Supplemental Information

Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and Polycomb proteins in genome organization

Matthew Denholtz *, Giancarlo Bonora*, Constantinos Chronis, Erik Splinter, Wouter de Laat, Jason Ernst, Matteo Pellegrini, and Kathrin Plath

Inventory

Supplemental Figures

Supplemental Figure 1
This figure is related to Figure 1 and provides further evidence of the reproducibility of 4C-seq data sets, quality control of 4C-seq data, and validation of 4C-seq defined chromatin contacts.

Supplemental Figure 2
This figure is related to Figure 2 and shows that individual 4C-seq replicate data sets per cell type correlate strongly, and revealing pluripotency-specific chromatin contacts in both *cis* and *trans*.

Supplemental Figure 3
This figure is related to Figures 3 and 4 and depicts interacting domains for all additional bait regions analyzed in ESCs by 4C-seq, principal component (PC) score distributions, and additional scatterplots of pseudo-bait versus Hi-C-defined mean PC scores.

Supplemental Figure 4
This figure is related to Figures 3 and 4 and provides a comparison of average PC1 score and average feature density within 4C-seq bait regions and their corresponding interactomes in ESCs, as well as the enrichment of genomic features and PC scores within ESC super enhancers.

Supplemental Figure 5
This figure is related to Figure 6 and shows individual feature contributions to the first principal component of the ESC+MEF PCA, as well as the 3C validation of ESC-specific chromatin interactions of the *Dppa2* bait region detected by 4C-seq.

Supplemental Figure 6
This figure is related to Figure 7 and provides additional validation and characterization of interaction preferences of PC3-positive, Polycomb protein enriched-, and PC3-negative bait regions in ESCs with and without Eed.

Supplemental Tables

Supplemental Table 1
This table summarizes the bait regions by cell type for all 4C-seq libraries in this study (worksheet one). It also provides the genomic positions of all significantly interacting domains for all bait regions and cell types (worksheet two). This table relates to data in figures 1-4, 6-7, and S1-S6.

Supplemental Table 2
This table lists the 4C-seq and 3C primers used in the study (worksheets one and two, respectively). This table relates to data in figures 1-4, 6-7, and S1-S6 for 4C primers, and figures S1 and S5 for 3C primers.

Supplemental Table 3
This table contains read distribution and quality control statistics for all 4C-seq libraries that passed quality control measures and were used in this study. This table relates to data in figures 1-4, 6-7, and S1-S6.

Supplemental Table 4
This table contains binomial test results for all baits, including the number of replicates/ simulations, $-\log_{10}$(P-value) threshold, number of significant HindIII sites (true positives), the number of false positives, and the false discovery rate (FDR), in both *cis* and *trans.* This table relates to data in figures 1-4, 6-7, and S1-S6

Supplemental Table 5
This table contains Spearman's *rho* statistic values for the rank correlation of the 4C hit percentages determined in this study and published Hi-C read counts corresponding to each 4C bait (Dixon et al., 2012) shown in figures 3 and 4.

Supplemental Table 6
This table provides a summary of the linear genomic feature data sets used for data analysis shown in figures 3, 4, 6, 7, and S3-S6, as well as source information for each data set.

Supplemental Table 7
This table lists the empirical background model parameters for each cell type, including the intercept and slope for the segmented linear (0 – 1 Mb around the bait) and log-log (1 – 8 Mb from the bait) regression models (*Supplemental Experimental Procedures*) used to identify interacting domains and generate binomial test p-values shown in figures 1-4, 6, 7, and S1-S6.

Additional Supplemental Information

Supplemental Experimental Procedures
Detailed descriptions of all experimental methods and data analyses.

Supplemental References
Includes the references cited in the Supplemental Information.

Supplemental Figures and Legends

Denholtz, Bonora Figure S1

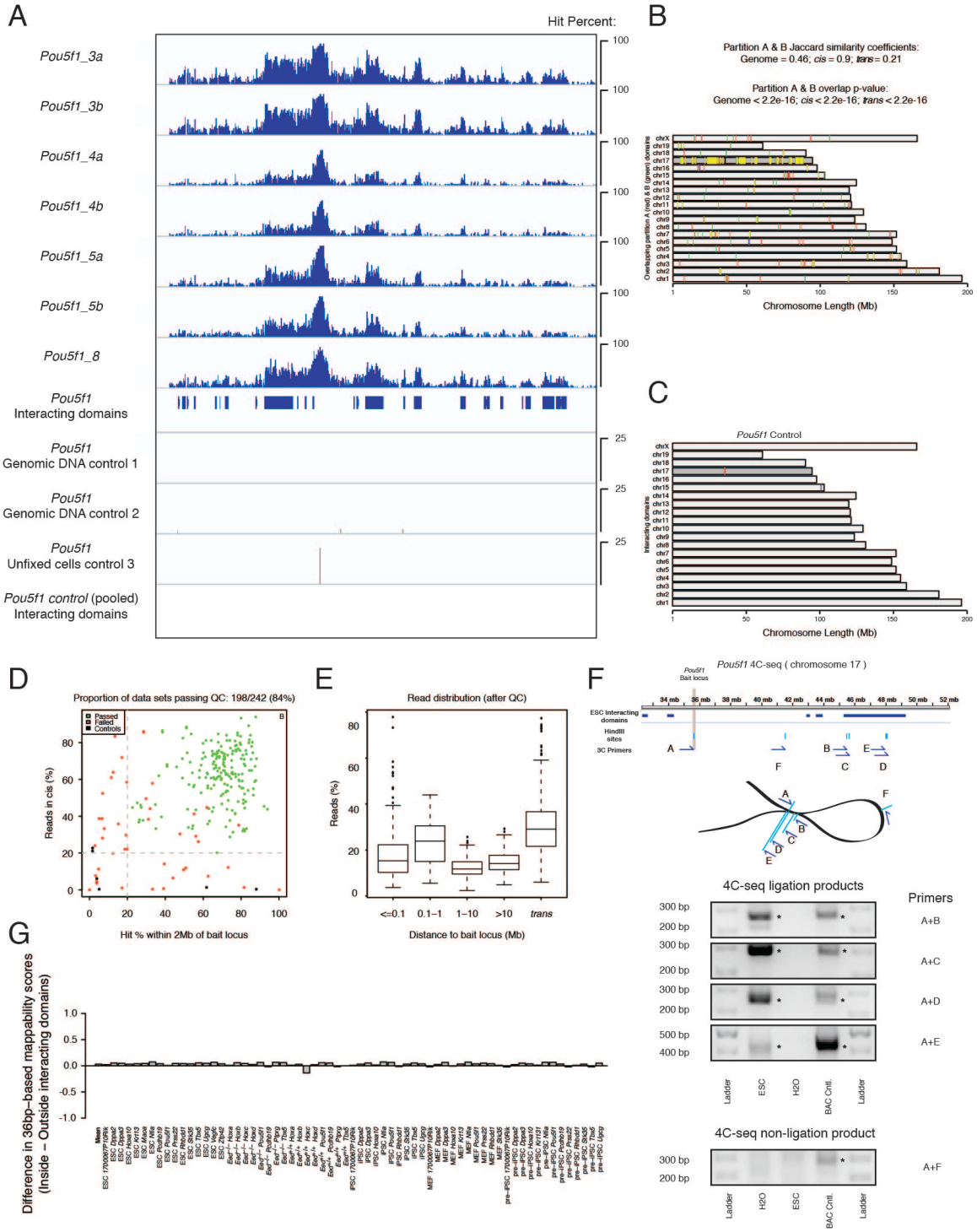


Figure S1. Reproducibility, validation and quality control of 4C-seq data (Related to Figure 1)

A, Integrative Genomics Viewer tracks demonstrating the reproducibility of *Pou5f1* 4C-seq hit percentage data for four biological and three technical replicate data sets. Biological replicates are designated with numbers, technical replicates with letters. The interacting domains, as identified by the binomial test for the pooled 4C-seq *Pou5f1* data, are given below the hit percentage tracks. 4C-seq experiments with genomic DNA and control libraries from unfixed ESCs follow, demonstrating the lack of significant interactions.

B, To determine the reproducibility of our data in both *cis* and *trans*, *Pou5f1* replicates were pooled (blue) or partitioned into two groups (A and B, red and green, respectively) and interacting regions were determined for each set and are shown on the genome-wide plot. Yellow regions are those that were called as interacting domains in both A and B partitions, and blue regions were called as interacting domains in only the pooled data set. Jaccard similarity coefficients are noted for the overlap of significant domains between A and B, genome wide, *cis* only, and *trans* only, and the significance of the overlap was determined by the hypergeometric test. The darker chromosome represents the *cis* chromosome.

C, 4C-seq experiments on pooled control libraries, here shown for the *Pou5f1* locus (*Pou5f1* Control) in ESCs, display no significant interactions genome-wide. The red mark denotes the bait locus, and the darker chromosome represents the *cis* chromosome.

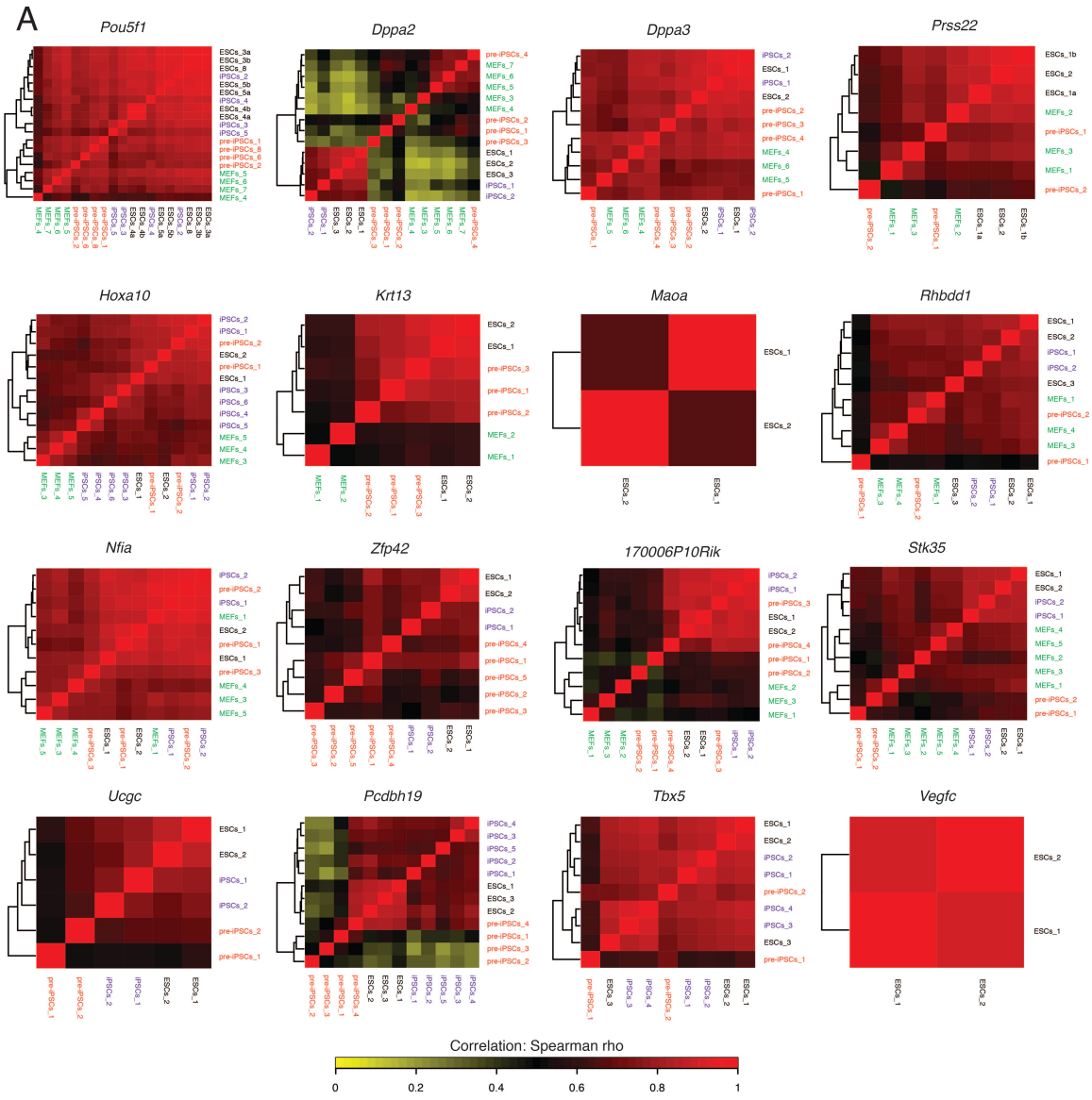
D, Quality control (QC) metrics. All 4C-seq data sets used in this study were required to have at least 20% of HindIII sites covered by reads within 2Mb of the bait locus (x-axis), as well as at least 20% of their reads occurring in *cis* to the bait locus (y-axis), among other quality control metrics (see *Experimental Procedures*). Only libraries that passed these QC metrics were analyzed further and are included in this study.

E, Read distribution as a function of distance from the bait locus for all data sets that passed QC. This analysis excludes the self-ligated and undigested products. Boxplots show the median value (line within box), and the first and third quartiles (lower and upper edges of boxes, respectively). Whiskers demarcate +/- 1.5 times the interquartile range.

F, 3C validation of *Pouf51* interactions. (i) Schematic representation of the 3C experimental design (performed as described in (Miele et al., 2006)) to confirm interactions between the *Pou5f1* locus (primer A) and distal regions of the same chromosome (in *cis*), including the four interacting fragments (primers B-E) and an intervening non-interacting fragment (primer F). (ii) 3C PCR results for the setup described in (i) confirming the presence of ligation products resulting from the juxtaposition of the genomic regions (B-E) with A, and an absence of ligation products between A and the intervening genomic region (F). An H_2O control and a BAC-generated positive control are shown for each PCR product. The asterisks mark the respective PCR products.

G, UCSC ENCODE mm9 36mer-based mappability scores (ftp://encodeftp.cse.ucsc.edu /pipeline/mm9/wgEncodeMapabilitywgEncodeCrgMapabilityAlign36mer.bigWig) were compared between the interacting domains (determined based on our 4C-seq analysis

pipeline) and the scores outside these domains. The graph displays the difference in the mean 36-mer-based mappability scores inside and outside of the intrachromosomal interaction domains for the indicated baits and cell types, as well as the average across all baits and cell types (labeled ‘MEAN’), demonstrating that our 4C-seq results are not biased by mappability.

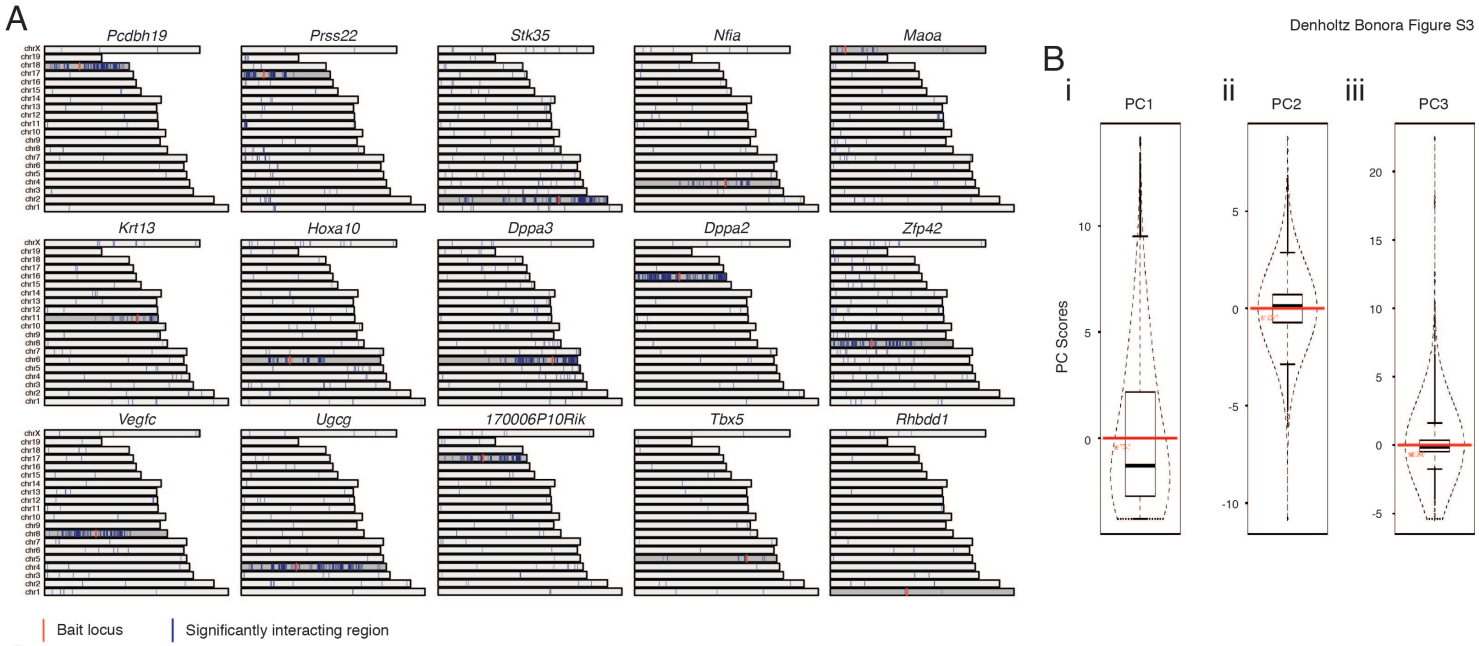


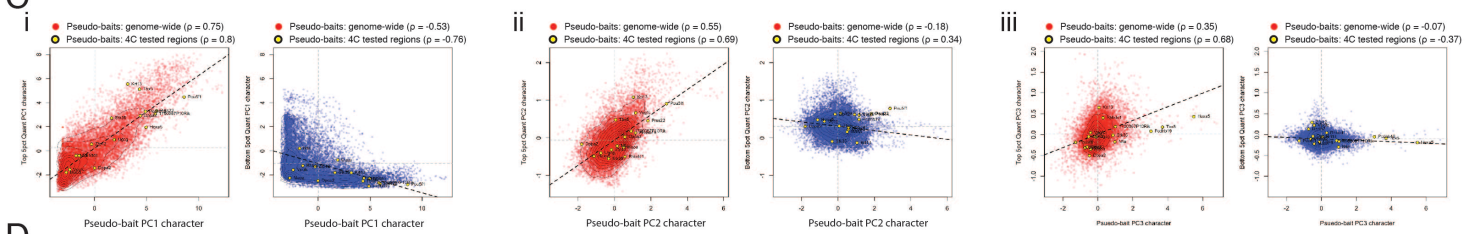
Denholtz, Bonora et al. Figure S2

Figure S2. 4C-seq replicate data sets cluster by cell type, revealing pluripotency-specific chromatin contacts (Related to Figure 2)

A, Spearman rho correlation matrixes based on the percentage of HindIII sites hit within 1000kb windows in *cis* for individual 4C replicates in wildtype ESCs (V6.5 line), iPSCs, pre-iPSCs, and MEFs for the noted baits. Note that the preferential clustering of individual replicates by cell type, and the high correlations of the data between ESCs and iPSCs. Correlation values are indicated by the key, and numbers listed next to cell types correspond to biological replicates, while technical replicates are designated with letters.

B, (i) Unsupervised hierarchical clustering of Spearman rank correlation values of the hit percentages within 200kb windows along the *trans* chromosomes of eight different bait loci (*Pou5f1*, *Stk35*, *1700067P10Rik*, *Nfia*, *Dppa3*, *Rhbdd1*, *Hoxa10,* and *Dppa2*) in wildtype ESCs (V6.5 line), MEFs, iPSCs, and pre-iPSCs, demonstrating the pluripotency-specific organization of interchromosomal chromatin contacts within the mouse genome. Color key defines Spearman *rho* values. (ii) Unsupervised hierarchical clustering of Jaccard similarity coefficients for the overlapping interacting domains *trans* in ESCs, iPSCs, pre-iPSCs, and MEFs, for the same eight bait loci as (A). Color key defines Jaccard similarity values.


A
Pcdbh19
Prss22
Stk35
Nfia
Maoa
Krt13
Hoxa10
Dppa3
Dppa2
Zfp42
Vegfc
Ugcg
170006P10Rik
Tbx5
Rhbdd1
Bait locus
Significantly interacting region
Denholtz Bonora Figure S3
B
i
PC1
ii
PC2
iii
PC3
PC Scores


C
i
Pseudo-baits: genome-wide (ρ = 0.75)
Pseudo-baits: 4C tested regions (ρ = 0.8)
Pseudo-baits: genome-wide (ρ = -0.53)
Pseudo-baits: 4C tested regions (ρ = -0.76)
ii
Pseudo-baits: genome-wide (ρ = 0.55)
Pseudo-baits: 4C tested regions (ρ = 0.69)
Pseudo-baits: genome-wide (ρ = -0.18)
Pseudo-baits: 4C tested regions (ρ = 0.34)
iii
Pseudo-baits: genome-wide (ρ = 0.35)
Pseudo-baits: 4C tested regions (ρ = 0.68)
Pseudo-baits: genome-wide (ρ = -0.07)
Pseudo-baits: 4C tested regions (ρ = -0.37)
Pseudo-bait PC1 character
Pseudo-bait PC2 character
Pseudo-bait PC3 character

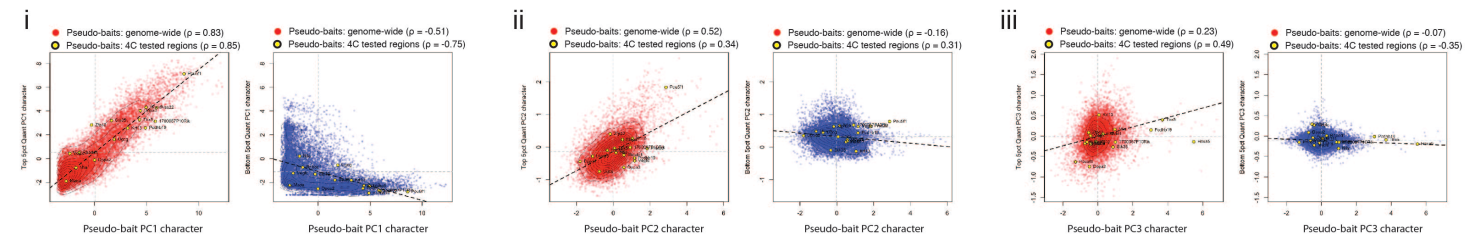

D
i
Pseudo-baits: genome-wide (ρ = 0.83)
Pseudo-baits: 4C tested regions (ρ = 0.85)
Pseudo-baits: genome-wide (ρ = -0.51)
Pseudo-baits: 4C tested regions (ρ = -0.75)
ii
Pseudo-baits: genome-wide (ρ = 0.52)
Pseudo-baits: 4C tested regions (ρ = 0.34)
Pseudo-baits: genome-wide (ρ = -0.16)
Pseudo-baits: 4C tested regions (ρ = 0.31)
iii
Pseudo-baits: genome-wide (ρ = 0.23)
Pseudo-baits: 4C tested regions (ρ = 0.49)
Pseudo-baits: genome-wide (ρ = -0.07)
Pseudo-baits: 4C tested regions (ρ = -0.35)
Pseudo-bait PC1 character
Pseudo-bait PC2 character
Pseudo-bait PC3 character

Figure S3. Genome-wide analysis of ESC interactomes (Related to Figures 3 and 4)

A, We obtained 4C-seq data for a total of 16 baits in the V6.5 ESC line. The interacting domains for the Pou5f1 4C-seq data set are shown in Figure 1, and the interacting domains identified for the remaining 15 baits are depicted here genome-wide, with their relative chromosomal locations. Blue marks represent significantly interacting windows, red marks the bait locus, and the darker chromosome denotes the *cis* chromosome.

B, Demonstration of the continuous nature of PC scores. (i) Box plot of genome-wide PC1 scores overlaid with a violin plot. (ii) As in (i), but for PC2. (iii) As in (i), but for PC3. Boxplots show the median value (line within box), and the first and third quartiles (lower and upper edges of boxes, respectively). Whiskers demarcate +/- 1.5 times the interquartile range.

C, Hi-C data analysis demonstrates that the preferential co-localization of genomic regions with similar genomic features identified by 4C-seq is a genome wide phenomenon for interactions in *cis*: Pseudo-4C analysis was performed on Hi-C data as described in Figures 3E/F. (i) (left) The mean PC1 score within the top 5% of 200kb windows (ranked by read count based on Hi-C data) along the cis-chromosome, excluding the 1Mb bait region, and the mean PC1 score within the respective 1Mb pseudo-bait region were determined, and plotted as red point on the scatterplot. If the bait region is one of the our baits analyzed by 4C-seq in Figure 3E-i, the data point was plotted in yellow (4C-bait loci). Correlations between bait and interactome PC1 scores are noted. The data are also summarized by the regression line in black, and the mean bait and interactome PC1 scores are demarcated by vertical and horizontal grey lines, respectively, and contour lines represent data density. The same data are shown in Figure 3G. (i) (right) Similar to (i), an analysis was performed for the 5% of windows that are the least likely to interact based on Hi-C read count (bottom 5%) and their mean PC1 score was plotted in blue against the mean score of the 1Mb bait regions. Again, if the bait region is one of the our baits analyzed by 4C-seq in Figure 3E-i, the data point was plotted in yellow (4C-bait loci). (ii) As in (i), except that the analysis was performed for PC2 scores, and the red scatterplot is repeated from Figure 4D. (iii) As in i,), except that the analysis was performed for PC3 scores, and the red scatterplot is repeated from Figure 4H.

D, As in (C), except that all 200kb windows within the 10Mb region around the bait loci were excluded from the determination of the mean PC scores within the top and bottom 5% of 200kb windows along the cis-chromosome. Importantly, the correlation values between interactome and bait character, with regards to PC scores, were not significantly changed upon exclusion of the 10Mb region around the bait loci, or of increasingly larger proximal interactions around the bait (up to 25Mb on either side of the bait region, data not shown), indicating that even distal interactions follow the logic that regions with similar PC character interact.

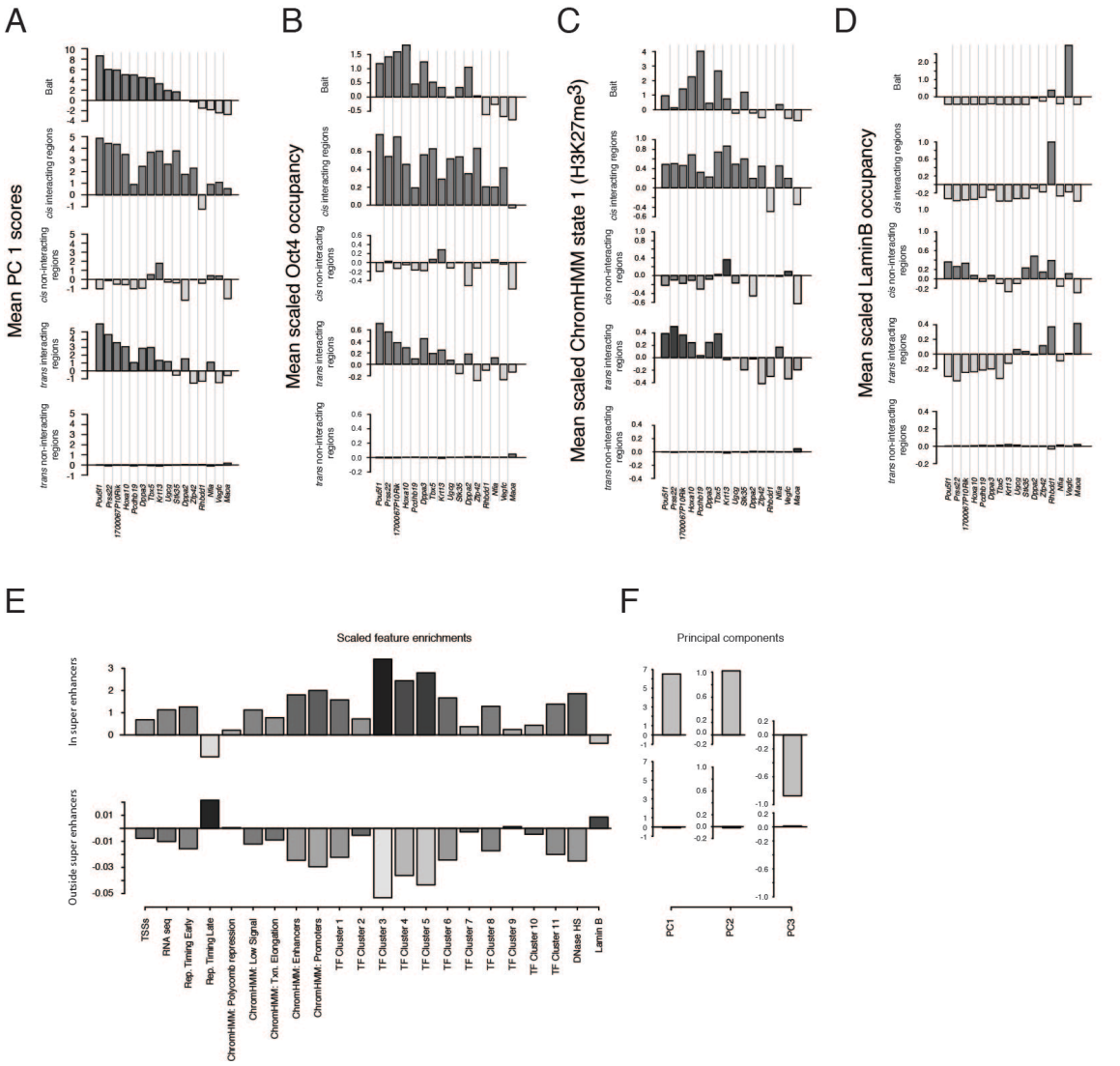


Figure S4. Relationship between PC1 and individual feature enrichment (Related to Figures 3 and 4)

A-D, Comparison of PC1 and feature scores between bait regions and their interactomes: **A,** Top to bottom: Mean PC1 score for the 1Mb region centered on each bait locus in ESCs; mean PC1 score within interacting regions in *cis* for each bait locus; mean PC1 score within non-interacting regions in *cis* for each bait locus; mean PC1 score within interacting regions in *trans* for each bait locus; mean PC1 score within non-interacting regions in *trans* for each bait locus. Spearman rho's give the correlation between PC1 bait character and interactome character across all analyzed baits in *cis* and *trans*. These results are duplicated from Figure 3E-i for illustration. **B,** As in (A), but showing the average, standardized Oct4 enrichment within the bait regions, and their interacting and non-interacting regions. **C,** As in (B), but for H3K27me3 (which is

captured by ChromHMM state 1). **D,** As in (B), but for LaminB enrichment. The data shown in (A-D) suggest that due to the strong enrichment of transcription factors, chromatin regulators, transcriptional machinery, Cohesin, and active chromatin states in genomic regions with high PC1 scores and their preferential co-localization in 3D space, their interactomes are also highly enriched for any of the features positively correlating with PC1. Indeed, the extent of Oct4 and H3K27me3 enrichment in a bait's interacting regions mirrors its PC1 character. Conversely, LaminB occupancy, which correlates with closed, PC1 negative chromatin character, shows the opposite relationship. As such, the enrichment of features positively correlating with PC1 in a given bait's interactome may not be indicative of specific mechanistic roles for each enriched feature but may be instead a consequence of the overarching nature of chromosomal conformation.
E, Relationship of ESC super enhancers to the genomic features tested in our study. Enrichment of chromatin states and transcription factor clusters within and outside of ESC super-enhancers.
F, Mean PC1, PC2, and PC3 scores within and outside of ESC super-enhancers.

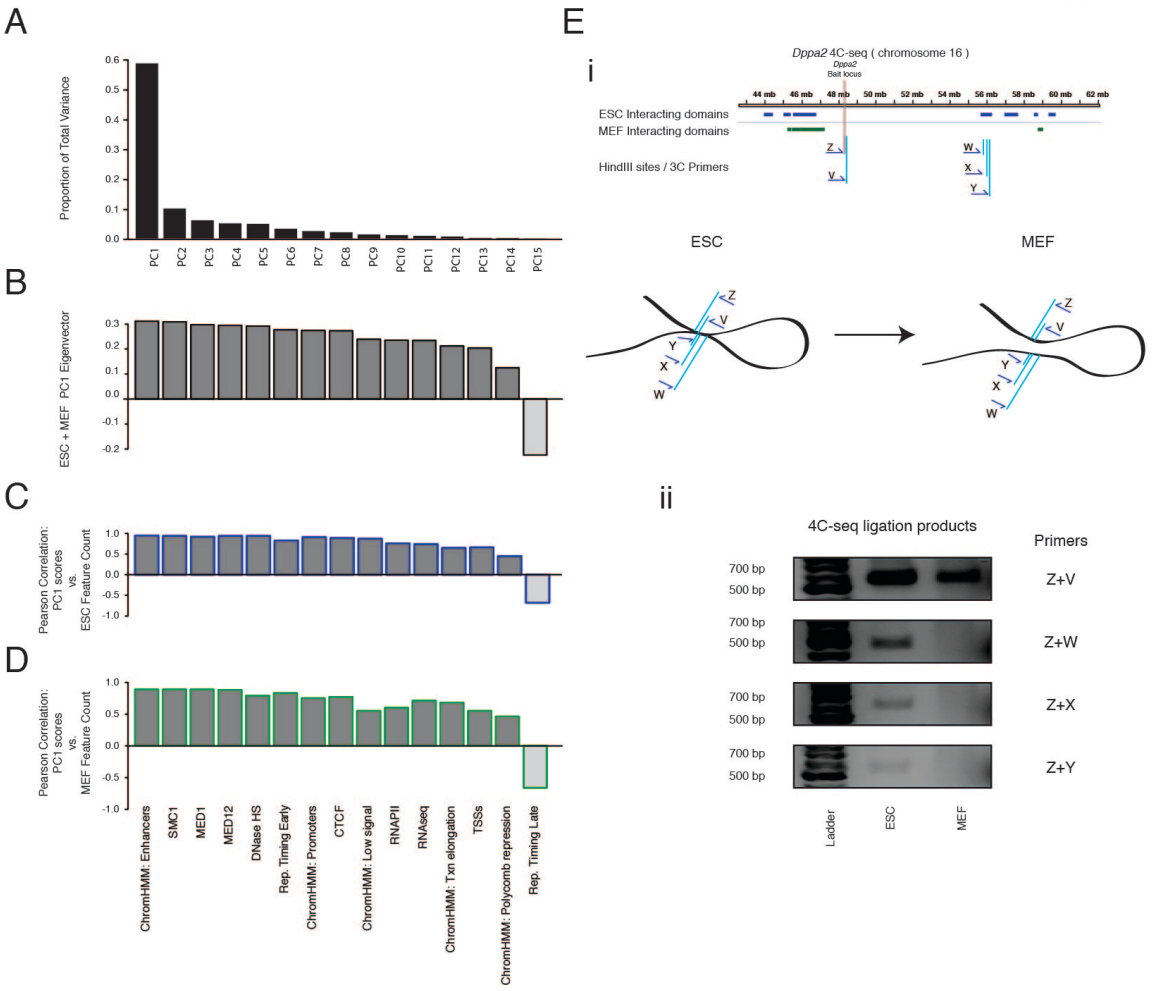


Figure S5. Comparison of the spatial interactomes between ESCs and MEFs (Related to Figure 6)

A, PCA was performed on genomic features across ESCs and MEFs. Proportion of total variance in genomic features described by each principal component for the ESC+MEF PCA.

B, Genomic feature contribution to the ESC + MEF feature PC1 eigenvector.

C, Correlation of ESC genomic feature density with PC1 scores.

D, Correlation of MEF genomic feature density with PC1 scores.

E, (i) Schematic representation of the 3C experimental design to confirm the cell type-specificity of interactions with the *Dppa2* locus. At the top, a portion of chromosome 16 harboring the *Dppa2* locus is depicted, and a subset of interacting domains in ESCs and MEFs as defined by our 4C-seq analysis is indicated. Three ESC-specific interaction sites were examined (primers W-Y) by 3C in both MEFs and ESCs. A positive 3C control

was designed to amplify a ligation product between the *Dppa2* locus (primer Z) and a proximal restriction site (primer V), which should be detectable in both ESCs and MEFs. (ii) 3C PCR results confirming the presence and cell type-specificity of the ESC-specific ligation products Z/W, Z/X, and Z/Y. Note the presence of the positive control PCR product (Z+V) in both the ESC and MEF 3C experiments.

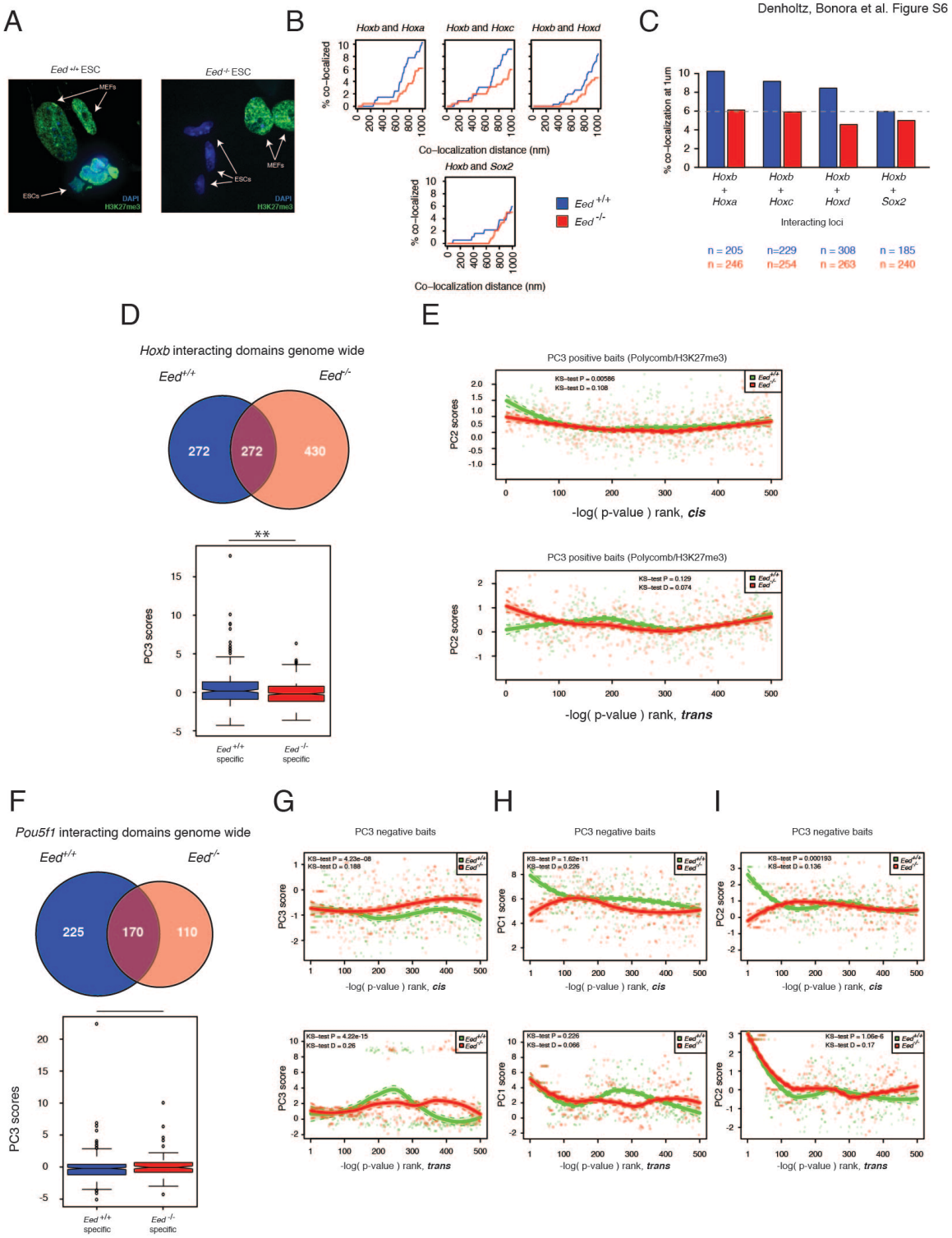

A
Eed +/+ ESC
Eed -/- ESC
MEFs
ESCs
DAPI
H3K27me3
B
Hoxb and Hoxa
Hoxb and Hoxc
Hoxb and Hoxd
Hoxb and Sox2
% co-localized
Co-localization distance (nm)
Eed +/+
Eed -/-
C
% co-localization at 1um
Hoxb + Hoxa
Hoxb + Hoxc
Hoxb + Hoxd
Hoxb + Sox2
Interacting loci
n = 205
n=229
n = 308
n = 185
n = 246
n=254
n = 263
n = 240
D
Hoxb interacting domains genome wide
Eed+/+
Eed-/-
272
272
430
**
PC3 scores
Eed +/+ specific
Eed -/- specific
E
PC3 positive baits (Polycomb/H3K27me3)
KS-test P = 0.00586
KS-test D = 0.108
PC2 scores
-log(p-value) rank, cis
KS-test P = 0.129
KS-test D = 0.074
-log(p-value) rank, trans
F
Pou5f1 interacting domains genome wide
225
170
110
G
PC3 negative baits
KS-test P = 4.23e-08
KS-test D = 0.188
PC3 score
KS-test P = 4.22e-15
KS-test D = 0.26
H
PC1 score
KS-test P = 1.62e-11
KS-test D = 0.226
KS-test P = 0.226
KS-test D = 0.066
I
KS-test P = 0.000193
KS-test D = 0.136
PC2 score
KS-test P = 1.06e-6
KS-test D = 0.17

Figure S6. Additional validation and characterization of interaction preferences in *Eed+/+* and *Eed-/-* ESCs (Related to Figure 7)

A, Immunostaining for H3K27me3 (green) in wildtype (*Eed*$^{+/+}$) and *Eed*$^{-/-}$ ESCs, grown on irradiated MEFs. DAPI staining in blue marks the nuclei. Note that Eed ablation leads to the loss of H3K27me3. Wildtype MEFs still stain positively for H3K27me3.

B. DNA FISH provides an independent confirmation of 4C-seq-defined differences in chromatin interactions between *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs**.** Cumulative distribution plots for interaction frequencies (y-axis) at different distances (x-axis) for the *trans* interactions measured between the *Hoxb* region and the individual *Hox* loci indicated in the figure (the sum of these individual plots is shown in Figure 7I-i), and, for comparison, between the *Hoxb* region and the *Sox2* region, which does not interact with *Hoxb* based on our 4C-seq data (same FISH analysis plot as shown in Figure 7I-i. The data for wildtype ESCs are shown in blue, and for mutant ESCs in red.

C, Co-localization frequencies at 1um for each pair of interactions listed in (B). The sum of these individual plots between *Hox* genes is shown in Figure 7I-ii, and the *Hoxb-Sox2* analysis plot is the same as that shown in Figure 7I-ii for comparison. n = FISH signal pairs analyzed for each cell type.

D, Top: Overlap of interacting domains of the *Hoxb3* bait region between *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs genome-wide. Bottom: PC3 score distribution for the *Eed*$^{+/+}$- and the *Eed*$^{-/-}$ ESC-specific interacting domains. These data demonstrate that interacting domains that are specific for wildtype ESCs have significantly higher PC3 scores genome-wide than those that are specific for *Eed*$^{-/-}$ ESCs. Boxplots show the median value (line within box), and the first and third quartiles (lower and upper edges of boxes, respectively). Whiskers demarcate +/- 1.5 times the interquartile range. Notches ≈ 95% confidence interval around medians. ** p-value <= 0.01.

E, The *cis*- and *trans* interactomes of the six PC3-positive (Polycomb/H3K27me3 enriched) target bait loci *Hoxa10, Hoxb3, Hoxc4, Hox12, Pcdhb19, Tbx5* were ranked by -log(p-value) for both *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs, and, in each case, the 500 top ranked sites plotted against their average PC2 scores in wildtype ESCs, indicating minor changes in interaction preferences with regards to PC2 scores upon *Eed* ablation.

F, As in (D), except for the non-Polycomb, PC3-negative bait *Pou5f1*. *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESC specific interactions do not show significant differences in PC3 scores.

G-I, The *cis*- and *trans* interactomes of the two PC3-negative target bait loci *Pou5f1* and *Ptprg* were ranked by -log(p-value) for both *Eed*$^{+/+}$ and *Eed*$^{-/-}$ ESCs, and the 500 top ranked sites in *cis* (top) or *trans* (bottom) plotted against their average (G) PC3; (H) PC1; and (I) PC2 scores in wildtype ESCs. These data show that non-Polycomb targets do not show major changes in interaction preferences with regards to the PC's. Kolmogorov-Smirnov test (KS) p-values shown with KS-test D value.

Supplemental Experimental Procedures

4C Library Preparation

4C libraries were prepared from mouse ESCs (V6.5 line), MEFs (wildtype of 129SvJae background), an iPSC line (described in Chin, Plath et al., manuscript in preparation), the pre-iPSCs (12-1 line) (Sridharan et al., 2013), and the *Eed* mutant ESC line (17Rn5-3354SB) and a sibling wild-type ESC line (Morin-Kensicki et al., 2001), essentially as described (Splinter et al., 2012). Specifically, 10^7 cells were trypsinized and filtered to single cell suspensions with 40um cell strainers. Following a PBS wash, cells were cross-linked in 1xPBS with 10% fetal bovine serum (FBS) and 2% formaldehyde. Ice-cold 1M glycine was added to the cells on ice to a final concentration of 0.13M to quench the crosslinking reaction. Cells were spun at 500g for 5 minutes at 4°C, resuspended in ice-cold lysis wash buffer (10mM Tris-HCl pH7.5, 10mM NaCl, 5mM $MgCl_2$, 0.1mM EGTA, and 1x protease inhibitors (Roche)), and re-pelleted at 4°C. Pellets were subsequently resuspended in 1ml ice-cold lysis buffer (lysis wash buffer with 0.2% IGEPAL (Sigma)), cells were lysed on ice for 30 minutes, and dounced using a tight piston for 10 strokes to isolate nuclei. Nuclei were spun down at 200g for 7 minutes at 4°C, washed with 1.2x buffer B (Roche), resuspended in 1ml ice-cold 1.2x buffer B, and transferred to non-stick tubes. Cells were brought to 37°C and 20ul 15% SDS were added to each tube, which were then incubated for 1 hour at 37°C while rotating end over end. 150ul Triton X-100 was added and tubes were allowed to incubate for another hour at 37°C. 800U of high concentration HindIII (Roche, cat# 10798983001) were added to each tube and the restriction digest reaction was run overnight at 37°C while tubes rotated. The restriction enzyme was inactivated at 65°C for 25 minutes and digest efficiency was determined as described (Splinter et al., 2012). Digested samples were transferred to 50ml falcon tubes, and 5.3ml H_2O and 700ul 10x ligation buffer (660mM Tris-HCl pH 7.5, 50mM $MgCl_2$, 50mM DTT, 10mM ATP) were added to each sample. 100U T4 ligase (Roche cat#10799009001) were added and samples ligated overnight at 16°C. Ligation efficiency was checked as described (Splinter et al., 2012). 30ul 10mg/ml proteinase K were added to efficiently ligated samples and the samples incubated overnight at 65°C. Subsequently, 30ul RNaseA (Invitrogen, cat# 12091-021) were added to each sample, and samples were incubated for a additional 45 minutes at 37°C. DNA was phenol-chloroform extracted, and precipitated by the addition of 7ml H_2O, 1ml 3M Na-acetate pH 5.6, 7ul glycogen (Roche, cat# 10901393001), and 35ml 100% ethanol,

followed by freezing at -80°C. Precipitated DNA was spun down at 8800g for 45 minutes at 4°C, washed with ice-cold 70% ethanol, and re-spun at 3000g for 15 minutes at 4°C. Upon drying, DNA was resuspended in 150ul 10mM Tris pH 7.5, 300ul H_2O, 50ul 10x DpnII restriction buffer (NEB), and 50U high concentration DpnII (NEB, cat# R0543M) were added to each tube and the DNA was digested again overnight at 37°C. DNA was phenol-chloroform purified, precipitated via addition of 50ul 3M Na-Acetate and 1ml ethanol, and re-dissolved in 100ul H_2O. The DpnII-digested DNA was then transferred to a falcon tube to which 12.5ml H_2O, 1.4ml 10x ligation buffer, and 200U high concentration T4 ligase were added for ligation overnight at 16°C. Following phenol-chloroform purification, samples were precipitated via the addition of 14ul glycogen and 35ml 100% ethanol at -80°C. The precipitated DNA was pelleted and washed as above, and resuspended in 150ul 10mM Tris-HCl pH 7.5. Residual salt was removed via Qiagen PCR purification columns.

4C library PCR amplification and Illumina high-throughput sequencing

Inverse PCR primers (Table S2) were designed to anneal to a bait locus HindIII/DpnII restriction fragment (selected with the criteria that it be longer than 300bp and be within 50kb of the indicated gene) and to amplify the unknown portion of the chimeric DNA circle generated during 4C library preparation. The resulting DNA circles consist of the bait locus restriction fragment and its interacting partner's restriction fragments. The six 3' nucleotides of the primers annealing to the HindIII side of the restriction fragment contained the HindIII restriction site when possible, or were generally within 4bp of the start of the HindIII site, to avoid uninformative reads upon sequencing. Primers on the DpnII side were allowed more positional flexibility, as sequencing data were not produced from the DpnII end of the restriction fragment.

200ng of DNA from the 4C library were used as template for the PCR amplification using the Expand Long Range PCR system (Roche). 5uM each of forward and reverse primer lacking Solexa sequencing adaptors were applied to amplify the interactome of interest in a 25ul reaction volume under the following PCR conditions: 1 cycle at 92°C for 2 minutes; (92°C 30 seconds; 58°C 1 minute; 68°C 1 minute) x 10 cycles; 1 cycle of 68°C 7 minutes. PCR products were run on an agarose gel, and amplicons between 100-500bp were isolated, gel extracted (Qiagen Gel purification system), and used as template for a second PCR reaction utilizing the same primers

with the addition of the Solexa adaptors in a 50ul volume as follows: 1 cycle of 92°C for 2 minutes; (92°C 30 seconds; 58°C 1 minute; 68°C 1 minute) x 10 cycles; (92°C 30 seconds; 68°C 1 minute +20 seconds/additional cycle; 68°C 1 minute) x 15 cycles; and 1 cycle at 68°C for 7 minutes. The PCR-amplified library was purified over GFX PCR DNA purification kit columns (GE Healthcare) to remove primer dimers, followed by a second purification with Qiagen MinElute Reaction Cleanup kit (Qiagen) to remove residual salt. Samples were quantified using the Quant-iT dsDNA BR assay kit quantification system (Invitrogen) with a Qubit fluorometer. Purified, PCR-amplified 4C-seq libraries were pooled in EB (Qiagen), 0.1% Tween-20 for multiplexed sequencing as primer distinctiveness allowed, and sequenced at the Broad Stem Cell Research Center at UCLA.

4C sequencing and read mapping

Two to seven 4C libraries were multiplexed and sequenced using the Illumina Genome Analyzer II to obtain 76 base pair (bp) reads or Illumina HI-seq-2000 to obtain 100 bp reads. Reads were parsed based on a unique, non-annealing two base pair bar code and/or unique bait-specific primer sequences. The resultant reads were mapped to the mouse genome (build mm9) using Bowtie software (Langmead et al., 2009). Only reads that aligned to a unique position in the genome with no more than two sequence mismatches were retained for further analysis. Reads that were successfully aligned to the genome were then remapped to the 736,199 unique HindIII sites along the genome by matching their respective loci. Because we were concerned about any potential biases created by differential mappability, we excluded all HindIII sites that do not precede a unique 50bp along both DNA strands. In other words, we only mapped reads to HindIII sites that are unique in the mm9 genome with respect to the 100bp centered on the hexamers that comprise the sites, as illustrated below:

```
(+)              AAGCTT--- 50bp ---
(-) --- 50bp ---TTCGAA
```

By only considering unique HindIII sites, we have restricted our 4C analysis to highly mappable regions of the genome

Table S1 provides a summary of all the bait regions for which 4C-seq libraries were generated for all cell types discussed in this study. Read distributions and statistics

for all individual data sets that passed quality control steps (see below) are contained in Table S3.

4C hit determination

To reduce potential clonal amplification effects inherent to PCR-based genomic approaches, we collapsed the raw 4C-seq read count at each HindIII site down to a 'hit' if the site met a read count threshold. The threshold was chosen so that at least 80% of all hits were intrachromosomal, i.e. at most 20% of our hits fell in *trans* (Figure S1E). For each library, the threshold and the number of sites that passed this threshold (i.e. the number of hits) are provided in Table S3. For the majority of libraries the count threshold for calling hits was a single read.

4C library quality control (QC)

We used three criteria adapted from van de Werken et al. (van de Werken et al., 2012) to estimate the quality of our individual 4C-seq libraries, as well as an additional two criteria of our own. First, we checked whether the library under consideration was comprised of at least 500,000 reads in total. Second, the *cis*/genome read count ratio (the number of mapped reads in *cis* over the total number of mapped reads) had to be at least 20%. Third, at least 20% of HindIIII sites within the 2Mb region around the bait had to be covered by at least one read. If a library passed all three of these de Laat group-inspired criteria, then it received one credit. Libraries received additional credit for passing each of the following two tests: 1) having a *cis:trans* hit ratio that was at least 4:1 (i.e. a maximum of 20% of thresholded hits could fall in *trans* for the library to be credited); and 2) having at least 1.5% of all sites along the *cis* chromosome covered by a hit. In summary, each library could achieve a maximum score of 3/3. Libraries that received a total score of 3/3 automatically passed QC, while those with 2/3 were subject to further scrutiny and only passed if they exhibited strong metrics for their two passing criteria. The libraries scoring less than 2 did not pass QC. Excluding three control libraries, 198/242 (84%) of the 4C-seq libraries passed the QC and were kept for further analysis (Figure S1D). Those libraries that passed QC are given in Table S3.

To ensure that those genomic regions within closed chromatin environments (negative PC1 scoring regions) were digested as efficiently as those regions in open chromatin surroundings, we required two things: 1) Our 4C-seq libraries had to exhibit at

least 20% HindIII site coverage within the 2Mb regions around the bait, as shown in Table S3 and Figure S1D. Notably for baits with negative PC1 scores, this demonstrates that closed chromatin is subject to proper digestion. 2) Regions proximal to the bait of high and low PC1 character had to show similar average hit probability (data not shown).

Pooling of replicate 4C-seq libraries

Replicate 4C-seq libraries for a single bait locus in a given cell type that passed QC (described above) were pooled by calculating the average number of times each site was called a hit (by the thresholding criteria described above) in all replicates. In essence, we determined the probability of a hit at each HindIII site along the genome for each bait and cell type. Table S1 summarizes the bait regions for which pooled 4C-seq libraries were produced in each cell type, while Table S3 lists all the replicate libraries that were pooled, as well as statistics pertaining to hit probability. In this study we considered 66 pooled 4C-seq data sets (Table S1).

Definition and calculation of 4C hit percentage

In order to obtain a smoother continuous signal at a scale that was compatible with our genomic feature and PCA data (see below), we determined the average hit probability within 200 kilobase (kb) windows tiled along each chromosome, referring to this as the 'hit percentage.' We observed a strong correlation between hit percentage and binned Hi-C read counts using equivalent window sizes (Table S5).

Binomial test analysis

To demarcate positively interacting regions of a bait locus along each chromosome, we sought to identify statistically significant clusters of HindIII sites that exhibited a high probability of being hit across replicates. We used R's binomial test function (Team, 2011) to calculate the probability of seeing the observed proportion of hits to HindIII sites, or 'hit percentage', within a 200kb window around each HindIII site along each chromosome, relative to the expected proportion obtained by modeling the average hit percentage as a function of distance from the bait locus across all data sets in a given cell line (see below for details on the background modeling) in *cis*, or relative to the average hit percentage for each respective *trans* chromosome. Using the observed hit probability within the 200Kb surrounding each site, we determine the number of hits that this represented, given the number of sites within the window: i.e. number of hits = hit

probability * number of sites with 200Kb window. We used the resulting number of hits as the value for the binomial test parameter representing the number of successes, with the number of HindIII sites being the number of trials, and the hit percentage obtained from the empirical background model (for a locus at the given distance from the bait locus in *cis*) or the average hit percentage in *trans* being the hypothesized probability of success. Only sites centered within windows containing at least ten HindIII sites were considered. A threshold of $-\log_{10}$(p-value) >= 1.8 was used to determine HindIII sites centered within windows showing significant clusters of interaction. This threshold was used as it resulted in a small false discovery rate (FDR; see determination below) in *cis*, while allowing us to pick up significant *trans* interactions. The binomial test results are given in Table S4. Adjacent and overlapping positive windows were concatenated into 4C-positive domains. Table S1 catalogs these significantly-interacting domains for pooled libraries ("interactome") as determined by our 4C-seq analysis pipeline.

In order to determine an FDR for each of our 4C-seq interactomes, we generated corresponding data sets of simulated hit probabilities. For each bait, we generated as many data sets of simulated hits as we had experimental replicates for that bait as follows: To simulate intrachromosomal hits we used the hit percentage specified by the empirical background model (described below) as the probability of sampling a hit at each site with respect to its distance from the bait. To simulate interchromosomal hits, the average hit probability for each chromosome in the experimental pooled data set was used as the probability of sampling a hit at each site along the chromosome. We pooled the resulting simulated replicates in the manner described above for our experimental replicates. Table S4 lists the number of significant windows and FDR in *cis* and *trans* for each of our 4C-seq data sets.

A 200kb window-size for the binomial test was chosen after having tested various window sizes and generated a multi-scale representation of the results (domainogram) to confirm the consistency of the binomial test p-values across window sizes (data not shown). This window size contained a sufficient number of HindIII sites (trials) to produce a robust binomial test result, while providing sub-megabase resolution.

To confirm that there was no strong bias for higher mappability within our libraries, illustrating that our selection of unique HindIII sites (as described above) had the desired

effect of only interrogating highly mappable regiions of the genome, we compared the average mappability within our 4C positive regions (4C domains) to that within 4C-negative regions, for intra-chromosomal interactions. We downloaded the UCSC ENCODE mm9 mappability scores based on 36bp alignment (ftp://encodeftp.cse.ucsc.edu/pipeline/mm9/wgEncodeMapability/wgEncodeCrgMapabilityAlign36mer.bigWig), using the 36mer mappability data because it reflects the typical length of our 4C-seq reads after trimming of the primer prefixes. The 36mer-based mappability scores are also the most stringent of the available mappability scores. All intra-chromosomal interactomes produced in our study are shown in the bar plot of the average difference in mappability scores (mean score of 4C domains (inside) - mean score of outside 4C domains (outside)) for all 66 of our baits (Figure S1G). We found that the average difference in mappability score for all baits (labeled 'MEAN') is small, supporting the conclusion that our results are not significantly biased by mappability. A similar lack of mappability bias was observed when genome-wide interactions were examined (data not shown). This result is consistent with the fact that we only considered data at unique HindIII sites in the genome.

Intrachromosomal empirical background model

To obtain an estimate of the expected probability of success for the binomial test, we used the average hit probability within 200kb windows from the bait locus across all baits to build a regression model of hit probability as a function of distance to the bait for each cell type. Adapting the approach used by Lieberman-Aiden et al. (Lieberman-Aiden et al., 2009) for their Hi-C data, we used the 1 – 8 Mb region proximal to the bait locus to fit a log-log regression model. The data in this region produced a much better fit than using the entire range of data points available (Table S7). We also noticed that within 1 Mb of the bait, the data deviated sharply from the power law scaling observed over the 1 – 8 Mb region, and that a linear regression model was more appropriate for this immediately proximal region around the bait locus. We therefore performed segmented regression analysis by partitioning the distance from the bait locus into the 0 – 1 Mb region and the 1 – 8 Mb region, using linear and log-log regression on these two regions, respectively. Model parameters and R-squared values are provided in Table S7. Figure 1A gives an example of the resulting empirical background model generated for the ESC *Pou5f1* pooled data set.

Data set correlation, overlap determination, and clustering

To compare interactomes across cell types, we performed Spearman's rho correlations on hit probabilities within 200kb windows tiled across each chromosome, and calculated Jaccard similarity coefficients using binary vectors representing 200kb tiled windows that overlapped 4C-positive domains (Figures 2 and S2). Dendrograms were obtained by converting the Spearman's *rho* statistics and Jaccard coefficients into distance measures and performing unsupervised hierarchical clustering using R's hclust function (Team, 2011) using the Ward agglomeration method.

RNA-seq

Strand-specific RNA-seq from V6.5 ECSs and wildtype 129SvJae MEFs (Table S6) was performed essentially as described in Parkhomchuk et al (Parkhomchuk et al., 2009), using 4 ug of total RNA as starting material. Reads were mapped to the mouse genome (mm9) using TopHat software (Trapnell et al., 2009) and only those reads that aligned with no more than two sequence mismatches were retained.

ChIP-seq

All histone modification data used for this study (Table S6) were determined using native ChIP (Wagschal et al., 2007) and will be described in detail in a separate manuscript (Chronis et al., in preparation). Briefly, nuclei were isolated from non-crosslinked V6.5 ESCs and 129SvJae MEFs by centrifugation through a sucrose cushion (1.2M sucrose, 60mM KCL,15mM NaCl, 5mM $MgCl_2$, 0.1mM Tris-HCl, 0.5mM DTT and protease inhibitor cocktail). Nuclei were then resuspended in Mnase-digestion buffer (0.32M sucrose, 50mM Tris-HCl, 4mM $MgCl_2$, 1mM $CaCl_2$, and protease inhibitor cocktail) and digested with 3 units of MNase (Roche) for 10 minutes at 37^{o}C. Soluble chromatin fractions were incubated with anti-H3K4me3 (Abcam; ab8580), anti-H3K4me2 (Abcam ab7766), anti-H3K4me1 (Abcam; ab8895), anti-H3K27me3 (Active Motif; 39155), anti-H3K27ac (Abcam; ab4729), and anti-H3K36me3 (Abcam; ab9050), respectively. Extracts were washed twice with wash buffer A (50mM Tris-HCl, 10mM EDTA, 75mM NaCl), wash buffer B (50mM Tris-HCl, 10mM EDTA, 125mM NaCl), wash buffer C (50mM Tris-HCl, 10mM EDTA, 250mM NaCl), once with LiCl buffer, and once with 1xTE. DNA extraction and library preparation as described (Wagschal et al., 2007).

Transcription factor binding data generated in this study (Table S6) were acquired using cross-linking ChIP and will be described in detail in a separate manuscript (Chronis et al., in preparation). V6.5 ESCs and 129SvJae MEFs were grown to a final concentration of $5x10^7$ cells for each sequencing experiment. Cells were chemically cross-linked by the addition of formaldehyde to 1% final concentration for 10 minutes and quenched with glycine at a final concentration of 0.125 M. Cells were then resuspended in buffer I (0.3M sucrose, 60mM KCl, 15mM NaCl, 5mM $MgCl_2$ 10mM EGTA, 15mM Tris-HCl, 0.5mM DTT, 0.2% NP-40, and protease inhibitor cocktail), and incubated on ice for 10 minutes. Nuclei were generated by centrifugation in a sucrose cushion (1.2M sucrose, 60mM KCL, 15mM NaCl, 5mM $MgCl_2$, 0.1mM Tris-HCl, 0.5mM DTT, and protease inhibitor cocktail). Isolated nuclei were resuspended in sonication buffer (50mM Hepes, 140mM NaCl, 1mM EDTA, 1% TritonX-100, 0.1% Na-deoxycholate, 0.1% SDS), and sonicated using a Diagenode Bioruptor. Subsequently, nuclear extracts were incubated overnight at 4C with one of the following antibodies: anti-Klf4 (R&D; AF3158), anti-Myc (R&D; AF3696), anti-Nanog (Cosmobio), anti-Oct4 (R&D; AF1759), anti-Sox2 (R&D AF2018), anti-p300 (SantaCruz; sc-585). Extracts were washed twice with RIPA, low salt buffer (20mM Tris pH 8.1, 150mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), high salt buffer (20mM Tris pH 8.1, 500mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS), LiCl buffer (10mM Tris pH 8.1, 250mM LiCl, 1mM EDTA, 1% deoxycholate, 1% NP-40), and 1xTE. Reverse cross-linking was performed by overnight incubation at 65C with 1% SDS and proteinase K. All protocols for Illumina/Solexa sequencing library preparation, sequencing, and quality control were performed as recommended by Illumina, with the minor modification of limiting the PCR amplification step to 10 cycles.

Reads were mapped to the mouse genome (mm9) using Bowtie software (Langmead et al., 2009) and only those reads that aligned to a unique position with no more than two sequence mismatches were retained for further analysis. Multiple reads mapping to the same location in the genome were collapsed to a single read to account for clonal amplification effects. ChIP-seq peaks were called using MACS software (Version 1.4.2) (Zhang et al., 2008) using a bandwidth parameter of 150 bp.

Chromatin states

Five chromatin states in ESCs and MEFs were identified at a resolution of 200bp as described by Ernst and Kellis (Ernst and Kellis, 2012) using the six histone modification ChIP-seq data sets listed in Table S6, plus a ChIP input dataset.

Transcription factor clusters

The genome was tiled into 1kb windows and the presence of transcription factor (TF) peaks from sixteen in-house and previously published ChIP-seq data sets for ESCs were used to define the TF clusters used in this analysis (Figure 3A and Table S6). For published data sets, we used peaks determined by the authors of the respective studies. The "Cohesin" data set represents the merging of peaks from the Smc1 and Smc3 data sets, and the "PRC2" data set represents the merging of peaks from the Eed, Ezh2, and Suz12 datasets (Table S6). This procedure resulted in a vector of binary data for each TF reflecting its absence or presence within 1kb windows across the genome. The windows represented by these vectors were then clustered using R's k-means function using the Hartigan-Wong method (Team, 2011) to obtain groups of windows exhibiting common combinatorial binding patterns across the genome. The number of centers (k=11) was chosen so as to substantially reduce the number of potential combinatorial TF groups (2^{16}-1), while ensuring that each cluster was represented by a significant number of windows (>7000 windows or ~0.25% of the genome). The eleven combinations of TFs found to co-bind within each window of a cluster are analogous in their combinatorial nature to the five chromatin states described above.

Principal Component Analysis

To compare our 4C interactome data to linear genomic features, including gene density, gene expression, replication timing, chromatin states, and transcription factor combinations in ESCs, we used Principal Component Analysis (PCA) to reduce the dimensionality of the 31 linear ESC genomic feature data sets (Table S6; discussed above and below). This allowed us to focus on weighted combinations of features, or principal components (PCs), that best characterized the genomic landscape of a cell type. Chromatin states and transcription factor clusters were used in order to capture the biologically important combinatorial nature of these features. The five described 200bp-based chromatin states and eleven 1 kb-resolution transcription factor cluster data were binned into 200kb windows resulting in a semi-quantitative profile of feature density across the genome. Density profiles for DNase and LaminB were obtained in a similar

manner by tallying the number of times they were present (at 1kb resolution) within 200kb windows. Replication timing data (Hiratani et al., 2010) were used to designate 1kb windows across the genome as either early (> 0.2) or late replicating (< -0.2) (Table S6). Again, the number of 1kb windows positive for either of these two replication timing states was tallied within larger 200kb windows to obtain vectors representing the density of early and late replicating regions of the genome, respectively. RNA-seq reads were binned in 200kb windows along the genome and the resulting read count totals were log-transformed to obtain a log-normal distribution more compatible with PCA. For gene density profiles, counts of unique transcription start sites from the UCSC mm9 refGene table within 200kb windows across the genome were obtained. This preprocessing step resulted in a 200 kb-resolution, 13,283 x 22 (windows x features) ESC feature matrix. This matrix of genomic feature data was passed to R's prcomp function (Team, 2011). Each column of data was scaled prior to performing the PCA. We used the top three PCs that best characterized the genomic landscape of ESCs for further analysis (PC1/2/3) (Figures 3 and 4).

We investigated whether mappability was particularly associated with any of the principal components considered in the study by including the average mappability within 200 Kb windows in the input matrix for our PCA of genomic features. We found that mappability does not contribute significantly to any of the top three components discussed in our study (data not shown).

To describe the linear genomic state of ESCs and MEFs, features available for both cell types were selected for PCA (Table S6). Vectors containing feature counts within 200kb windows for each cell type were concatenated, allowing PCA to be conducted on the combined genomic features. The resulting 26,566 x 15 (concatenated windows x common features) ESC+MEF feature matrix was passed to R's prcomp function to conduct the PCA. Only PC1 was considered for the ES+MEF features (Figure 6A, S5A-D).

Bait versus interactome comparisons based on PC score enrichment

To obtain a PC score enrichment value for each 4C bait, the mean PC score within the 200kb bait window and the four flanking windows (for a total of five 200kb bait windows = bait region) was calculated. The PC enrichment within the bait's interactome was

calculated in two ways: 1) as a Spearman's *rho* statistic by correlating the vector of 4C hit percentages (average hit probabilities) within 200kb windows across the chromosome to the corresponding vector of PC scores; and 2) by calculating the mean PC score within 200kb windows that overlapped 4C positive domains (as determined by the binomial test) by at least 25%. The five bait windows were excluded from the interactome enrichment. The rankings of the bait and interactome PC score enrichment values were then correlated to obtain the Spearman *rho* statistics shown in Figures 3E, 4C, and 4G.

Curve fitting

We used R's loess (Team, 2011) function to perform local fitting of a curve to PC data as a function of $-\log_{10}$(p-value) rank (Figure 7J-M, S6E, G-I), and subsequently R's predict.loess function to predict a loess fit and estimated standard error for each predicted value. An estimated 95% confidence interval was obtained by drawing a band +/- 2 s.e. on either side of the fitted curve.

Hi-C comparisons

Mouse ESC, normalized, Hi-C interaction matrices based on 40kb bins were downloaded from the Ren Lab website (http://chromosome.sdsc.edu/mouse/hi-c/download.html) (Dixon et al., 2012) and re-binned into 200kb bins to match the resolution of our 4C data. The row/columns within the Hi-C interaction matrix were extracted.

The rebinned Hi-C data (as described above) were also used for conducting "pseudo-4C", the Hi-C equivalents of the 4C-seq-based bait-interactome Spearman rank correlations in Figures 3E, 4C and 4G**.** Specifically, the row/columns within the Hi-C interaction matrix corresponding to each of the sixteen 4C baits analyzed in our ESC study were extracted. Bait PC score enrichments were calculated as described above for the 4C data sets by taking the mean PC score within the bait window and four flanking windows (five 200kb bait windows = bait region). The mean PC score within the top and bottom 5% of windows based on read count in each row of the rebinned Hi-C matrices were similarly calculated. To ensure that the interactome enrichment was not driven by the strong contacts centered around the bait locus, we excluded regions around the bait locus ranging in size from 1 to 50Mb to show that the enrichment was robust across the

length of the chromosome (Figure S3C/D, and data not shown). Furthermore, the vector of Hi-C read counts rebinned within the 200kb windows, besides the 5 bait windows, was Spearman rank correlated to the PC scores across the respective chromosome to obtain an enrichment value in terms of the *rho* statistic (Figure 3H, Table S5).

For the genome-wide bait-interactome PC score enrichment correlations (Figure 3G, 4D/H, S3C/D), each 200kb window of rebinned Hi-C data was treated as a pseudo-bait. Bait and interactome PC score enrichment values were calculated as described above for specific baits.

Fluorescence *in situ* hybridization (FISH)

Cells were grown on glass coverslips, washed for 30 seconds with ice cold cytoskeletal (CSK) buffer (100mM NaCl, 300mM sucrose, 3mM $MgCl_2$, 10mM PIPES, pH6.8), 30 seconds with CSK buffer containing 0.5% Triton X-100, and again for 30 seconds with CSK buffer, fixed in 4% PFA in 1xPBS for 10 minutes at room temperature, transferred to 70% ethanol, and stored at 4°C. Cells were dehydrated through a series 5 minute incubations in ice-cold 85%, 95%, and 100% ethanol, rehydrated in 2xSSC for 5 minutes, incubated in 2xSSC with 100ug/ml RNaseA (Invitrogen) at 37°C for 30 minutes, and washed three times for 5 minutes each in 2xSSC. DNA was denatured for 20 minutes at 80°C in 2xSSC with 70% deionized formamide (CalBiochem), followed by immediate quenching in ice-cold 70% ethanol and a second dehydration series (performed as above). DNA FISH probes were denatured at 95°C for 5 minutes and allowed to pre-hybridize for 1 hour at 37°C before being added to dry slides. Probes were then allowed to hybridize with cellular DNA at 37°C for 16-48 hours in a humid chamber containing 50% formamide in 2x SSC. Following hybridization, cells were washed three times in each of the following solutions, pre-warmed to 42°C: 2x SSC/50% formamide, 2x SSC, and 1x SSC. The second 1x SSC wash contained 100 ng/ml DAPI to visualize nuclei. Slides were mounted in aquapolymount (Polysciences) and allowed to set overnight. FISH probes were generated from bacterial artificial chromosome (BAC) DNA (Figure 1c: *Pou5f1* locus – RP23-213M12; A – RP23-98F21; B – RP23-106C23; C – RP23-85E24; D – RP23-2B8; Figure S6c/d: *Hoxa* - RP24-283F1; *Hoxb* - RP23-290I2; *Hoxc* - RP23-473J19; *Hoxd* - RP24-398B4; *Sox2* - RP23-2B8) by incorporation of fluorescently labeled nucleotides (Cy3-dCTP, Perkin Elmer and Alexa 488-dUTP, Invitrogen) via biopriming (Invitrogen).

3D-FISH image acquisition and analysis

3D images were constructed from a series 0.2um z-stacks through selected individual ESCs or ESC colonies. 3D-distance measurements between FISH signal centers were acquired using the *Smart FISH3D* plugin for ImageJ (Gue et al., 2005; Schneider et al., 2012). Distance distribution statistics were calculated in R (Team, 2011).

Immunofluorescence and image analysis

V6.5 mouse ESCs were grown on glass coverslips, washed for 30 seconds with ice cold cytoskeletal (CSK) buffer (100mM NaCl, 300mM sucrose, 3mM $MgCl_2$, 10mM PIPES, pH6.8), incubated for 1 hour with CSK buffer containing 0.7% Triton X-100, and washed again for 30 seconds with CSK buffer, fixed with 1xPBS/4% PFA for 10 minutes at room temperature, washed for 5 minutes in PBS/0.2% tween, incubated in blocking buffer (1x PBS, 10% goat serum, 0.2% fish skin gelatin, 0.2% tween) for 30 minutes, and incubated overnight in primary antibody diluted in blocking buffer at 4°C (anti-Nanog [eBioscience 14-5761-80]; anti-H3K27me3 [Active Motif 39155]; anti-RNAPII-S5 [Millipore 05-623]). Following primary antibody incubation cell were washed three times in PBS/0.2% tween and incubated in secondary antibody in blocking buffer. Cells were then wash three times in PBS/0.2% tween, with the second wash containing 1ug/mL DAPI, and mounted in Polyaquamount.

Image acquisition was performed on a spinning disc confocal microscope. Line intensities in Figure 5 were determined using Slidebook software from 3i. Images were exported as tif-files, and subsequent analysis was performed with customized R software (Team, 2011). Nuclear signal was identified via the removal of blue pixels whose intensities were below the 20^{th} percentile of all blue pixel intensities for a given capture. Red and green signal outside of these preliminarily defined 'nuclear' signals were discarded. To remove nucleoli from the analysis 'nuclear' pixels with 0 values for either red or green pixels were similarly dropped from analysis. The remaining red and green pixels were quantile normalized, and the relationship between the remaining normalized red and green pixel positions and intensities were analyzed as shown in Figure 5.

3C

3C libraries were prepared in an analogous manner to 4C libraries, ending with the first ligation step. 3C primers are listed in Table S2, and PCRs were run with the following parameters: Primers A+B/C/D/E/F 95C 2 minutes, (95C 30 seconds, 60C 45 seconds, 72C 1 minute) x30, 72C 2 minutes, 1:5 dilution used as template for nested PCR with same parameters; Primers Z+V/X/Y/W - 95C 2 minutes, (95C 30 seconds, 68C 45 seconds, 72C 1 minute) x35, 72C 2 minutes.

Supplemental References

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.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature methods *9*, 215-216.

Gue, M., Messaoudi, C., Sun, J.S., and Boudier, T. (2005). Smart 3D-FISH: automation of distance analysis in nuclei of interphase cells by image processing. Cytometry A *67*, 18-26.

Hiratani, I., Ryba, T., Itoh, M., Rathjen, J., Kulik, M., Papp, B., Fussner, E., Bazett-Jones, D.P., Plath, K., Dalton, S.*, et al.* (2010). Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome research *20*, 155-169.

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 biology *10*, R25.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.

Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J., and Dekker, J. (2006). Mapping chromatin interactions by chromosome conformation capture. Current protocols in molecular biology / edited by Frederick M Ausubel [et al] *Chapter 21*, Unit 21 11.

Morin-Kensicki, E.M., Faust, C., LaMantia, C., and Magnuson, T. (2001). Cell and tissue requirements for the gene eed during mouse gastrulation and organogenesis. Genesis *31*, 142-146.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic acids research *37*, e123.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods *9*, 671–675.

Splinter, E., de Wit, E., van de Werken, H.J., Klous, P., and de Laat, W. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. Methods.

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M.*, et al.* (2013). Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1gamma in reprogramming to pluripotency. Nat Cell Biol *15*, 872-882.

Team, R.D.C. (2011). R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria.).

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

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

Wagschal, A., Delaval, K., Pannetier, M., Arnaud, P., and Feil, R. (2007). Chromatin Immunoprecipitation (ChIP) on Unfixed Chromatin from Cells and Tissues to Analyze Histone Modifications. CSH protocols *2007*, pdb prot4767.

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 biology *9*, R137.