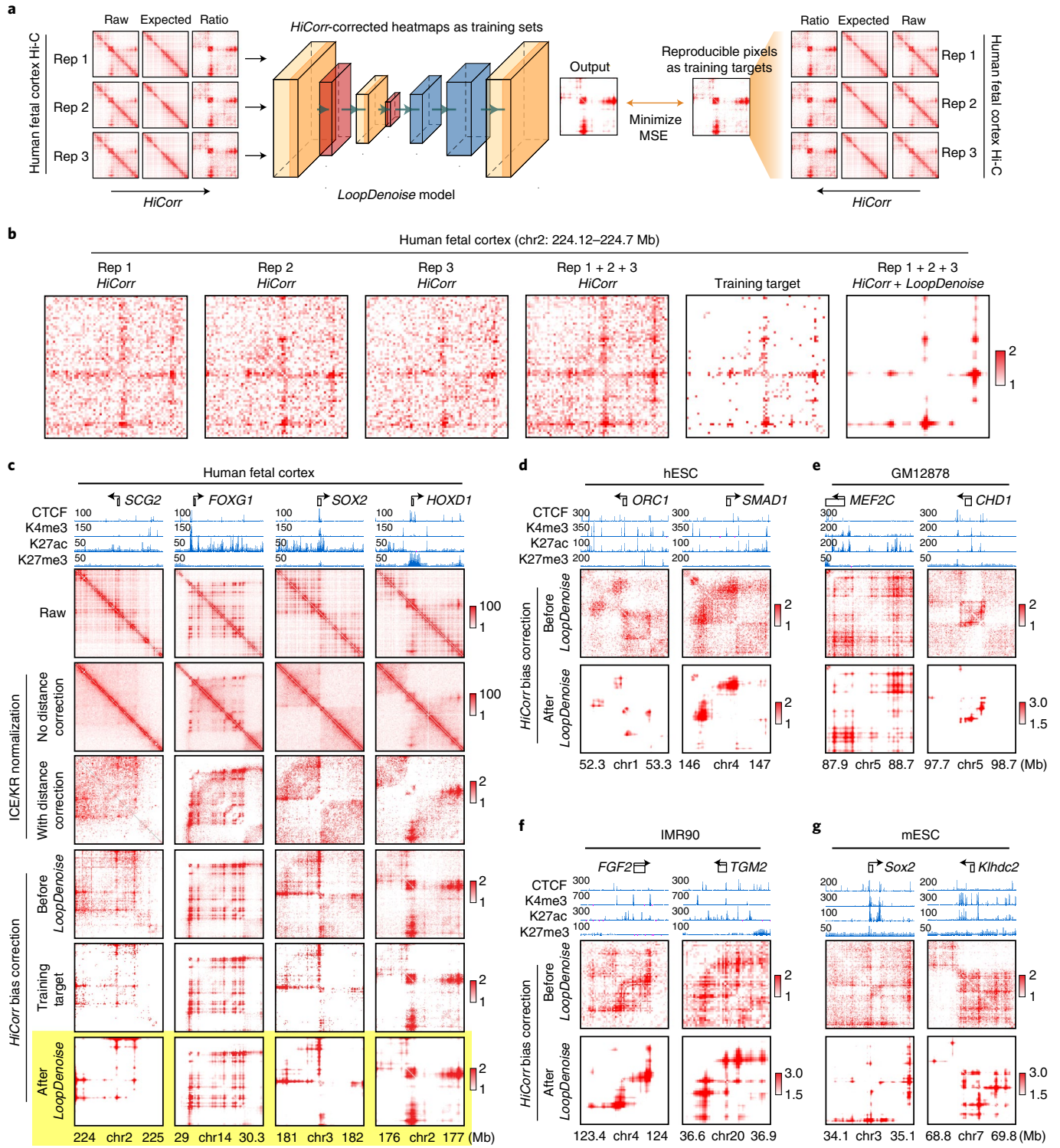


Fig. 1 | *HiCorr* and *LoopDenoise* reveal chromatin loops from noisy Hi-C datasets. a, Scheme showing the *LoopDenoise* model architecture and training. The three *HiCorr*-corrected human fetal brain datasets are used as training sets. Training targets are the reproducible pixels in the heatmaps from pooled data. **b**, Example heatmaps from human fetal cortex Hi-C data, including three *HiCorr*-corrected replicates (Rep), pooled data, training target and output from *LoopDenoise*. **c**, *LoopDenoise* performance in the training human fetal cortex Hi-C data at four loci. Heatmaps of raw and various processed data are compared. Highlighted row is *LoopDenoise* output. **d**–**g**, Heatmaps showing the application of *LoopDenoise* to four independent Hi-C datasets in hESC (**d**), GM12878 (**e**), IMR90 (**f**) and mESC (**g**). ChIP-seq tracks show raw reads pile-up. See Methods for information on how to determine the color scale of each heatmap.

(Fig. 3d). Finally, we called 17,500 micro-C loops at 5-kb resolution using a standard KR-Hi-C computational unbiased peak search (HICCUPS) pipeline, then performed ROC analyses using these loops as true positives. *DeepLoop*-enhanced, low-depth (50 million) Hi-C data performed better than all other pipelines, even better than KR-processed full-depth data (Fig. 3e).

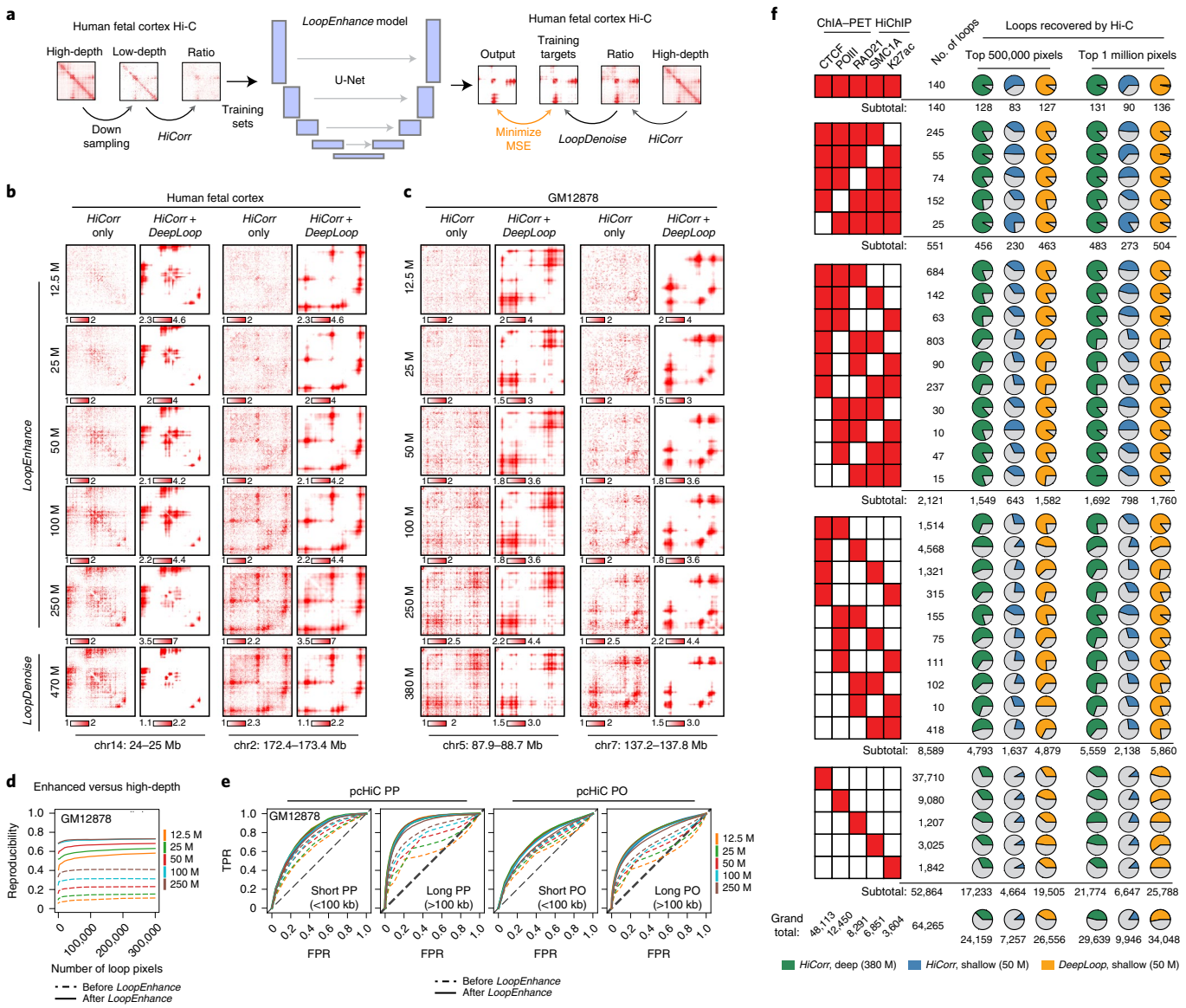


Fig. 2 | *LoopEnhance* enables sensitive and robust loop calling from low-depth Hi-C data. a, Scheme showing the architecture and training of the *LoopEnhance* model. Left: downsampling from high-depth human fetal cortex Hi-C data as training sets after *HiCorr* correction. Middle: U-Net architecture of *LoopEnhance*. Right: training targets are high-depth human fetal brain data following both *HiCorr* and *LoopDenoise*. **b**, Heatmap examples showing the outputs of *LoopEnhance* when applied to downsampled human fetal cortex data (training sets) at variable depth. Two loci are shown. The bottom row is *LoopDenoise* output using the full dataset (training target). **c**, Heatmap examples showing the application of *LoopEnhance* to downsampled independent GM12878 data. The full GM12878 data were analyzed with *LoopDenoise* (bottom row). **b**,**c**, Sequencing depth on the left indicates the numbers of mid-range (<2 Mb) *cis* contacts (M, million). **d**, Reproducibility, the fraction of overlapped loop pixels, between downsampled and full-depth GM12878 data when the same numbers of loop pixels were called. For comparison, *LoopDenoise* was used on full-depth GM12878 data. Solid lines: *HiCorr* and *LoopEnhance* applied to downsampled data; dashed lines: only *HiCorr* was applied. **e**, ROC curves showing the recovery of GM12878 pcHi-C loops with enhanced low-depth Hi-C data. Significant ($P < 0.01$, three-parameter Weibull distribution) pcHi-C interactions (PP and PO) in GM12878 are considered as true positives. Solid lines: *HiCorr* + *LoopDenoise*; dashed lines: *HiCorr* only. **f**, Loops identified from five published ChIA-PET and HiChIP studies in GM12878 were grouped by their recurrence among these experiments. Loop number and subtotal for each 'recurrence' group are listed. Pie charts indicate the percentage of each group of loops recovered by the Hi-C map when calling the top 500,000 or 1 million loop pixels. Green: 380-million-depth *HiCorr* map; blue: 50-million-depth *HiCorr* map; orange: 50-million-depth *DeepLoop*-enhanced map. TPR, true positive rate; FPR, false positive rate.

Micro-C is expected to reveal more small loops (<50 kb) than either six- or four-cutter Hi-C with the standard HICCUPS pipeline[45,47,48]. We found that, with four-cutter Hi-C, *DeepLoop* recovered most micro-C small loops and the recovery rate was only slightly lower than for large loops (Extended Data Fig. 5b). However, because *DeepLoop*-enhanced six-cutter Hi-C missed most small micro-C loops, this indicates that *HindIII* hits a hard limit for small loop detection due to large fragment size: sufficent numbers of restriction sites need to be cut between the two anchors to discern a small loop. Notably, micro-C may find even smaller loops at higher resolution[45,48]. Improvement of *DeepLoop* resolution will be an interesting future direction. Regardless, *DeepLoop* achieves better cross-platform convergence between Hi-C and micro-C.

Application of *DeepLoop* to sparse and single-cell Hi-C data. We firs enhanced published sparse Hi-C data in 14 human tissues

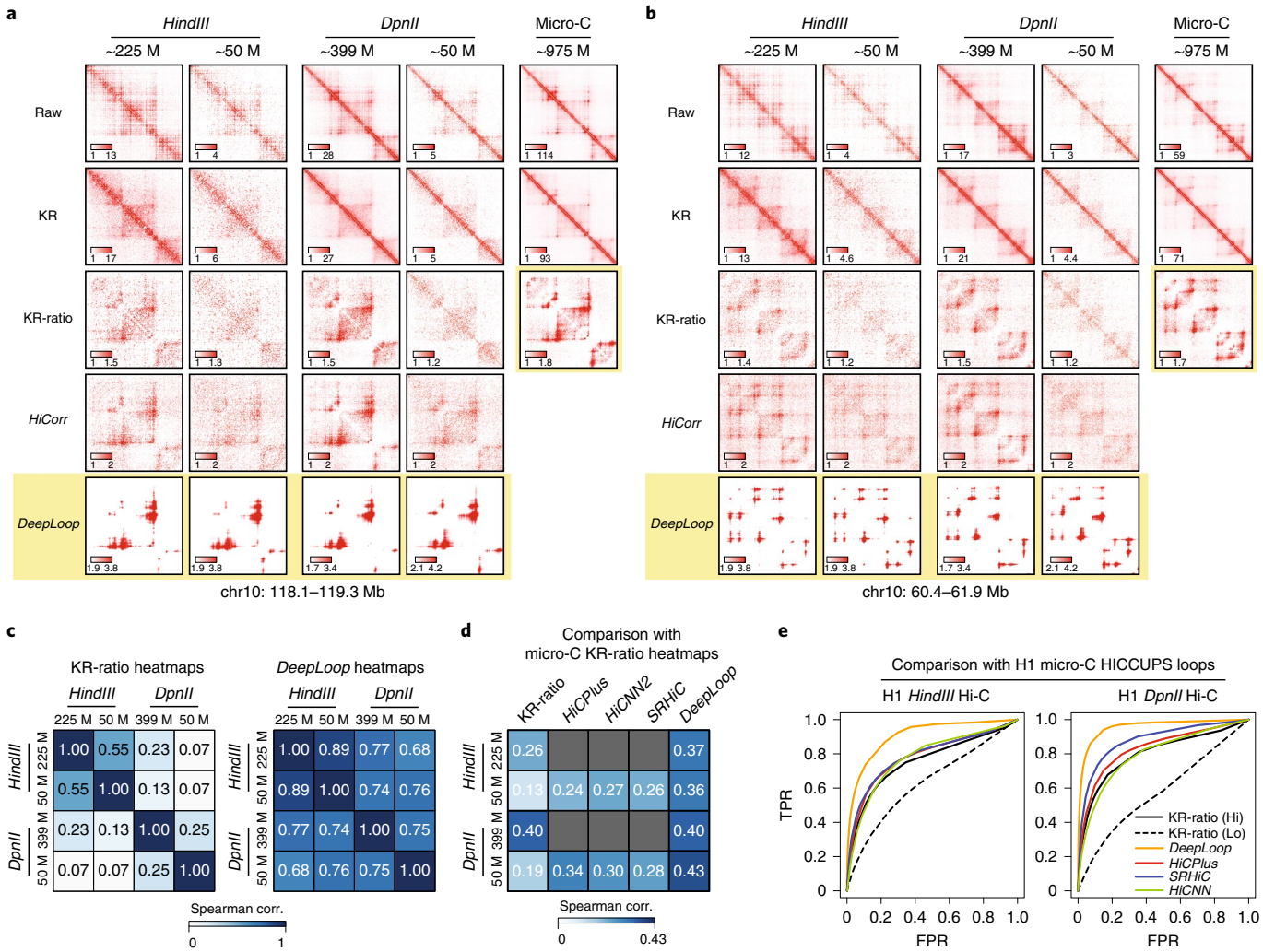


Fig. 3 | *DeepLoop* outputs convergent Hi-C loop profiles regardless of read depth and digestion resolution. a, Heatmap examples showing the outputs of different pipelines with full depth or downsampled (50M) *HindIII*- or *DpnII*-based Hi-C data in H1 hESCs. The last column shows KR-processed heatmaps from ultradeep micro-C data. **b**, Similar to **a** at different locus. **c**, Left: Spearman correlations between Hi-C experiments with different restriction enzymes and read depths when KR-ratio contact heatmaps were compared at the pixel level. Right: same, except that *DeepLoop* outputs were used in the comparison. **d**, Spearman correlations between micro-C KR-ratio heatmaps and outputs of various pipelines with *HindIII*- or *DpnII*-based Hi-C data. **e**, ROC curves comparing the performance of different enhancing pipelines in recovery of micro-C loops when applied to *HindIII*- or *DpnII*-based H1 hESC Hi-C data. For all Hi-C analysis pipelines, loop pixels were called from ratio heatmaps after ranking by intensity. Pixels in micro-C HICCUPS loops (after KR normalization) were treated as true positives. KR-ratio heatmaps from full-depth (solid black curve) or downsampled Hi-C (dashed black curve) were plotted as reference.

(depth, ~7–53 million mid-range *cis* contacts)[29,30,49]. We observed specific loop interactions near many tissue-marker genes after enhancement, such as *ALB* (liver), *MYOZ2* (aorta, left and right ventricles) and *ADD2* (cortex, hippocampus, CP and GZ) (Fig. 4a and Extended Data Fig. 6). Quantitively, pixel-level correlation between related tissues improved markedly after enhancement (Fig. 4b).

We next applied *DeepLoop* to a mESC single-cell Hi-C dataset[50]. The average depth of this dataset was ~58,000 mid-range *cis* contacts per cell. To test the lower limit of cell numbers required for loop analysis, we ranked all 4,098 cells by sequencing depth and generated a series of matrices (depth ~973,000–33,000,000) after pooling 92 of the deepest single cells. We pooled the remaining 4,006 cells into a bulk dataset (depth 203 million) and used the top 300,000 loop pixels from the denoised 4,006-cell data as 'true positives'. *DeepLoop* heatmaps become stable with near-perfect ROC curves when either cell numbers reached ~10–41 or read depth reached ~10 million (Fig. 4c,d and Supplementary Fig. 2a). The enhanced data consistently recovered a large fraction of promoter interactions identified from an independent pcHi-C dataset[10] using CHiCAGO[51] (Supplementary Fig. 2b).

Finally we applied *DeepLoop* to a single-nucleus methyl-3C sequencing (sn-m3C–seq) dataset in human prefrontal cortex (PFC) in which the identities of 14 cell populations were previously resolved by DNA methylation profiles[52]. Most cell populations have at least 100 cells and 10-million read depth, which is adequate for direct observation of population-specific loop profiles. For example, the *RORB* loop signal is restricted in layer 4/5 neurons but not in layers 2/3/6, which is highly consistent with the DNA hypomethylation signal (Fig. 4e,f). Similar observations were also made for tissue-specific genes *SATB2* (layers 2/3/4/5), *MBP* (ODC/OPC/MG) and *APOE* (astrocytes) (Extended Data Fig. 7).

***DeepLoop* nominates allelic loops at imprinting or differentially methylated region loci.** The remainder of this manuscript focuses on resolving human allele-specific (AS) chromatin loops, which is a difficult task due to the sparse and uneven distribution of hetero-

zygous single-nucleotide polymorphisms (SNPs). Specifically, the GM12878 genome has ~1.7 million heterozygous SNPs (or one SNP per ~1.5 kb), which enforces a hard limit for data resolution because only those reads overlapping SNPs are usable. Starting from 4.5 billion GM12878 in situ Hi-C reads[7], only 337 million (~7.5%) could be assigned to either the maternal or paternal genome (Fig. 5a): each haploid has ~56 million mid-range *cis* contacts. We applied *DeepLoop* to maternal and paternal data independently at 5-kb

Fig. 4 | *DeepLoop* identifies chromatin interactions from low-depth and single-cell Hi-C data. a, Contact heatmaps of exemplary marker genes in 14 published low-depth human tissue Hi-C data. High-depth CP and GZ are also included for comparison with brain tissue maps. The numbers of mid-range *cis* contacts are indicated for each tissue. **b**, Reproducibility refers to the fraction of overlapped loop pixels between the top 100,000 loop pixels from every pair of tissues, which are used for tissue clustering before and after signal enhancement. **c**, Analysis of single-cell Hi-C data. After pooling different numbers of single mESC cells (read depth indicated for each pool), raw, *HiCorr*-corrected and enhanced heatmaps are shown. Heatmaps for the pool of 4,006 diploid cells are shown in the far-right column. **d**, ROC curves for each enhanced Hi-C map using the top 300,000 loop pixels from the 4,006-cell dataset (*LoopDenoise* output) as true positives. **e**, Single cells from human PFC sn-m3C-seq data were split into 14 populations based on cell type. Data from the same population were pooled and processed with *DeepLoop*. The heatmaps are at the *RORB* locus; numbers in parentheses indicate the number of cells per population. **f**, Left: *t*-distributed stochastic neighbor-embedding (*t*-SNE) plot showing cell type identification by methylation profile; right: methylation levels of *RORB* for every cell are visualized on the same *t*-SNE plot; mCG, CpG methylation.

resolution and called the top 300,000 loop pixels from each haploid genome. After enhancement, R^2 between two homologs improved substantially, from 0.216 to 0.628 (Fig. 5b), which allows much more robust allelic analyses.

The best-known example of AS loops is at the *H19/IGF2* imprinting locus. Early studies using allelic chromosome conformation capture polymerase chain reaction (3C–PCR)[53–55] and. more recently, allelic circular chromosome conformation capture (4C–seq)[56] showed that, in mouse cells, a paternally methylated, gametic differentially methylated region (DMR) blocks CTCF binding and loop formation (insulator model). We therefore examine the 3,736 loop pixels anchored on all 992 DMRs previously defined in human GM12878 cells[57] (colored dots in Fig. 5c). Only three loci, *H19*, *MEST* and *MRPL28*, have DMR and AS loops, consistent with the idea that the 'insulator model' is not a common mechanism for imprinting control[58]. For *H19/IGF2*, the AS loops are barely observable from KR-normalized heatmaps at 25-kb resolution and the ambiguity is worse at 5-kb resolution (Fig. 5d, first and second columns). *HiCorr* clearly improved 5-kb resolution bias correction and allowed *DeepLoop* to output clean maps of AS loops consistent with maternal-specific CTCF binding at *H19* DMR (Fig. 5d,h,k).

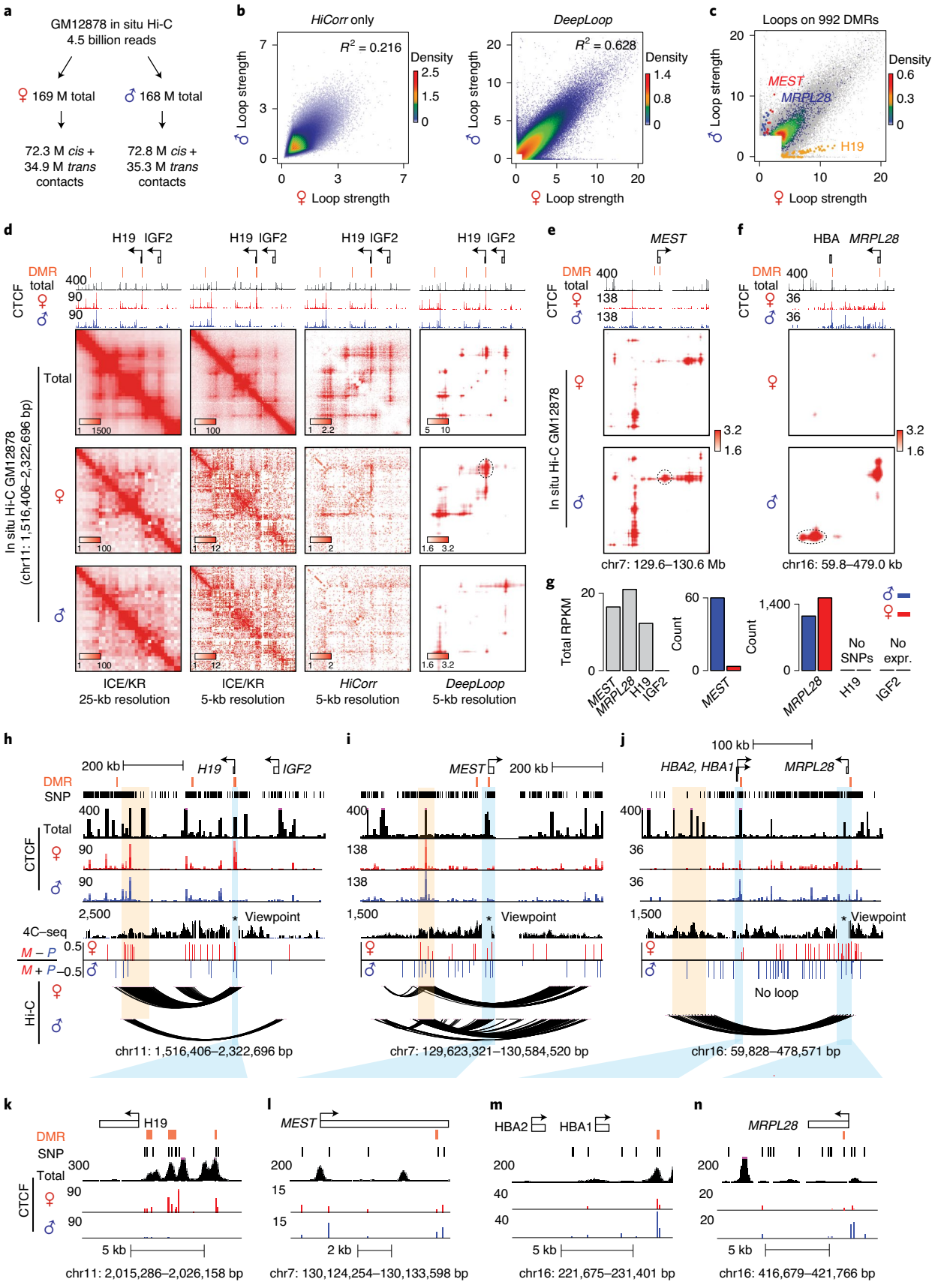
We performed 4C–seq using DMRs as viewpoints and confirmed the allelic imbalance of the AS loops (Fig. 5h–j). We also examined the allelic imbalance of the CTCF ChIP–seq data at *MEST* and *MRPL28* loci. *MEST* is a well-known paternally imprinted gene[59] (Fig. 5g). The *MEST* DMR is close to two CTCF peaks that form a paternal-specific loop (Fig. 5e,i,l); one peak is ~480 base pairs (bp) distant from its closest heterozygous SNP that supports a paternal CTCF binding with marginal significance (ten versus four reads, $P=0.09$; Fig. 5l). The *MRPL28* transcriptional allele specificity is weak but the loop is highly specific (Fig. 5f,g). There is a strong CTCF peak near *MRPL28* DMR that presumably anchors the paternal-specific loops (Fig. 5f,j,n). Although the allele specificity of this CTCF peak is unknown due to the lack of informative SNPs, a small CTCF peak in this region is highly paternal specific (23 versus four reads, $P=1.6\times10^{-4}$; Fig. 5n). Another CTCF peak at the *HBA1/2* DMR is also paternal specific (60 versus 20 reads, $P=0.007$[4]; Fig. 5m). In fact, the entire region between *HBA1/2* and *MRPL28* is decorated with stronger paternal CTCF signals (Fig. 5j). It should be noted that we are still unsure whether *MRPL28* is an imprinting locus because it is unclear whether *MRPL28* or *HBA1/2* DMRs are gametic DMRs.

***DeepLoop* reveals chromatin loops that escape X-inactivation.** Allelic Hi-C analyses at low resolution in both human and mouse cells have reproducibly recorded the loss of TAD domains and the formation of megadomain and ultradistal superloops in the inactivated X chromosome (X_i)[2,7,60–62]. However, the architectures of X_i and X_a (active X chromosome) have not been compared at the sub-TAD loop level. In human GM12878 cells the paternal chrX is inactive. *DeepLoop* called 3,550 and 806 loop pixels from X_a and X_i, respectively (Fig. 6a), indicating that most chromatin loops are repressed by X-inactivation. Most chrX genes are monoallelic except 17 escape genes, including the X-inactivation center (XIC) genes *XIST* and *JPX* (cutoff P/(M+P) >0.2; M, maternal expression; P, paternal expression; see Methods; Fig. 6b). As expected, escape loop pixels (present in both X_i and X_a) are enriched near the escape genes (Fig. 6c, with examples in Fig. 6e).

We next examined the relationship between chromatin loops and high-order megadomain or superloop structures in X_i. *DXZ4* is at the boundary of the megadomain (Fig. 6d) and also forms a superloop with the downstream *FIRRE* locus[7,63]. The gene bodies of both *DXZ4* and *FIRRE* gain CTCF binding in X_i, which may function to anchor X_i to the nucleolus[61,63,64]. Interestingly, we found that the two loci responded differently to X-inactivation. At the *DXZ4* locus the chromatin loops, CTCF peaks and ATAC–seq peaks were invariant between X_a and X_i, suggesting that this locus had escaped X-inactivation (Fig. 6e,f). In contrast, although the X_i *FIRRE* gained much-strengthened loop pixels within its own gene body, all loops connecting *FIRRE* to surrounding regions were lost (Fig. 6e,g), indicating that the *FIRRE* locus is X-inactivated. Consistent with these observations, *FIRRE* gained CTCF and ATAC–seq signals in its gene body but lost CTCF and ATAC–seq signals at the promoter (Fig. 6e,g). Notably, *FIRRE* is predominantly expressed from Xa (Fig. 6b), also indicating that it is X-inactivated[65].

Because both *DXZ4* and *FIRRE* form superloops but only *DXZ4* is at the megadomain boundary, our observation suggests that the escape loops near *DXZ4* (presumably mediated by cohesin and loop extrusion) are mechanistically coupled to the formation of the megadomain but not the superloop; other mechanisms (for example, colocalization to the nucleolus) may result in superloops. These results agree very well with a recent study showing that loss of cohesin disrupts the *Dxz4* megadomain but enhances the *Dxz4–Firre* superloop in mouse cells[66]. Taken

Fig. 5 | Homolog-specific chromatin interactions are associated with imprinting and DMR. a, Reads summary of allele-resolved in situ Hi-C data in GM12878 cells. **b**, Scatterplots comparing the loop strength of all anchor pairs between two haploid genomes. Left: *HiCorr* only; right: after *DeepLoop*. **c**, Heat scatter showing all loop pixels overlapping 992 DMRs. Loop pixels at three loci with highest allele specificity are highlighted in different colors. Background scatterplots are a union of the top 300,000 loop pixels from both haploid genomes. **d**, Contact heatmaps of the *H19/IGF2* locus. **e**,**f**, Contact heatmaps of genes *MEST* (**e**) and *MRPL28* (**f**) after *DeepLoop*. **g**, Gray bar plot on the left: reads per kilobase million (RPKM) of four genes in GM12878 showing their expression level; bar plots on the right: RNA read counts on the two alleles for each gene. Note that, although *H19* is expressed, its messenger RNA sequence does not contain heterozygous SNPs for allelic analysis. **h**–**j**, Browser tracks for the three loci in **d**–**f**, respectively; 4C–seq tracks showing chromatin interactions with the DMR region as viewpoint. Tracks of allelic 4C–seq analysis are included to show the maternal (red) or paternal (blue) preference of the 4C–seq signal. Light blue, DMR anchoring allelic loops; light orange, the other anchor of the allelic loop. **k**–**n**, Zoomed-in track views of **h**–**j**, respectively, showing regions with DMR. The height of browser tracks shows ChIP–seq read count pile-up.



together we propose that, in contrast to their names, the mega-domain uses a cohesin-dependent looping mechanism while the superloop does not.

***DeepLoop* functionally characterizes large heterozygous structure variants.** We were intrigued to see many loop pixels showed extreme allele specificity (>tenfold difference, $P<0.01$) after *DeepLoop* but

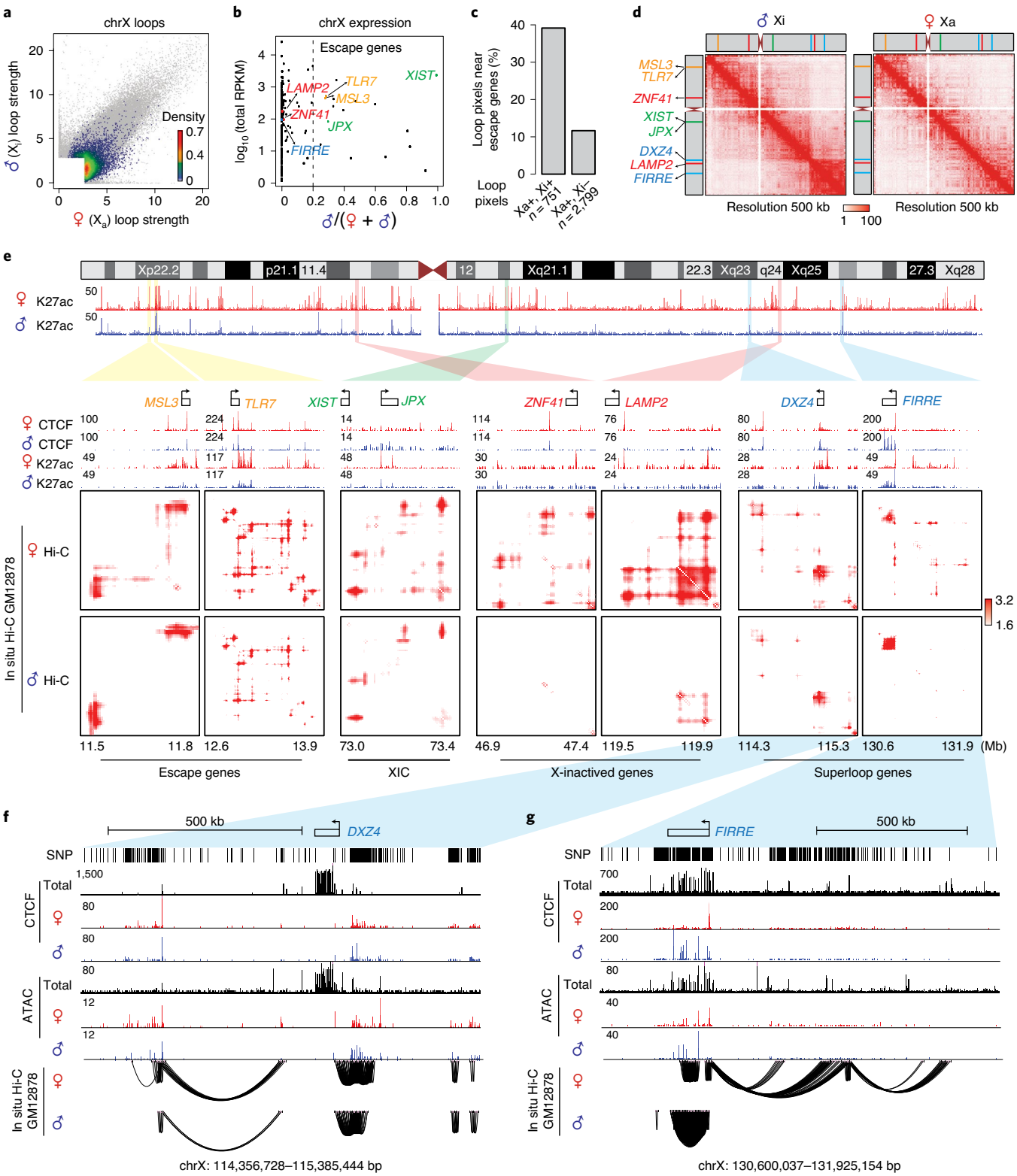


Fig. 6 | Homolog-specific chromatin interactions are associated with X-inactivation. a, Heat scatterplot of all chrX loops, shown in color. Gray background dots represent the union of the top 300,000 loops in both haploid genomes. **b**, Scatterplot showing gene expression from the two chrX copies. *X* axis: fraction of RNA reads on paternal alleles from total of both alleles; *y* axis: RPKM of total expression in log scale; genes of interest in **d**,**e** are highlighted in different colors. Dashed vertical line indicates cutoff to define escape genes. **c**, Bar plot showing percentages of 'escape loop pixels' (present in both X_a and X_i) and 'inactivated loop pixels' anchored to the 17 escape genes (transcriptional start site ±100 kb) defined in **b**. **d**, chrX heatmaps with KR normalization at 500-kb resolution showing the megadomain. **e**, *DeepLoop*-enhanced Hi-C heatmaps for two homologs at seven representative loci, including escaping loci (yellow), XIC (green), X-inactivated loci (red) and X_i megadomain or superloop loci (blue). **f**,**g**, Genome browser tracks at *DXZ4* (**f**) and *FIRRE* (**g**) loci. ChIP–seq tracks show raw reads pile-up.

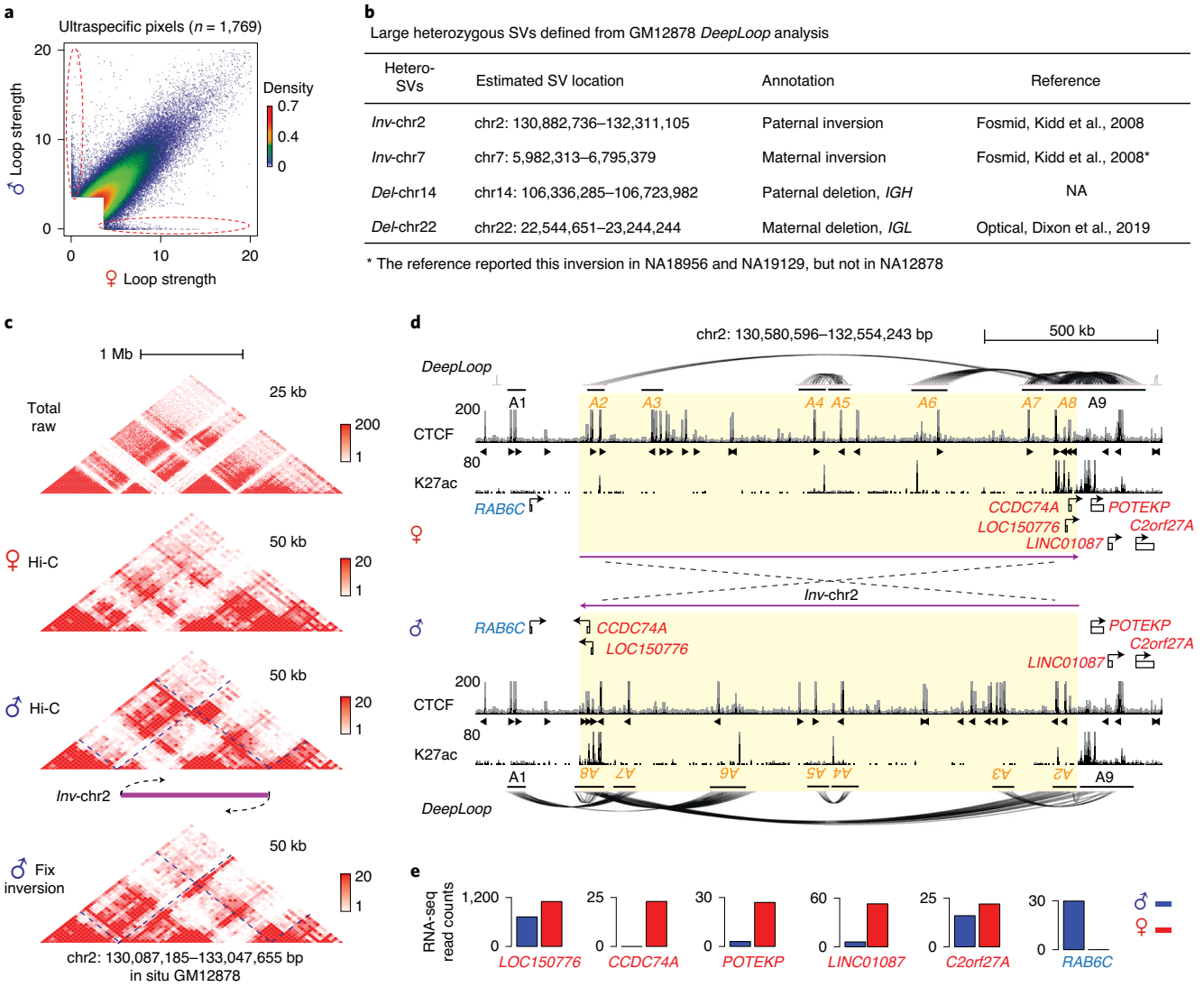


Fig. 7 | Allelic *DeepLoop* maps detect and functionally characterize large heterozygous SVs. a, Scatterplot showing ultraspecific loops (bounded by red ellipses). **b**, Four large heterozygous SVs containing the majority of ultraspecific loops. **c**, Raw contact heatmaps of the *Inv*-chr2 locus. The 'corrected' raw heatmap of the inverted paternal allele is included ('fix inversion'). Resolution level shown at right. **d**, Genome browser track of the *Inv*-chr2 locus showing CTCF and H3K27ac binding and chromatin loops in the uninverted maternal allele and inverted paternal allele. **e**, Bar plots showing allelic expression of genes highlighted in **d** at inversion boundaries. NA, not applicable.

not before enhancement (Fig. 5b; circled scatter points in Fig. 7a). Interestingly, 1,533 of these 1,769 (87%) ultraspecific pixels are in four regions. Based on the patterns of maternal and paternal contact heatmaps[67], we concluded that these regions harbor large heterozygous deletions and inversions (Fig. 7b,c and Extended Data Fig. 8a–c). *Del*-chr14 (~300 kb) and *Del*-chr22 (~600 kb) are large heterozygous deletions at the *IGH* and *IGL* immunoglobulin loci, respectively, consistent with the allele-exclusive V(D)J recombination process in B lymphocytes (Fig. 7b and Extended Data Fig. 8b). The two inversions are even bigger (*Inv*-chr2, ~1.4 Mb; *Inv*-chr7, ~900 kb; Fig. 7b). This extreme allele specificity is apparently due to incorrect distance bias correction when using the reference genome for structure variant (SV) alleles.

The detection of heterozygous SVs, especially large inversions, is notoriously difficult[68–70]. We looked up the four heterozygous SVs in published GM12878 data using various SV-detection tools[53,57,58] (Fig. 7b) and found that (1) neither short- nor long-read whole-genome sequencing detected any of the four SVs[67,71]; (2) optical mapping detected *Del*-ch22 at the *IGH* locus[67]; (3) a previous Hi-C analysis did not detect any of these SVs because the study assumed a homozygous genome and performed analysis only at 1-Mb resolution[67]; (4) the conventional fosmid subcloning-based method detected *Inv*-chr2 but showed nothing about its heterozygosity[72]; and (5) the fosmid method detected *Inv*-chr7 in two independent NA18956 and NA19129 genomes but not in NA12878, suggesting that *Inv*-chr7 is a recurrent SV in the human population[72]. Taken together, allelic *DeepLoop* analysis appears to be a promising approach for detection of large heterozygous SVs.

To correctly map chromatin loops affected by inversions, we adjusted the orientation of the inverted allele using the annotated inversion coordinate[72] and repeated *DeepLoop* enhancement (Fig. 7c,d and Extended Data Fig. 8c,d). For *Inv*-chr2, paternal inversion broke up an enhancer cluster at the 3' boundary that was heavily interconnected in the maternal genome (A7–9 in Fig. 7d). Genes connected by this enhancer cluster, including *LOC150776*, *CCDC74A*, *POTEKP*, *LINC01087* and *C2orf27A*, were all downregulated in the inverted paternal genome (Fig. 7d,e). On the other hand, *Inv*-chr2 moved half of the 3' boundary enhancer cluster

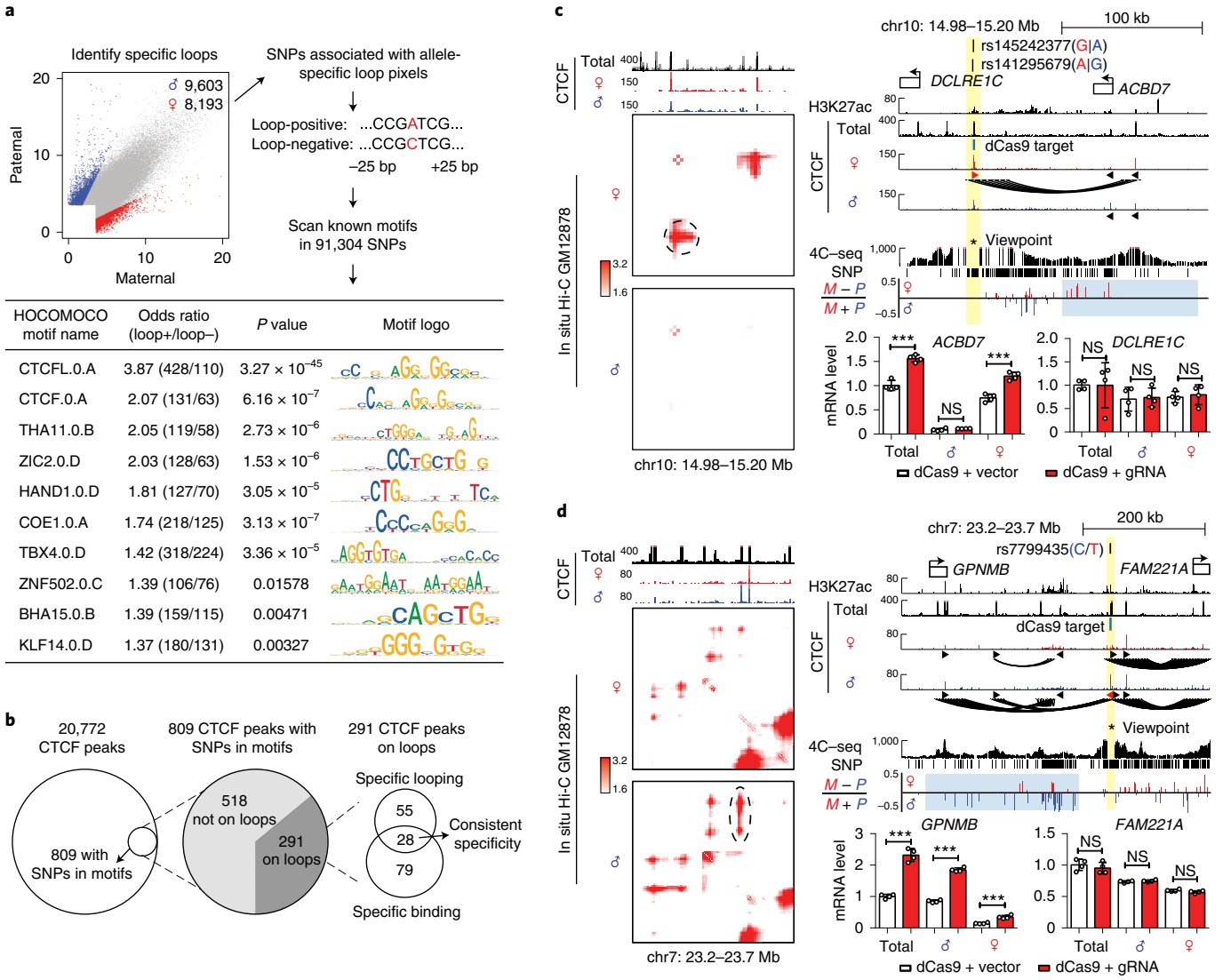


Fig. 8 | Allelic *DeepLoop* maps pinpoint common SNPs that affect chromatin loops. a, Flowchart showing de novo motif findings associated with AS chromatin loops, highlighted in the scatterplot. The 51-base sequences (25 bp up/down) around SNPs were used to scan for motifs significantly enriched in loop-positive alleles. Fisher's exact test was performed to measure motif enrichment. **b**, Summary of the procedure used to identify causal SNPs for AS CTCF loops and TF occupancy. **c**, Using 'insulator epigenome editing' to validate the transcription regulatory functions of two selected allelic-specific CTCF loops. Both contact heatmaps and genome browser tracks are included to show the locations of SNPs, specific CTCF peaks and DNA loops. 4C-seq tracks show chromatin interactions with the SNP region as viewpoint (highlighted in yellow). Tracks of allelic 4C-seq analysis show the maternal (red) or paternal (blue) preference of the 4C-seq signal. Regions of interest are highlighted in light blue. Bar plots show changes in nearby gene expression with AS using quantitative PCR with reverse transcription following CTCF blocking with dCas9. $n = 2$ biologically independent experiments. All data are presented as means ± s.e.m. from four replicated experiments. $^{***}P < 0.001$; two-sided Wilcoxon test. NS, no significant difference (additional results in Extended Data Fig. 10). **d**, Similar to **c** at different locus.

(A7–8) to the 5' boundary and new loops formed across the 5' boundary between A1 and the inverted A7–8 anchors (Fig. 7d). These new loops help explain paternal expression of the *RAB6C* gene (Fig. 7e). Similarly, *Inv*-chr7 also rewired the DNA loops, which explains the paternal-specific *CCZ1* expression (Extended Data Fig. 8d,e). These results demonstrate that *DeepLoop* can detect and predict the regulatory effects of large heterozygous SVs that may link to diseases or phenotypes[73,74].

***DeepLoop* pinpoints SNPs that affect loops and transcription.** Last, we investigated the impacts of heterozygous SNPs on chromatin loops. After exclusion of AS loop pixels associated with imprinting, X-inactivation and SVs, we used a simple twofold cutoff and called thousands of AS loop pixels at 1,959 loci (Fig. 8a). These loop pixels contained 91,304 heterozygous SNPs for which 'loop-positive' and 'loop-negative' alleles could be unambiguously defined. *CTCFL* and *CTCF* were the top two motifs enriched in loop-positive alleles, proving the feasibility of resolving the genetics of loops with *DeepLoop*. Other motifs were also enriched, such as COE1.0.A bound by the B lymphocyte-specific transcription activator *EBF1* and one motif KLF14.0.D bound by Kruppel-like factors that have been shown to regulate loops in other cell types[75–77] (Fig. 8a). Further studies are necessary to verify the loop-regulatory functions of individual SNPs and their cognate transcription factors (TFs).

We next sought to map causal SNPs of CTCF AS loops. In GM12878 cells, 809 (3.9%) of all 20,772 CTCF peaks had

heterozygous SNPs in their cognate motifs (Fig. 8b), from which we narrowed this down to 28 highly credible AS-CTCF peaks (involving 30 SNPs in 26 loci) anchoring consistent AS loops. For two selected loci we confirmed their allele specificity with 4C–seq (Fig. 8c,d). Snapshots of the remaining 24 loci are shown in Extended Data Fig. 9.

We also used a dCas9-based insulator editing approach[78,79] to test whether AS-CTCF loops affect transcription in *cis*. With single-guide RNAs precisely targeting cognate CTCF motifs, both dCas9 and dCas9-KRAB proteins abolished the CTCF loops of interest (Extended Data Fig. 10a,d). In the first example (Fig. 8c and Extended Data Fig. 10a–c), the maternal alleles of rs141295679 and rs145242377 (both SNPs are within the same CTCF motif) were associated with stronger CTCF binding and a maternal loop encompassing the *ACBD7* gene. Blocking of this loop increased the maternal expression of *ACBD7* but did not affect a control gene outside the loop (*DCLRE1C*). In the second example (Fig. 8d and Extended Data Fig. 10d–f), the paternal allele of rs7799435 formed a strong CTCF loop encompassing the *GPNMB* gene. Blocking of the paternal CTCF loop also increased paternal *GPNMB* expression but did not affect the *FAM221A* gene from a different neighborhood. These examples demonstrate that allelic *DeepLoop* analysis can pinpoint common SNPs that directly regulate gene expression by influencing DNA looping.

Discussion

DeepLoop is a novel framework that enhances Hi-C ratio heatmaps (rather than contact heatmaps) without distance effects. Because bias correction and signal enhancement are carried out in two independent modules, each module can be modified or upgraded without affecting the other. *DeepLoop* is a universal tool that can be applied to different Hi-C data types if *HiCorr* has been properly adjusted. The lower limit of read depth is ~10 million mid-range *cis* contacts, which typically can be obtained from about 50–100 million total reads. Nearly all published Hi-C datasets have adequate reads for *DeepLoop* reanalysis. Existing single-cell Hi-C technologies can yield sufficient reads from a few dozen cells. *DeepLoop* allowed us to map the human AS loops and revealed the genetic and epigenetic determinants of chromatin loop variations. We have set up a public webapp to visualize *DeepLoop*-enhanced heatmaps for around 40 datasets mentioned in this study. In summary, *DeepLoop* makes Hi-C a robust and affordable approach to revealing genome organization at sub-TAD loop level.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01116-w.

Received: 15 April 2021; Accepted: 30 May 2022;
Published online: 11 July 2022

References

1. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
2. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
3. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
4. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30**, 1357–1382 (2016).
5. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065 (2011).
6. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
7. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
8. Forcato, M. et al. Comparison of computational methods for Hi-C data analysis. *Nat. Methods* **14**, 679–685 (2017).
9. Lu, L. et al. Robust Hi-C maps of enhancer-promoter interactions reveal the function of non-coding genome in neural development and diseases. *Mol. Cell* **79**, 521–534 (2020).
10. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
11. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
12. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
13. Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
14. Mumbach, M. R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).
15. Fang, R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.* **26**, 1345–1348 (2016).
16. Hu, M. et al. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
17. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
18. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
19. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069 (2019).
20. Zhang, Y. et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat. Commun.* **9**, 750 (2018).
21. Liu, T. & Wang, Z. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. *Bioinformatics* **35**, 4222–4228 (2019).
22. Hong, H. et al. DeepHiC: a generative adversarial network for enhancing Hi-C data resolution. *PLoS Comput. Biol.* **16**, e1007287 (2020).
23. Li, Z. & Dai, Z. SRHiC: a deep learning model to enhance the resolution of Hi-C data. *Front. Genet.* **11**, 353 (2020).
24. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
25. 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).
26. Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
27. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
28. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
29. Hawkins, R. D. et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
30. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
31. Shen, Y. et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
32. Lettice, L. A. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet* **12**, 1725–1735 (2003).
33. Li, Y. et al. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS ONE* **9**, e114485 (2014).
34. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
35. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
36. Won, H., Huang, J., Opland, C. K., Hartl, C. L. & Geschwind, D. H. Human evolved regulatory elements modulate genes involved in cortical expansion and neurodevelopmental disease susceptibility. *Nat. Commun.* **10**, 2396 (2019).
37. de la Torre-Ubieta, L. et al. The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* **172**, 289–304 (2018).
38. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).

39. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
40. Heidari, N. et al. Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905–1917 (2014).
41. Tang, Z. et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
42. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
43. Li, G., Chen, Y., Snyder, M. P. & Zhang, M. Q. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.* **45**, e4 (2017).
44. Liu, T. & Wang, Z. HiCNN2: enhancing the resolution of Hi-C data using an ensemble of convolutional neural networks. *Genes (Basel)* **10**, 862 (2019).
45. Krietenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* **78**, 554–565 e7 (2020).
46. Reiff, S. B. et al. The 4D Nucleome Data Portal: a resource for searching and visualizing curated nucleomics data. *Nat. Commun.* **13**, 2365 (2022).
47. Akgol Oksuz, B. et al. Systematic evaluation of chromosome conformation capture assays. *Nat. Methods* **18**, 1046–1055 (2021).
48. Hsieh, T. S. et al. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell* **78**, 539–553 (2020).
49. Schmitt, A. D. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
50. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
51. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
52. Lee, D. S. et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* **16**, 999–1006 (2019).
53. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev.* **20**, 2349–2354 (2006).
54. Murrell, A., Heeson, S. & Reik, W. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.* **36**, 889–893 (2004).
55. Kurukuti, S. et al. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc. Natl Acad. Sci. USA* **103**, 10684–10689 (2006).
56. Lleres, D. et al. CTCF modulates allele-specific sub-TAD organization and imprinted gene activity at the mouse Dlk1-Dio3 and Igf2-H19 domains. *Genome Biol.* **20**, 272 (2019).
57. Kuleshov, V. et al. Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* **32**, 261–266 (2014).
58. Barlow, D. P. & Bartolomei, M. S. Genomic imprinting in mammals. *Cold Spring Harb. Perspect. Biol.* **6**, 952–965 (2014).
59. Kobayashi, S. et al. Human PEG1/MEST, an imprinted gene on chromosome 7. *Hum. Mol. Genet.* **6**, 781–786 (1997).
60. Deng, X. et al. Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* **16**, 152 (2015).
61. Giorgetti, L. et al. Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575–579 (2016).
62. Minajigi, A. et al. Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **17**, 349 (2015).
63. Horakova, A. H., Moseley, S. C., McLaughlin, C. R., Tremblay, D. C. & Chadwick, B. P. The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. *Hum. Mol. Genet.* **21**, 4367–4377 (2012).
64. Yang, F. et al. The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol.* **16**, 52 (2015).
65. Fang, H. et al. Trans- and cis-acting effects of Firre on epigenetic features of the inactive X chromosome. *Nat. Commun.* **11**, 6053 (2020).
66. Kriz, A. J., Colognori, D., Sunwoo, H., Nabet, B. & Lee, J. T. Balancing cohesin eviction and retention prevents aberrant chromosomal interactions, Polycomb-mediated repression, and X-inactivation. *Mol. Cell* **81**, 1970–1987 (2021).
67. Dixon, J. R. et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
68. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
69. Chaisson, M. J., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
70. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
71. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
72. Kidd, J. M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
73. Puig, M., Casillas, S., Villatoro, S. & Caceres, M. Human inversions and their functional consequences. *Brief. Funct. Genomics* **14**, 369–379 (2015).
74. Giner-Delgado, C. et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* **10**, 4222 (2019).
75. Schoenfelder, S. et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).
76. Di Giammartino, D. C. et al. KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nat. Cell Biol.* **21**, 1179–1190 (2019).
77. Wei, Z. et al. Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. *Cell Stem Cell* **13**, 36–47 (2013).
78. Tarjan, D. R., Flavahan, W. A. & Bernstein, B. E. Epigenome editing strategies for the functional annotation of CTCF insulators. *Nat. Commun.* **10**, 4258 (2019).
79. Jia, Z. et al. Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection. *Genome Biol.* **21**, 75 (2020).

Methods

No ethical approval was needed.

Experiments. *Hi-C on H9 cells.* H9 cells (WiCell, no. WA09) were maintained in mTeSR1 medium (StemCell Technologies, no. 05850) on plates coated with hESC-Qualified Matrigel (Corning, no. 354277) before harvesting for Hi-C. After hand-picked removal of differentiated colonies, cells were digested to single cells with Accutase (Innovative Cell Technologies, no. AT104) and then fixed with 1% formaldehyde. Hi-C was performed according to a published protocol[6]. First, fixed cells were lysed with cell lysis buffer containing 10 mM Tris-Cl pH 8.0, 10 mM NaCl, 0.2% NP-40 and 1× protease inhibitor cocktail (Roche, no. 11873580001) with douncing in between. Nuclei were then collected and digested with *HindIII* (NEB, no. R3104M) in 1× cutsmart buffer (NEB, no. B7204S) overnight at 37 °C. Digested fragment ends were then labeled with Biotin-14-dCTP (ThermoFisher, no. 19518-018) using DNA polymerase I, large fragment (NEB, no. M0210L). After biotin labeling, nuclei were subjected to proximity ligation using T4 DNA ligase (Invitrogen, no. 15224-090) in a large volume (7.5 ml). Ligated nuclei were then collected by spinning down at 2,500*g* for 5 min, followed by DNA extraction with phenol/chloroform after reverse linking with proteinase K overnight. Purified DNAs were first quantified with the Qubit dsDNA HS assay kit (Invitrogen, no. Q32854) then treated with T4 DNA polymerase (NEB, no. M0203L) to remove unligated DNAs. To generate fragments that can be sequenced, DNAs were subjected to sonication with a Covaris S2 sonicator under the following conditions: duty cycle ten, intensity four, cycles/burst 200 for 55 s. The resulting DNAs were end repaired using the DNA End-Repair kit (Lucigen, no. ER81050). An 'A' was then added to the ends of each fragment using Klenow fragment (3'→5' Exo-) (NEB, no. M0212L). Fragments of 300–500 bp were then selected using homemade Sera-Mag beads. C1 Streptavidin Beads (Invitrogen, no. 650.02) were used to pull down biotin-labeled ligates. After pulling down, beads were washed three times with 400 µl of 1× binding buffer (5 mM Tris-Cl pH 8.0, 0.5 mM EDTA and 1 M NaCl) followed twice with 100 µl of 1× ligation buffer (NEB, no. B0202S). Illumina Truseq adapters were then ligated using T4 DNA ligase (NEB, no. M0202L); 6 pmol of paired-end adapters was used for 1 µg of DNA. The resulting DNAs were then PCR amplified using short primers (Supplementary Table 7). Final libraries were sequenced on the Illumina HiSeq 3000 platform.

4C–seq. The 4C–seq procedure was performed following a published protocol[80]. First, 3–5 million cells were harvested and fixed with 2% formaldehyde then quenched with 125 nM glycine. Fixed cells were then lysed with a cell lysis buffer containing 50 mM Tris-Cl pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100 and 1× protease inhibitor cocktail (Roche, no. 11873580001) for 20–30 min on ice. After lysing, nuclei were collected by spinning down at 2,500*g* for 5 min at 4 °C, followed by washing once with 1× restriction enzyme buffer. Nuclear pellets were then resuspended in 1× restriction enzyme buffer and treated with 0.3% SDS for 1 h at 37 °C under shaking, followed by a further 1 h with 2.5% Triton X-100. Chromatin digestion was then done by incubation of samples with the designated restriction enzyme at the correct temperature overnight while rotating in an airbath. The restriction enzymes used for each locus are listed in Supplementary Table 7. After digestion, heat inactivation at 65 °C was applied to inactivate the enzymes, and nuclei were then subjected to ligation with 50 µl of T4 DNA ligase (Invitrogen, no. 15224-090) in a 7 ml ligation solution at 16 °C overnight. Reverse linking was then performed by treating samples with proteinase K to yield proximity-ligated DNA. Purified DNAs were quantified and subjected to secondary restriction enzyme digestion (roughly one unit of restriction enzyme per 1 µg of DNA) at the suggested temperature overnight. After inactivation of restriction enzymes, samples were then self-ligated with T4 DNA ligase. Ligated DNAs were recovered with sodium acetate and ethanol and quantified with a Qubit dsDNA HS assay kit (ThermoFisher, no. Q32851). The 4C templates were then amplified with designed primers to generate libraries for sequencing. We modified the primer system to make it compatible with the Illumina Nextera system using two sequential PCRs. Locus-specific inverse PCR primers are listed in Supplementary Table 7. For each locus, the 4C templates were amplified with locus-specific primers using 200 ng of template in each reaction, and products from five parallel amplifications were pooled to generate the final 4C library. PCR product aliquots (50 µl) were purified with homemade Sera-Mag beads. One-fifth of the purified DNAs was used for the second PCR using primers N7xx and N5xx, which are the same as Illumina Nextera sample preparation primers. The final products were then purified and subjected to sequencing. Reads for the first cutting site were used for data analysis.

Cloning. For the guide RNA expression vector we used a pX332-original plasmid gifted from the laboratory of J. Wysocka (Standford)[81], which contains an mCherry expression cassette. The dCas9 and dCas9-KRAB expression vectors described in this study were generated on a backbone of Cas9 expression vector—pX330 plasmid (Addgene, plasmid no. 42230) using the In-Fusion cloning method. Both dCas9 and dCas9-KRAB genes were amplified from pHAGE EF1α dCas9-KRAB (Addgene, plasmid no. 50919) with PCR and cloned separately into the AgeI and EcoRI sites of the pX330 plasmid, replacing the Cas9 open reading frame. Detailed information on primers can be found in Supplementary Table 7. All sgRNAs in this study were designed on the CCTop-CRISPR/Cas9 target online predictor[82,83] and manually picked.

GM12878 cell culture and nucleofection. GM12878 cells were maintained in RPMI 1640 medium (Gibco, no. 11875-085) supplemented with 15% FBS (Gibco, no. 16000-044) and 1% penicillin/streptomycin (Gibco, no. 10378-016). Cells were split and seeded at 300,000 ml^{-1} in fresh medium the day before nucleofection. About 4 million cells were prepared for each nucleofection. Briefly, cells were pelleted by centrifugation at 90*g* for 5 min and then resuspended in nucleofection reagent as suggested in the manufacturer's manual (Lonza, SF cell line 4D-Nucleofector X kit, no. V4XC-2024). For each reaction, about 5–7 µg of designated plasmids (dCas9 or dCas9-KRAB combined with pX332-gRNAs, each around 2–4 µg) was applied. Nucleofection was carried out on a 4D Lonza nucleofector with program CM-137. Cells were then left to stand and recover for 24 h in the cell culture incubator before harvesting for RNA extraction or 3C analysis.

RNA extraction and quantitative PCR with reverse transcription. RNA was extracted with Trizol from nucleofected cells following the standard protocol. Complementary DNAs were generated by reverse transcription using M-MLV Reverse Transcriptase (Invitrogen, no. 28025013) following the manufacturer's manual. Quantitative PCR (qPCR) was performed in triplicate.

3C–qPCR. After nucleofection, cells were harvested for 3C assay by fixing with 1% formaldehyde. Cells were lysed using a cell lysis buffer (10 mM Tris-Cl pH 7.5, 10 mM NaCl, 0.2% NP-40 and 1× proteinase inhibitor cocktail) with douncing 30× in between, on ice, for about 20 min. Cell nuclei were pelleted by centrifugation at 2,500*g* for 5 min at 4 °C, then nuclei were digested overnight with *MboI* (NEB, no. R0147M; 400 U for about 4 million cells) at 37 °C. After heat inactivation of *MboI*, proximity ligation was performed overnight with T4 DNA ligase (Invitrogen, no. 15224-025) at 16 °C. Proximity-ligated chromatins were reverse linked by treatment with proteinase K at 65 °C overnight and then purified by phenol/chloroform. To generate random ligation control for 3C–qPCR, we picked BAC clones covering the two anchors of the loop of interest (a list of BAC clones is provided in Supplementary Table 7) and performed the 3C procedure on DNA prepared from BAC clones.

Sequencing data analysis. *Hi-C data mapping, filtering and normalization.* Conventional Hi-C. Because some conventional Hi-C libraries are sequenced with paired-end 36 bp (for example, human tissue datasets), for the purposes of consistency and convenience we trimmed all conventional Hi-C data to 36 bp. Each end of the raw reads was mapped separately to the hg19 (for human) or mm10 (for mouse) reference genome using bowtie (v.1.1.2)[84]. Sam files were then paired with an in-house script. After removal of PCR duplications, we first discarded reads with both ends mapped to the same *HindIII* fragments as invalid pairs. All remaining read pairs then represented two different *HindIII* fragments in *cis*. Because cut-and-ligation events are expected to generate reads within 500 bp upstream of *HindIII* cutting sites due to size selection ('+' strand reads should be within 500 bp upstream of a *HindIII* site and '−' strand reads should be within 500 bp downstream of a *HindIII* site), we retained only read pairs with both ends satisfying these criteria. We next split all remaining reads into three classes based on their strand orientation ('same-strand', 'inward' or 'outward'). We retained inward read pairs if the distance between two reads was >1 kb, and outward read pairs if the distance between two reads was >25 kb. We then merged the filtered inward, filtered outward and same-strand as the *cis* reads pair. The *HiCorr* '*HindIII*' mode was used to acquire bias-corrected 5-kb anchor loop files from *cis* and *trans* fragment read pairs.

In situ Hi-C and micro-C. Full-length reads (150 bp for in situ GM12878) were used for alignment to enable more reads overlapping SNPs for allele-resolved analysis. After removal of PCR duplicates and read pairs classification, we filtered out the outward read pairs with distance <5 kb and inward read pairs with distance >1 kb. The filtered read pairs were then mapped to *Mbol* fragment pairs, with the *HiCorr* '*Bam-process-DpnII*' mode used for bias correction. H1 micro-C processing followed similar steps: we used 5-kb bins to map read pairs and Juicebox (v.1.18.08)[85] '*pre*' to convert 5-kb bin pairs to '*hic*' format and ran KR normalization. We then dumped the contact pairs and performed distance correction with in-house scripts. In brief, we split all contact pairs within 2 Mb by loop distance into 400 groups with 5 kb as the interval. In each distance group, the KR-normalized value was normalized by average values within the same group. Here, we called the normalized value from KR normalization and distance correction as KR-ratio.

Single-cell Hi-C preprocessing. Processed *DpnII* fragment contact files for 4,098 mouse embryonic stem cells were downloaded from the original study (Supplementary Table 1). Fragment pairs were then mapped from mm9 to mm10 using the liftover tools from UCSC. The number of *cis* contacts within 2 Mb was used to rank cells. We took the top-ranked cells of a certain number (~1–92) and merged the fragment contacts files for *cis* and *trans* separately and mapped them to ~10-kb anchor pairs. *HiCorr* '*DpnII*' mode was used to correct bias at the anchor

level. The 'contact read pairs' files for human PFC sn-m3C-seq and cell type labels identified from the methylation profiled in the same cell were downloaded from the original study (Supplementary Table 1). We aggregated cells from the same cell type, filtered reads pairs as in situ Hi-C steps and further mapped read pairs to *DpnII* fragment pairs. Due to the sparsity and limited depth of each cell type, we further converted fragment pairs to ~10-kb anchor pairs. For each cell type, the merged *cis* anchor contact file and *trans* anchor pairs were taken as input to run *HiCorr DpnII* mode.

4C–seq. Data for 4C–seq were analyzed using pipe4C (v.1.1.3)[80] to generate bam and wig files for visualization.

AS mapping for Hi-C, ChIP–seq, RNA sequencing and 4C–seq. We first masked the hg19 reference genome with SNPs downloaded from the original study (Supplementary Table 1) and built an index for bowtie2 (v.2.2.6)[86] and Hisat2 (v.2.1.0)[87].

Hi-C. Each end of the raw reads with the full length (150 bp) was mapped separately to the masked hg19 genome by bowtie2 (v.2.2.6). SNPsplit (v.0.3.4)[88] was utilized to assign mapped reads in bam files to two alleles using the SNP information. The read pairs filtering step was the same as for in situ Hi-C (In situ Hi-C and micro-C). *HiCorr DpnII* mode was used for bias correction. The *LoopEnhance* model trained by 50 million data was used to enhance the two 5-kb-resolution contact data from the two alleles. The top 300,000 loops from two datasets were combined and then loops with at least twofold difference between the enhanced loop strength of the two alleles were defined as AS loops; ultraspecific loops were defined by a tenfold difference.

ChIP–seq. FASTQ files were mapped to the masked hg19 genome by bowtie2 (v.2.2.6). SNPsplit (v.0.3.4) was used to assign mapped reads in bam files to two alleles using the SNP information. *macs2* (v.2.2.7.1)[89] was used to call peaks.

RNA sequencing. FASTQ files were mapped to the masked hg19 genome by Hisat2 (v.2.1.0), SNPsplit (v..3.4) was used to assign mapped reads in bam files to two alleles using the SNP information. We used FeatureCounts (v.1.6.1) to summarize the mapped reads for each gene across samples. Reads on the same allele from different samples were merged. A binomial test was performed to calculate *P* values comparing expression levels between two alleles for each gene (background possibility, 0.5). X-inactivation resulted in an imbalance of gene activity between X_a (maternal) and X_i (paternal) genomes; escape genes were defined as those with a ratio >0.2:

$$\text{ratio} = \text{expr}_{\text{Xi}}/\text{expr}_{\text{Xi}} + \text{expr}_{\text{Xa}}$$

Where expr_{Xi} is the paternal (X_i) expression of the gene and expr_{Xa} is the maternal (X_a) expression of the gene.

4C–seq. Data for 4C–seq were analyzed using pipe4C (v.1.1.4) to generate bam and wig files for visualization. We further converted bam files to bed format and extracted the reads of overlapping SNPs and split them into maternal and paternal bed files. For each SNP, we summarized the overlapped reads on maternal and paternal genomes, calculated allele imbalance using the formula in equation (2) and visualized it on a UCSC genome browser:

$$\text{Allele imbalance} = (M - P)/M + P$$

Where *M* is 4C reads assigned to the maternal genome on each SNP and *P* is 4C reads assigned to the maternal genome on each SNP.

Data representation and model structure in *DeepLoop*. *Data representation.* To train deep learning models on Hi-C contact matrices, we need to represent the data in a way that is more computationally tractable than holding each full chromosome matrix in memory. We took each full chromosome matrix and split it into nonoverlapping, equally sized submatrices lying within the 2-Mb band. For a single genome using our selected submatrix size of 128×128, we used on average ~18,000 unique submatrices per replicate when training a model, although we used random cropping and shifting to further augment the training dataset. Once the model was trained, each of these submatrices was passed into the model separately and the full chromosome matrix was reconstructed from the outputs of the trained model.

LoopDenoise. Denoising autoencoders. A convolutional autoencoder[90] is a type of neural network that consists of an encoder function and a decoder function. The encoder maps an input vector to a lower-dimensional latent representation using successive convolution layers combined with some form of dimensionality reduction, such as pooling layers or strided convolutions. The decoder then maps this representation to a reconstructed vector using transpose convolutions or some other form of upsampling. Autoencoders can be thought of as a function f_θ parameterized by θ, which maps each input vector X_i from a given dataset to a reconstructed vector $f_\theta(X_i)$. Classical autoencoders try to learn an approximation to the identity function using the input vector as the training target[91]. That is, for dataset *X* the model tries to minimize the loss between each input vector and the reconstructed output. Mean squared error is commonly used as the loss function:

$$\theta^* = \text{argmin}_\theta \left[\frac{1}{n} \sum_{i=1}^{n} \left(f_\theta \left(X_i \right) - X_i \right)^2 \right]$$

Denoising autoencoders are a specific type of autoencoder that attempts to learn a mapping from noisy input vectors to clean, ground truth targets[92]. Contrary to classical autoencoders, these denoising models attempt to minimize the loss between target vector $\widehat{X}_i$ and the reconstructed output:

$$\theta^* = \text{argmin}_\theta \left[\frac{1}{n} \sum_{i=1}^{n} \left(f_\theta \left(X_i \right) - \hat{X}_i \right)^2 \right]$$

This target vector has some desirable properties, such as being noise free and having higher resolution than the input vector. Building a denoising autoencoder usually involves starting from clean ground truth data as the target vectors and corrupting them to generate the input vectors. If the goal of the model is to be robust to noise, we could corrupt ground truth data by adding random noise; however, in the case of Hi-C contact matrices, the data already contain noise and thus training a convolutional autoencoder to denoise Hi-C data requires a more desirable training target. We obtain cleaner training targets by statistically filtering out insignificant signals from high-depth data using biological replicates.

Training set. For model training, we picked a published *HindIII*-based Hi-C dataset in human fetal cerebral cortex[24]. The data were generated for three donors, each of which has one library from CP and one from GZ. All six libraries have roughly the same sequencing depth, and the pooled data of all six libraries have ~470 million mid-range *cis* contacts (Supplementary Table 2). We disregarded the difference between CP and GZ and split the Hi-C data into three biological replicates, each replicate having ~140–150 million mid-range *cis* contacts combining CP and GZ libraries from the same donor. We applied *HiCorr* to each of the three replicates and extracted ~18,000 submatrices at ~5–10-kb resolution (within the 2-Mb range) from every replicate as training sets

Training target. The training target for *LoopDenoise* should contain significant and reproducible signals with as little noise as possible. To generate these targets, we pooled all libraries and applied *HiCorr*; the heatmaps from pooled data will thus be less noisy due to higher sequencing depth (Fig. 1c). *HiCorr* provides *P* values for every pixel in the heatmaps from individual replicates and the pooled data. We then removed pixels from the pooled heatmaps with $P > 0.05$ due to lack of signal enrichment. We then required the remaining pixels to be significant ($P < 0.05$, negative binomial test) in at least one of the biological replicates. The resulting pixels were used as the ground truth training target in our convolutional autoencoder. All remaining pixels were assigned a zero value, indicating no interaction. Even though these training targets were not completely noise free, results show that our model is able to learn a meaningful latent representation for the true loop signals and also to output Hi-C submatrices that are even cleaner than the training target used. This is probably because the model is forced to learn an average of noise-free matrices that could explain the noisy observation, rather than learning the perfect mapping to our training target, which is not noise-free.

Model structure. The encoder of *LoopDenoise* (Fig. 1a and Extended Data Fig. 1a) consists of two instances of a convolution layer followed by a rectified linear unit (*ReLU*) activation function and a maximum pooling layer. The decoder half of *LoopDenoise* consists of two transpose convolutions followed by a final convolution layer and *ReLU* activation. Each convolution layer has eight filters except for the final layer, which has only one, to return the correct number of output channels. The convolution layers in both the encoder and final convolution layer use a filter size of 13×13 while the transpose convolutions in the decoder use a filter size of 2×2. Because the maximum pooling layers act on a 2×2-region, after each pooling layer in the encoder the size of the input is halved. For each transpose convolution layer the size of the input is doubled, giving us the same size output as the input. We applied zero-padding to the edges of each input submatrix to ensure that the output size of each convolution or transpose convolution remained unchanged. The output of each convolution layer with *ReLU* activation was computed as follows:

$$h_i \left(x \right) = \max \left(0, w_i * x + b_i \right)$$

where we define the discrete convolution operation * as the weighted sum of neighboring pixels using weights w_i as the convolution kernel, b_i as the bias and x as the input matrix—either a Hi-C submatrix for the first layer or the output of a previous layer for subsequent layers. This operation was performed at every pixel of the input matrix using a stride value of 1 to move the convolution window across the input space one pixel at a time. In the transpose convolutions we performed the same mathematical operation but we transformed the input by inserting padding between the input values to simulate a fractional stride value, which therefore maps

each pixel to multiple different values, increasing the size of the input matrix to perform the upsampling necessary in the decoder.

Model training. The model was trained by minimization of the mean squared error (MSE) of the reconstructed outputs and the combined targets using the Adam[93] optimizer, with a learning rate of 0.001 and default hyperparameters. We used a submatrix size of 128×128 and a batch size of four training for 50 epochs. Three normalized CP–GZ merged replicates were used for training, and chromosomes X and Y were ignored during training. When training this autoencoder architecture, MSE did not reach zero; this would indicate that our model is overfitting to our training targets and has memorized only mapping from inputs to targets without learning a useful latent representation that generalizes to novel examples. To avoid this, we used GM12878 replicates as a validation dataset and monitored both loss and reproducibility on this validation set to ensure that the model would successfully generalize.

Hyperparameter exploration. To find the optimal model for denoising we trained multiple models with different hyperparameters on human fetal brain datasets and validated the model using GM12878 replicates. We tested different filter sizes to determine whether the inclusion of more information from neighboring regions would lead to improved performance. We evaluated reproducibility among the training and validation replicates and determined the optimal filter size as 13×13. We also tested performance when using a stride value of 2 in the convolution layers of the encoder rather than in the pooling layers. This would perform the same amount of dimensionality reduction, but each convolution would potentially give us less information than when using pooling layers. We found that maximization of pooling layers slightly improved reproducibility on our training and validation datasets. This makes sense, because using a stride value of 2 means that some pixels are never convolved with their neighbors before the dimensionality reduction step and thus the model loses information about certain regions. Compared with a convolution of stride value of 1 followed by maximized pooling, we captured the full relationship between each pixel and its surrounding region then selected the maximum value among a small group of these pixels. The latter method is more specific in regard to the information that is forgotten when performing dimensionality reduction whereas the former, using a stride value of 2 without pooling, randomly loses information based on the location of each pixel.

LoopEnhance. Model structure. The U-Net architecture (Fig. 2a) is a fully convolutional network similar to, but much larger than, the convolutional autoencoder used in the denoise model. It contains an encoder and a decoder, with the main addition being skip connections that concatenate feature maps from each stage of the encoder to each corresponding stage of the decoder. The goal of these skip connections is to maintain the localization and different scales of features when upsampling during the decoder path. Since the receptive field of the convolutions at the final layer of the encoder is very large compared with the size of our input submatrices, we found that deep convolutional autoencoders without these skip connections produce very cloudy/blurry signals, whereas concatenating feature maps across the different depths of the model yield more precise signals in the output. The encoder of *LoopEnhance* contains ten convolution layers with four pooling layers. Our model has a depth of four because it has four 'blocks' of convolutions followed by dimensionality reduction steps. The input is a Hi-C submatrix of size 128×128. We successively applied two convolution layers with *ReLU* activation followed by a pooling layer to produce final feature maps with dimensionality 64×8×8=4,096. Since we use U-Net architecture, we also retain the feature maps at each depth of the network. The convolution layers in the first block of the model use four filters, and this number of filters is doubled at each depth, eventually reaching the $2^6=64$ filters found in the final convolution layer. The decoder of *LoopEnhance* consists of 13 convolution layers with four upsampling layers. The upsampling layers are instances of an upconvolution function that simply turns each pixel into a 2×2-region of identical values, then applies a convolution layer with *ReLU* activation. In practice, this is very similar to a transpose convolution. However, in deep networks transpose convolutions can propagate padding artifacts to the output of the model. Following each upsampling layer, we applied two convolutions with *ReLU* activation. The number of filters is now halved after each upsampling layer, starting at 64 filters following the latent encoding and eventually reaching four filters. After the final upsampling layer and its following two convolutions, we applied one final convolution layer with one filter and *ReLU* activation to obtain an output with a single channel.

Model training. The input to the model is a low-depth-normalized Hi-C submatrix, and the training target is the corresponding denoised high-depth-normalized submatrix obtained using the denoise model. This is the main distinction between our model and previous works such as *HiCPlus*[20] and *HiCNN*[21]. Zhang et al.[20] note that training a neural network to map low- to high-depth Hi-C data assumes that the high-depth target used is the ground truth. Although many deep learning models are able to distinguish between noise and true signals, natural variation among Hi-C replicates introduces multiple valid explanations for each low-depth input. The increased replicate reproducibility achieved by *LoopDenoise* facilitates training of *LoopEnhance* using a ground truth target with less noise and variation. Our model minimizes MSE between the enhanced output and denoised high-depth targets. We also used a larger submatrix size of 128×128 compared with *HiCPlus* and *HiCNN*, which use 40×40. This larger submatrix size allows our model to map each input submatrix to a richer scale of features while still using minimal padding in the convolution layers. Because our model is a fully convolutional network, once trained it can enhance submatrices of any size, although we recommend using the same size used for training because padding artifacts are possible with small submatrix sizes.

Hyperparameter exploration. To determine the optimal model for enhancement of low-depth contact matrices, we trained multiple models with different hyperparameters on the 10% downsampled CP–GZ merged replicates and validated the model using downsampled GM12878 replicates. We tested different filter sizes to determine whether the inclusion of more information from neighboring regions would lead to improved performance. Like *HiCPlus*[20], we found that larger filters do improve performance to an extent: filters larger than 9×9 showed no substantial improvements, so we decided on a final filter size of 9×9.

Hi-C data visualization. Heatmaps were used to visualize Hi-C contact profiles. The color scales for heatmaps (raw, expected, ratio) were selected based on the contact matrix. Because the brightness of pixels in raw, ratio and *DeepLoop* heatmaps represents different things, we use different strategies to determine color scales:

(1) Raw heatmaps represents read counts; the brightest red color indicates the 98th percentile of the contact matrix. Color is proportionally scaled down to one read (white).
(2) *HiCorr* heatmaps represents ratios; the brightest red color indicates at least twofold enrichment. Color is proportionally scaled down to onefold (no enrichment).
(3) *DeepLoop* heatmaps output 'transformed fold change' that represents only relative levels of signal enrichment (that is, a value of onefold may no longer be the real cutoff for no enrichment). We therefore set the brightest red color as the lower limit of the top 300,000 pixels genome wide. Color is proportionally scaled down to half of that lower limit or onefold, whichever is higher.

Loop curves in the figures were sourced from the UCSC Genome Browser by uploading the top 300,000 loops in the format 'biginteract'. Triangle heatmaps were sourced from the UCSC Genome Browser[94] by uploading the '*hic*' file generated by Juicebox.

Statistics. All statistical methods and tests used in this paper are described in the main text, figure legends, Methods and Supplementary Information as appropriate.

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

Data availability

Accession numbers for third-party data used in this study can be found in Supplementary Table 1. The raw data of H9 Hi-C and 4C–seq generated in this study, and reanalyzed published data, can be found at accession no. GSE167200. The 40 Hi-C datasets analyzed by *DeepLoop* can be found at https://hiview.case.edu/public/DeepLoop/.

Code availability

The code is available is available at Zenodo (https://doi.org/10.5281/zenodo.6495831) and github (https://github.com/JinLabBioinfo/DeepLoop).

References

80. Krijger, P. H. L., Geeven, G., Bianchi, V., Hilvering, C. R. E. & de Laat, W. 4C-seq from beginning to end: a detailed protocol for sample preparation and data analysis. *Methods* **170**, 17–32 (2020).
81. Gu, B. et al. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**, 1050–1055 (2018).
82. Labuhn, M. et al. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res.* **46**, 1375–1385 (2018).
83. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS ONE* **10**, e0124633 (2015).
84. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
85. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

88. Krueger, F. & Andrews, S. R. SNPsplit: allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res.* **5**, 1479 (2016).
89. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
90. Xiao, X. et al. Endogenous reprogramming of alpha cells into beta cells, induced by viral gene therapy, reverses autoimmune diabetes. *Cell Stem Cell* **22**, 78–90 (2018).
91. Gondara, L. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* 241–246 (IEEE, 2016).
92. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. *Citeseer* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.569.2442&rep=rep1&type=pdf (2008).
93. Kingma, D. P. & BA, J. *Adam: a method for stochastic optimization* https://doi.org/10.48550/arXiv.1412.6980 (2014).
94. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Acknowledgements

This work is supported by grants from the National Institutes of Health (nos. R01HG009658 to F.J. and R01DK113185 to Y.L.) and Mount Sinai Health Care Foundation (nos. OSA510113 to F.J. and OSA510114 to Y.L.). F.J. is also supported by a subaward from the University of Miami (no. NIH U01AG072579) and a Cancer Data Sciences pilot grant from Case Comprehensive Cancer Center Support Grant (no. NIH P30CA043703). J.L. is supported in part by National Science Foundation grant nos. CCF-2006780 and CCF-1815139. D.P. is supported by a NIH training grant (no. T32HL007567) and a fellowship from the Callahan Foundation. This work made use of the High-Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

Author contributions

F.J., J.L. and Y.L. designed the study. S.Z. and D.P. performed analyses. L.L. and J.C. performed validation experiments. W.X., X.L. and N.P. helped twith Hi-C data analyses. M.W., J.S., D.S. and P.F. helped analyze mESC pcHi-C data. S.Z., D.P. and F.J. wrote the manuscript with help from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

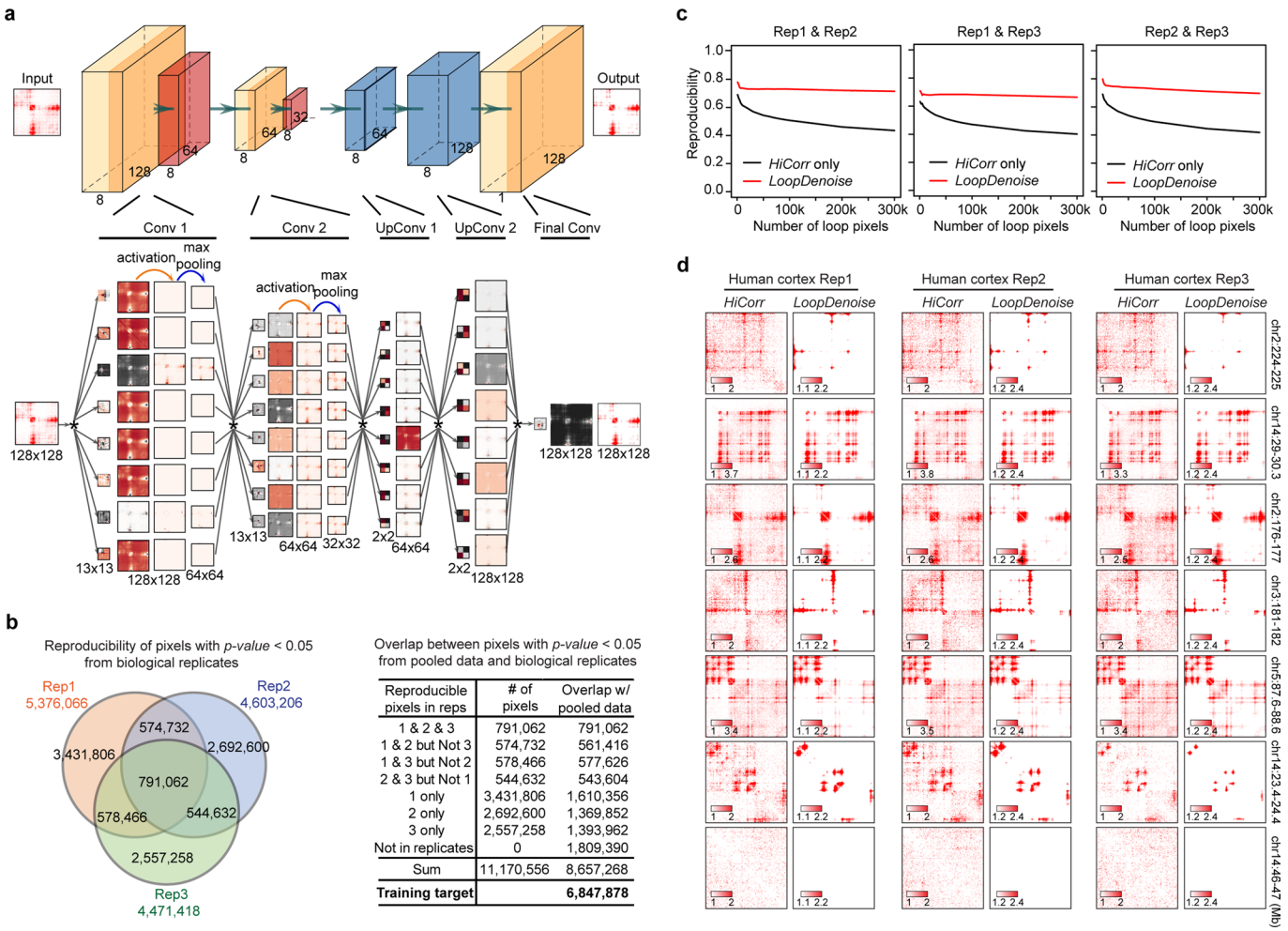
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01116-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01116-w.

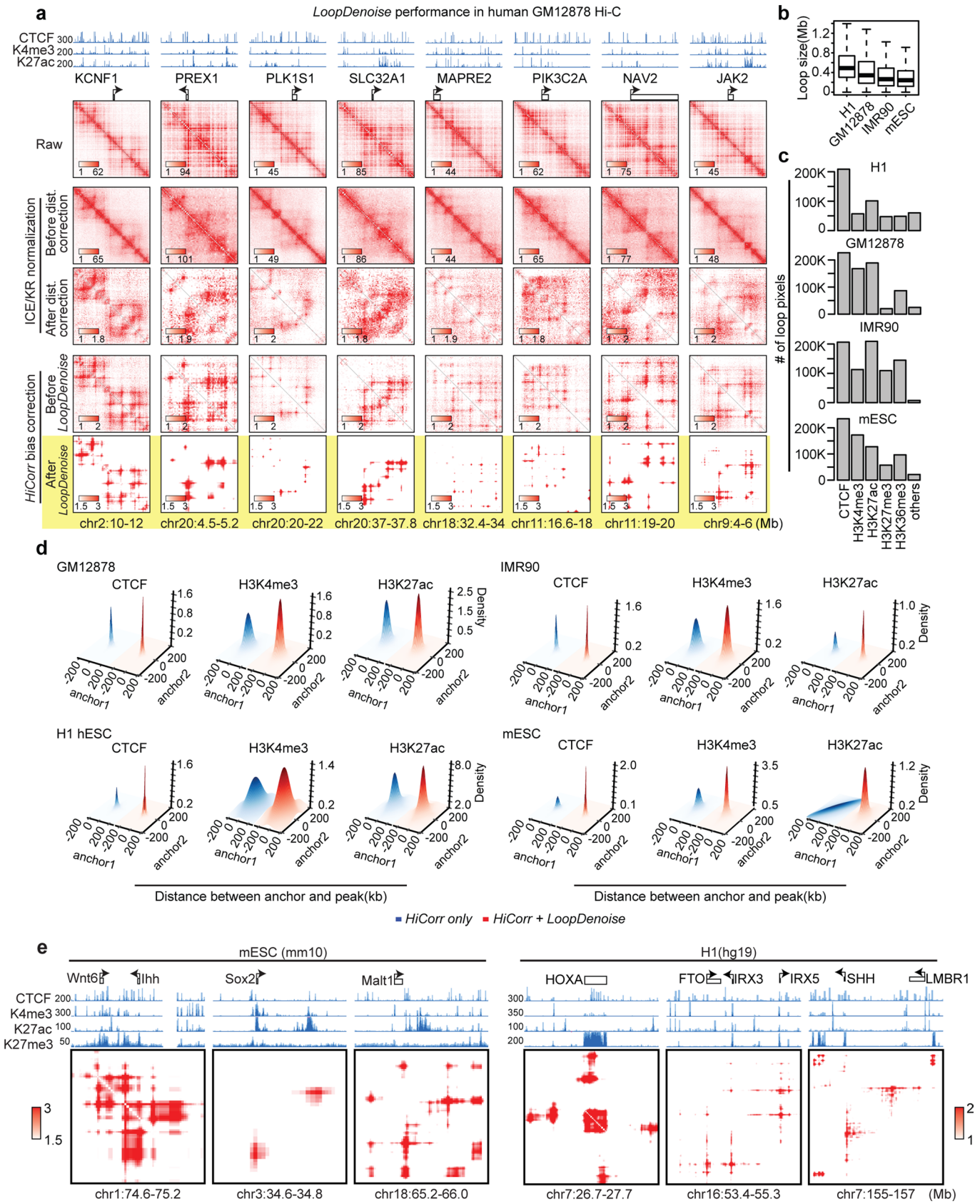
Correspondence and requests for materials should be addressed to Yan Li, Jing Li or Fulai Jin.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

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

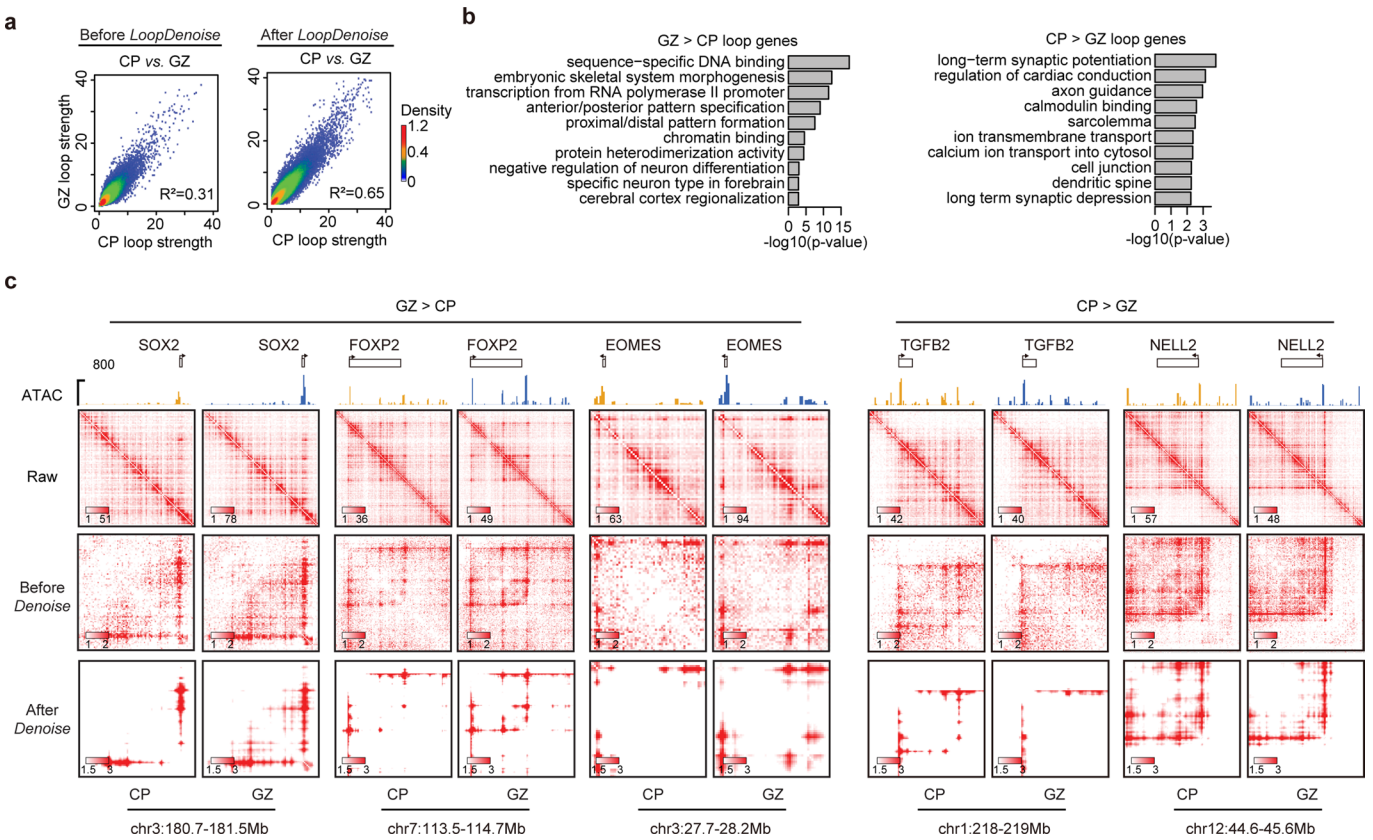


Extended Data Fig. 1 | *LoopDenoise* training procedure, performance and visualization. a, Detailed *LoopDenoise* convolutional autoencoder model architecture showing five convolution layers, two in the encoding path using eight 13×13 filters, two transpose convolution layers in the decoding path using eight 2×2 filters and one final convolution layer using a single 13×13 filter. The matrices dimensions of each layer output were also shown. Each layer is visualized by the filters used, the output of convolving the input with this filter, the result of applying *ReLU* activation and the result of max pooling. The convolution operation is denoted by *. **b**, Venn diagram showing the reproducible loop pixels between three human fetal brain replicates. The table showing the number of overlapped pixels between significant pixels in the pooled data and each part of pixels shown in the Venn diagram. The pixels that are significant in both pooled data and at least one of the three replicates are the training target in the *LoopDenoise* model ($P < 0.05$, negative binomial test). The significance of loop pixels come from the negative binomial test wrapped in *HiCorr* package. **c**, Pairwise reproducibility at pixel level (defined as the fraction of common ones when calling the same number of loop pixels from two datasets) between biological replicates of human fetal cortex Hi-C data, when the same numbers of the loop pixels were called. **d**, The heatmap examples from 7 locus in three human fetal brain replicates, and *LoopDenoise* output showing more reproducible contact patterns.

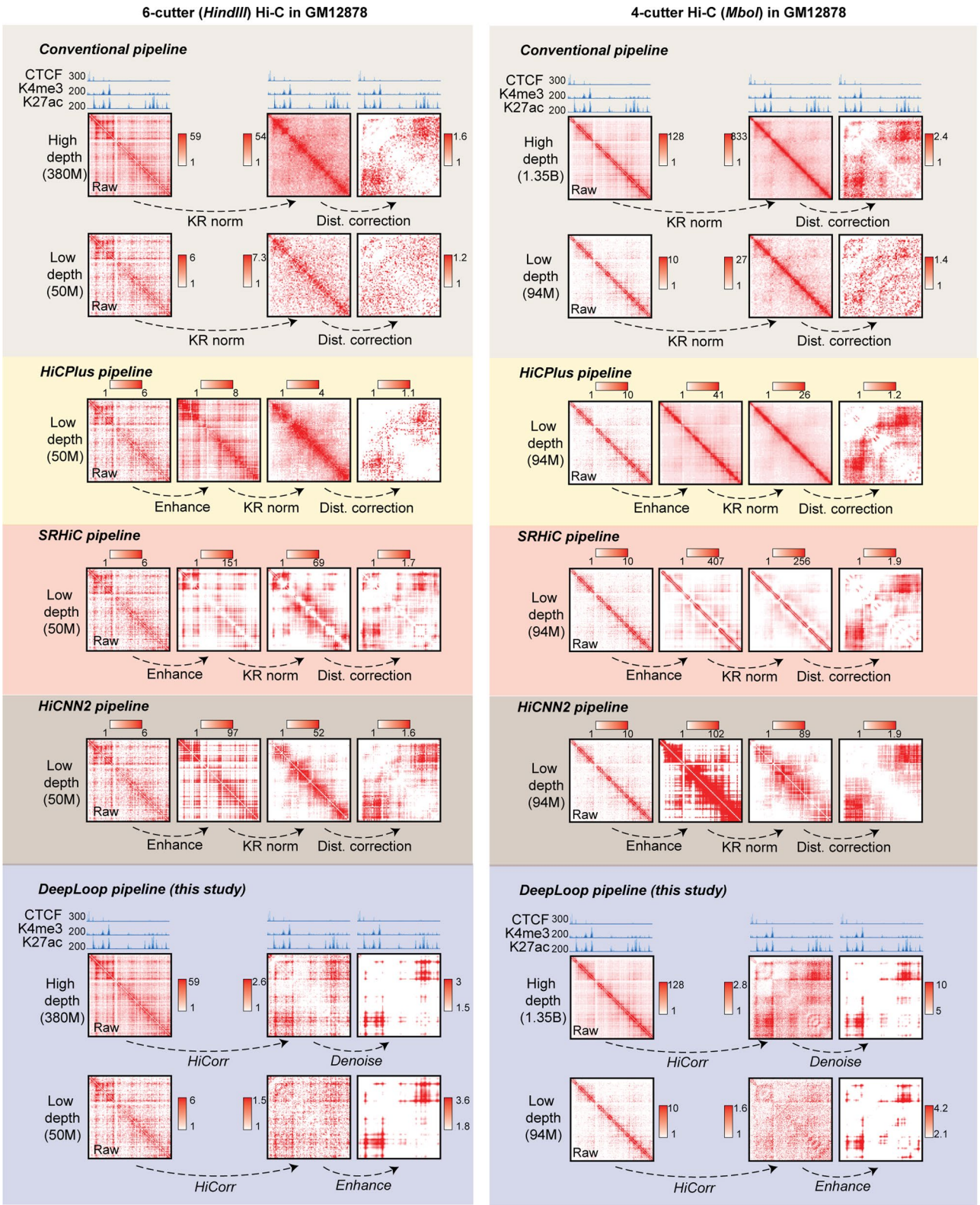


Extended Data Fig. 2 | See next page for caption.

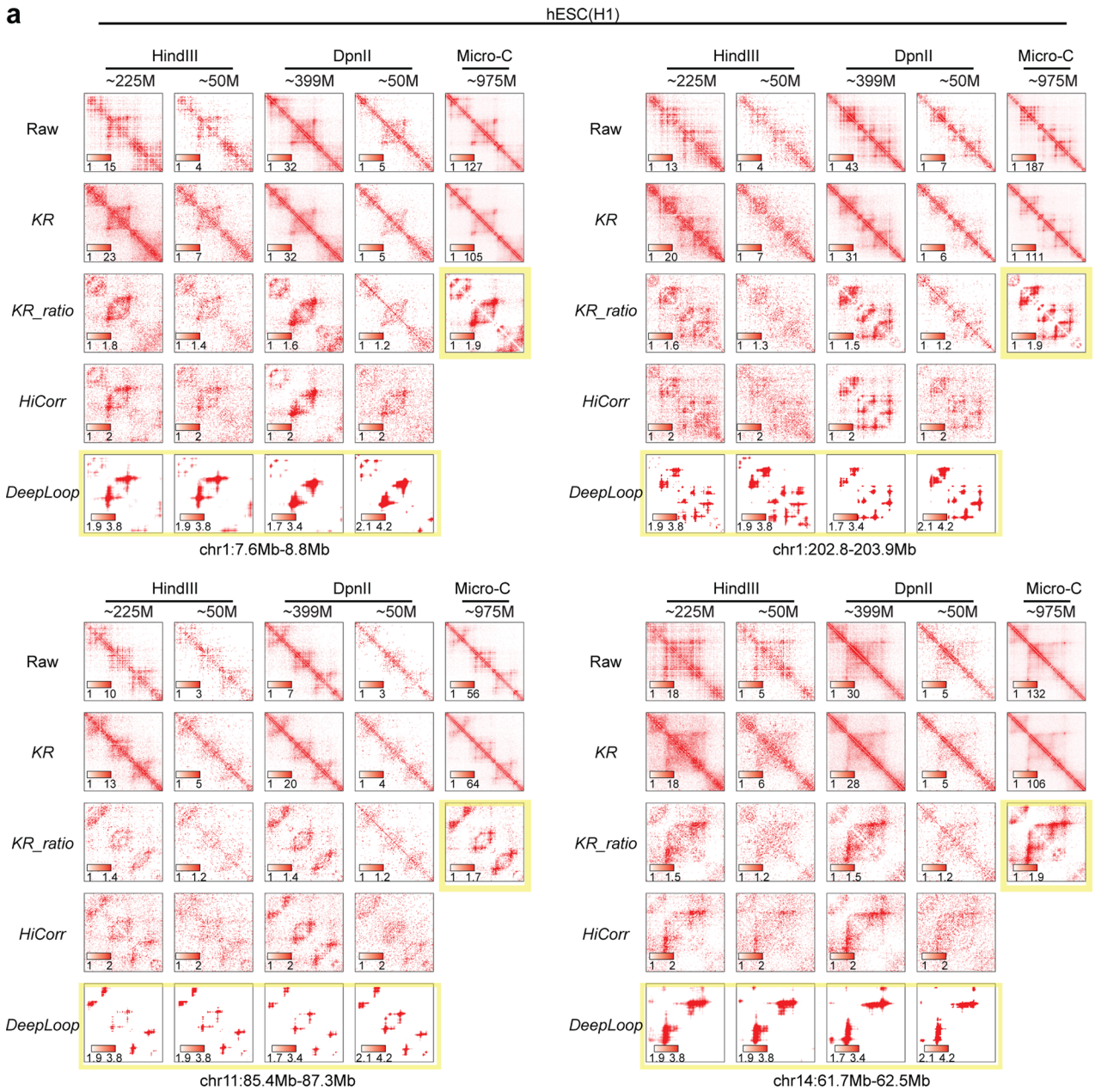
Extended Data Fig. 2 | *LoopDenoise* generalization across cell types and species. a, Eight heatmap examples in GM12878, the highlight row is the output from *LoopDenoise*. **b**, The distance distribution of top 300K pixels in H1(hESC), GM12878, IMR90 and mESC. Upper and lower limits of boxes indicate interquartile ranges, center lines indicate median values, whiskers indicate values with a maximum of 1.5 times the interquartile range and outliers indicate values beyond 1.5 times the interquartile range. **c**, The number of loops pixels with at least one anchor overlapped with ChIP-seq peaks out of top 300K pixels. **d**, Density plots show the distribution of distances between loop anchors (top 100K loop pixels used) and their nearest ChIP-seq peaks in GM12878, IMR90, H1(hESC) and mESC. **e**, The heatmap examples of six loci with known long-range gene regulation. The height of browser tracks indicating the raw counts of ChIP-seq.



Extended Data Fig. 3 | *LoopDenoise* enables the quantitation of dynamic chromatin interactions. a, Scatterplots showing the pixel-level correlation between CP and GZ sample in human fetal cortex before and after *LoopDenoise*. The R-square values were also shown in the plots. **b**, GO analyses of genes associated with GZ- or CP-specific loops. Fisher's Exact test was used to measure the gene-enrichment in annotation terms. **c**, The contact heatmaps of selected gene loci with top GZ- or CP-specific loop pixels. ATAC-seq tracks in CP (yellow) and GZ (blue) are also included for comparison. The height of browser tracks indicating the raw counts of ATAC-seq.



Extended Data Fig. 4 | Compare the performance of different pipelines on 6-cutter and 4-cutter Hi-C data in GM12878 cells. For 4-cutter Hi-C datasets, we chose a 94M down-sampled dataset (1/16 of the original depth) used in *HiCPlus*, *HiCNN2* and *SRHiC* studies, and the 1.35 billion full-depth as reference. For 6-cutter Hi-C datasets, we chose a 50M down-sampled dataset and the 380M full-depth as reference. For locus chr5:87,964,000-88,764000, the left side showed the contact heatmaps from 6-cutter (*HindIII*) GM12878 Hi-C processed by different pipelines (colored in background). The right side showed the 4-cutter (*MboI*) GM12878 Hi-C. The height of browser tracks indicating the raw counts of ChIP-seq.

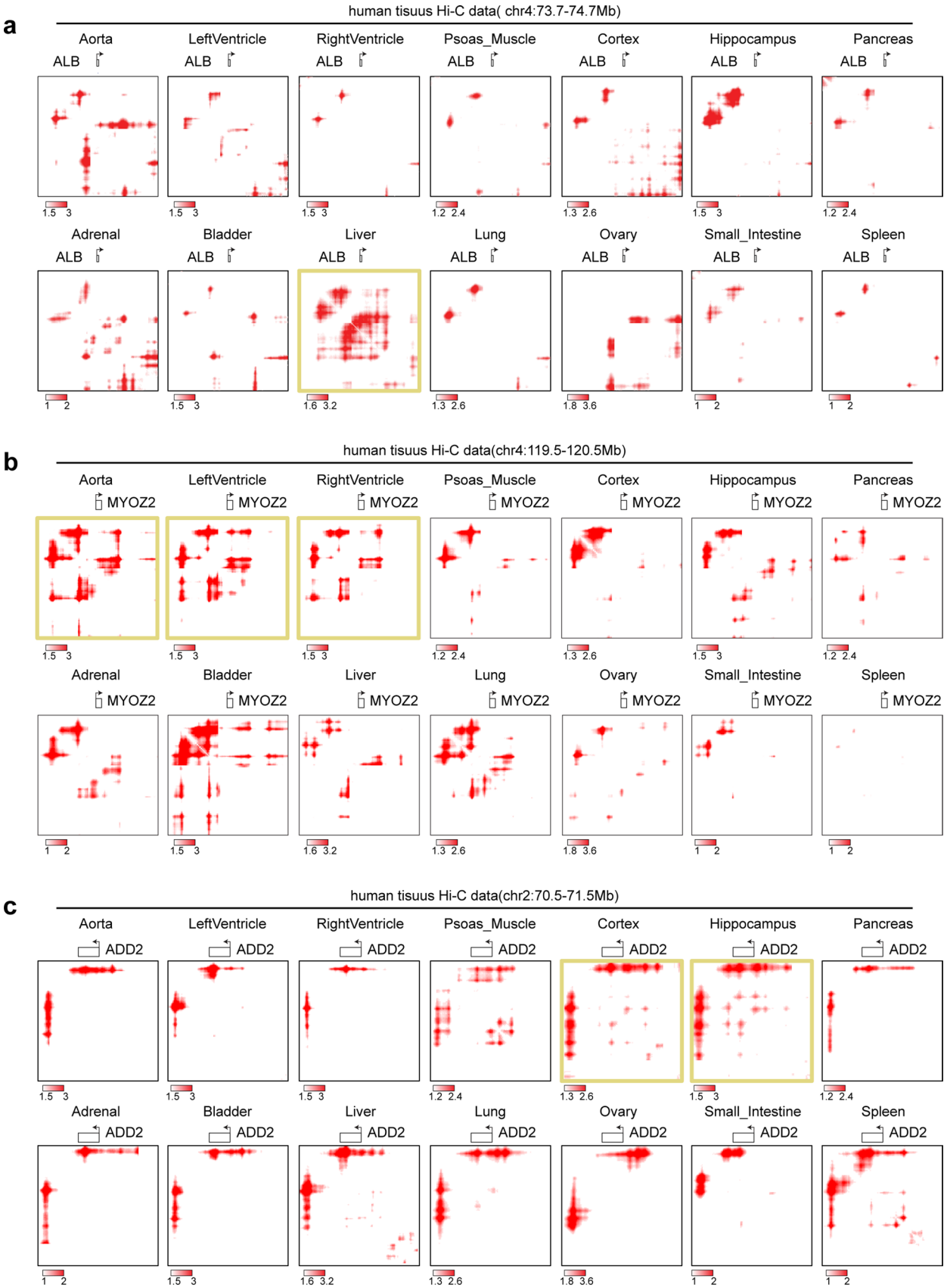


b

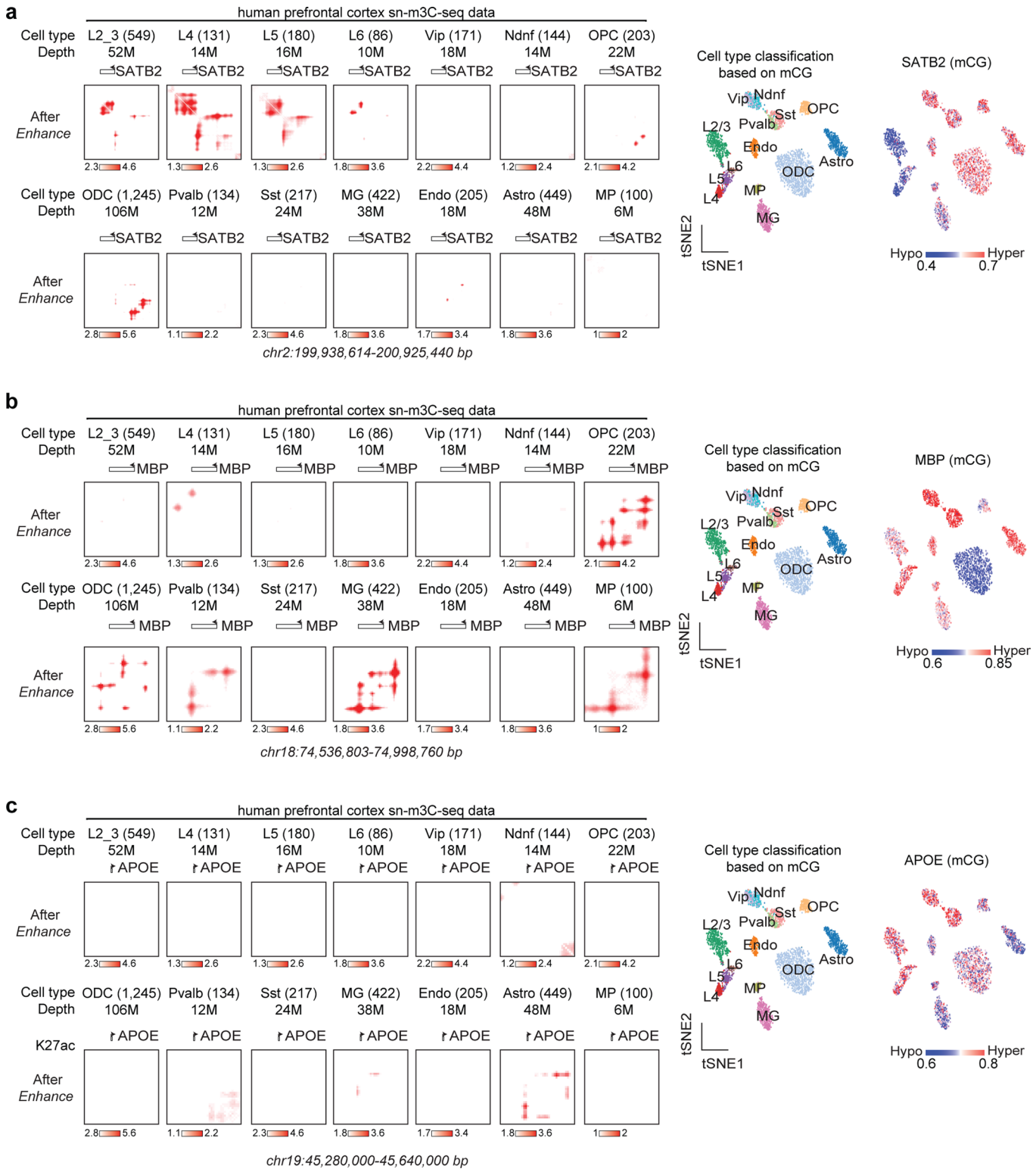
Table: Recovery of Micro-C HiCCUPs loops with DeepLoop enhanced Hi-C data

Loop size	<50kb	50~100kb	>100kb
Total Micro-C HiCCUPs loops	3,489	3,424	10,587
Recovered by 50M *HindIII*-Hi-C after DeepLoop	762 (22%)	1,276 (37%)	6,188 (58%)
Recovered by 50M *DpnII*-Hi-C after DeepLoop	1,809 (52%)	2,090 (61%)	7,656 (72%)

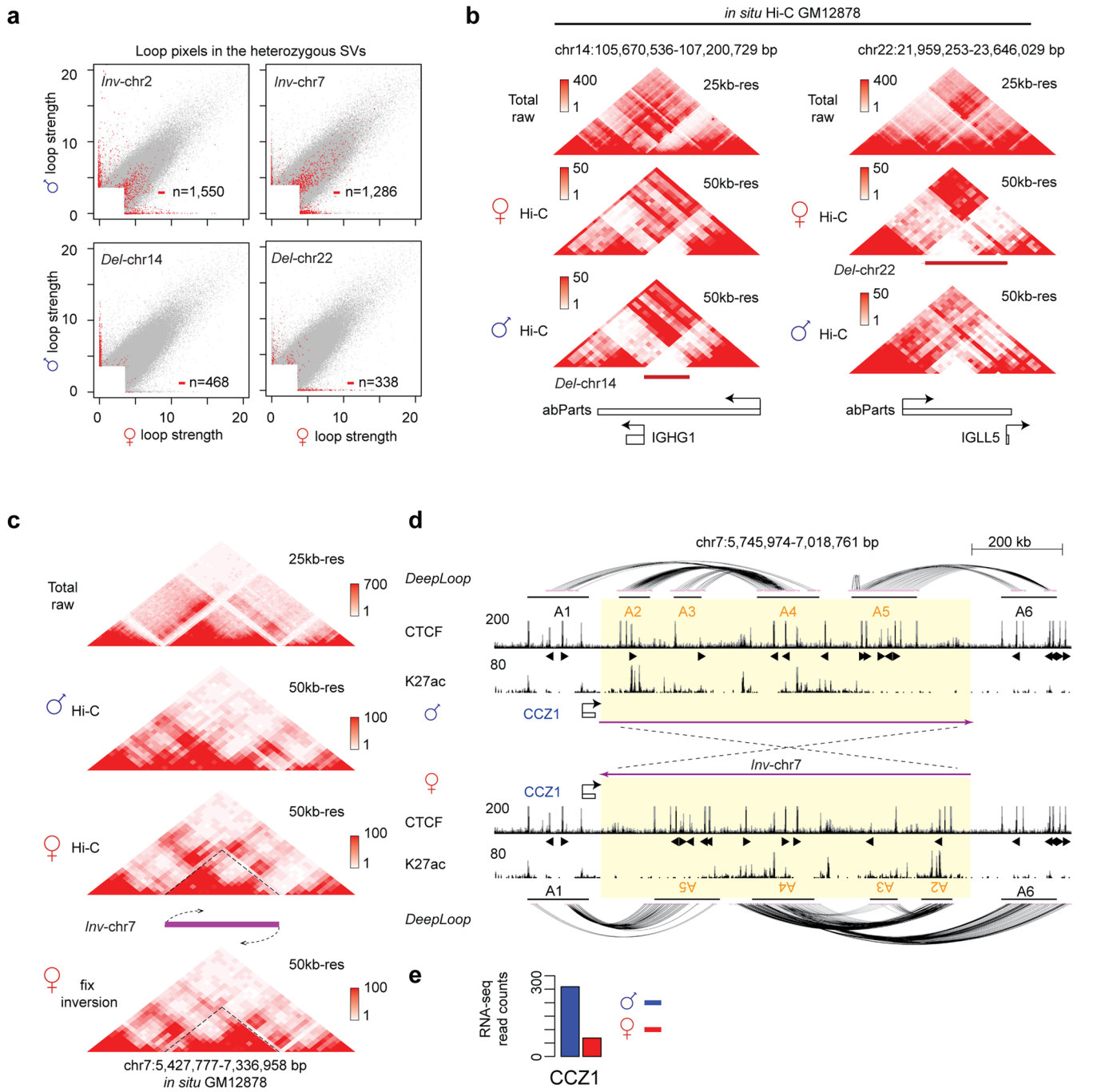
Extended Data Fig. 5 | Compare the consistency of Hi-C and Micro-C in H1. a, Similar to Fig. 3a, b, more heatmap examples at 4 loci. **b**, Size breakdown of recovered micro-C HICCUPS loops by 50M deep *HindIII*- or *DpnII*- Hi-C after enhancement.



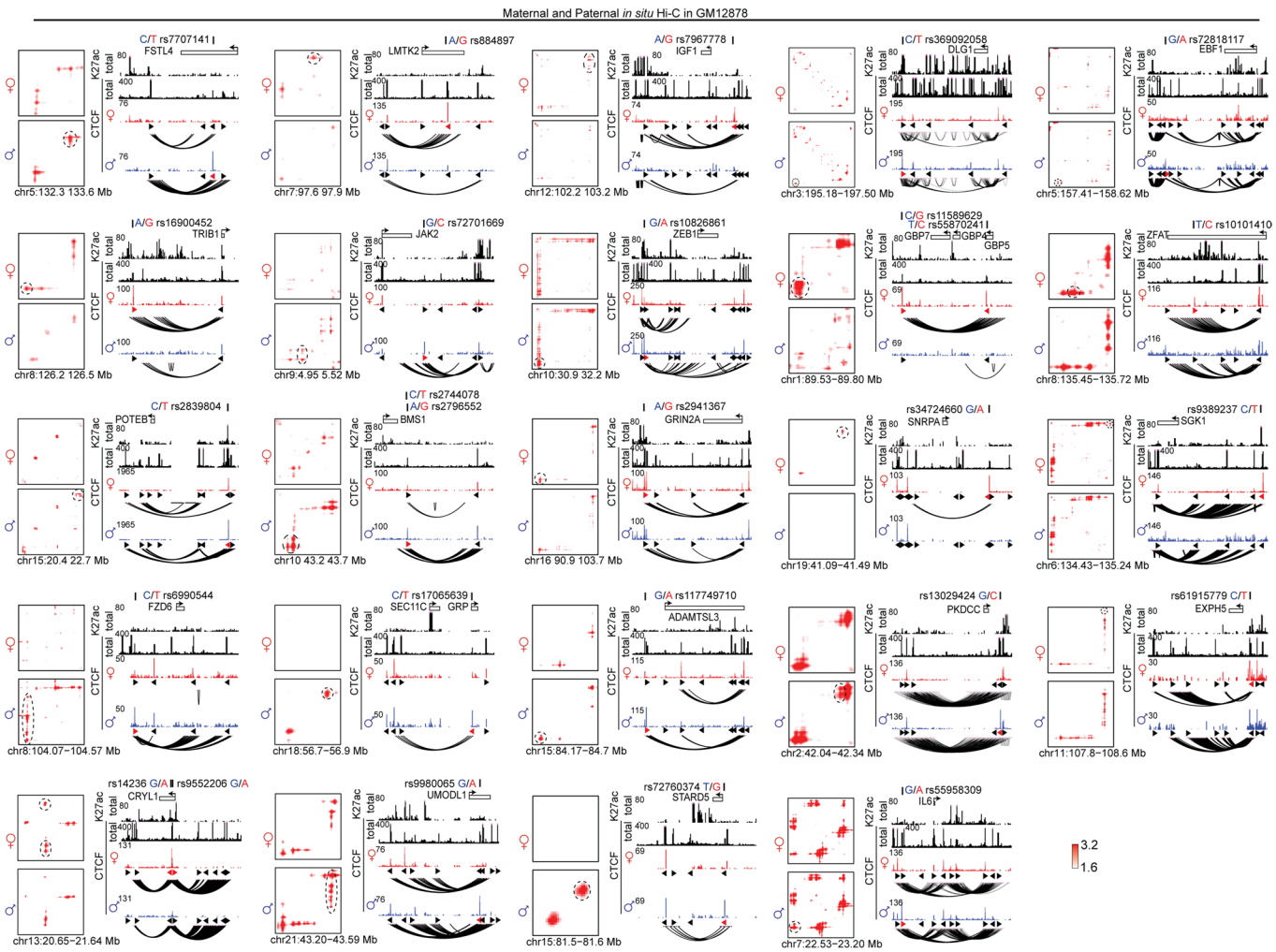
Extended Data Fig. 6 | *DeepLoop* reveals tissue-specific loop interactions for low-depth Hi-C data. Applying *LoopEnhance* to low depth Hi-C data from 14 human tissues. Contact heatmaps of three tissue-specifically expressed genes in all the tissues were shown. **a**, *ALB*, highly expressed in liver. **b**, *MYOZ2*, highly expressed in heart tissues. **c**, *ADD2*, highly expressed in brain tissues.



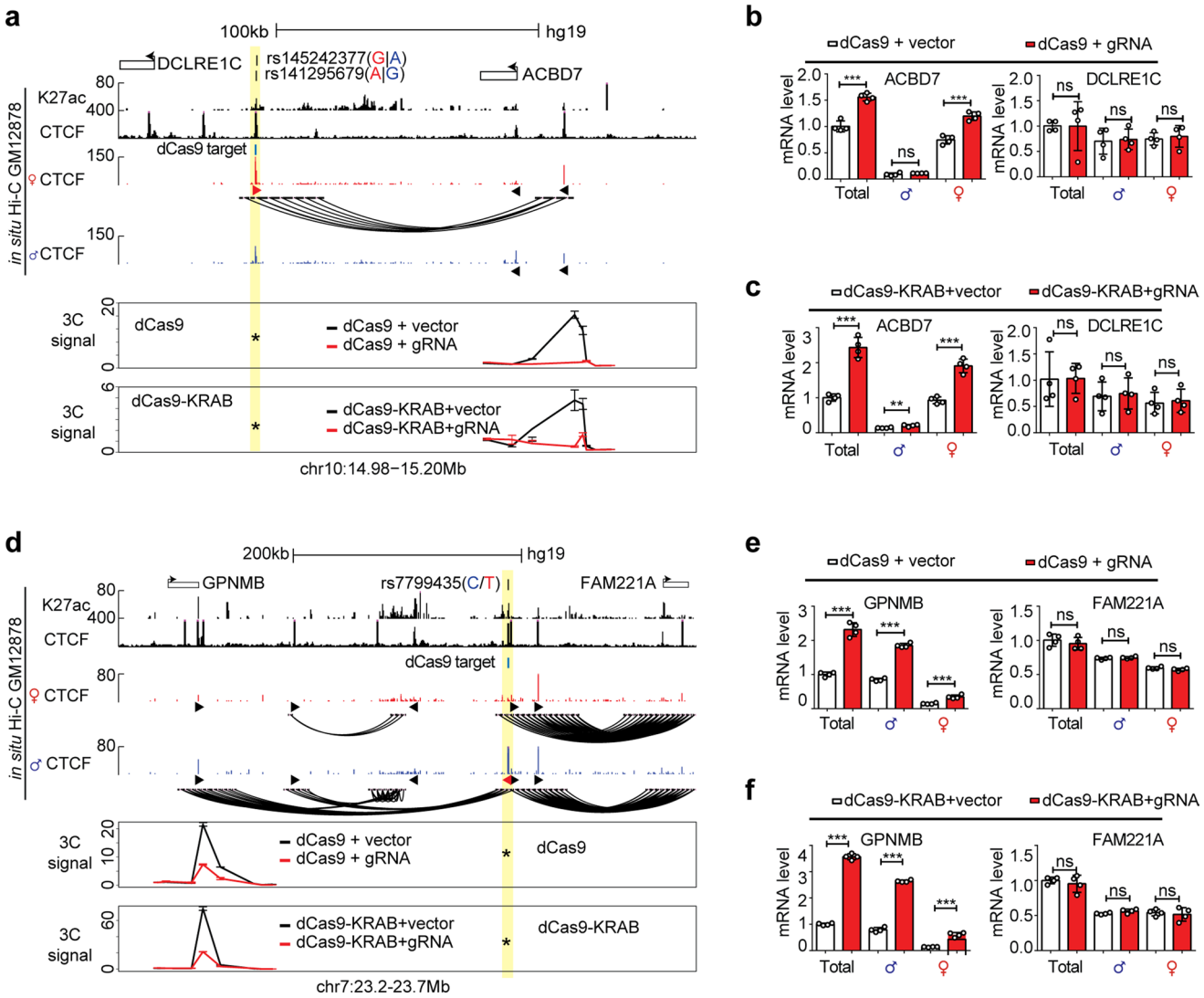
Extended Data Fig. 7 | *DeepLoop* reveals cell type specific loop interactions from sn-m3C-seq data. Same as Fig. 4e,f, single cells from the same cell type are pooled and enhanced by *DeepLoop*. The tSNE plots show the identities of each cell population (left) and the methylation level at the locus of interest (right).



Extended Data Fig. 8 | Large heterozygous deletions and inversions detected by allelic *DeepLoop* analysis. a, The scatterplots highlight the loop pixels within the entire four SVs region (two inversions and two deletions). **b**, The contact heatmaps of paternal deletion *Del-chr14* and maternal deletion *Del-chr22*. **c**, The contact heatmaps of *Inv-chr7*. **d**, The genome track of *Inv-chr7* shows the chromatin interactions, CTCF and H3K27ac binding on the un-inverted allele and 'inversion-fix' allele. In this region, the un-inverted paternal genome has A1-A4 and A5-A6 cross-boundary CTCF loops. The maternal inversion created new A1-A5 and A4-A6 cross-boundary loops due to the inverted orientation the CTCF motifs. Note that in paternal genome, the A1-A4 loop encompass multiple enhancers, while in the inverted maternal genome the A1-A5 loop lack enhancers. **e**, The gene expression level of gene CCZ1 in two alleles. The height of browser tracks indicating the raw counts of ChIP-seq.



Extended Data Fig. 9 | The contact heatmaps and browser snapshots of 24 loci containing 27 SNPs associated with both allelic CTCF binding and allelic DNA looping. For each SNP, the paternal (blue) and maternal (red) genotypes are included. The allelic loops are circled in the heatmaps. The CTCF motif orientation are indicated with triangles. The height of browser tracks indicating the raw counts of ChIP-seq.



Extended Data Fig. 10 | Allele-specific chromatin loops regulate gene expression. a, 3C assays showing the loss of chromatin loop between the SNP (highlight in yellow) and *ACBD7* locus after displacing CTCF binding with either dCas9-KRAB or dCas9 protein. **b**,**c**, Bar plots showing the changes of allelic gene expression upon blocking CTCF loops with dCas9 or dCas9-KRAB. **d**–**f**, CTCF blocking experiments at *GPNMB* locus. $n = 2$ biologically independent experiments. All data are presented as means ± SEM from 4 replicated experiments. $^{**}P < 0.01$, $^{***}P < 0.001$. NS, no significant difference. Two-sided Wilcoxon test. The height of browser tracks indicating the raw counts of ChIP-seq.

nature portfolio

Corresponding author(s): Fulai Jin, Jing Li, Yan Li

Last updated by author(s): Mar 14, 2022

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We used Python 3.6.8, Perl v5.16.3, R 3.6.0, Bowtie 1.1.2, pipe4C 1.1.4, samtools 1.3.1, SNPsplit-0.3.4, bedtools v2.25.0, hisat2 2.1.0, juicer 1.18.08, HOCOMOCO v11, DAVID 6.8, corrplot 0.92, Primer3 0.4.0, CCTop, Tensorflow 2.3.1, Keras 2.2.4, Scipy 1.7.3, Numpy 1.20.3, Matplotlib 3.5.1, Pandas 1.1.5, macs2 2.2.7.1, featureCounts 1.6.1, bowtie2 2.2.6, fimo 4.11.2, pipe4C 1.1.3. The code is available in GitHub is available on GitHub DOI: 10.5281/zenodo.6495831 at https://github.com/JinLabBioinfo/DeepLoop.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw data generated in this study and processed data from published studies are provided in GEO accession number GSE167200, the heatmap visualization can be accessed through https://hiview.case.edu/public/DeepLoop/. The details of published data we reanalyzed can be found in Supplementary Table 1.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No sample size calculations were performed. For H9 Hi-C data generated in this study, we included 10 biological replicates for library complexity and deep sequencing.
Data exclusions	No data was excluded in the analyses.
Replication	The DeepLoop models were trained with 3 replicates and many batches of data to avoid over fitting in this study
Randomization	We randomly down-sampled deep Hi-C data when training LoopEnhance models and some performance comparison.
Blinding	No blinding was performed.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	H9 hESC: WiCell, #WA09; GM12878 (catalog ID:GM12878)
Authentication	No further authentication was performed.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commercially misidentified cell lines were used