Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions

Dixon et al.

Supplemental Table of Contents

I. Public Datasets analyzed

II. Supplemental Methods

III. Supplemental References

IV. Supplemental Figures

V. Supplemental Tables

Table S1 – Sequencing Library Information

Table S2 – Directionality Index Correlation Table

Table S3 - Domains in mESC, mouse Cortex, hESC, IMR90.

Table S4 - Boundaries in mESC, mouse cortex, hESC, IMR90.

Table S5 - GO Terms for Genes at Dynamic Interactions.

Table S6 – Dynamic interactions determined between mESC and mouse cortex

Table S7 - GO Terms for Genes at Boundary Regions.

Table S8 - Common boundary regions between mESC and mouse cortex.

Table S9 - Common boundary regions between hESC and IMR90 cells

Table S10 - Boundary regions in mESC annotated with CTCF binding sites or house keeping genes.

I. Public Datasets analyzed

Dataset	Figure	Accession	Reference
Lymphoblastoid Hi-C	Supplemental Figure 7	GSE18199	Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289-93 (2009).[31]
H3K4me3, H3K4me1, H3K27ac, p300, CTCF, ChIP-seq, mESC and cortex RNA-seq	Figures 1-4, Supplemental Figures 5,10,20-23		Shen, Y. et al. A Map of cis-Regulatory Sequences in the Mouse Genome. *in submission* (2012).[32]
Lung Fibroblast 5C	Supplemental Figure 4		Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120-4.[33]
Med1, Med12, Smc1, Smc3,	Supplemental Figure 5, 20-22	GSE22557	Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 430-5.[34]
mESC 2D-FISH	Figure 1, Supplemental Figure 6		Eskeland, R. et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell* 38, 452-64.[35]
Cortex H3K9me3	Figure 2	GSE33722	Xie, W. et al. Base-resolution analysis of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* 148 (4), 816-831.[36]
IMR90 H3K4me3, hESC H3K9me3, IMR90 H3K9me3	Figure 2, 4	SRP000941	Hawkins, R.D. et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 6, 479-91.[37]
mESC Lamina DAM-id	Figure 2, Supplemental Figure	GSE17051	Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* 38, 603-13.[38]
mESC Replication Timing	Supplemental Figure 14, 16	GSE18019	Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* 20, 155-69.[39]
H3K9me2 (LOCK) Domain ChIP-Chip	Supplemental Figure 15	GSE13445	Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* 41, 246-50 (2009).[40]
mESC H3K27me3, H4K20me3	Supplemental Figure 20-22	GSE12241	Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553-60 (2007).[41]

mESC H3K36me3, H3K79me2, Oct4, Sox2, Nanog	Figure 4, Supplemental Figure 20-22	GSE11724	Marson, A. et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134, 521-33 (2008).[42]
mESC H3K9me3	Figure 2, 4	GSE18371	Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B. & Young, R.A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* 23, 2484-9 (2009).[43]
mESC Jarid2, Jarid1a, Suz12, Ezh2	Supplemental Figure 20-22	GSE18776	Peng, J.C. et al. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* 139, 1290-302 (2009).[44]
mESC PolII Serine 5, PolII Serine 2, NelfA, Ctr9, Spt5	Supplemental Figure 20-22	GSE20530	Rahl, P.B. et al. c-Myc regulates transcriptional pause release. *Cell* 141, 432-45.[45]
DNase I HS	Supplemental Figure 20-22		Schnetz, M.P. et al. CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. *PLoS Genet* 6, e1001023.[46]
GRO-Seq	Figure 4	GSE27037	Min, I.M. et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* 25, 742-54.[47]
bioGPS database	Figure 4		Lattin, J.E. et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res* 4, 5 (2008).[48]

II. Supplemental Methods

Mapping

We mapped the data using BWA using default parameters. We consider only uniquely mapping reads (mapping quality > 10). We remove PCR duplicate reads using Picard (http://picard.sourceforge.net).

Interaction Matrices

The interaction matrices were calculated as previously described[31] at bin sizes ranging from 10kb to 1Mb.

Normalization

We normalized the Hi-C data as previously described by Yaffe and Tanay[49]. However, we did not perform linear weight smoothing and BFGS non-linear optimization

and the normalization is still effective at removing restriction enzyme bias (see Supplemental Figures 1 and 2).

Heat Maps and Visualization of Data

To visualize the high-resolution interaction data, we generated 2D heat-maps that were overlaid with publicly available ChIP-Seq data sets visualized in a genome browser (Figure 1a). Interaction frequencies were calculated as above. Interaction frequencies between any two loci can be visualized by identifying the point off the axis where diagonals originating from each locus intersect, in a manner similar to a linkage disequilibrium plot.

The heat maps in Supplementary Figure 4 are made differently. This is to correspond to the method used in (ref. 33) so we can accurately compare the interaction frequencies between our Hi-C data and the published 5C data from Wang et al. The interaction matrix is generated as follows. The 120kb HoxA locus is split into 30 segments using a 30kb sliding window with sliding in 3kb intervals. For each interaction between two 30kb windows *i* and *j*, we identify all possible HindIII cut sites in *i* and *j* and all possible HindIII cut sites interactions between these bins *i* and *j*. The interaction score between two segments of the heatmap is the mean frequency of interactions among all possible HindII cut site combinations between the two bins. The data for the Wang et al. 5C heatmaps was downloaded from the accompanying supplemental data[33].

Estimate of Intermolecular Ligation Rates

We estimated the intermolecular ligation rate between any two loci in the genome by analyzing the number of reads that map from a nuclear chromosome (chr(N)) to the

mitochondrial chromosome (chrM). As random intermolecular interactions will depend on the concentration of molecules in solution, the number of random interactions between the nuclear and mitochondrial chromosomes should be proportional to the amount of nuclear and mitochondrial DNA in solution during the ligation step of the protocol. As the number of mitochondria can vary between cell types, we use an estimated number of mitochondria of 40 based on previous experiments in the literature to test the number of mitochondria in mouse ES cells[50]. The total amount of "interacting space" between the mitochondrial genome and the nuclear genome is the product of the amount of mitochondrial DNA in solution (roughly 16kb/mitochondria * 40 mitochondria/cell) and the size of DNA in solution (roughly 5.1 Gigabases per diploid nucleus). By dividing the total number of chrM to chr(N) reads by this "interacting space," we can get an estimate of the number of reads/kbp^2 for any interaction in the genome. Our estimate suggest that for any two 40kb bins, there would be on average 0.015 reads per bin due to intermolecular ligations in the mouse ES cell HindIII original library and 0.079 reads /40kb interaction in the mouse ES cell replicate library. This is detailed in Supplemental Figure 27.

We would note that there are two potential pitfalls of this method. First, this requires an estimate of the number of mitochondria in a given cell type, which may not be available for any particular cell type of interest and can potentially vary by orders of magnitude. A second potential pitfall is that for the NcoI restriction enzyme, there are no mappable NcoI cut sites in the mitochondrial chromosome. Therefore, this method of analysis is not amenable to all restriction enzymes that could be used in a Hi-C experiment.

Correlation Between Experiments

We calculate the correlation between two experiments as follows: The set of all possible interactions I_{ij} for two experiments *A* and *B* were correlated by comparing each point in interaction matrix I_A from experiment *A* with the same point I_B from experiment *B*. Because the interaction matrix is highly skewed towards proximal interactions, we restricted the correlation to a maximum distance between points *i* and *j* of 50 bins. We use R to calculate the Pearson correlation between the two vectors of all point in I_A and I_B.

Directionality Index, Domain and Boundary Calling

We noted that the regions at the periphery of the topological domains are highly biased in their interaction frequencies. In other words, the most upstream portion of a topological domain is highly biased towards interacting downstream, and the downstream portion of a topological domain is highly biased towards interacting upstream. We reasoned that by identifying such biases in interaction frequency in the genome, we would be able to identify the locations of topological domains and boundaries in the genome.

To determine the directional bias at any given bin in the genome, we developed a Directionality Index (DI) to quantify the degree of upstream or downstream bias of a given bin. The directionality index is calculated in equation 1, where A is the number of reads that map from a given 40kb bin to the upstream 2Mb, B is the number of reads that

map from the same 40kb bin to the downstream 2Mb, and E, the expected number of reads under the null hypothesis, is equal to (A + B)/2.

Eq. 1

$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

The directionality index is based on the chi-squared test statistic, where the null hypothesis is that each bin is equally likely to interact with the regions upstream and downstream of it. Bins that show a directional bias have a directionality index proportional to the degree of bias, with more biased bins having a higher magnitude of directionality index. We use a 40kb bin size and a 2Mb because these parameters maximize the reproducibility of the DI and the domain calls while retaining a sufficiently high resolution to identify domains and boundary regions.

To generate a random directionality index, we randomized the direction either upstream or downstream of every read pair that mapped to a given bin and calculated the directionality index with the randomized directions. Bins with large random directionality indexes are virtually absent by chance, with less than 1% of the absolute value of random DI being greater than 6.57.

We consider the directionality index as an observation and believe that the "true" hidden directionality bias (DB) can be determined using a hidden Markov model (HMM). The HMM assumes that the directionality index observations are following a mixture of Gaussians and then predicts the states as "Upstream Bias", "Downstream Bias" or "No

Bias" (See Supplementary Figure 28 for a mathematical representation of our Hidden Markov Model).

Describing the observed directionality index as Y's $[Y_1,Y_2..Y_n]$, the hidden true directionality biases as Q's $[Q_1,Q_2..Q_n]$ and the mixtures as M's $[M_1,M_2..M_n]$. The probability $P(Y_t|Q_t = i,M_t = m)$ is represented using a mixture of Gaussians for each state *i*. The Conditional probability distribution [CPDs] of Y_t and M_t nodes are,

$P(Y_t = y_t|Q_t = i,M_t = m) = N(y_t;\mu_{i,m},\Sigma_{i,m})$

$P(M_t =m|Q_t =i) = C(i,m)$, where C encodes the mixture weights for each state i.

We used Baum-Welch algorithm [EM] to compute maximum likelihood estimates and the parameter estimates of transition and emission (characterized by mean, covariance and weights). The posterior marginals were then estimated using the Forward-backward algorithm.

For each chromosome, we allowed 1 to 20 mixtures and chose the mixture with best goodness of fit using the AIC criterion, $AIC = 2k - 2\ln(L)$, k is the number of parameters in the model and L being the maximum likelihood estimate. Matlab was used to perform the HMM.

As a post-processing step, we estimated the median posterior probability of a region, defined as a stretch of same state, and believed only in regions having a median posterior marginal probabilities ≥ 0.99 or a region that is at least 80kb long.

Domains and boundaries are then inferred from the results of the HMM state calls throughout the genome. A domain is initiated at the beginning of a single downstream

biased HMM state. The domain is continuous throughout any consecutive downstream biased states. The domain will then end when the last in a series of upstream biased states are reached, with the domain ending at the end of the last HMM upstream biased state. We term the regions in between the topological domains as either "topological boundaries" or "unorganized chromatin." We defined unorganized chromatin to be these regions that are > 400kb, and the topological boundaries to be less than 400kb. We would note that the topological boundaries, though defined as regions less than 400kb, are mostly quite small, with 76.33% being less than 50kb in size (mESC data).

Transcription Factor and Histone Modification Enrichment Analysis

We collected histone modification ChIP-Seq datasets from a variety of publically available databases. For mouse, each dataset was mapped using Bowtie[51] to the NCBI Build 37/mm9 reference genome. For humans, the data was mapped using Bowtie to NCBI Build 36/hg18. Peaks were called using MACS[52]. We performed post-processing of the MACS peaks by filtering out peaks with less than a 2-fold enrichment in signal compared to matched input or less than an absolute difference in RPKM of 1. The peak or binding sites frequency was then calculated for every 10kb bin in the genome. For generating the average peak frequency plots, the mid-point of each boundary region was identified, and peak frequency was calculated in 10kb bins for +/- 500kb from the boundary mid-point. For block like factors (H3K9me3, H3K27me3, H3K36me3, and H3K79me2), we did not use MACS peak calling and each 10kb bin score was simply the log2 ratio of the total ChIP-seq signal over the 10kb window divided by the input signal

of the window. The data were either averaged for the enrichment graphs (Figure 4, Supplementary Figure 20) or were plotted as heatmaps (Figure 2).

For determining which boundaries are associated with a given factor, we considered a boundary to be associated with a factor if there were a binding site called by MACS (for chromatin factors like CTCF) or if there were a locus (for example, the transcription start site of a housekeeping gene) within +/- 20kb of the boundary. The 20kb window is chosen because this reflects the inherent uncertainty in the exact position of the domain calls due to 40kb binning. The analysis shown in the pie chart in Figure 4e is performed as follows: First, boundaries with CTCF were identified. Second, boundaries with housekeeping genes were identified. If a boundary was not associated with a housekeeping gene, yet is associated with a non-housekeeping gene according to entropy scores, that is shown as a "other gene" associated boundary.

For the analysis of the patterns of H3K9me3 and Lamina DamID signal surrounding the boundary regions shown in Figure 2, we used k-means clustering to cluster the data. For Figure 2d, k-means clustering is performed on the hESC and IMR90 data simultaneously. Likewise, the mESC and cortex data were also clustered simultaneously.

GO Terms Enrichment analysis

GO terms enrichment analysis was performed using the DAVID tool. In figure 4, we display only non-redundant GO terms with a Benjamini corrected p-value less than 10^{-3}.

Dynamic Interactions

Differential interactions between mESCs and cortex were modeled as a Binomial distribution. For this analysis we combined the data from two pairs of replicates together (mouse ES cell versus cortex). We performed a binomial test for each possible interaction in the genome up to a distance of 5Mbp. The total number of trials (n) is equal to the sum of the reads in the two mESC replicates plus the sum of the reads in the two cortex replicates that map between two 20kb bins (I_{ij}) at a distance (*d*) ($n = I_{ij\text{-mESC}} + I_{ij\text{-cortex}}$). The expected ratio (p) of the mESC to cortex read ratio is equal to the ratio of the sums of all reads in the two mESC replicates between bins at distance (*d*) throughout the genome compared to the sum of the reads total reads between bins at distance *d* ($p = \Sigma I_{mESC}/n$ at distance *d* or $p = \Sigma I_{cortex}/n$). Therefore, deviations in the ratio of the number of interactions in mouse ES cells ($I_{ij\text{-mESC}}$) to the number of interactions in cortex ($I_{ij\text{-cortex}}$) will result in a significant p-value. We would note that this method accounts for the differences in sequencing depth between the two libraries by considering the expected ratio (p), which is proportional to the total sequencing depth. To model the extent to which noise or variability could contribute to dynamic interacting regions, we performed the same analysis but randomly permuted the combination of data. Specifically, under random permutation 1, we combine the mouse ES replicate 1 with the cortex replicate 1 and compared this to the combination of mouse ES replicate 2 with cortex replicate 2. For random permutation 2, we combined the mouse ES replicate 1 with the cortex replicate 2 and compared this to the combination of mouse ES replicate 2 with cortex replicate 1. Under a null hypothesis that the mouse ES cell and cortex Hi-C data sets are the same, we would expect a similar number of dynamic interactions when the actual groupings were considered (mESC1+mESC2 vs. cortex1+cortex2) as we would under the

random permutation (mESC1+cortex1 vs mESC2+cortex2 or mESC1+cortex2 vs. mESC2+cortex1). This also allows for an estimate of the number of dynamic interactions that would be observed to due random chance or noise, allowing us to calculate the False Discover Rate (FDR) of identifying dynamic interaction regions (the FDR is equal to the number of observed dynamic interactions in the randomly permutated data divided by the number of observed interaction in the actual data). For the dynamic interaction analysis, we only considered data from Hi-C experiments using the HindIII restriction enzyme to eliminate restriction enzyme effects as a possible confounding factor.

Housekeeping and Tissue Specific Gene Expression

"Housekeeping" and "Tissue Specific" genes were identified based on gene expression data from the bioGPS gene atlas database[48]. Specifically, the normalized probe intensities are used as a measure of absolute gene expression, with gene x being expressed at a level x_i in a given tissue or cell type *i*. The probability of expression p_i in a given cell *i* type is calculated as:

$$p_i = \frac{x_i}{\sum_1^N x}$$

and the entropy score for a given gene x is calculated as:

$$H(x) = -1 * \sum_1^N p_i \log2(p_i)$$

High entropy scores (> 6.12, corresponding to uniform expression in >70/96 tissues) have relatively uniform expression patterns and are considered to be "housekeeping" genes, while low entropy scores (<4.9) have highly variable expression patterns and are considered tissue specific (uniform expression in < 30/96 tissues). We exclude genes with entropy score between 4.9 and 6.12 as these are not well categorized as either "tissue specific" or "housekeeping."

Boundary Correlation Between and Across Cell Types

To correlate the boundaries both between and across cell types, we calculated the Spearman correlation coefficient of the directionality index between two cells. Specifically, if a boundary was called by the HMM in either cell type, we would identify the center of that boundary and correlate a vector of directionality indexes +/- 10 bins from the center of the boundary between two experiments of interest. For random correlation, we randomly selected 20 bins from each of the two cell types and calculated the spearman correlation between the two vectors. We repeated the randomization 10,000 times to achieve the random distribution of spearman correlation coefficients. Boundaries were called as "cell type specific" if the boundary regions was identified by the HMM domain calling in only one cell and lacked a significant correlation in the directionality index between the two cell types.

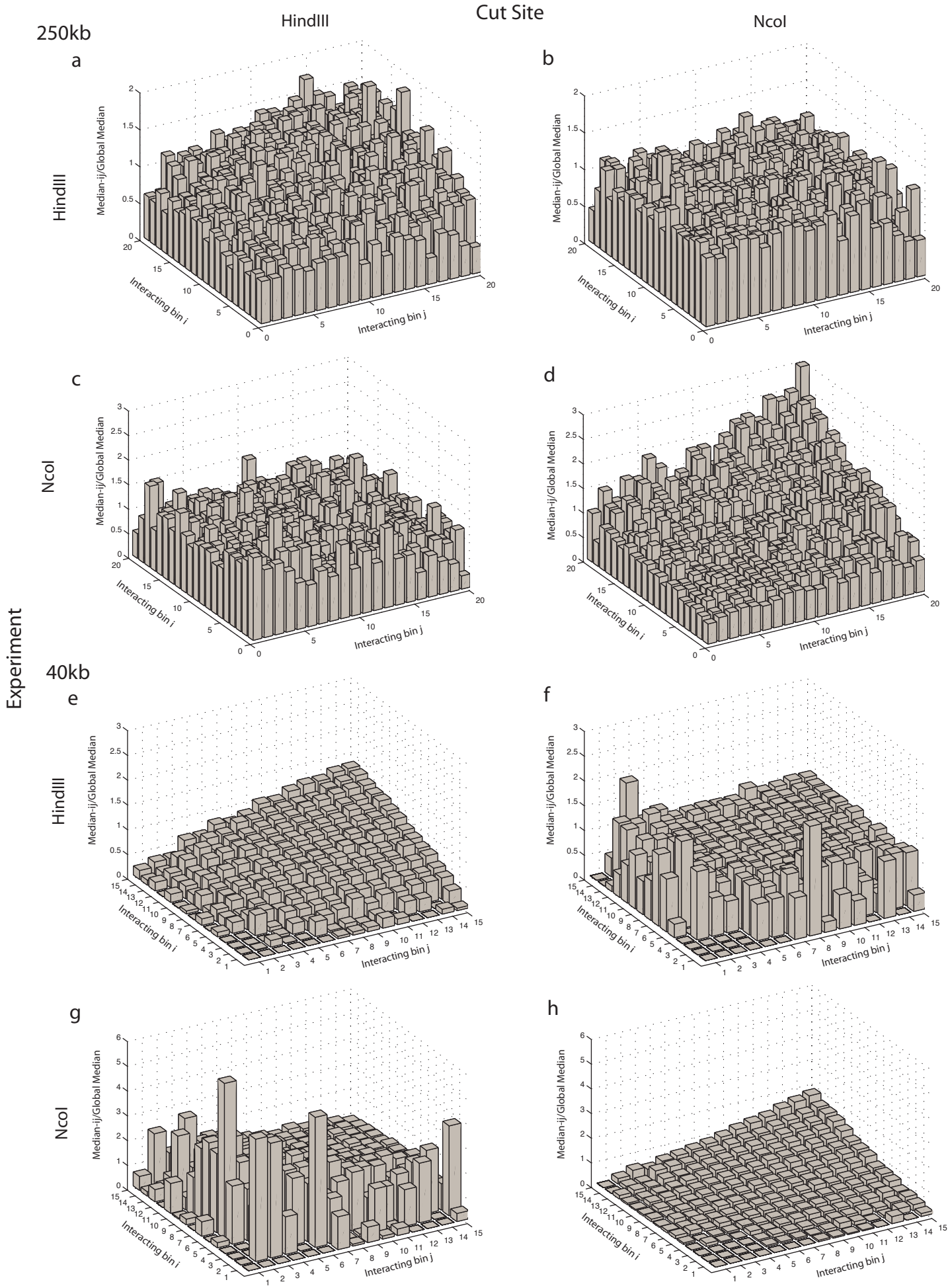
Boundary Conservation Across Species

Boundaries were lifted over using the UCSC Liftover tool[53] from species1 to species2 and the overlap between species1to2:species2 and species2to1:species1 were estimated. This overlap was compared with the random boundaries. The random boundaries were constrained on the distribution of boundary lengths and distribution of chromosomal occurrence.

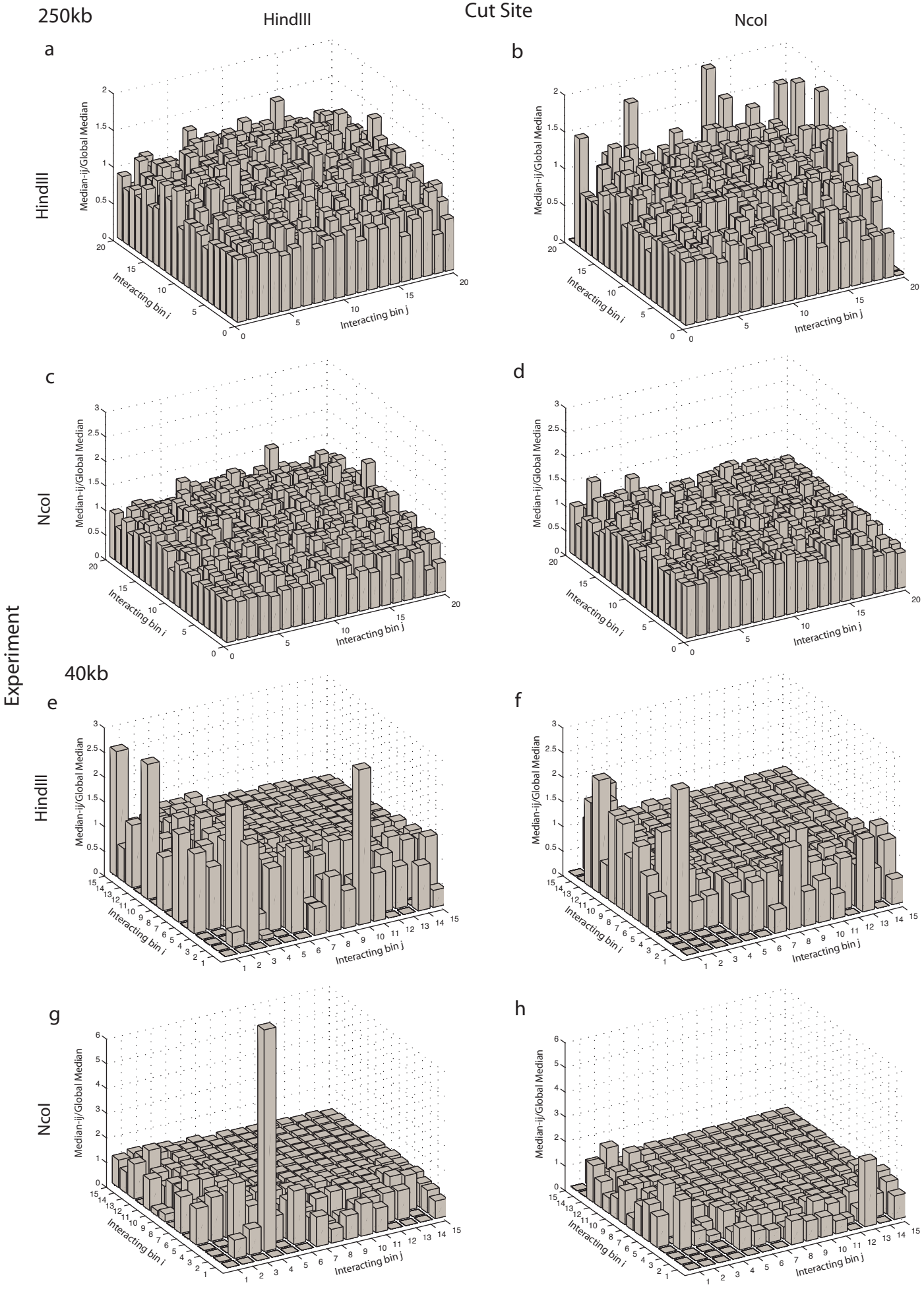
III. Supplemental References

31. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
32. Shen, Y. et al. A Map of cis-Regulatory Sequences in the Mouse Genome. *in submission* (2012).
33. Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-4.
34. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-5.
35. Eskeland, R. et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell* **38**, 452-64.
36. Xie, W. et al. Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. *Cell* **148**, 816-31.
37. Hawkins, R.D. et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-91.
38. Peric-Hupkes, D. et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-13.
39. Hiratani, I. et al. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* **20**, 155-69.
40. Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* **41**, 246-50 (2009).
41. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-60 (2007).
42. Marson, A. et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-33 (2008).
43. Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B. & Young, R.A. SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. *Genes Dev* **23**, 2484-9 (2009).
44. Peng, J.C. et al. Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139**, 1290-302 (2009).
45. Rahl, P.B. et al. c-Myc regulates transcriptional pause release. *Cell* **141**, 432-45.
46. Schnetz, M.P. et al. Genomic distribution of CHD7 on chromatin tracks H3K4 methylation patterns. *Genome Res* **19**, 590-601 (2009).
47. Min, I.M. et al. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**, 742-54.
48. Lattin, J.E. et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res* **4**, 5 (2008).
49. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-65.

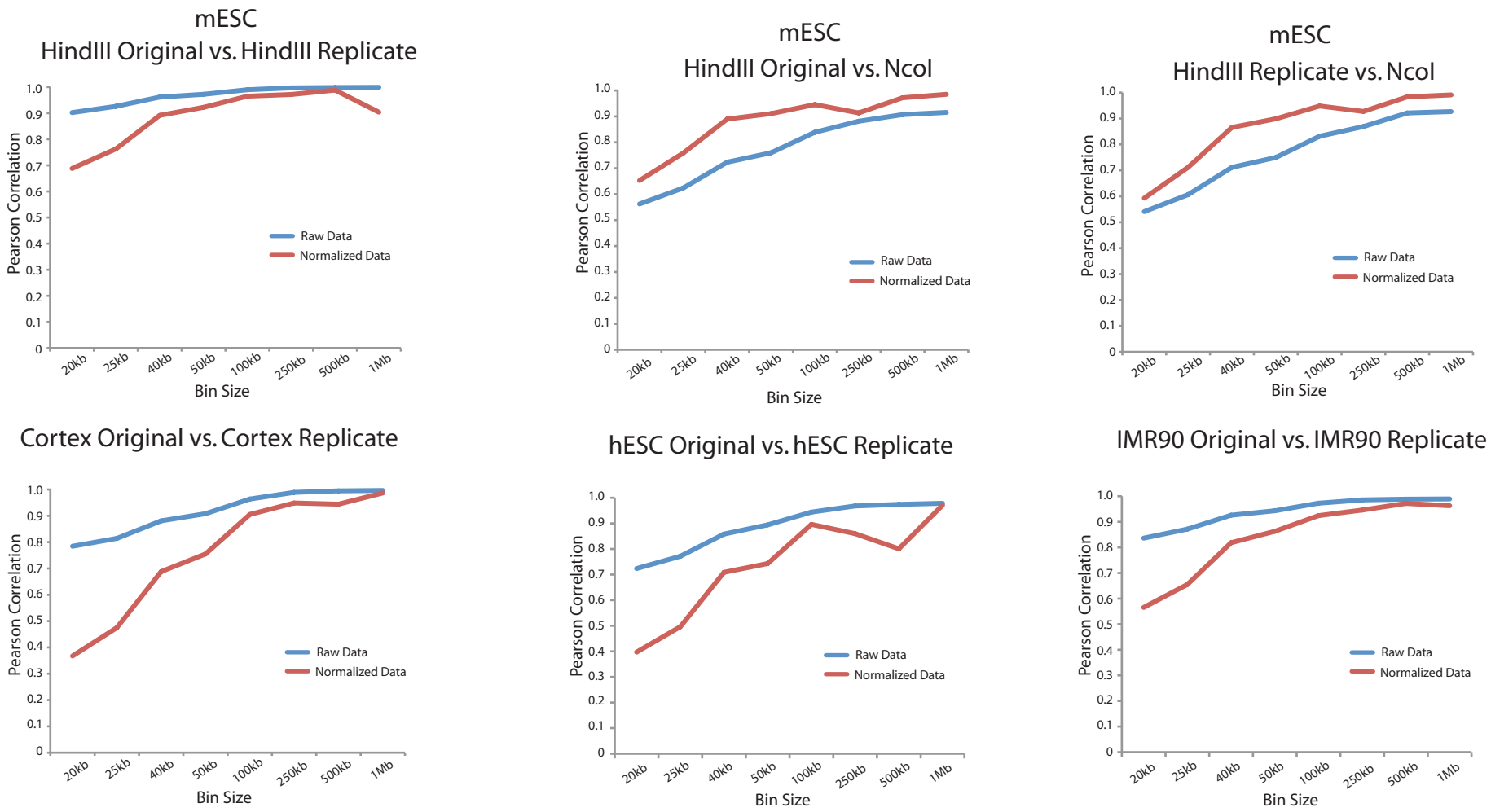
50. Facucho-Oliveira, J.M., Alderson, J., Spikings, E.C., Egginton, S. & St John, J.C. Mitochondrial DNA replication during differentiation of murine embryonic stem cells. *J Cell Sci* **120**, 4025-34 (2007).
51. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
52. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
53. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).


Cut Site
HindIII
NcoI
250kb
a
b
c
d
40kb
e
f
g
h
Experiment
HindIII
NcoI
Median-ij/Global Median
Interacting bin i
Interacting bin j

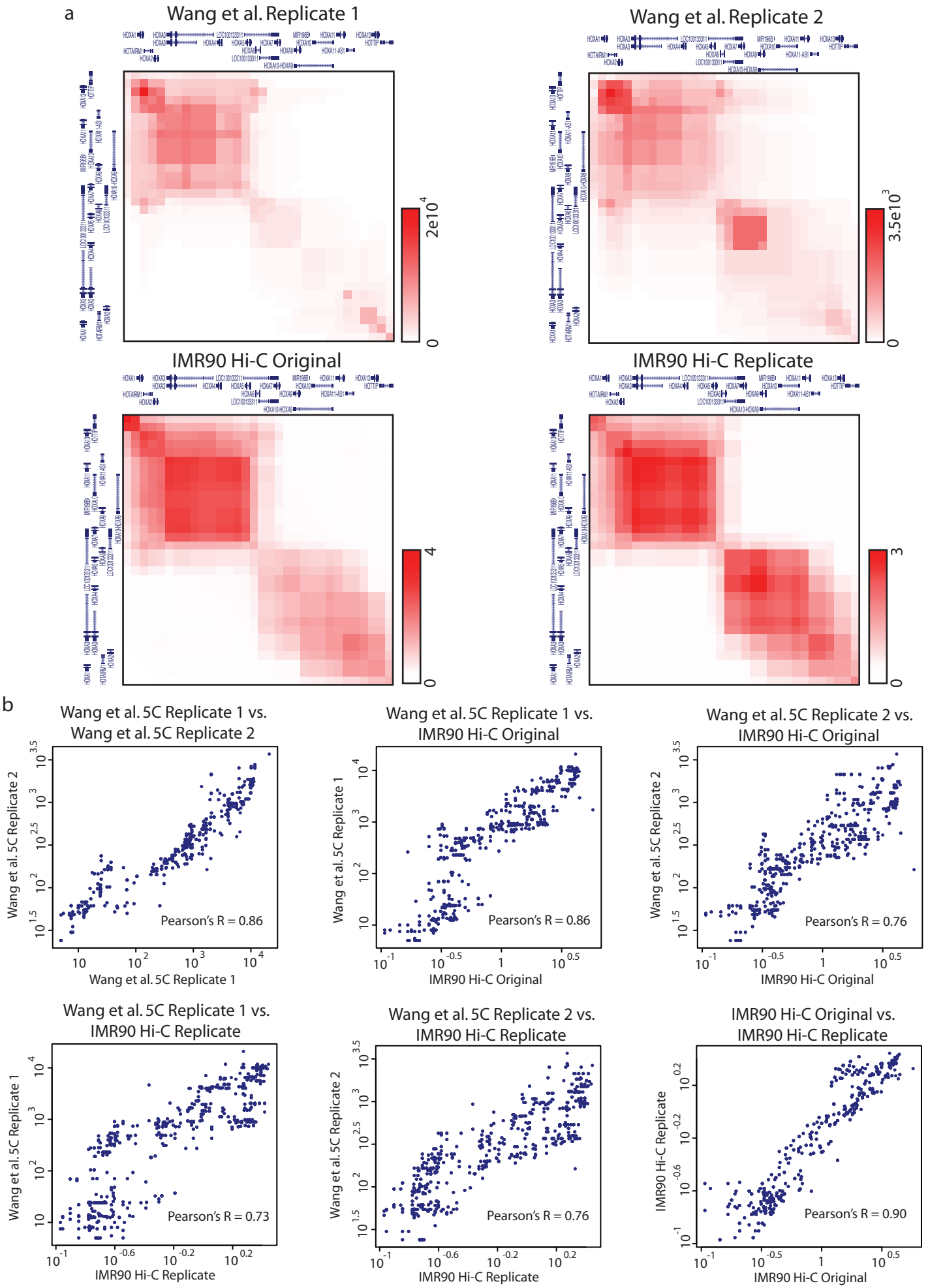
Supplementary Figure 1. Raw Hi-C Data and Restriction Enzyme Bias. a-d, Bias plots showing the correlation between restriction enzyme cut site frequency and Hi-C interaction frequency using a bin size of 250kb at a distance of 1Mb. For a-d, all 250kb bins were grouped into 20 equal sized groups based on increasing restriction enzyme frequency. The two horizontal axes correspond to the restriction enzyme group of each of the two bins, *i* and *j*, involved in an interaction $I_{ij.}$ The vertical axis shows the median of all interactions I_{ij} divided by the global median. Perfectly unbiased data should have all values roughly equal to 1. a, Comparison of HindIII restriction enzyme frequency with HindIII Hi-C data. b, Comparison of NcoI restriction enzyme frequency with HindIII Hi-C data. c, Comparison of HindIII restriction enzyme frequency with NcoI Hi-C data. d, Comparison of NcoI restriction enzyme frequency with NcoI Hi-C data. Note the correlation between the restriction enzyme cut site frequency and the Hi-C interaction frequency is only present when considering the restriction enzyme used in the Hi-C experiment. e-h, Similar to a-d, but using a bin size of 40kb and a distance of 80kb. The horizontal axis in e-h are the number of cut sites/40kb bin.


250kb
Cut Site
HindIII
NcoI
a
b
c
d
e
f
g
h
Experiment
HindIII
NcoI
40kb
Median-ij/Global Median
Interacting bin i
Interacting bin j

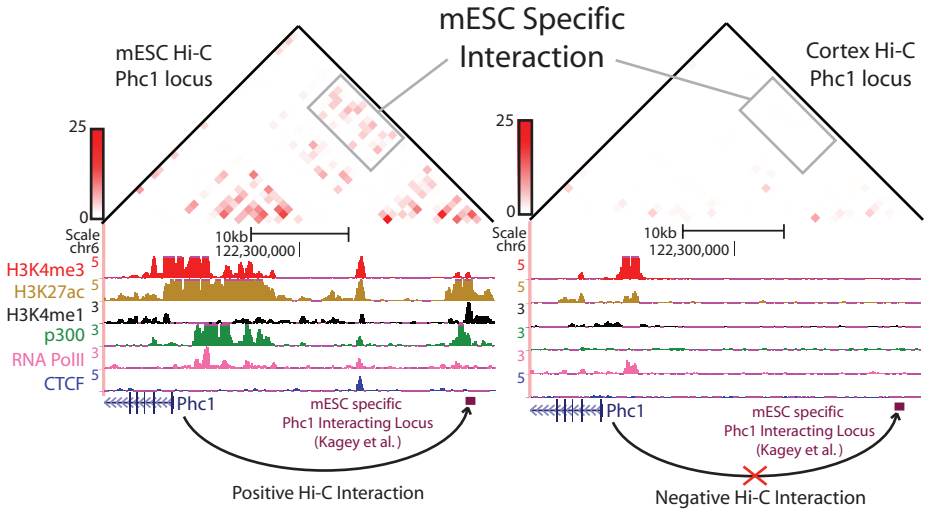
Supplementary Figure 2. Normalized Hi-C data shows no restriction enzyme bias. Identical to Supplementary Figure 1, yet using the normalized Hi-C data. Note that most values are roughly equal to 1, regardless of bin size or restriction enzyme, demonstrating that the restriction enzyme bias has been eliminated with normalization.



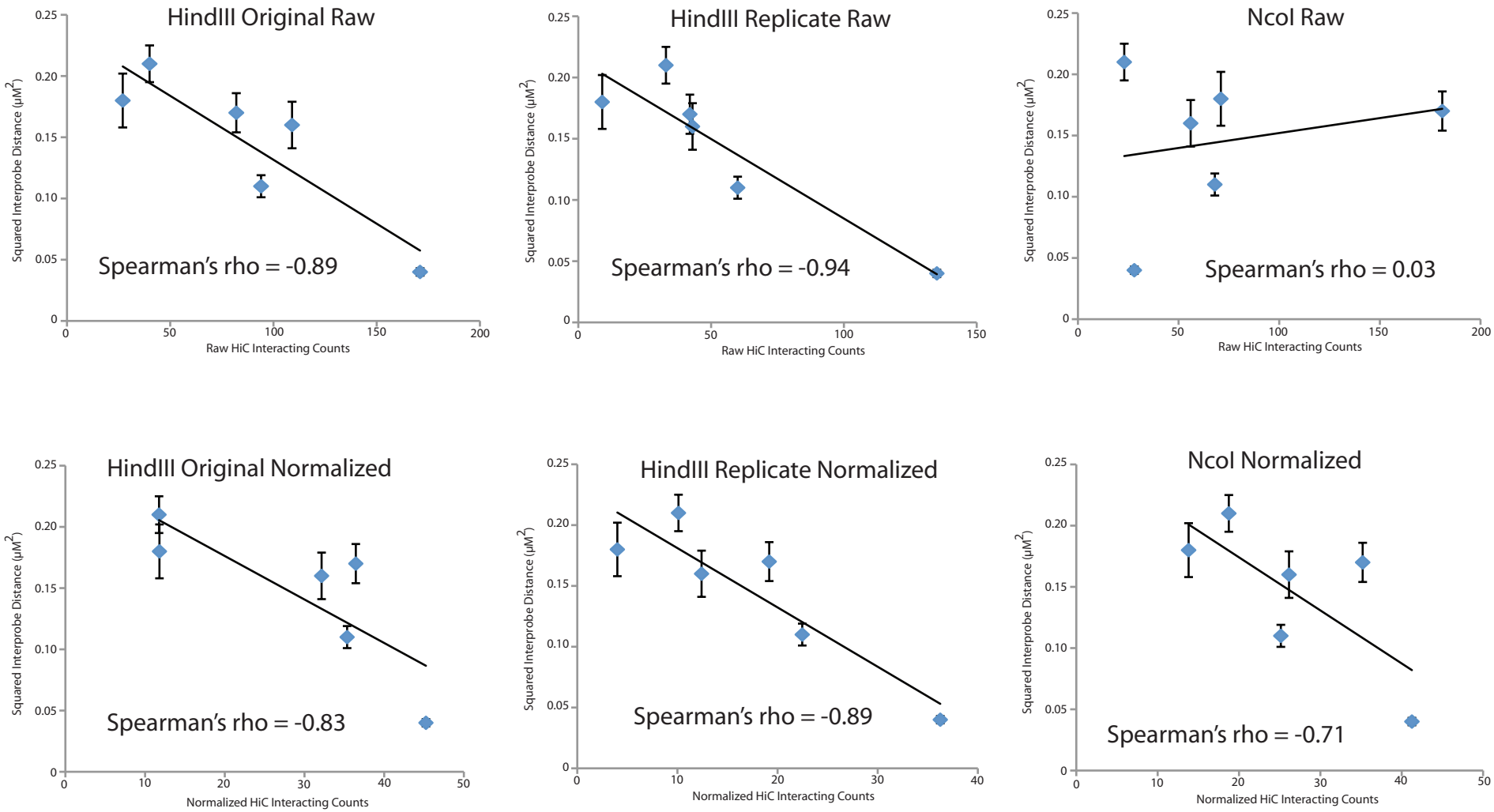
Supplementary Figure 3. Pearson Correlation between replicates. The Pearson correlation was calculated between each Hi-C replicate at varying bin sizes. The non-normalized data are shown in blue. The normalized data are shown in red.


a
Wang et al. Replicate 1
Wang et al. Replicate 2
IMR90 Hi-C Original
IMR90 Hi-C Replicate
b
Wang et al. 5C Replicate 1 vs. Wang et al. 5C Replicate 2
Pearson's R = 0.86
Wang et al. 5C Replicate 1 vs. IMR90 Hi-C Original
Pearson's R = 0.86
Wang et al. 5C Replicate 2 vs. IMR90 Hi-C Original
Pearson's R = 0.76
Wang et al. 5C Replicate 1 vs. IMR90 Hi-C Replicate
Pearson's R = 0.73
Wang et al. 5C Replicate 2 vs. IMR90 Hi-C Replicate
Pearson's R = 0.76
IMR90 Hi-C Original vs. IMR90 Hi-C Replicate
Pearson's R = 0.90
Wang et al. 5C Replicate 1
Wang et al. 5C Replicate 2
IMR90 Hi-C Original
IMR90 Hi-C Replicate

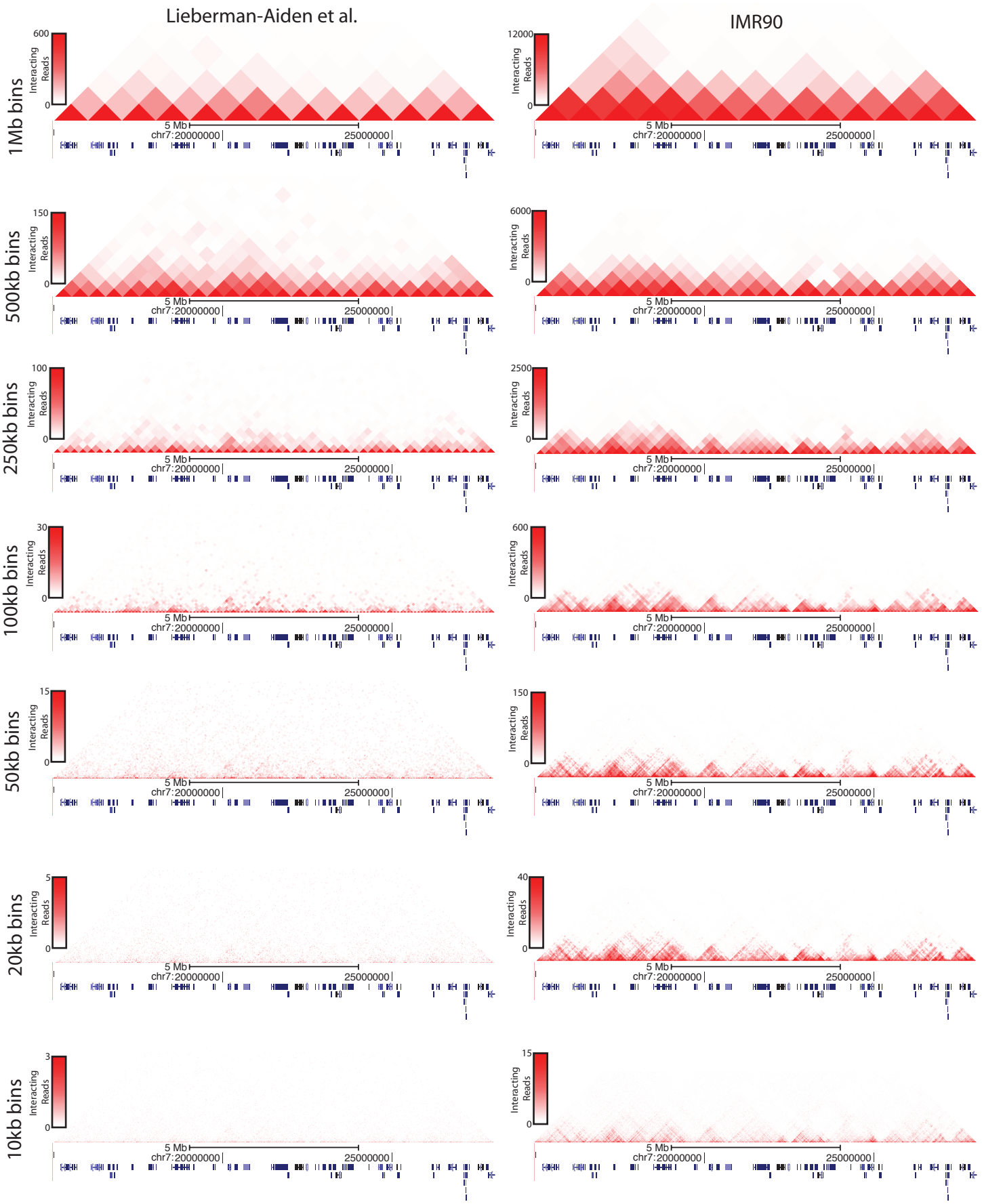
Supplementary Figure 4. Comparison with Previous 5C. a, Heat maps over the HoxA locus of 5C data from lung fibroblasts as reported previously[33] and the IMR90 Hi-C data generated in this report. Visually, there are two separate clusters of interactions in the upper left and lower right portions of the heat map. b, Scatter plots showing the correlations between 5C replicates and Hi-C data. In all cases, the correlation is > 0.73, demonstrating a high degree of correlation between IMR90 Hi-C data and existing 5C data a similar cell type.



Supplementary Figure 5. Comparison with Previous 3C data. 2D heatmap of Hi-C interactions at the Phc1 locus. The Phc1 promoter was previously shown to interact with a nearby enhancer by 3C, indicated by the arrow and red box. Gray boxes indicate the mESC specific Hi-C interactions.

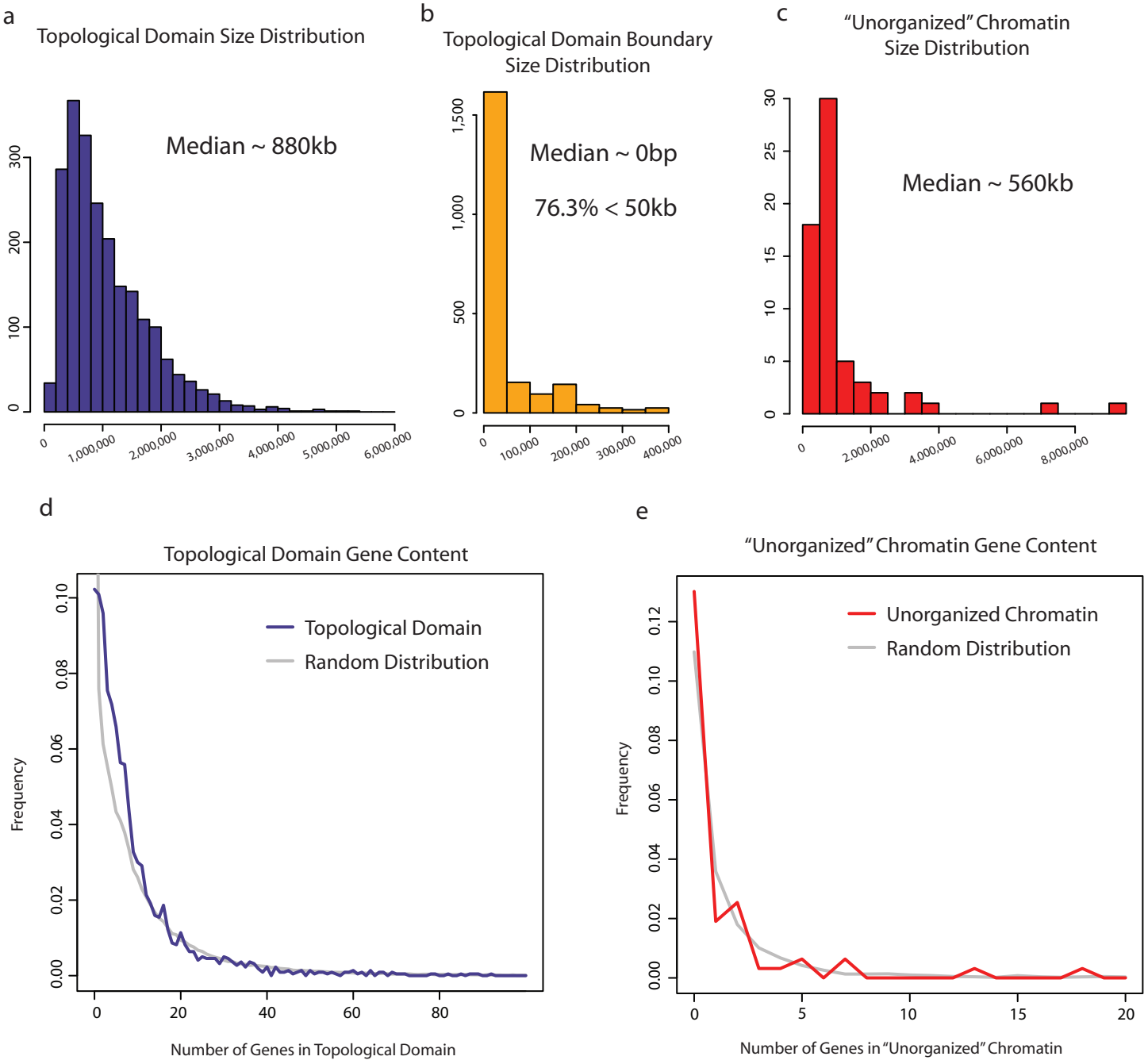


Supplementary Figure 6. Hi-C interaction frequency and mean spatial distance. The raw and normalized Hi-C interaction frequencies were compared with the mean nuclear separation as measured by 2D-FISH between six loci. The 2D-FISH data are from ref. 35.

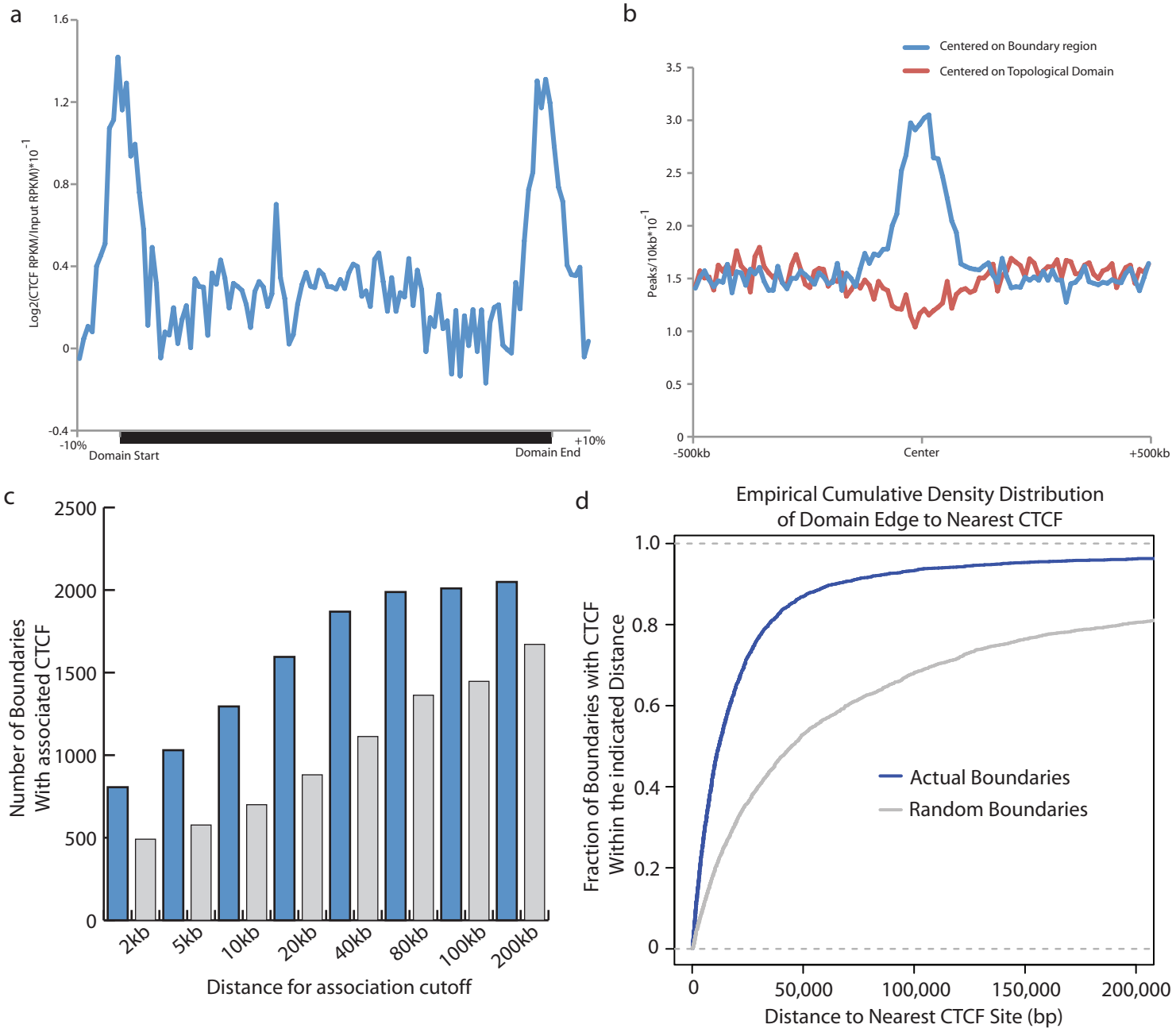

Lieberman-Aiden et al.
IMR90
1Mb bins
500kb bins
250kb bins
100kb bins
50kb bins
20kb bins
10kb bins
Interacting Reads
600
150
100
30
15
5
3
12000
6000
2500
600
150
40
15
0
5 Mb
chr7:20000000
25000000

Supplementary Figure 7. Hi-C interaction heat maps at varying bin sizes. Hi-C interaction frequencies are displayed as 2D heatmaps using differing bin sizes over a single locus on chromosome 7. Not the presence of the "triangles" on the heat map at a bin size or resolution of 100kb or less. A comparison with the data from the original Hi-C report is also shown for comparison[31].

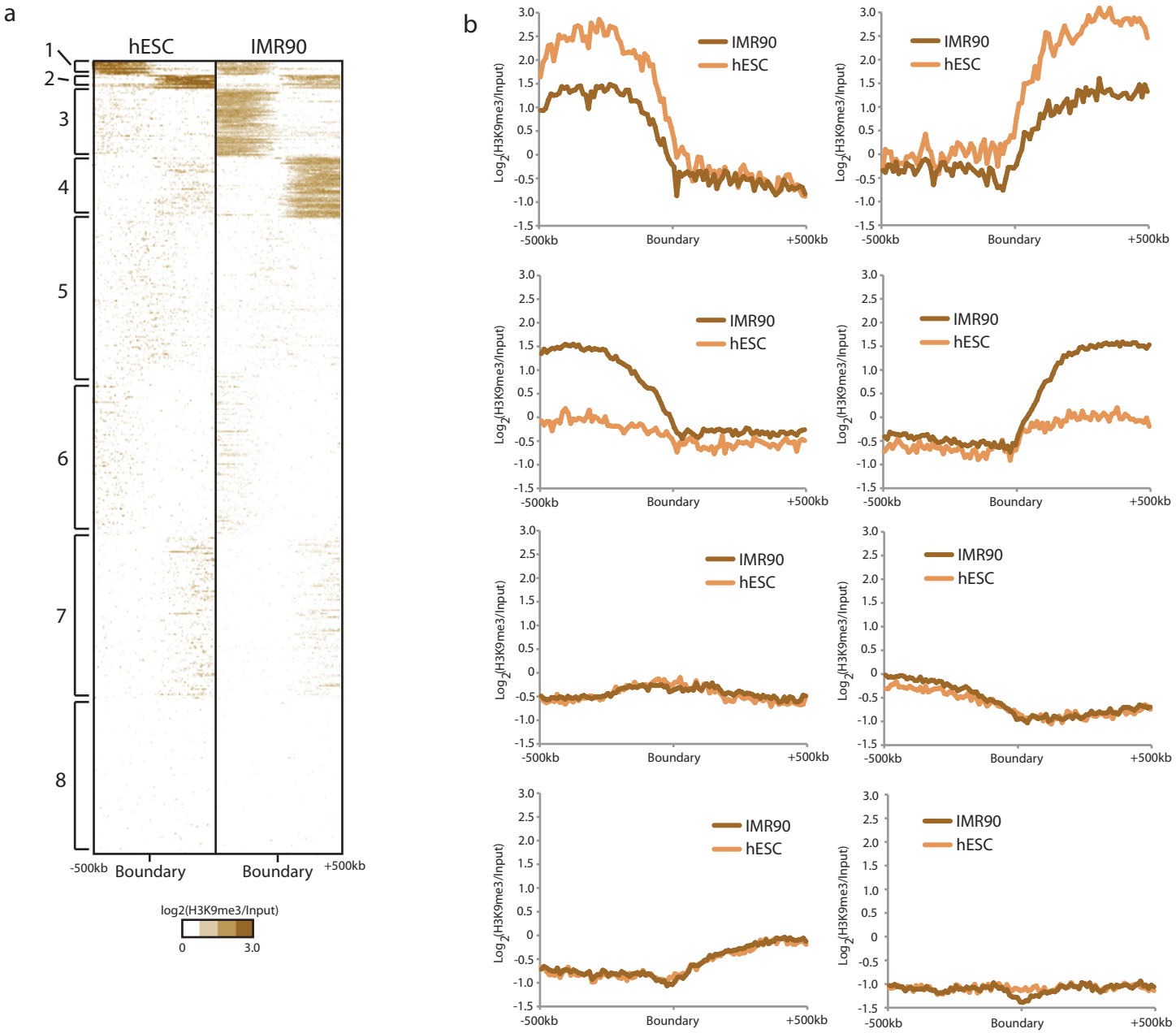
Supplementary Figure 8. Overlap of Topological Domain Boundaries between Hi-C replicates. Venn-diagrams comparing the amount of overlap between the topological domain boundaries called in each pair of Hi-C replicates.



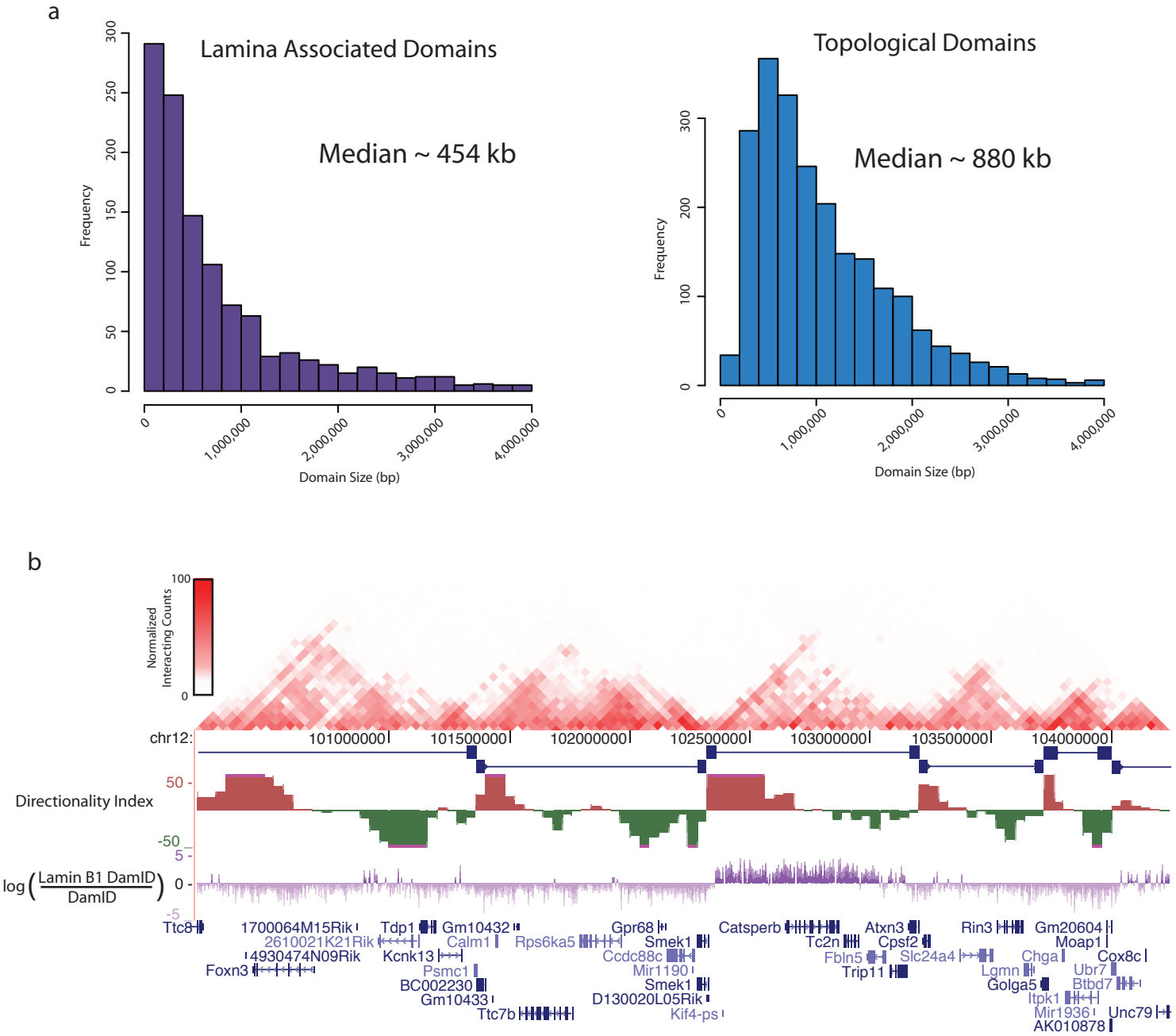
Supplementary Figure 9. Size distribution and gene content of topological domains, boundaries, and unorganized chromatin. a-c, Histograms of the sizes of topological domains (a), topological boundaries (b), and unorganized chromatin (c). d,e, Distribution of the gene content of topological domains and unorganized chromatin. Shown in gray is the gene content for randomly chosen regions of the genome with the same size distribution. Neither topological domains nor unorganized chromatin appear to differ from what is expected at random in terms of the distribution of their gene content.



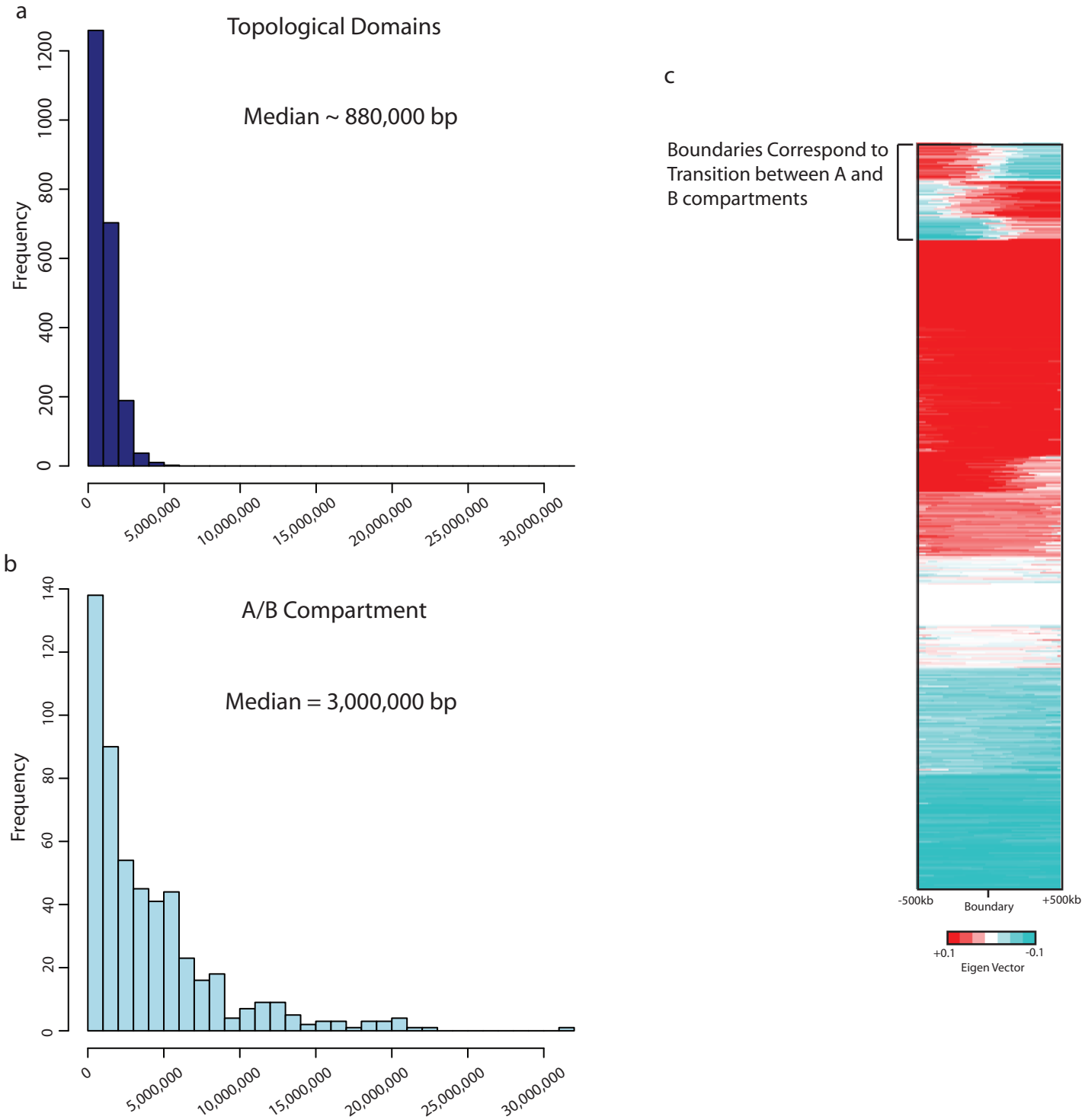
Supplementary Figure 10. CTCF enrichment at topological boundary regions. a, Average enrichment plot of CTCF over topological domains. Each topological domain was divided into 100 equally sized bins (+/- 10 bins from each end of the domain as well). The log2 ratio of CTCF RPKM over Input was calculated for each bin and shown as an average. CTCF appears to be enriched at the edges of each topological domain. b, Average enrichment plot of CTCF as shown in peaks/10kb bin. Shown in blue is the CTCF signal when centered on the topological boundary region. Shown in blue is the CTCF signal when centered on the middle of a topological domain, showing no enrichment. c, The number of boundaries with an associated CTCF site is shown for varying window size cut offs. For each distance *D*, the number of boundaries with a CTCF within +/- *D* are shown in blue. Shown in gray is the number expected at random at the same distance cut-off. d, The empirical cumulative density distribution of the distance between the domain border and the nearest CTCF binding site (in bp). The distance between the actual boundaries and the nearest CTCF site is shown in blue. The distance to randomized boundaries is shown in grey.



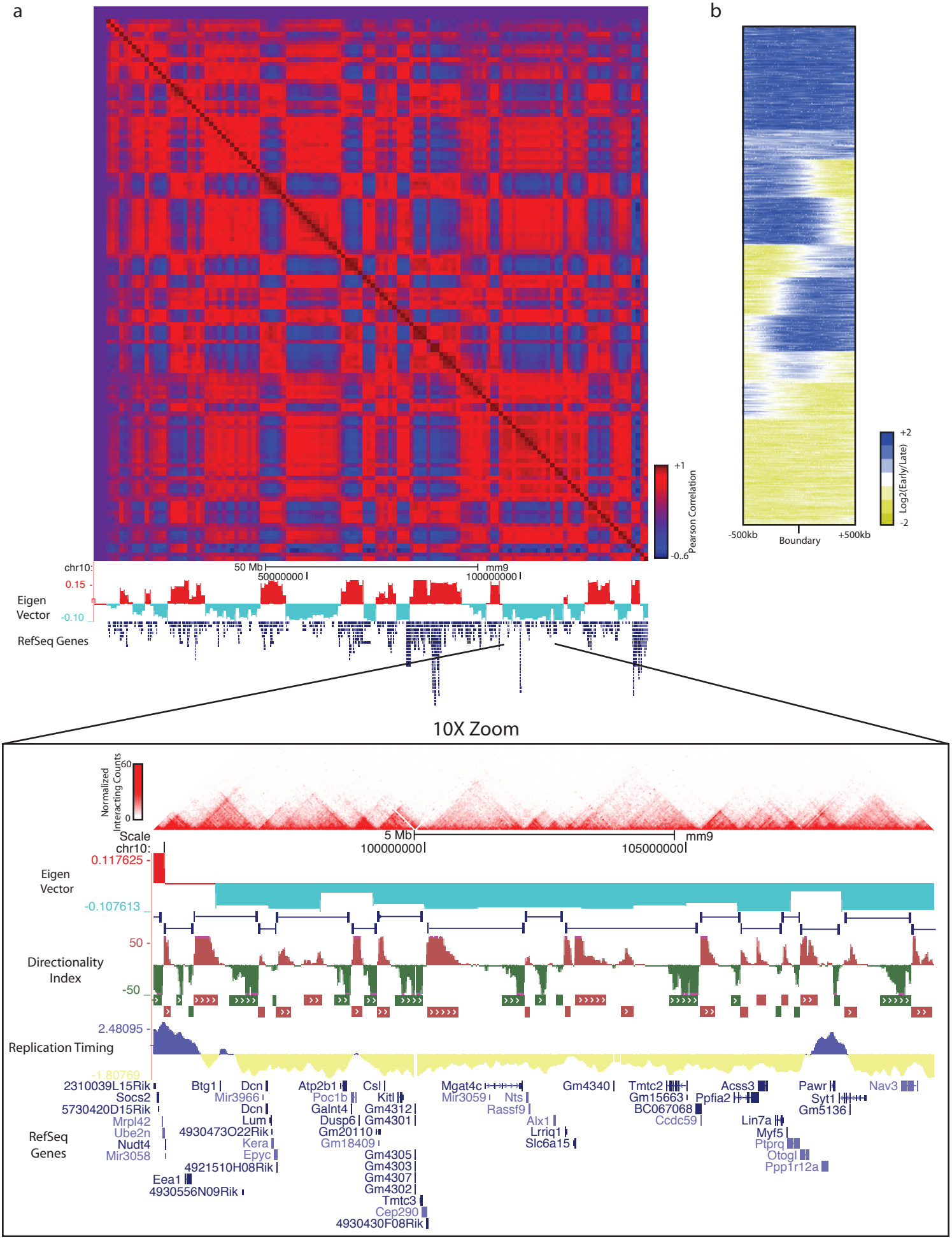
Supplementary Figure 11. Average Enrichment Plots of H3K9me3 surrounding the boundaries. a, Identical to Figure 2d in the main text, but labeled with cluster names 1-8 based on k-means clustering. b, The average enrichment plots of H3K9me3 for clusters 1-8 from panel a. Clusters 1-4 show clear enrichment of H3K9me3, and the transition from enriched to depleted H3K9me3 regions coincides with the location of the topological boundaries.



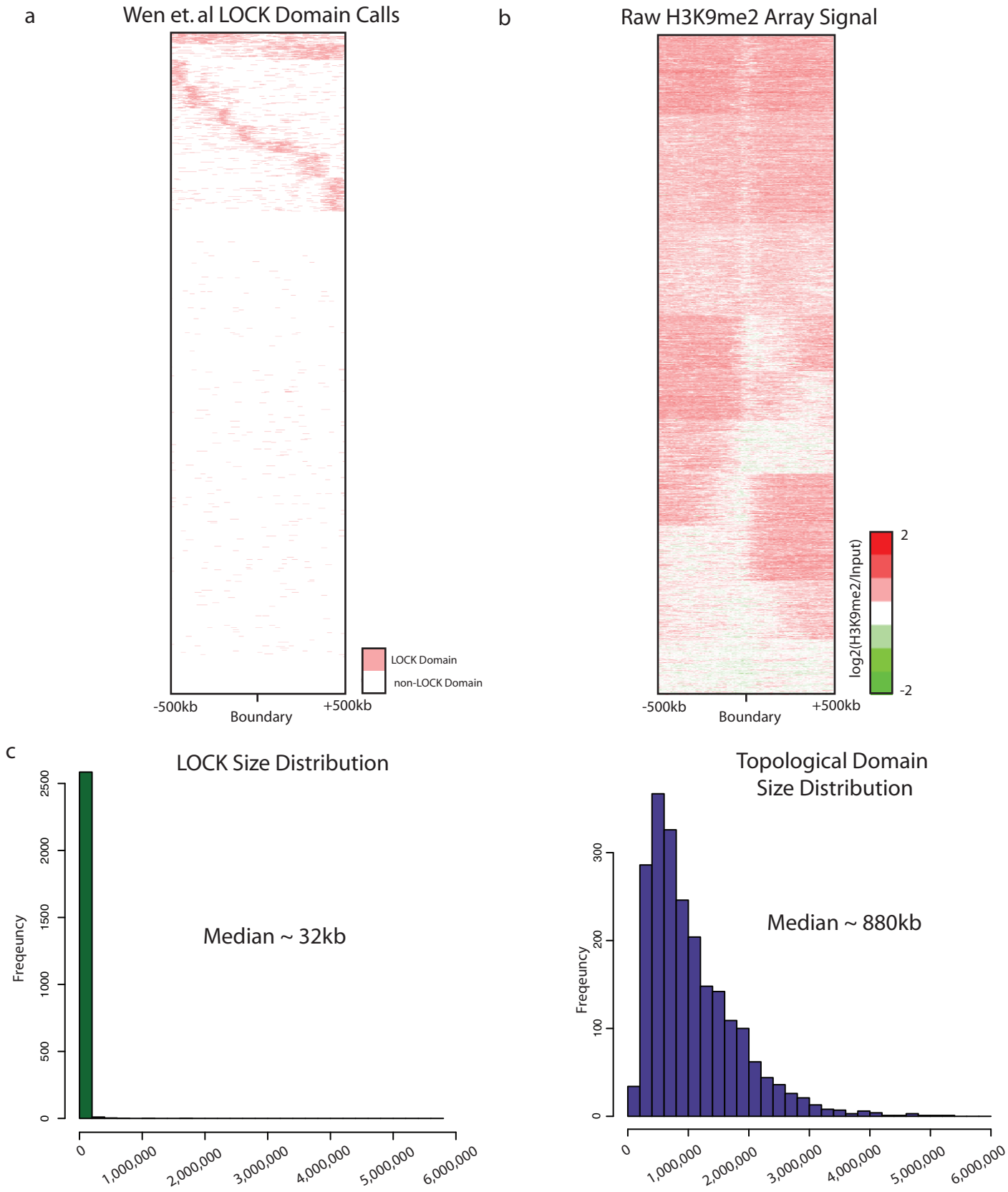
Supplementary Figure 12. Comparison of Topological Domains with Lamina Associated Domains (LADs). a, Histogram showing the size distribution of the topological domains and the LADs. Generally, LADs are smaller in size than topological domains. b, Genome browser shot showing a region on chromosome 12 with multiple topological domains, one of which appears to be entirely lamina-associated, with the remainder are non-lamina associated.



Supplementary Figure 13. Comparison of A and B compartments with topological domains in mouse ES cells. a,b, Histograms showing the size distributions of topological domains (a) and A and B compartments (b). Generally, the A and B compartments are larger than the topological domains. c, Heat map of the Eigen Vector values used to determine the A and B compartments at the topological boundary regions in mouse ES cells. The subset of boundaries that mark the transition between an A and B compartment are marked.

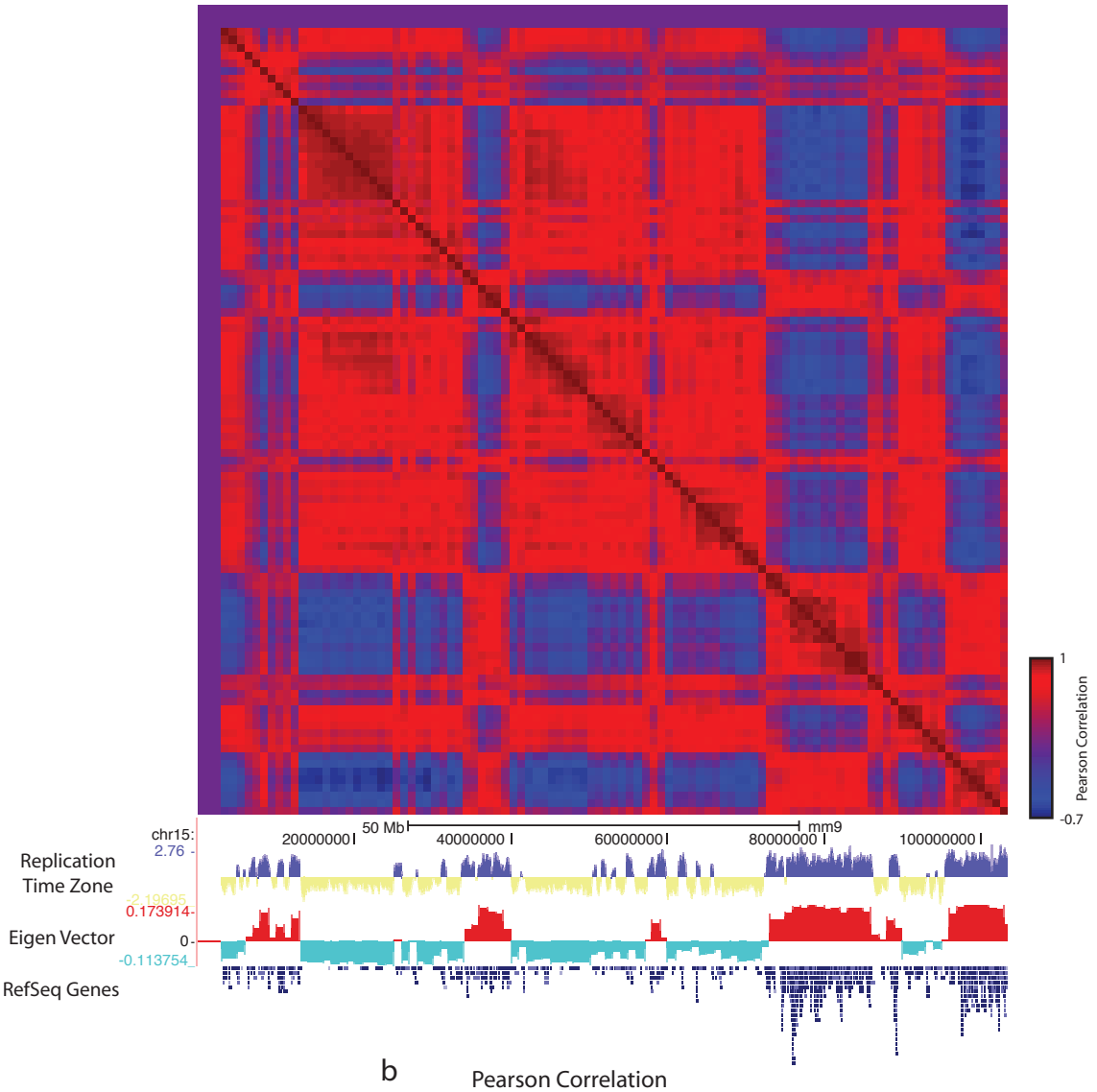

a
b
+1
Pearson Correlation
-0..6
chr10:
50 Mb
mm9
50000000
100000000
0.15
Eigen Vector
-0.10
RefSeq Genes
+2
Log2(Early/Late)
-2
-500kb
Boundary
+500kb
10X Zoom
Normalized Interacting Counts
60
0
Scale
chr10:
5 Mb
mm9
100000000
105000000
0.117625
Eigen Vector
-0.107613
50
Directionality Index
-50
2.48095
Replication Timing
-1.80769
RefSeq Genes
2310039L15Rik
Socs2
5730420D15Rik
Mrpl42
Ube2n
Nudt4
Mir3058
Btg1
Dcn
Mir3966
Dcn
Lum
4930473O22Rik
Kera
Epyc
4921510H08Rik
Eea1
4930556N09Rik
Atp2b1
Poc1b
Galnt4
Dusp6
Gm20110
Gm18409
Csl
Kitl
Gm4312
Gm4301
Gm4305
Gm4303
Gm4307
Gm4302
Tmtc3
Cep290
4930430F08Rik
Mgat4c
Mir3059
Nts
Rassf9
Alx1
Lrriq1
Slc6a15
Gm4340
Tmtc2
Gm15663
BC067068
Ccdc59
Acss3
Ppfia2
Lin7a
Myf5
Ptprq
Otogl
Ppp1r12a
Pawr
Syt1
Gm5136
Nav3

Supplementary Figure 14. Comparison of Topological Domains with A and B compartments and Replication Time Zones. a, Pearson correlation interaction heat map over chromosome 10. Shown in the blow up is a 10X zoom on a region entirely within the "B" compartment with multiple topological domains present in the region. b, Heat map of the replication time zone microarray data (ref. 39), surrounding the topological boundary regions.



Supplementary Figure 15. Comparison of Topological Domains with LOCK domains. a,b, Heat maps showing the enrichment of LOCK domains surrounding the topological boundary regions. Shown in (a) are the called LOCK domains[40], displayed as either LOCK in red or non-LOCK in white. Shown in (b) is the raw microarray data. c. Histograms showing the size distribution of LOCK domains and topological domains.

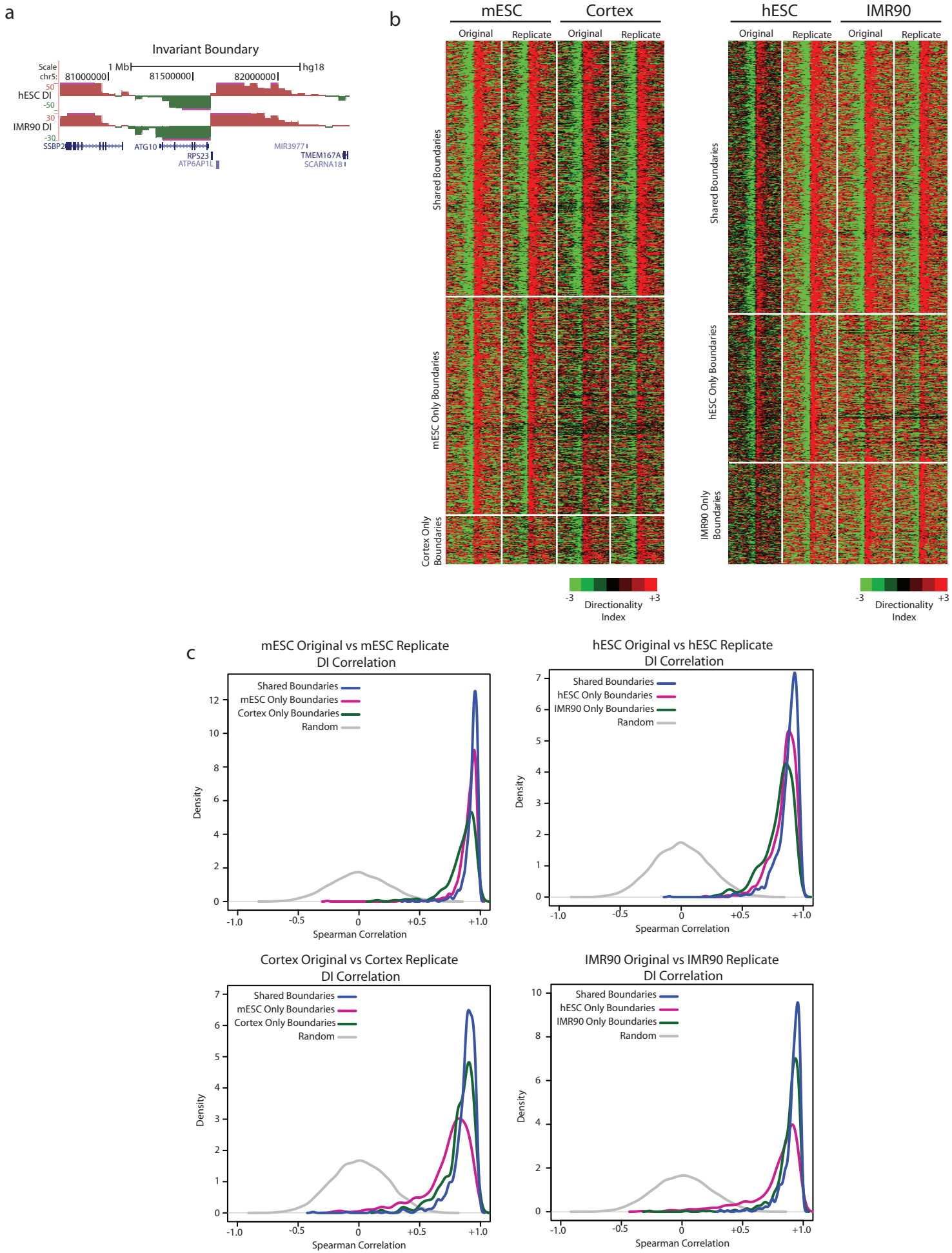
a



b

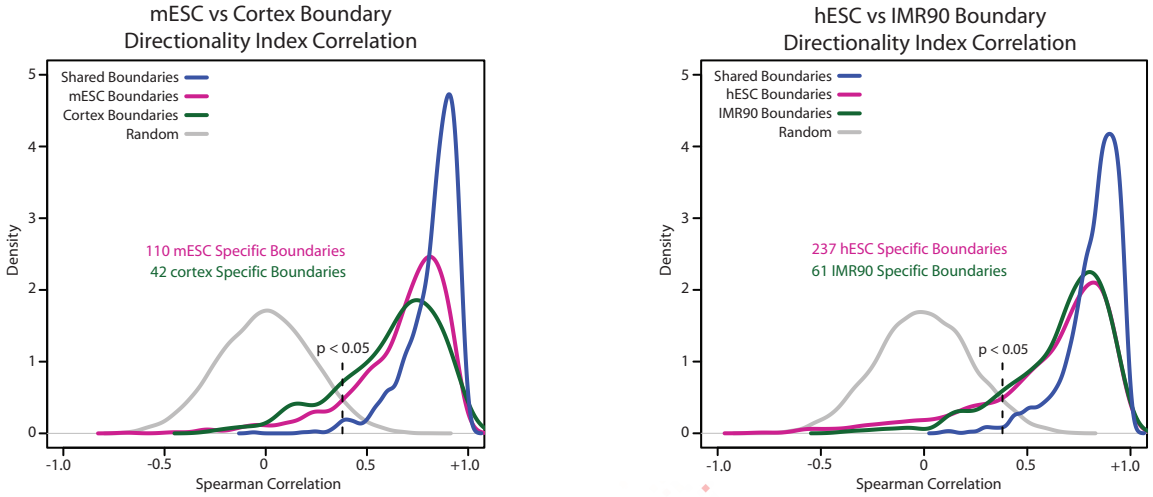
	Pearson Correlation
chr1	0.8851848
chr2	0.8840393
chr3	0.8435305
chr4	0.8936122
chr5	0.8861238
chr6	0.8416694
chr7	0.867437
chr8	0.8644409
chr9	0.8491173
chr10	0.872831
chr11	0.8923897
chr12	0.8651408
chr13	0.7615307
chr14	0.8289586
chr15	0.9228144
chr16	0.8513164
chr17	0.8556229
chr18	0.8418625
chr19	0.8943993

Supplementary Figure 16. Correlation of A and B compartments and replication time zones in mouse ES cells. a, Pearson correlation interaction heat map across chromosome 15. Below the heat map is the genome browser view of the Eigen vector used to determine the A or B compartments and the replication timing microarray data b, Pearson correlation coefficients of the Eigen vector values and the average probe intensity for replication timing data in 1Mb bins over each chromosome.

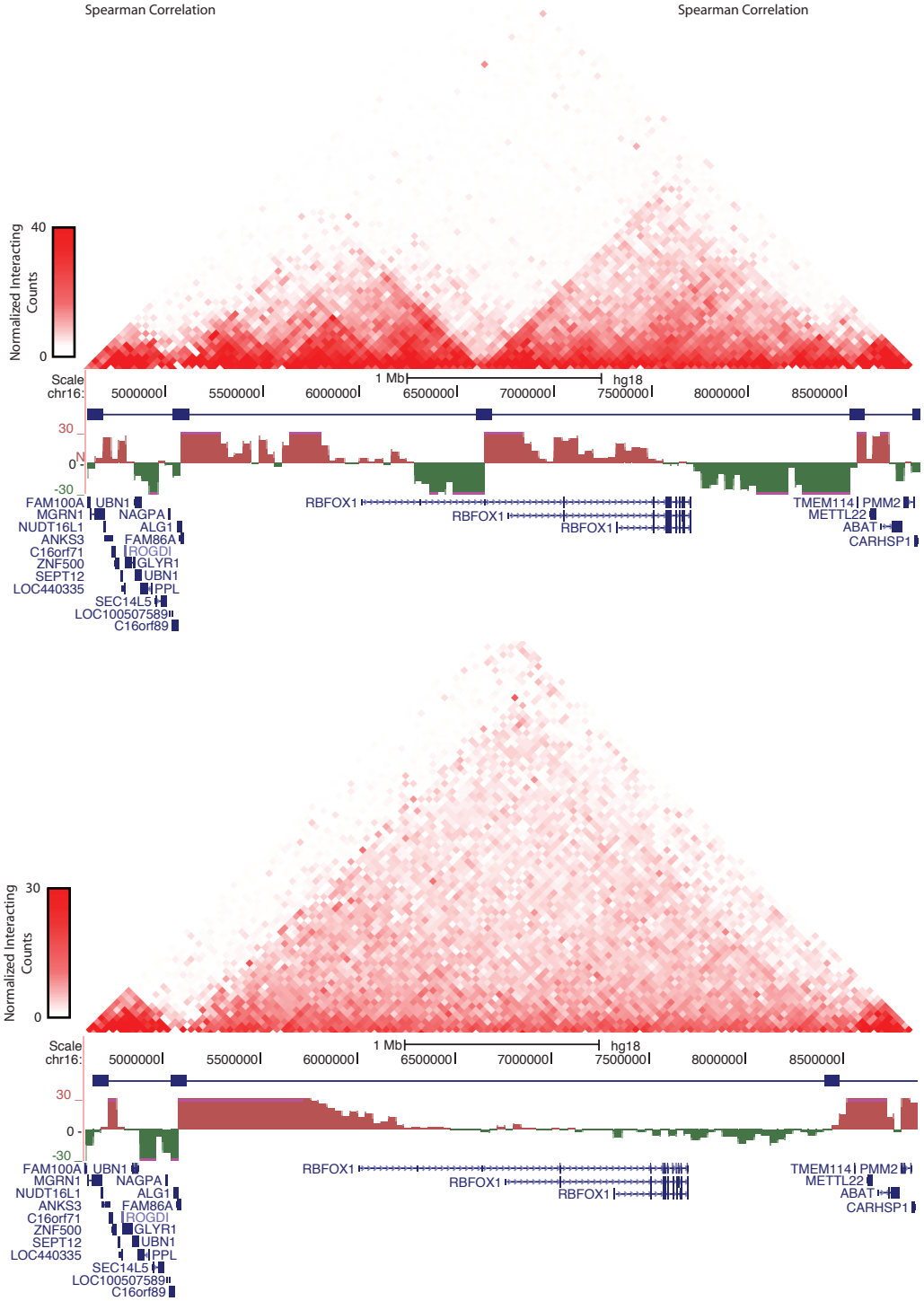

a
Invariant Boundary
Scale
chr5:
1 Mb
hg18
81000000
81500000
82000000
hESC DI
IMR90 DI
50
-50
30
-30
SSBP2
ATG10
RPS23
ATP6AP1L
MIR3977
TMEM167A
SCARNA18
b
mESC
Cortex
hESC
IMR90
Original
Replicate
Shared Boundaries
mESC Only Boundaries
Cortex Only Boundaries
hESC Only Boundaries
IMR90 Only Boundaries
-3
+3
Directionality Index
c
mESC Original vs mESC Replicate DI Correlation
hESC Original vs hESC Replicate DI Correlation
Cortex Original vs Cortex Replicate DI Correlation
IMR90 Original vs IMR90 Replicate DI Correlation
Shared Boundaries
mESC Only Boundaries
Cortex Only Boundaries
hESC Only Boundaries
IMR90 Only Boundaries
Random
Density
Spearman Correlation
-1.0
-0.5
0
+0.5
+1.0

Supplementary Figure 17. Domains are largely stable between cell types. a, Genome browser shot of an invariant boundary between hESC and IMR90 and the DI surrounding the boundary regions. b, Heat maps showing the directionality index surrounding the topological boundary regions. The heat maps are divided into three regions. Shared boundaries, boundaries called in cell type A and boundaries called in cell type B. c, Density plot of the Spearman correlations between the directionality indexes between Hi-C replicates at the topological boundary regions. Shown in blue are the shared boundaries. Shown in red is the boundaries called in ES cells (human or mouse) and shown in green are boundaries called in differentiated cells (human or mouse). Shown in grey are randomly generated spearman correlations. The replicates are all highly correlated at the boundary regions, regardless of whether the boundaries are called as shared or cell type specific.

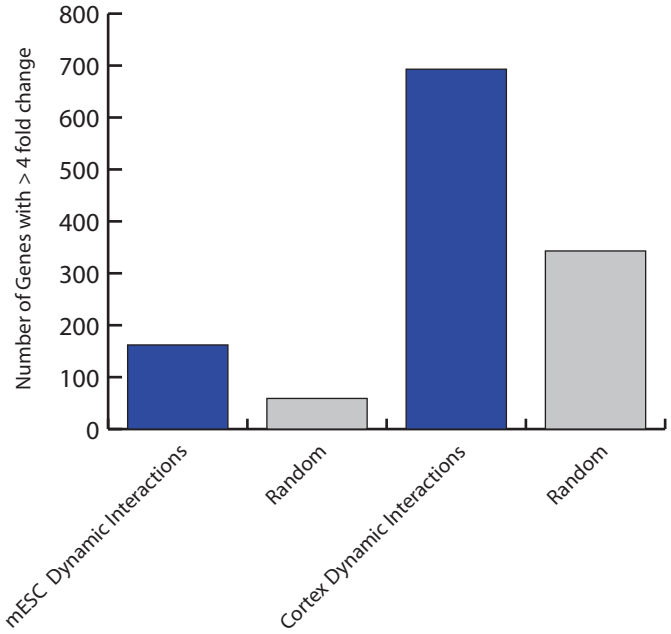
a


mESC vs Cortex Boundary
Directionality Index Correlation
Shared Boundaries
mESC Boundaries
Cortex Boundaries
Random
110 mESC Specific Boundaries
42 cortex Specific Boundaries
p < 0.05
Density
Spearman Correlation
hESC vs IMR90 Boundary
Directionality Index Correlation
Shared Boundaries
hESC Boundaries
IMR90 Boundaries
Random
237 hESC Specific Boundaries
61 IMR90 Specific Boundaries
p < 0.05
Density
Spearman Correlation

b

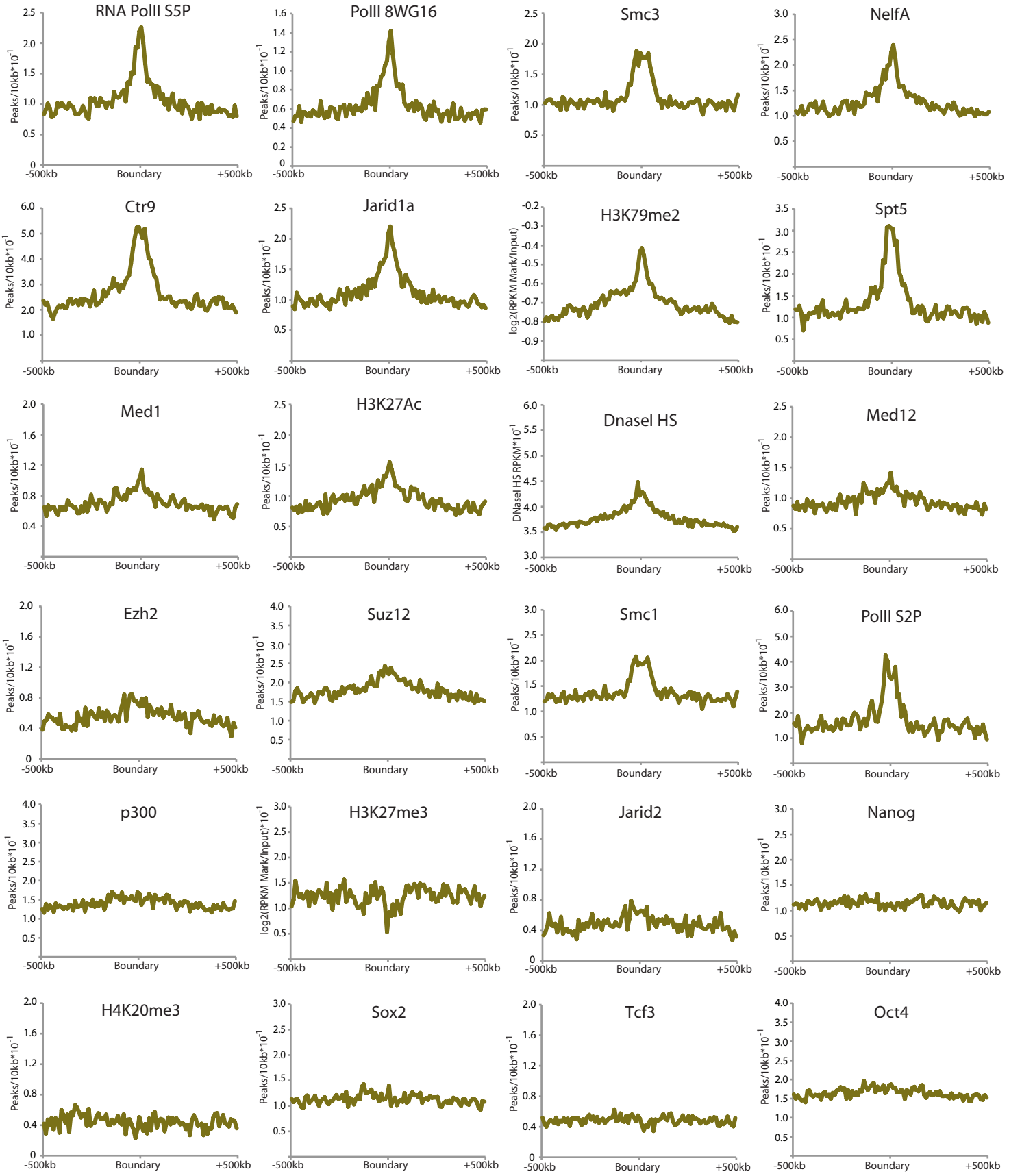

Normalized Interacting Counts
Scale
chr16:
1 Mb
hg18
5000000
5500000
6000000
6500000
7000000
7500000
8000000
8500000
FAM100A
MGRN1
NUDT16L1
ANKS3
C16orf71
ZNF500
SEPT12
LOC440335
UBN1
NAGPA
ALG1
FAM86A
ROGDI
GLYR1
UBN1
PPL
SEC14L5
LOC100507589
C16orf89
RBFOX1
TMEM114
PMM2
METTL22
ABAT
CARHSP1

Supplementary Figure 18. Cell type specific domains. a, we determined cell type specific domains between cell types by calculating the spearman correlation coefficient between the DI at each boundary called in a cell types. The DI at most boundaries is still well correlated in different cell types. We call a boundary as cell type specific if the boundary is called by HMM in only one cell type and the spearman correlation of the directionality index is not significant when compared to a random distribution of spearman correlations. A minority of boundaries are actually called as cell types specific. b, A genome browser shot of a cell type specific domain on chromosome 16. The domain is called in hESCs and is not called in IMR90.

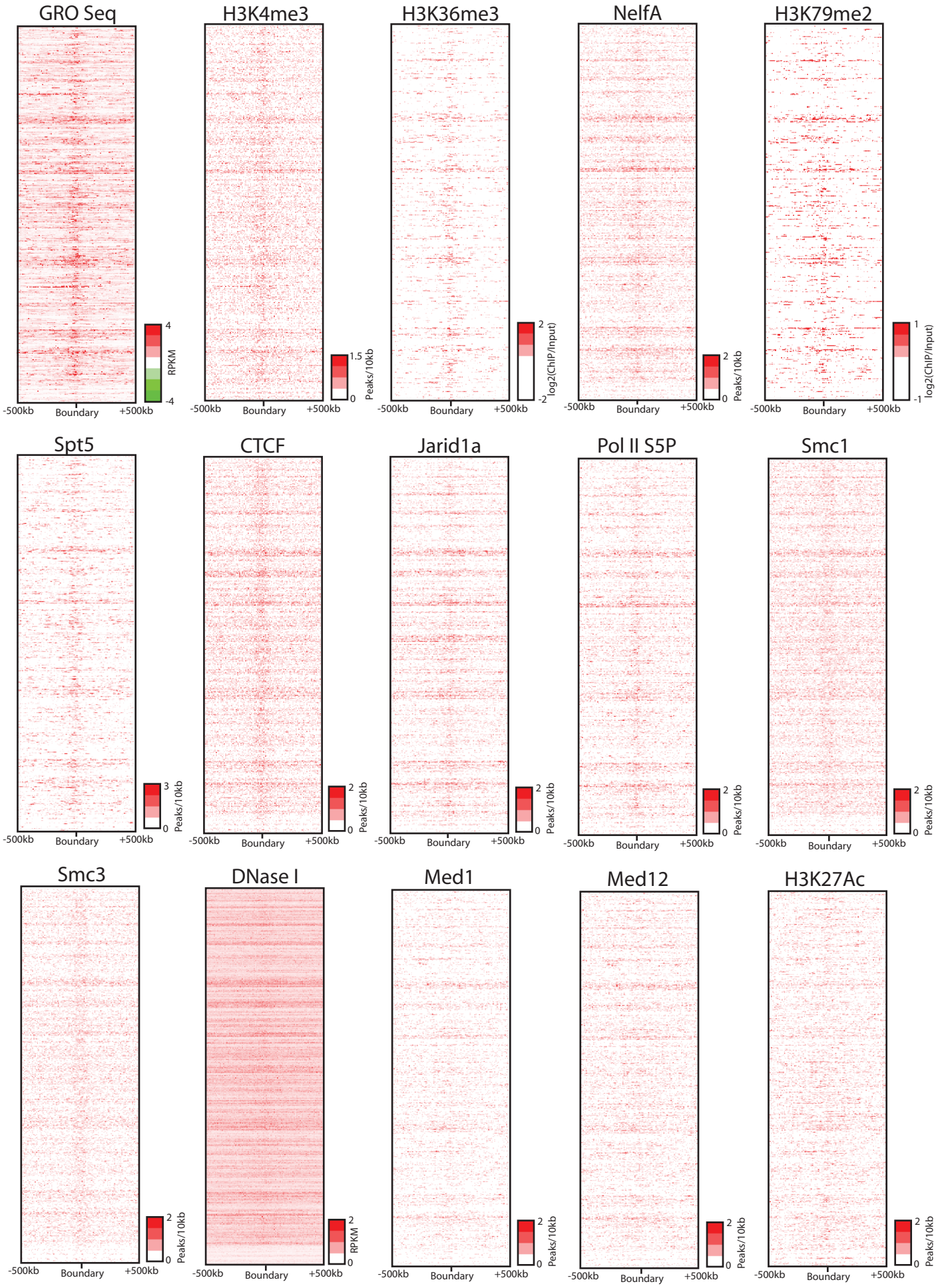


Supplementary Figure 19. Enrichment of Differentially Expressed genes at dynamic interacting regions. The number of genes with a > 4-fold change in gene expression are that are found in a dynamic interacting region in either mouse ES cell or cortex are shown. Shown in grey is the number of > 4-fold changed gene expected using randomly permuted dynamic interacting regions.

mESC Histone Modificatioons, Chromatin Binding Proteins, and Transcription Factors at Topological Boundaries



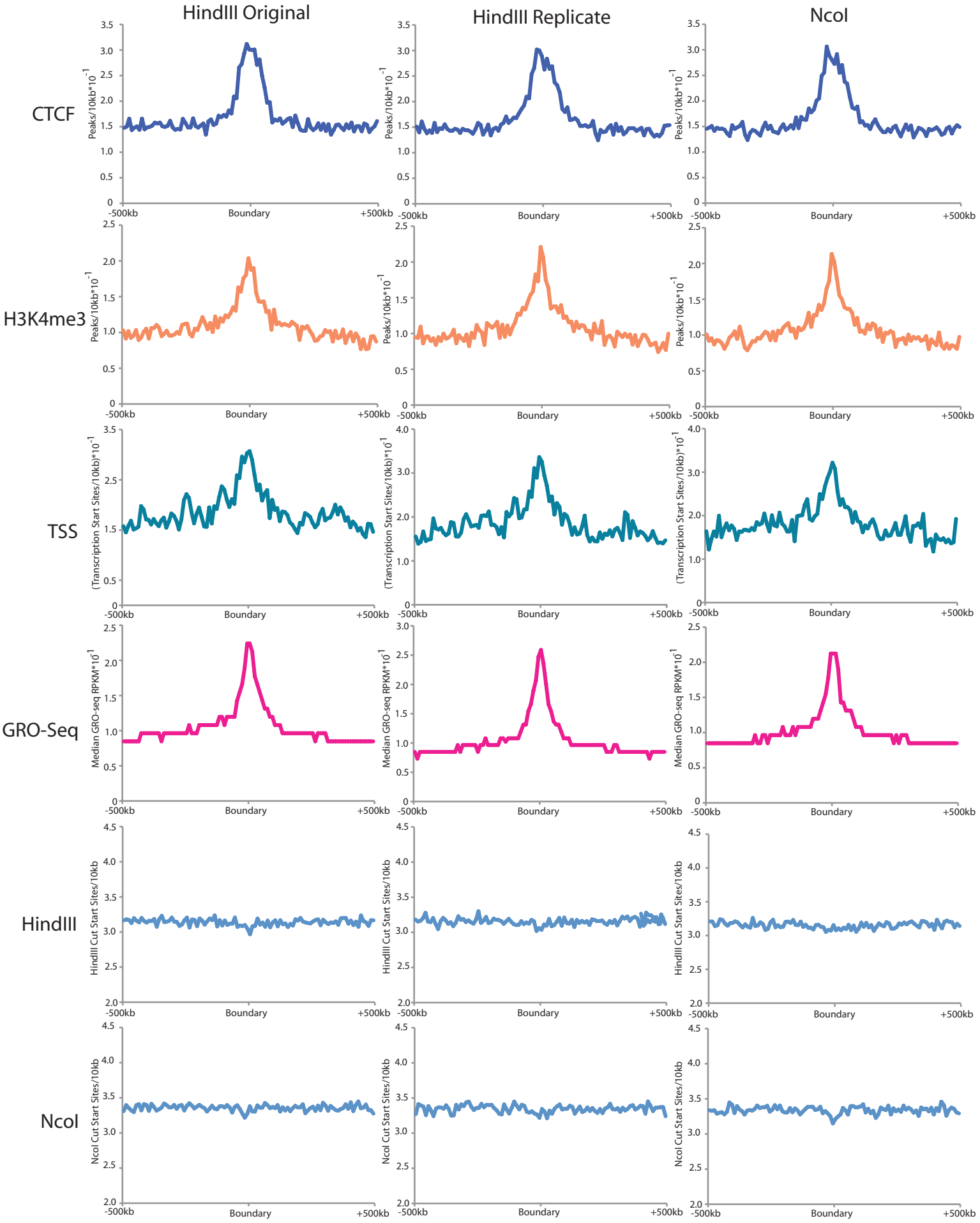
Supplementary Figure 20. Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions. Average enrichment plots for factors surrounding boundary regions called in mESC. For most marks, the signal is shown as the frequency of peaks or binding sites per 10kb. For "block like" marks, such as H3K27me3 and H3K79me2, the signal shown is the log2(ChIP/Input) over 10kb windows.


GRO Seq
H3K4me3
H3K36me3
NelfA
H3K79me2
Spt5
CTCF
Jarid1a
Pol II S5P
Smc1
Smc3
DNase I
Med1
Med12
H3K27Ac
-500kb
Boundary
+500kb
RPKM
Peaks/10kb
log2(ChIP/Input)

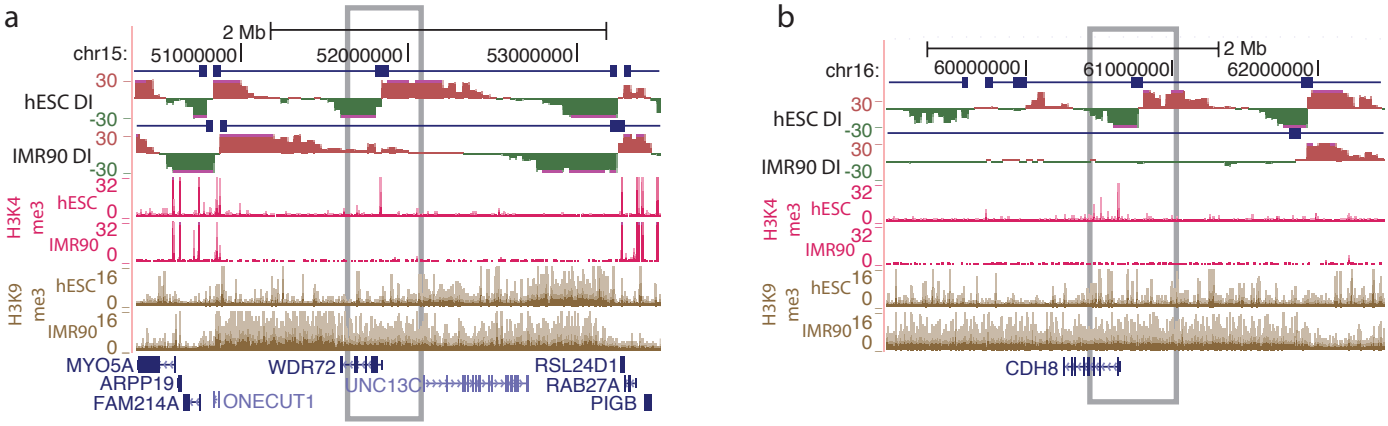
Supplementary Figure 21. Heat maps of boundary enrichment of Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions. Raw heat maps of each signal at the boundary region of a subset of marks from Supplemental Figure 20.

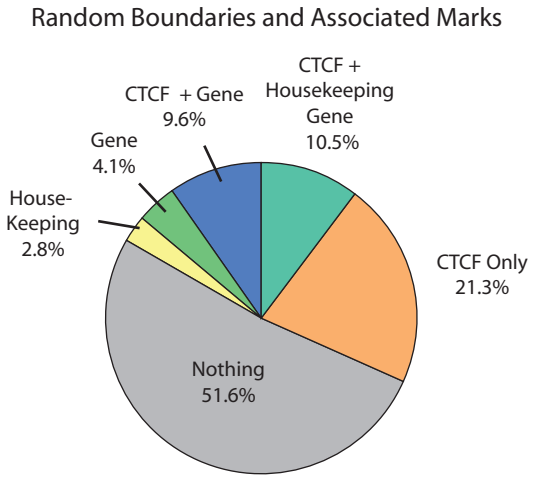
Nanog
Jarid2
H4K20me3
H3K4me1
H3K27me3
Ezh2
Ctr9
Oct4
p300
PolII 8WG16
PolII S2P
Sox2
Suz12
Tcf3
TSS
-500kb
Boundary
+500kb
Peaks/10kb
TSS/10kb
log2(ChIP/Input)

Supplementary Figure 22. Heat maps of boundary enrichment of Histone modification, chromatin binding protein, and transcription factor enrichment near boundary regions. Raw heat maps of each signal at the boundary region of the remainder of marks from Supplemental Figure 20 not shown in Supplementary Figure 21.


HindIII Original
HindIII Replicate
NcoI
CTCF
H3K4me3
TSS
GRO-Seq
HindIII
NcoI
Peaks/10kb*10^-1
(Transcription Start Sites/10kb)*10^-1
Median GRO-seq RPKM*10^-1
HindIII Cut Start Sites/10kb
NcoI Cut Start Sites/10kb
-500kb
Boundary
+500kb

Supplementary Figure 23. Marks enriched at boundaries in each mouse ES cell replicate. The enrichement plots for CTCF, H3K4me3, transcription start sites, and GRO-seq signal were calculated similarly to Supplementary Figure 20 for each of the three mouse ES cell replicates. Also calculated and plotted is the average enrichment of HindIII and NcoI cut sites at the boundary regions.

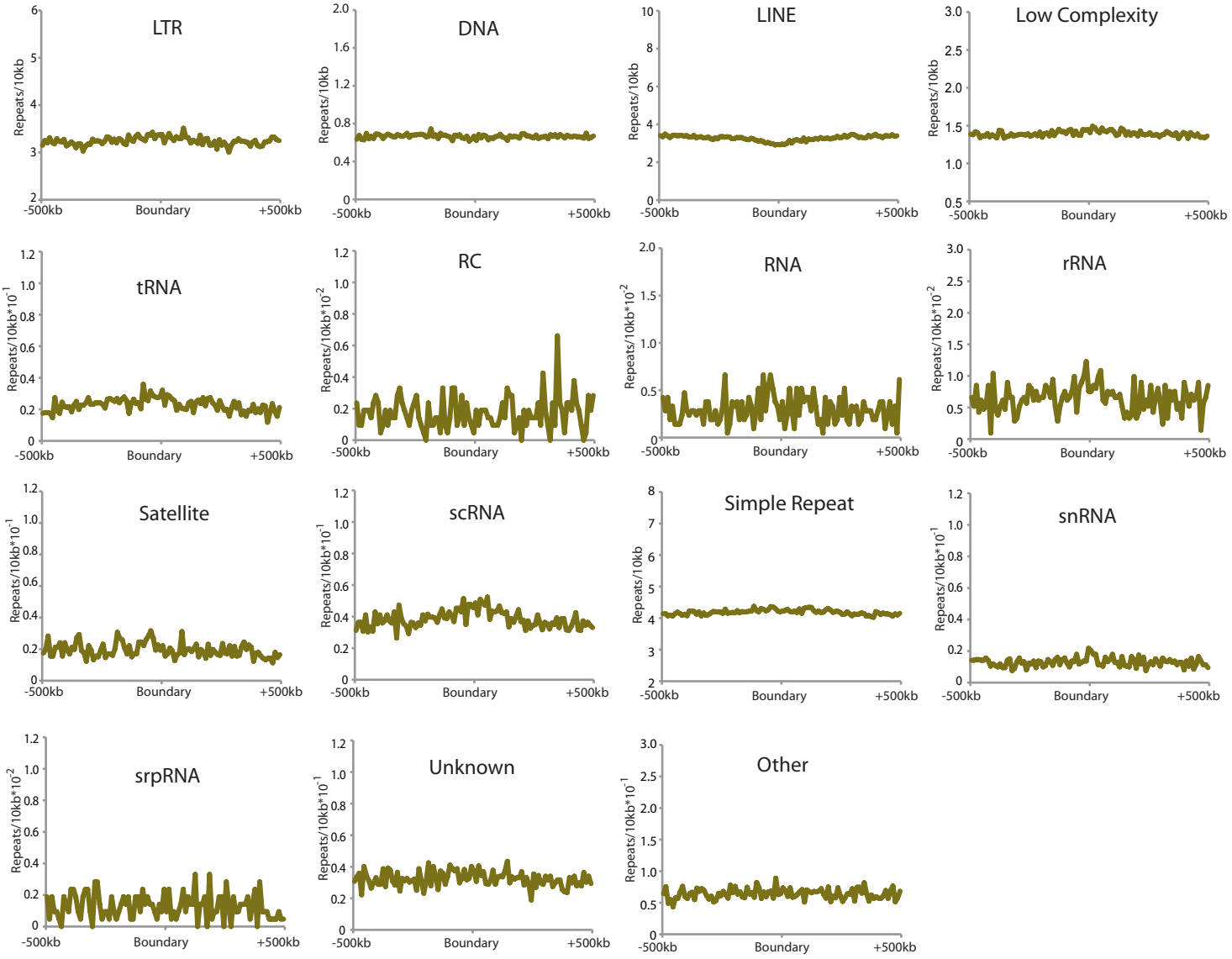




Supplementary Figure 24. Random assocation of CTCF and housekeeping genes in mESCs. a,b, Cell type specific boundaries between hESC and IMR90 that show associated changes in H3K4me3 near the boundary. c, Analogous to Figure 4e, pie chart showing the expected proportion of boundaries associated with CTCF, housekeeping genes, or other genes in mouse ES cells based on randomly generated boundaries.

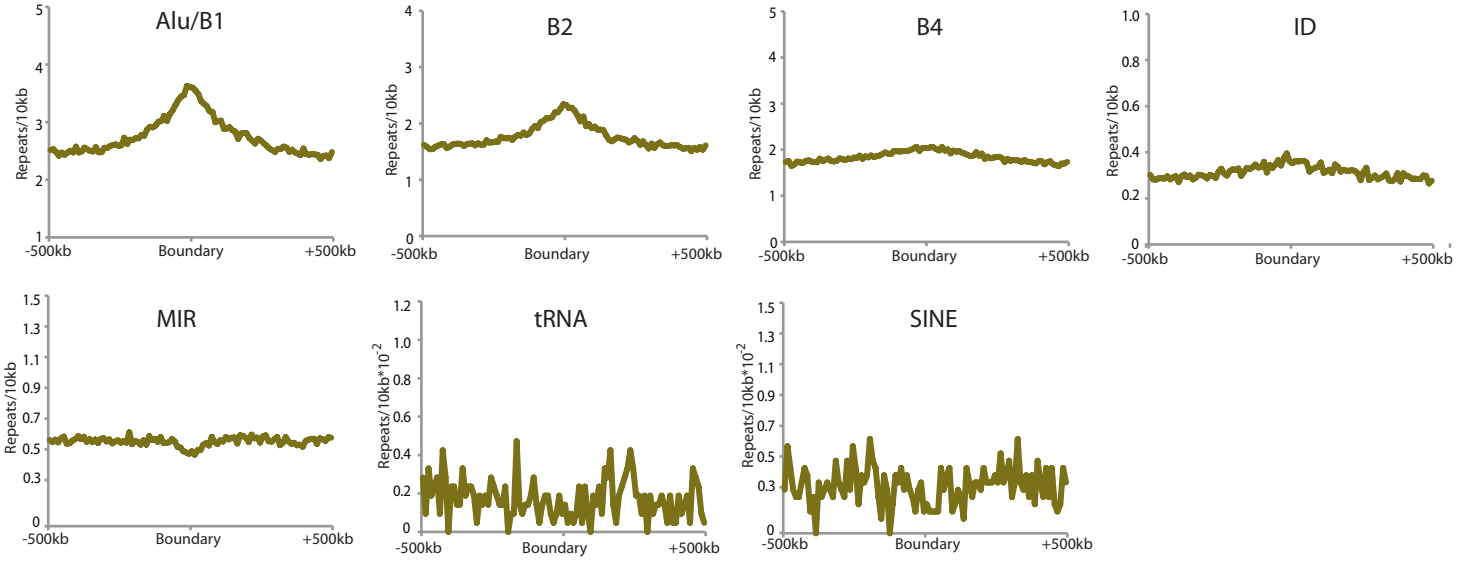
a

Repeat Masker Repeat Class Frequency at Topological Boundaries

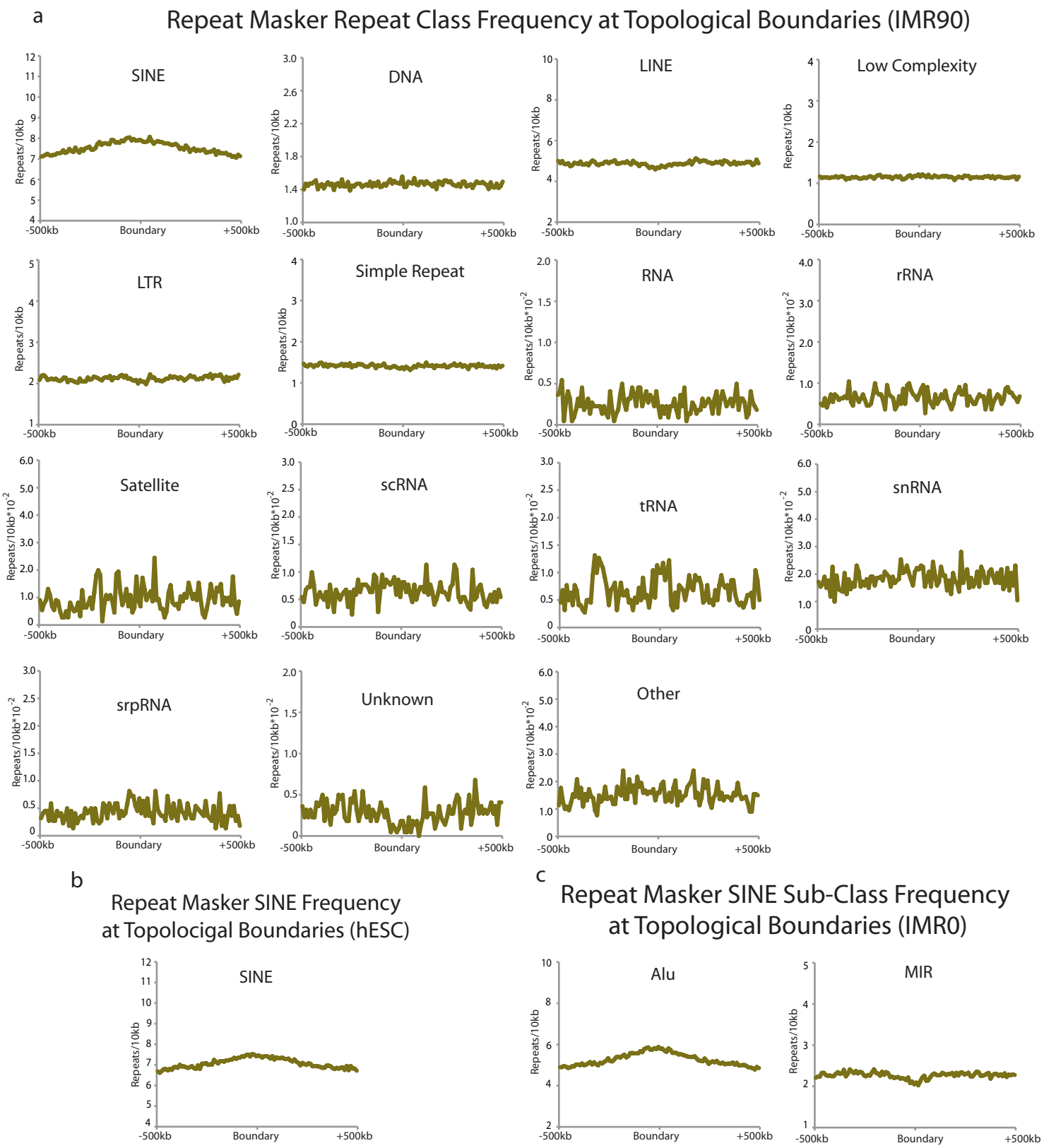


b

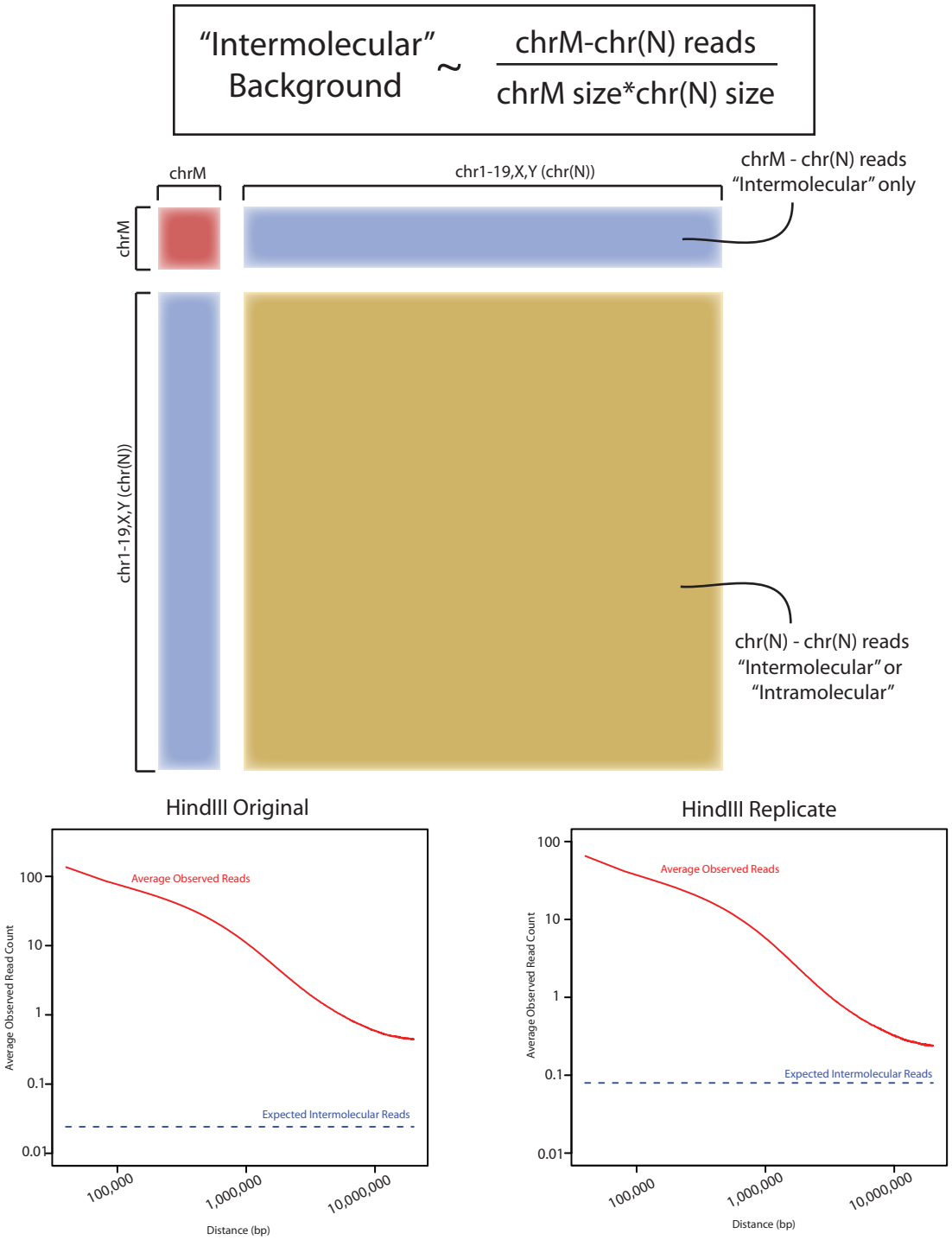
Repeat Masker SINE Sub-Class Frequency at Topological Boundaries



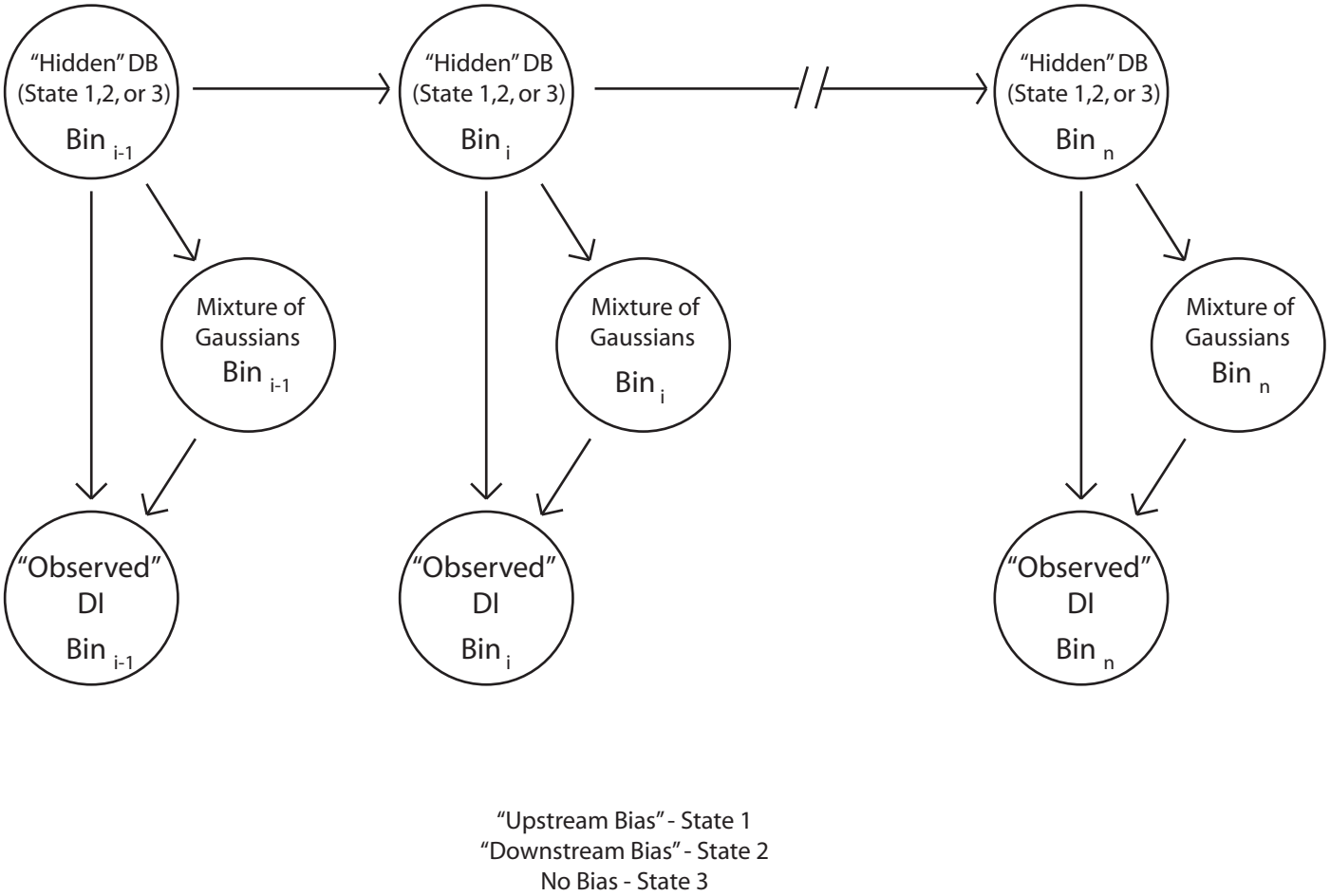
Supplementary Figure 25. Repeat Content at mouse ES cell boundaries. a, The frequency of repeats from UCSC Repeat Masker was calculated near the boundary regions. Only SINE element, shown in Figure 4a, show any enrichment at boundary regions. b, SINE subclass frequency at the topological boundary regions in mouse ES cells using UCSC Repeat Masker.



Supplementary Figure 26. Repeat Content at human boundaries. a, The enrichment of different classes of repeats at the IMR90 boundaries was calculated using the UCSC Repeat Masker data. b, Enrichment of SINE element frequency at boundaries in human ES cells. c, Enrichment of SINE element subclasses at the topological boundary regions in IMR90.



Supplementary Figure 27. Expected Intermolecular Ligations. To model the expected number of interactions between two loci in the genome due to random intermolecular ligation events, we calculated the expected number of reads per kbp^2 between the nuclear and mitochondrial chromosomes. As the nuclear and mitochondrial genomes are in different organelles, these reads can only occur due to random intermolecular ligations. We assume that the expected number of intermolecular reads between any two bins is constant, regardless of whether the two bins are nuclear or mitochondrial. Therefore, the number of intermolecular reads per bin between the nuclear and mitochondrial chromosomes should be equal to the number of intermolecular reads between any two bins both located on the nuclear chromosomes. Also shown is the number of reads at each distance (in red) for 40kb bins along the same chromosome. The number of random intermolecular reads is on average < 2% of what is actually observed for bins on the same chromosome less than 2 Mbp apart.



Supplementary Figure 28. HMM with mixture of Gaussian output. Each 40kb bin *i* along a chromosome having n bins has an observed Directionality Indexes ("Observed" DI) and a hidden Directionality Biases ("Hidden" DB, shown in the figure as states 1, 2, or 3 for simplicity). Assuming that the observed DI's are a mixture of Gaussians, we determine DB state (1, 2 or 3) at bin *i*.