Supplementary information, Data S1

EXTENDED EXPERIMENTAL PROCEDURE

4C-seq and Data Analysis

4C-seq experiments and analysis were performed as described previously (van de Werken et al., 2012a). Briefly, 5 million cells were cross-linked with 2% formaldehyde for 10 min at room temperature (RT) and quenched by adding 125 mM Glycine with 5 min additional incubation at RT. Cells were lysed and nuclei were isolated and digested with Csp6I (Thermo Scientific) or Dpn II (NEB) overnight (o/n). Enzyme was inactivated by heat at 65 ℃ for 20 min. The digested chromatin was subjected for ligation for 16 h with T4 ligase (Life Technologies). DNA was then purified with phenol/chloroform extraction and ethanol precipitation before the second digestion with NlaIII (NEB) or BfaI (NEB) at 37 ℃ o/n. After enzyme inactivation, a second ligation was performed at 16 ℃ for 4 h and DNA was purified, of which 4.8 μg in total was used for PCR amplification using 4 different pairs of primers (**Table S5**) which were designed compatible for illumina Hiseq 2500 sequencer.

Sequencing data were analyzed using a custom pipeline '4Cseq' as previously described (van de Werken et al., 2012b) with all default parameters. For consistency, all sequencing data involving E14 and DKO genomes in this work were mapped to mouse mm9 reference genome. Contact frequency was visualized for genomic regions in a 300 kb window that includes both SOX2 gene and SOX2 SE (chr3: 34,448,927-34,765,152).

Lentiviral Packaging and shRNA Knock-Down

Lentiviral particles were prepared using the Lenti-X single shot packaging system (Clontech), according to the manufacture's guidelines. Control shGFP (Addgene #30323) and murine shRad21 (Sigma SHCLNV_NM009009 (TRCN0000176084)) plasmids (Target sequences see **Table S5**) were transfected to Lenti-X 293 cell line (Clontech) and viral supernatant were concentrated using the Amicon Ultra-15 100 kDa centrifugal filters (Millipore). For shRNA knockdown, E14 cells were plated on a 6-well dish coated with 0.1% gelatin the day before transduction (2×10^5 cells per well). Concentrated viral supernatant were added to mouse ES medium containing 8 μg/mL polybrene (Sigma). Media containing lentiviruses were replaced with fresh media 24 h post-infection. Infected cells were selected by puromycin (1 μg/mL) for 72 h.

Chromatin Immunoprecipitation Followed by Sequencing (ChIP-seq) and Data Analysis

The ChIP-seq has been carried out as previously described (Jolma et al., 2013; Tuupanen et al., 2012; Yan et al., 2013). Briefly, 2 million cells were crosslinked with 1% formadehyde for 10 min at RT. The reaction was quenched by adding 125 mM of Glycine and incubating for 5 min at RT. Cells were lysed in RIPA buffer (10 mM Tris-HCl pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) supplemented with protease inhibitor (Roche). And chromatin was sonicated into short fragments (300-700 bp). The fragmented chromatin was incubated with antibodies (**Table S6**) to pull down the specific DNA bound TFs or histones. After intensive wash, DNA was purified and prepared as sequencing library using illumina Truseq LT kit. Several samples with different indexes were pooled together for 50 or 100 cycles of single read sequencing with illumina Solexa sequencer or Hiseq 2500.

Sequencing reads were mapped to mouse mm9 reference genome using bowtie (Langmead et al., 2009). PCR duplicates were removed and peaks were called with MACS (Zhang et al., 2008) using input chromatin as control, with the parameter –m 5, 50 otherwise default. RPKM was calculated for each peak by dividing the number of reads overlapping peak with the length of the peak and multiply 1000, and the resulted value would be multiplied with a scale factor as to normalize the total number of reads to 1 million. The RPKM of each peak for samples was subtracted with the RPKM of that peak for input. If the subtracted number is less than 0, the RPKM of the peak for that sample will be assigned as 0.

In order to define Mll3/4 dependent H3K4me1 peaks, we merged H3K4me1 peaks from both E14 and DKO cells and extended all peaks to 2 kb wide. We merged two close peaks if they were partially overlapping. According to the RPKM described above, we calculated RPKM for all the 2-kb peaks for both E14 and DKO cells. We sorted the peaks according to the difference of the RPKM between E14 and DKO. If the RPKM of a given peak in E14 is over 0.6 larger than DKO, that peak will be classified to ‘Decreased’. Similarly, if the difference is over 0.6 smaller than DKO, that peak will be classified as ‘Increased’. Other peaks will be classified as ‘Non-differential’.

In order to compute the ChIP coverage centered by H3K4me1 peaks, we used HOMER suite peak annotation function (Heinz et al., 2010). In brief, both ChIP and input control were used to calculate the Tag coverage with the following parameters: "annotatePeaks.pl -size 4000 -hist 100 -ghist".

The sequences from the top 300 decreased peaks were submitted to AME (MEME-suite (Bailey et al., 2009)), using the bottom 300 increased peaks as background.

GO analysis for different categories of H3K4me1 peaks was performed with GREAT (McLean et al., 2010).

RNA-seq and Data Analysis

Total RNA from ES cells was extracted with Trizol® according to protocol (Thermo Scientific, 15596-026). PolyA+ RNA was purified with the Dynabeads mRNA purification kit (Life Tech.). The mRNA libraries were prepared for strand-specific sequencing using illumina TruSeq Stranded mRNA Library Prep Kit Set A (illumina, RS-122-2101) or Set B (illumina, RS-122-2102). Libraries were sequenced with illumina Hiseq 2500 for 100 cycles of single reads.

The single cell RNA-seq was carried out using Chromium™ Single Cell 3’ v2 Library (10XGenomics). For each time point, 2000-5000 cells were analyzed. For single cell RNA-seq, cells were harvested after trypsin treatment and washed with PBS. After washing pellets were reuspended in PBS including 0.04% BSA and concentration was counted using a hemocytometer. Cell concentration was adjusted to 1000 cells/ul and 5,000 -10,000 cells were used as input for library preparation on the GemCode™ Single Cell Platform (10x Genomics) to generate GEMs containing single cells (Gel bead in emulsion). Single-cell RNA-seq libraries were constructed using the Chromium™ Single Cell 3’ v2 Library kit (10x Genomics). Poly-A mRNA are captured within the GEMs by primers containing an illumina R1 sequence, a 16 bp cell barcode, 10 bp of a unique molecular identifier and a oligo-dT seuquence. Reverse transcription and other amplification steps were carried out on a T100 thermal cycler (Bio-Rad). After reverse transcription, GEMs were broken and single stranded cDNA was cleaned up using MyOne Silane Beads (Thermo Fisher Scientific). Next, single stranded cDNA was PCR-amplified for 12 cycles and purified using SPRIselect Reagent Kit (Beckman Coulter). After cleanup, cDNA was quantified using the Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific) to calculate the required cycle number for final library amplification. cDNA was enzymatically fragmented followed by a double size selection selection with SPRIselect Reagent Kit (200-700bp, 0.6x and 0.8x, Beckman Coulter). Subsequently, adaptors were ligated and libraries were constructed by PCR. Finally, libraries were double size selected using SPRIselect Reagent Kit (200-700bp, 0.6x and 0.8x, Beckman Coulter). Libraries were using the Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific) and size distribution was confirmed using Tapestation (High Sensitivity D1000, Agilent). Average library size was 500 bp. Molarity was calculated using the concentration assessed by Qubit® dsDNA HS Assay Kit (Thermo Fisher Scientific) and the average size of library fragments. The libraries were loaded at a concentration of 13 pM and sequenced with illumina Hiseq2500 with the following settings: Read 1 26 cycles; Index 1 8 cycles; Read 2 98 cycles. For each individual library, 16-20 million reads were sequenced, respectively. Single cell RNA-seq data sets were analyzed using Cell Ranger Single-Cell Software (version 2). Reads were aligned to the mouse reference genome mm10 using STAR (Dobin et al., 2013; Zheng et al., 2017). To identify valid cellular barcodes, all UMI counts assigned to a barcode were summed up. For downstream analysis all barcodes with less than 10% of UMIs detected in the 99 percentile of expected number of recovered cells (3000 cells) were filtered. Next, UMI counts were normalized by dividing UMI counts by the total number of UMis in the cell and multiplication of the median UMI count in all cells. Principal Component Analysis (PCA) was used for dimensionality reduction of the normalized and filtered gene-cell barcode matrix. This produces a projection of each cell onto the first N principal components (N=10). Next, t-SNE was used to visualize the cells in the 2D space (van der Maaten and Hinton, 2008). To visualize normalized UMI values for *Nanog*, *Vimentin* and *Hoxd13* the Cell Ranger R Kit (10x Genomics) was used. The basic quality control statistic for each library is listed in Supplemental **Table S7**.

Sequencing was mapped to mouse mm9 reference genome with Tophat (Trapnell et al., 2009). The differential expression was analyzed with Cuffdiff (Trapnell et al., 2012). We plotted the differentially expressed genes if the fold change of adjusted fpkm value between E14 and DKO is larger than 2.

Gene Ontology analysis was carried out using DAVID release 6.7 with default parameters (Huang da et al., 2009).

RNA Extraction and qPCR

Total RNA was isolated from harvested cells using the RNeasy columns (Qiagen) according to the manufacturer's instructions. cDNA were synthesized from 400 ng of total RNA using High Capacity cDNA Reverse Transcription kit (Applied Biosystems). qPCR was performed in triplicates using SYBR FAST qPCR master mix (KAPA biosystems) on the LightCycler 480 (Roche). Two independent sets of qPCR primers for mouse Actb and Sox2 were used: customer synthesized primers (**Table S5)** and commercially available primer sets for mouse Actb (Qiagen, catalogue no. PPM02945B) and Sox2 (Qiagen, catalogue no. PPM04762E), and a set of primers for Rad21 (Qiagen, catalogue no. QT00141204).

Nucleosome Assembly and Pull down assay

Histones are expressed using E.coli strain BL21 (DE3) transformed with cDNA of wild type H2A, H2B, H3, H4 and a mutant H3 C110A K4C contruct (a generous gift from Dr. M. Carey). In order to make methyl-lysine analogs, we used a previously described protocol (Simon et al., 2007). Briefly, 5 mg of H3 was incubated and mixed with (2-hal-oethyl) amines under reducing conditions, followed by being quenched with β-mercaptoethanol. The methylated histone was dialyzed against water overnight, and spun to remove precipitant. Equimolar amounts of histones were mixed under denaturing conditions and dialyzed overnight to assemble octamers followed by size selection (Luger et al., 1999).

Biotin tagged double stranded 601λ positioning DNA sequence was prepared as previously described (Dyer et al., 2004). The mono-nucleosomes were produced via serial salt dialysis (Carruthers et al., 1999). The H3 lysine 4 methylation was tested by western blotting with antibodies specifically recognizing various H3K4me states.

The different modified mono-nucleosomes were immobilized to streptavidin-coated beads (Invitrogen MyOneT1) as per manufacturers instructions and used as baits in following binding studies. Briefly, three micrograms of mono-nucleosomes were pre-bound to MyOneT1 beads. Immobilized nucleosomes were incubated with rotation with HeLa Nuclei Extract (200 µl of ~5 mg/ml) for 1 hour at room temperature. Beads were washed 3 times with wash buffer containing 250mM NaCl, 25mM Tris pH 8.0, 1mM EDTA, 0.2% NP40, and 1mM DTT and resuspended in equal volume of 2X Laemmli Sample Buffer (Bio-Rad). Binding was tested via western blotting using antibodies listed in **Table S6**.

Hi-C FIRE Score Clustering

We used R function 'hclust' with the complete linkage to carry out hierarchical clustering analysis. In specific, we first perform log2 transformation to the FIRE score, and then calculated the Euclidean distance between any two samples.

Topological Associating Domain and Boundary calling

Topological domains were called based on the directionality index (DI) score using a Hidden Markov Model (HMM) as previously described (Dixon et al., 2012). The software used can be downloaded at Hi-C Domain Caller. According to the domain patterns, the genome is partitioned as follows: domains are marked as domains; gaps between domains that are larger than 100 kb were marked as unstructured regions; gaps between domains that are smaller than 100 kb were marked as boundaries; if two domains are consecutive, the 10 kb window centered at the boundary is marked as a boundary.

Support Vector Machine for FIRE Classification

We first partitioned the genome into non-overlapping bins of the same length (10kb) and filtered those with poor mappability. Bins with z-score greater than 1.65 ($P <$ 0.05) were selected as positive hits, same amount of bins with smallest z-score were chosen as negative set. With 8 features, Support Vector Machine (SVM) implemented by R package ‘1071’ was applied to classify positive FIRE bins from the negative ones with default model setting (gamma=1, epsilon=0.1 and radical kernel), prediction performance was evaluated by AUC (Area Under ROC curve) using 5-fold cross validation. To further evaluate the importance of each variable, we made prediction based on the same setting but using one feature each time. AUC estimated by 5-fold cross validation for each feature reflects its decimation power.

Neural Progenitor Cell Differentiation

NPC differentiation protocol is adapted from previously published methods with modification (Bibel et al., 2004; Hon et al., 2014; Wang et al., 2012). Briefly, mouse embryonic stem cell line WT and DKO cells were grown on γ-irradiated Mouse Embryonic Fibroblast feeder cells before seeding. Cells were split and seeded to 10-cm petri-dish coated with 0.2% Gelatin type-A one day before differentiation in ES culture medium supplemented with LIF. On Day 0, LIF was deprived from the culture medium and cells were continued to be cultured for 24 hours. From Day 1 to Day 3, cells were cultured in LIF-deprived ES medium supplemented with 5 μM retinoic acid (RA). Cells were harvested every 12 hours and aliquoted for further assays. One million cells were collected for RNA-seq, two millions cells were collected for in situ Hi-C experiments, fixed 1% formaldehyde (sigma). Five million cells were collected for 4C-seq and ChIP-seq, fixed with 2% and 1% formaldehyde, respectively.

Super-enhancer Analysis

Super-enhancer is defined using H3K27ac ChIP-seq data, similar to (Hnisz et al., 2013). Briefly, MACS was used to call narrow peaks of H3K27ac with input as controls. The peak file was then used as a guide file to define Super-enhancer, using published algorithm ROSE (Hnisz et al., 2013; Loven et al., 2013). For each ChIP-seq library including H3K27ac and input at each time points, the number of uniquely mappable reads of 15 million was set as a minimum requirement for super-enhancer call.

HiCNormCis and FIRE calling

We developed a novel computational approach, named as HiCNormCis (Schmitt et al, manuscript under minor revision at Cell Reports), to remove systematic biases in total cis intra-chromosomal interactions. We first filtered out all intra-chromosomal interactions with 15kb since they are very likely to be self-ligation artifacts. Next, we divided the mouse reference genome (mm9) into 10kb bins, and for each 10kb bin, calculated the total cis intra-chromosomal interactions within 200kb. Consistent with the previous study (Yaffe and Tanay, 2011), we observed that the raw total cis intra-chromosomal interactions contain biases from three local genomic features, including restriction enzyme fragment length, GC content and mappability score. We applied a Poisson regression approach to remove three systematic biases. In specific, let y_i represent the raw total cis intra-chromosomal interactions at the i th 10kb bin. In addition, let F_i, GC_i and M_i represent the effective fragment size, GC content and mappability score at the i th 10kb bin, respectively. The definition of these three local genomic features is described in our previous work (Hu et al, 2012). Assume y_i follows a Poisson distribution with mean θ_i, we fitted a Poisson regression model $\log \theta_i = \beta_0 + \beta_F F_i + \beta_{GC} GC_i + \beta_M M_i$, where β_F, β_{GC} and β_M are regression coefficients of the effective fragment size, GC content and mappability score, respectively. Here β_0 is a Poisson offset to account for total sequencing depth. After fitting this Poisson regression model, we obtained the estimate of the unknown parameters $\hat{\beta}_0, \hat{\beta}_F, \hat{\beta}_{GC}, \hat{\beta}_M$. Next, for each i th 10kb bin, we defined residual $res_i = y_i / \exp\{\hat{\beta}_0 + \hat{\beta}_F F_i + \hat{\beta}_{GC} GC_i + \hat{\beta}_M M_i\}$ as the FIRE score. We further converted FIRE score into z-score and corresponding one-sided p-value. 10kb bins with one-sided p-value less than 0.05 are determined as FIRE bins. We have shown that the raw total cis intra-chromosomal interactions y_i show strong correlation with three local genomic features F_i, GC_i and M_i (**Supplemental Figure S3E**). After applying HiCNormCis, the FIRE score res_i show negligible correlation with these genomic features (**Supplemental Figure S3F**), indicating that HiCNormCis has successfully removed such biases. As a comparison, we also implemented two popular matrix balancing based Hi-C data normalization approaches, Vanilla Coverage (Rao et al, 2014) and ICE (Imakaev, et al, 2012). However, both Vanilla Coverage and ICE normalized total cis intra-chromosomal interactions still show high correlation with F_i, GC_i and M_i (**Supplemental Figure S3G**, **S3H**). Therefore, we conclude that VC and ICE are not optimal for correcting biases buried in total cis intra-chromosomal interactions, and decide to use HiCNormCis for Hi-C data normalization.

REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic acids research *37*, W202-208.

Bibel, M., Richter, J., Schrenk, K., Tucker, K.L., Staiger, V., Korte, M., Goetz, M., and Barde, Y.A. (2004). Differentiation of mouse embryonic stem cells into a defined neuronal lineage. Nature neuroscience *7*, 1003-1009.

Carruthers, L.M., Tse, C., Walker, K.P., 3rd, and Hansen, J.C. (1999). Assembly of defined nucleosomal and chromatin arrays from pure components. Methods in enzymology *304*, 19-35.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dyer, P.N., Edayathumangalam, R.S., White, C.L., Bao, Y., Chakravarthy, S., Muthurajan, U.M., and Luger, K. (2004). Reconstitution of nucleosome core particles from recombinant histones and DNA. Methods in enzymology *375*, 23-44.

Heinz, S., Benner, C., Spann, N., Bertolino, E., *et al*. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell *38*, 576-589.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.

Hon, G.C., Song, C.X., Du, T., Jin, F., Selvaraj, S., Lee, A.Y., Yen, C.A., Ye, Z., Mao, S.Q., Wang, B.A., *et al.* (2014). 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. Molecular cell *56*, 286-297.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., *et al.* (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327-339.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Loven, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320-334.

Luger, K., Rechsteiner, T.J., and Richmond, T.J. (1999). Preparation of nucleosome core particle from recombinant histones. Methods in enzymology *304*, 3-19.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Simon, M.D., Chu, F., Racki, L.R., de la Cruz, C.C., Burlingame, A.L., Panning, B., Narlikar, G.J., and Shokat, K.M. (2007). The site-specific installation of methyl-lysine analogs into recombinant histones. Cell *128*, 1003-1012.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Tuupanen, S., Yan, J., Turunen, M., Gylfe, A.E., Kaasinen, E., Li, L., Eng, C., Culver, D.A., Kalady, M.F., Pennison, M.J., *et al.* (2012). Characterization of the colorectal cancer-associated enhancer MYC-335 at 8q24: the role of rs67491583. Cancer genetics *205*, 25-33.

van de Werken, H.J., de Vree, P.J., Splinter, E., Holwerda, S.J., Klous, P., de Wit, E., and de Laat, W. (2012a). 4C technology: protocols and data analysis. Methods Enzymol *513*, 89-112.

van de Werken, H.J., Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B.A., *et al.* (2012b). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods *9*, 969-972.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J Mach Learn Res *9*, 2579-2605.

Wang, C., Lee, J.E., Cho, Y.W., Xiao, Y., Jin, Q., Liu, C., and Ge, K. (2012). UTX regulates mesoderm differentiation of embryonic stem cells independent of H3K27 demethylase activity. Proceedings of the National Academy of Sciences of the United States of America *109*, 15324-15329.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307-319.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., *et al.* (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell *154*, 801-813.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. Nat Commun *8*, 14049.