Supplementary Information

Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts

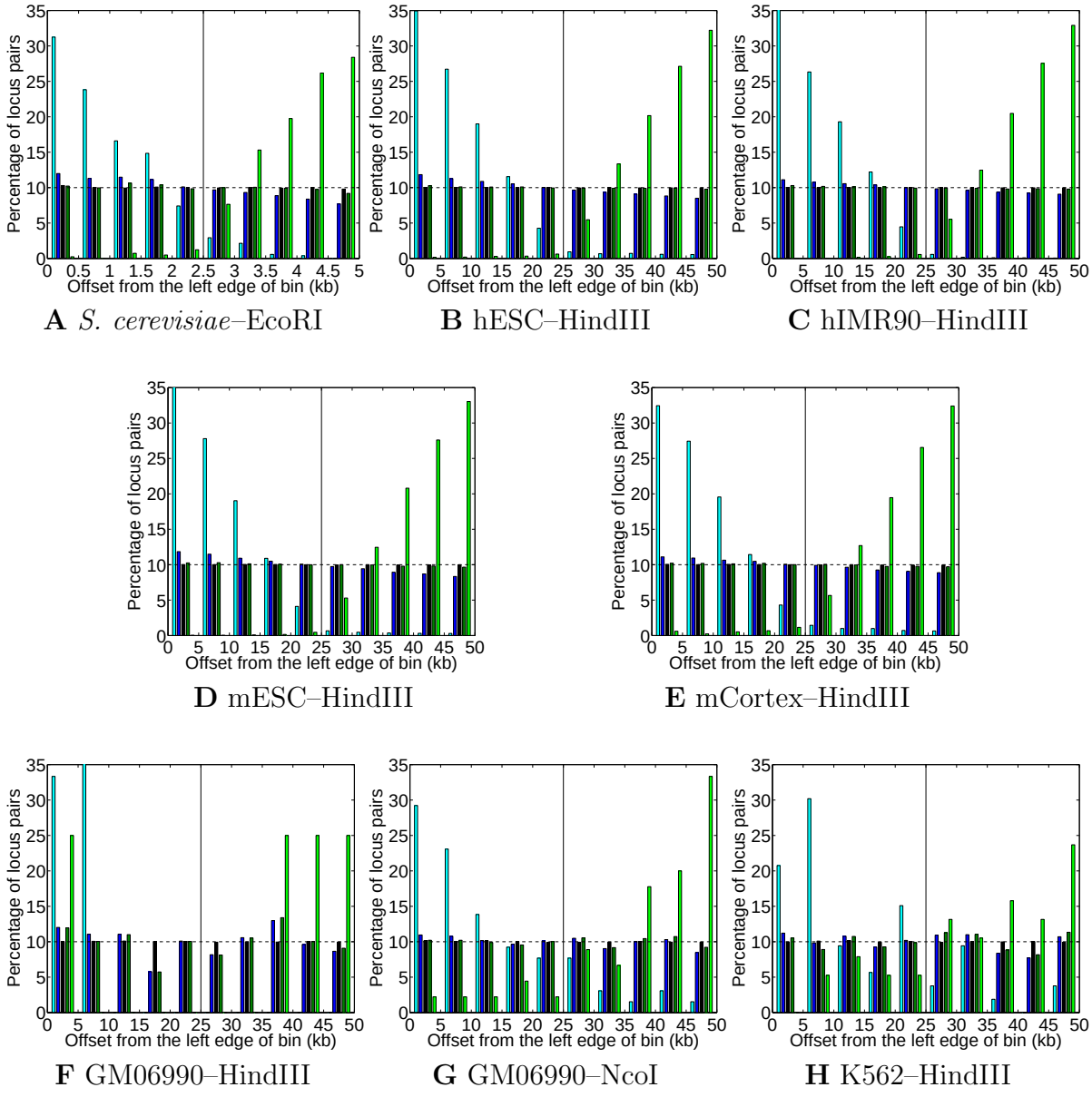
Ferhat Ay, Timothy L. Bailey, William Stafford Noble

Supplementary Figures

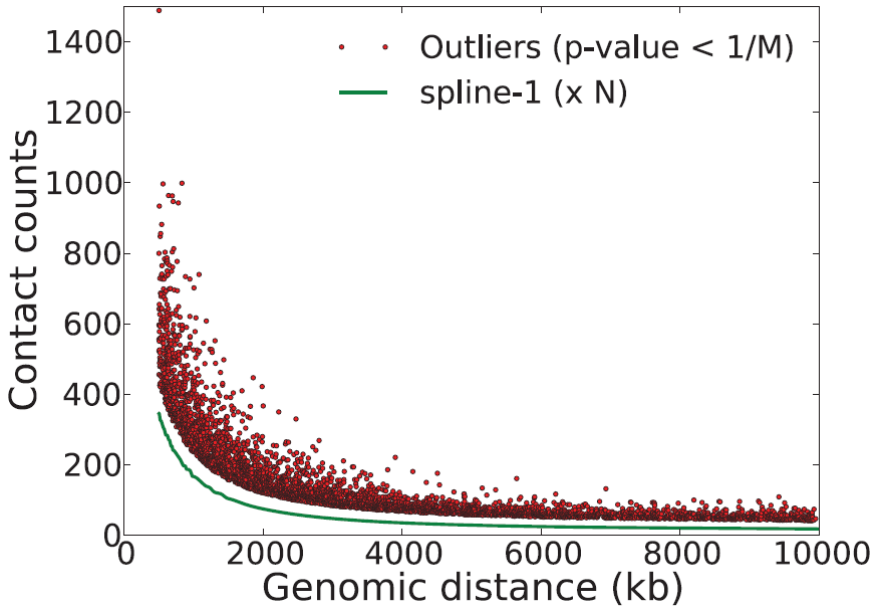
Supplementary Tables

Supplementary Notes

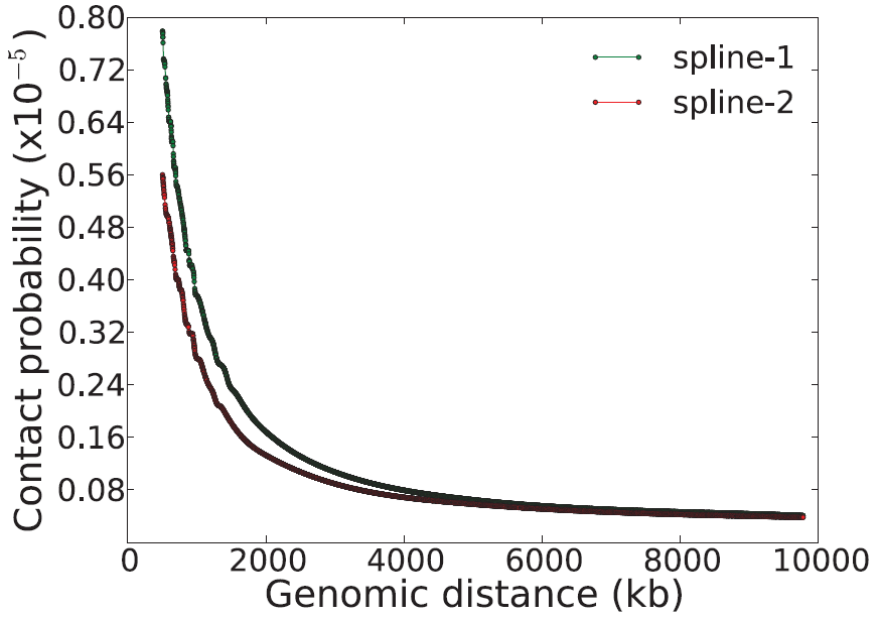
Supplementary Figures



Supplementary Figure 1: **Spline fitting corrects for binning artifacts.**
Each subfigure is similar to Figure 3B in the main text and plots the bias introduced by the binning method for each of the remaining eight cross-linked libraries we analyzed. For **(A)** we use an FDR threshold of 1% and a resolution of 1 RE fragment.[1] For **(B–E)** we use an FDR threshold of 1% and a resolution of 10 RE fragments.[2] For **(F–H)** we use an FDR threshold of 5% and a resolution of 50 RE fragments.[3] Each figure is generated using contact maps corrected by the ICE method[4] (Methods). Figures generated using raw contact maps (i.e., before ICE) are very similar for each library (data not shown).

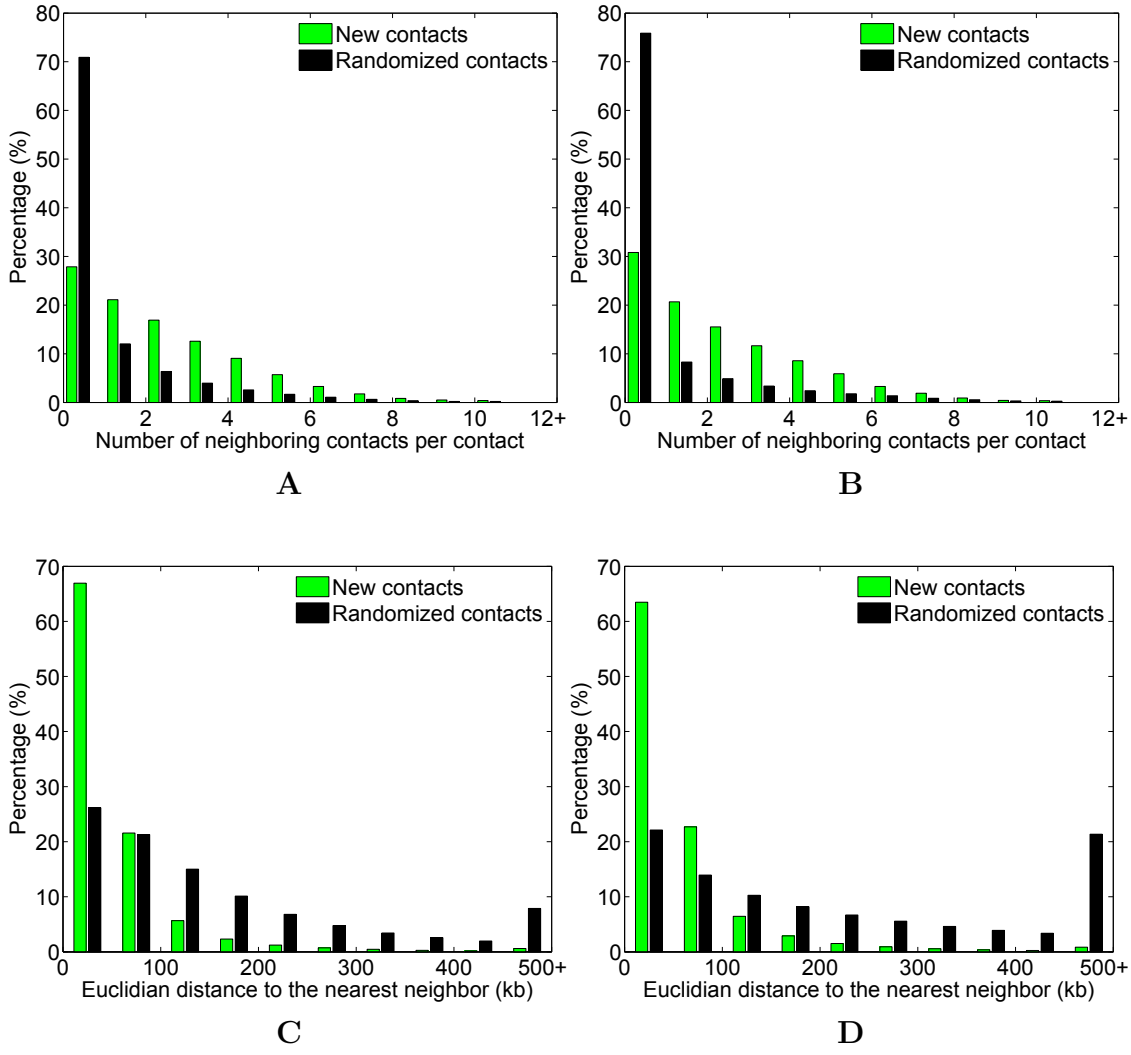


A

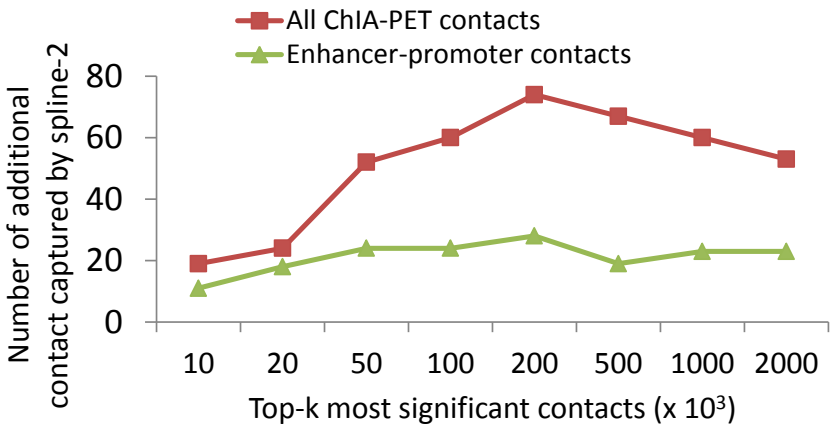


B

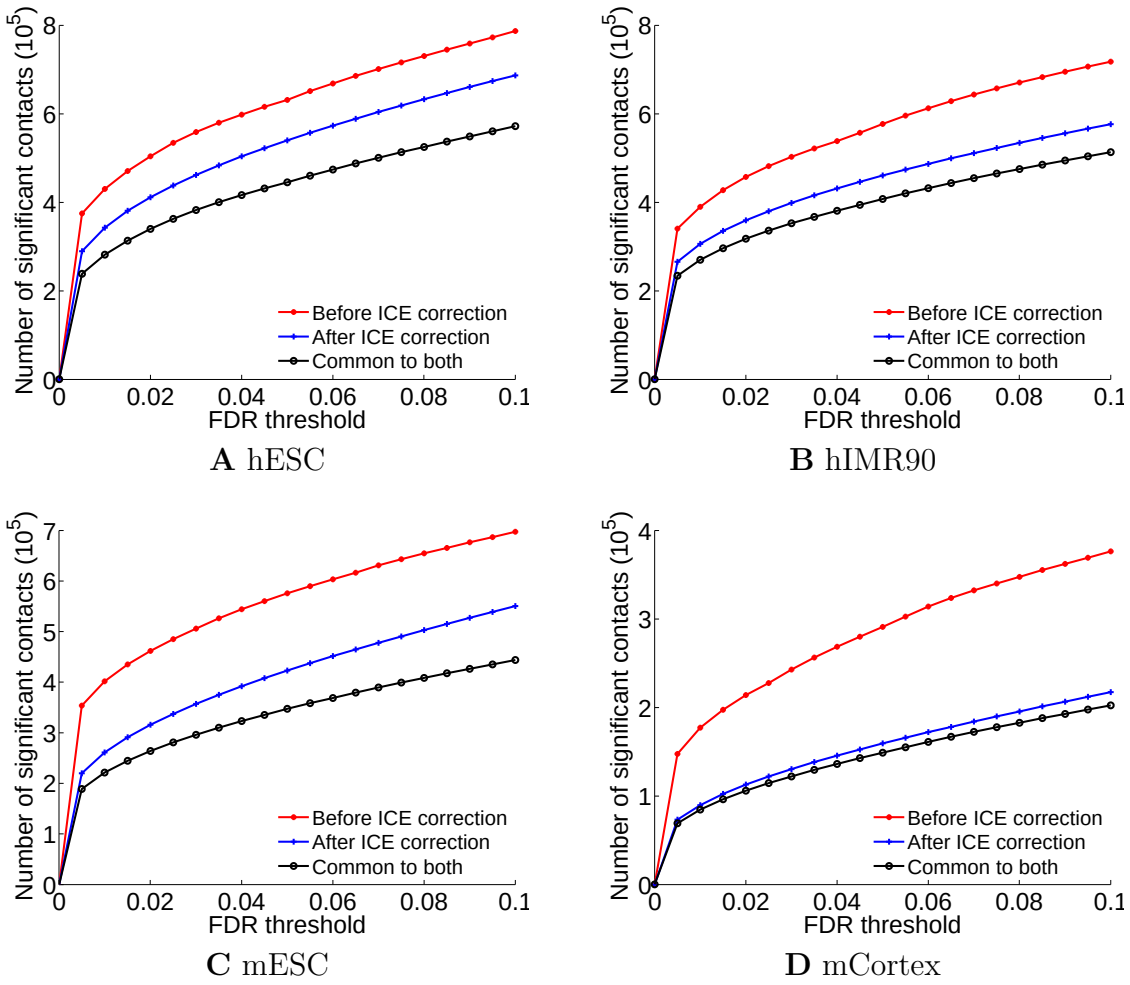
Supplementary Figure 2: **Refinement of the null model by outlier removal.**
(A) Initial spline fit to raw contact map at a resolution of 50 RE fragments for human fibroblast cell line (hIMR90) from Dixon et al.[2] Locus pairs with significantly higher contact counts compared to expectation are marked as outliers. We multiplied the contact probability curve of spline-1 with the total number of all mid-range reads (i.e., N) to obtain the green line that denotes the expected contact counts. **(B)** Spline fits before (spline-1) and after (spline-2) removing the outliers from the null (i.e., refinement) for the same library.



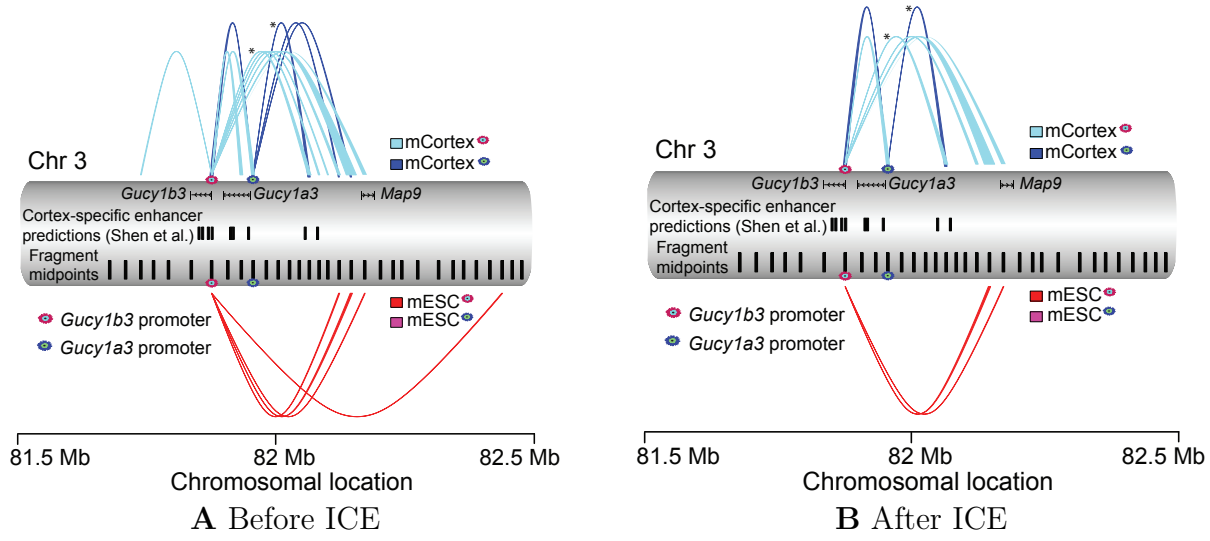
Supplementary Figure 3: **The set of "new contacts" are spatially consistent with the set of "old contacts."**
(A–B) The distribution of the number of neighboring old contacts that lie within a square that has 100 kb-long edges and is centered on the two-dimensional coordinate of a new contact at FDR 1% for **(A)** hESC and **(B)** hIMR90 Hi-C data sets. The two-dimensional coordinate for each contact is defined by the genomic coordinates of the two loci that are linked by that contact. **(C–D)** The distribution of the distance from a new contact to its nearest old contact at FDR 1% for **(C)** hESC and **(D)** hIMR90 Hi-C data sets. We define distance between a pair of contacts as the Euclidean distance between the two-dimensional coordinates defined by these contacts. To obtain the distribution of neighbor counts and minimum distances for randomized contacts we sample a random contact set that is 100 times larger than the set of new contacts (see Supplementary Note 1). Each figure is generated using contact maps at a resolution of 10 RE fragments and corrected by the ICE method[4] (Methods). Figures generated using raw contact maps (i.e., before ICE) are very similar for each library (data not shown).



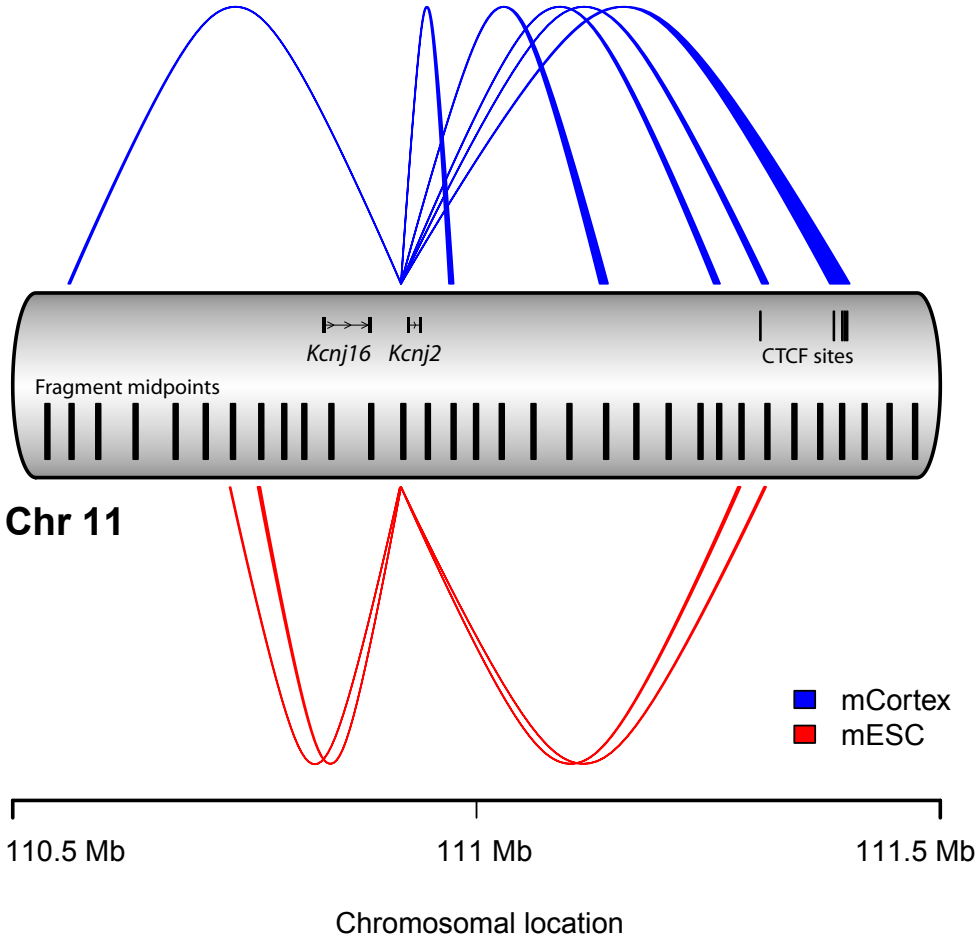
Supplementary Figure 4: Fit-Hi-C captures more of ChIA-PET contacts compared to discrete binning.
Number of additional RNAPII mediated ChIA-PET contacts within genomic distance range of (50 kb, 5 Mb] captured by Fit-Hi-C compared to discrete binning method of Duan et al.[1] when applied to mESC Hi-C data. In order to control for the number of predictions made by each method we only focus on the contact confidence rankings and consider k most significant contacts for k ranging from 10,000 to 2,000,000. We put an upper bound on k to avoid including contacts with p-values equal to 1 which would lead to random ties. Red and green tracks denote number of additional contacts identified by our method at each top-k among all ChIA-PET contacts and only contacts between enhancer-promoter pairs.



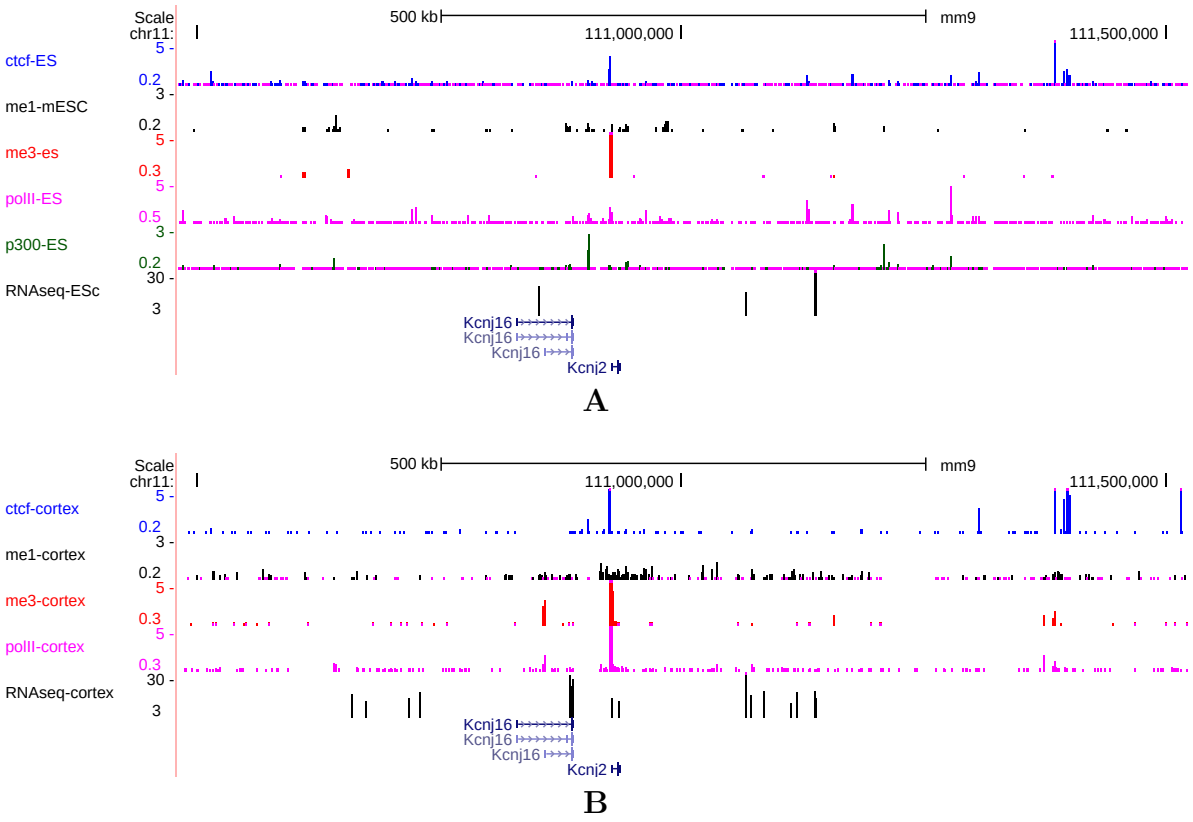
Supplementary Figure 5: **Number of significant contacts for raw and corrected contact maps.** Comparison of the number of contacts deemed significant using raw maps (before ICE correction), corrected contact maps (after ICE correction) and in both settings (common to both) for four cell lines from Dixon et al.[2] (Methods).



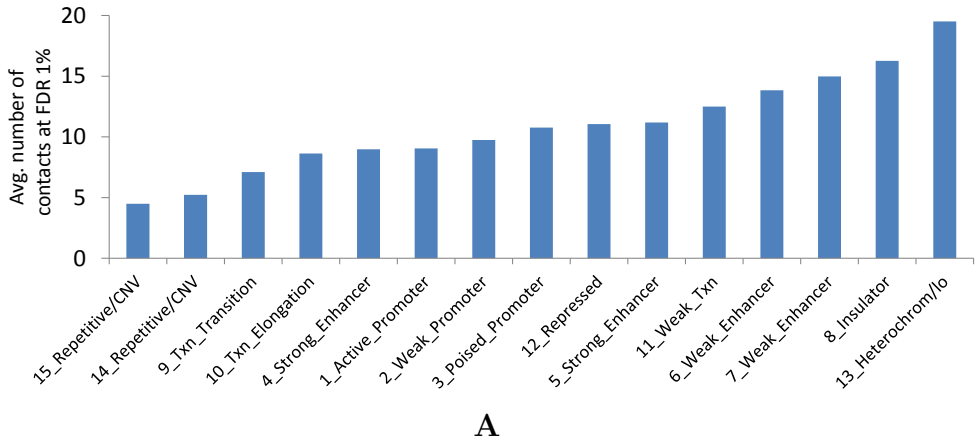
Supplementary Figure 6: Comparison of contact profiles for a locus of interest before and after integrating ICE biases.
(A) This figure is generated identically to Figure 4D in the main text but before integrating biases in confidence estimation. **(B)** This figure is identical to Figure 4D and only included here for ease of side by side comparison.



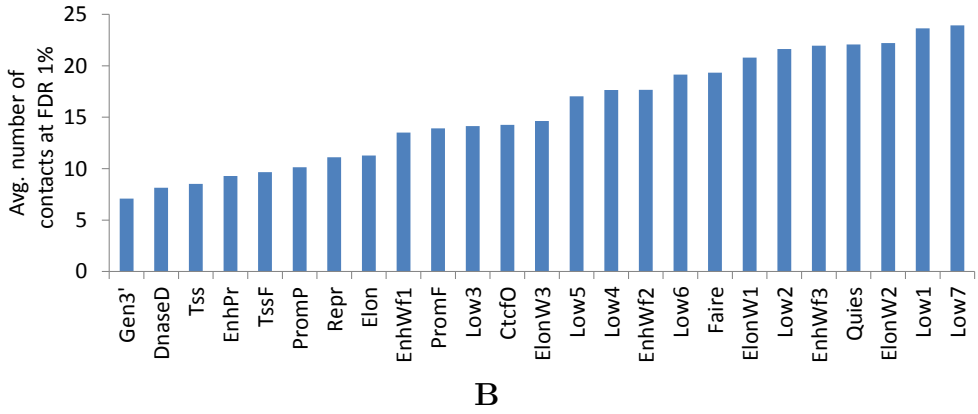
Supplementary Figure 7: **Cell line-specific contacts between gene promoters and distal CTCF sites correlate with cell line-specific expression in mouse.**
Contact profile of the locus that contains promoters of two inward rectifier potassium channel genes (*Kcnj16* and *Kcnj2*). Each connector represents a significant contact at FDR 1% for contact maps after ICE correction with thickness proportional to the minus log(p-value) of the contact. Distal CTCF sites that are around 470 kb away from the locus of interest and midpoints of each 10 consecutive restriction fragments are shown as two separate tracks. *Kcnj16* and *Kcnj2* are important for establishing the resting membrane potential of several cell types in the cortex including neurons and astrocytes[5] and presumably are less important in the functioning of stem cells compared to cortex cells where they are expressed.



Supplementary Figure 8: **Cell line-specific contacts between gene promoters and distal CTCF sites correlate with cell line-specific expression in mouse (continued).**
Measured gene expression (RNA-seq), protein binding (RNA polymerase II (polII), CTCF, p300) and histone modification (H3K4me1 (me1) and H3K4me3 (me3)) levels for **(A)** mESC and **(B)** mCortex cells[6] for the locus that contains promoters of *Kcnj16* and *Kcnj2* genes.

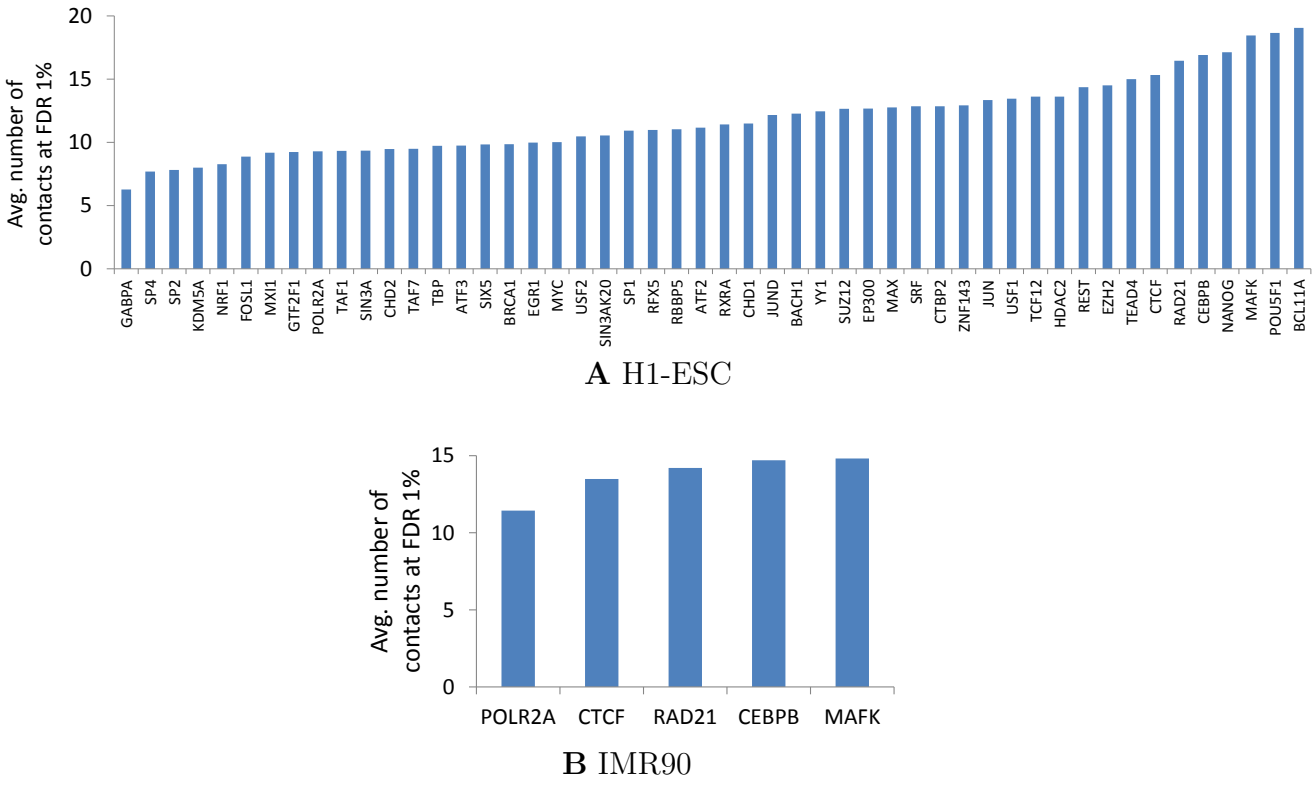


A

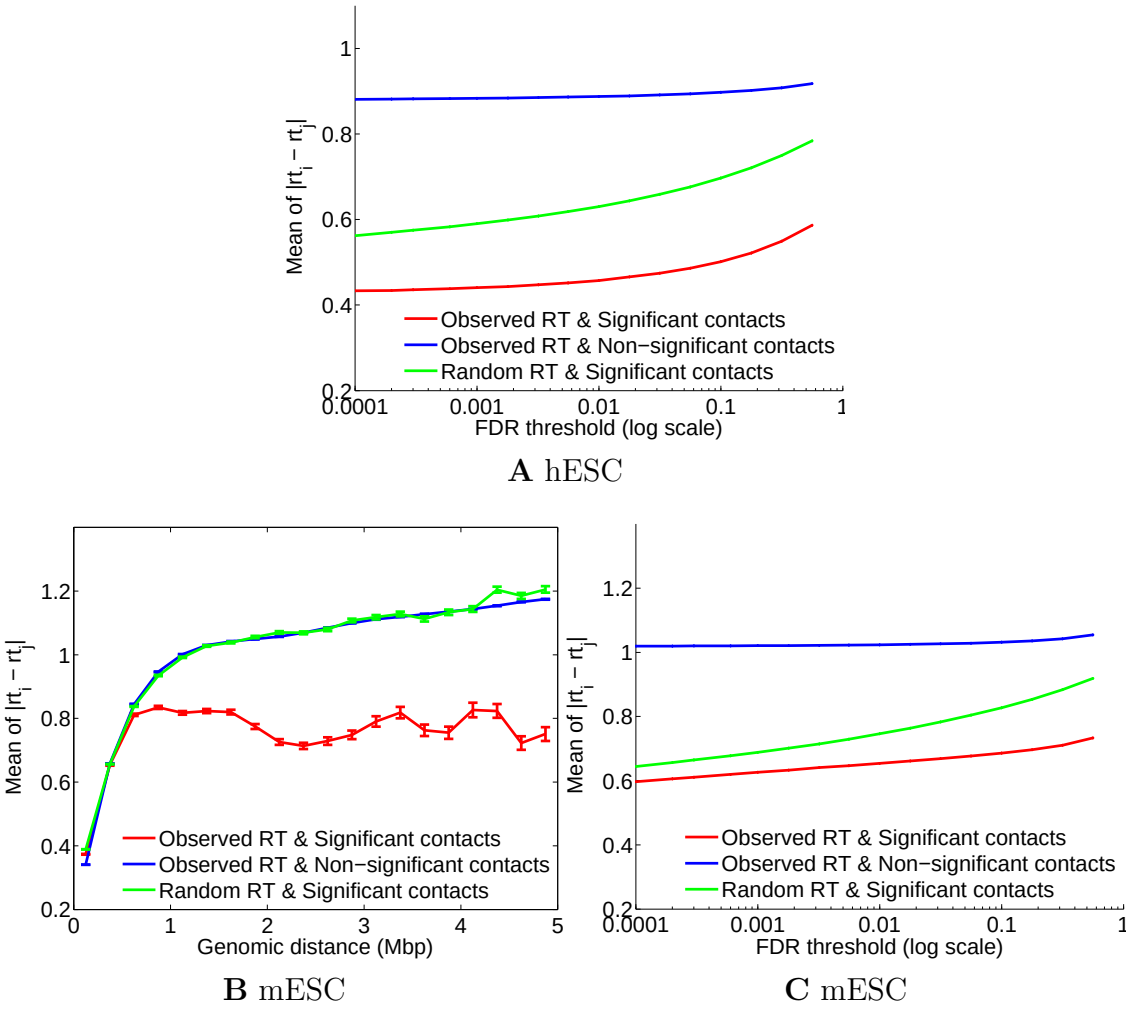


B

Supplementary Figure 9: **Number of high-confidence contacts per locus largely varies with the functional annotation of the locus.**
These figures are generated identically to Figures 5A and 5B in the main text but using the whole set of **(A)** 15-labels from ChromHMM and **(B)** 25-labels from Segway. For more information about semantics of each label and how they are assigned see Ernst et al.[7] and Hoffman et al.[8]



Supplementary Figure 10: **Regions with binding peaks of pluripotency factors engage in a large number of high-confidence contacts in human ECSs**
Average number of high-confidence (FDR 1%) contacts identified by Fit-Hi-C for each set of transcription factor (TF) binding peaks for two human cell lines **(A)** H1-ESC and **(B)** IMR90. Contact confidences are assigned similar to done in Figure 5 in the main text. We map each TF binding peak to the window used in Fit-Hi-C analysis with which it has the most overlap. We use binding profiles of 50 TFs for H1-ESC and 5 TFs for IMR90 cell lines which were made available by ENCODE consortium.[9]

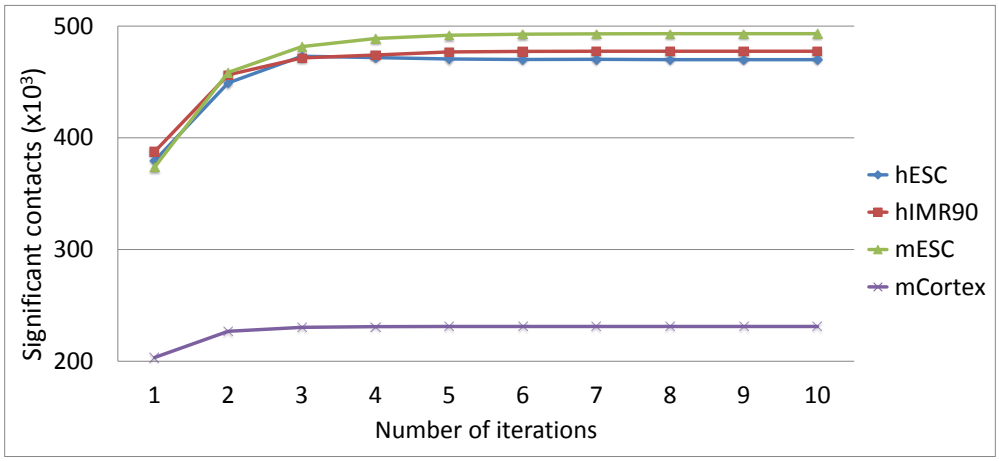


A hESC

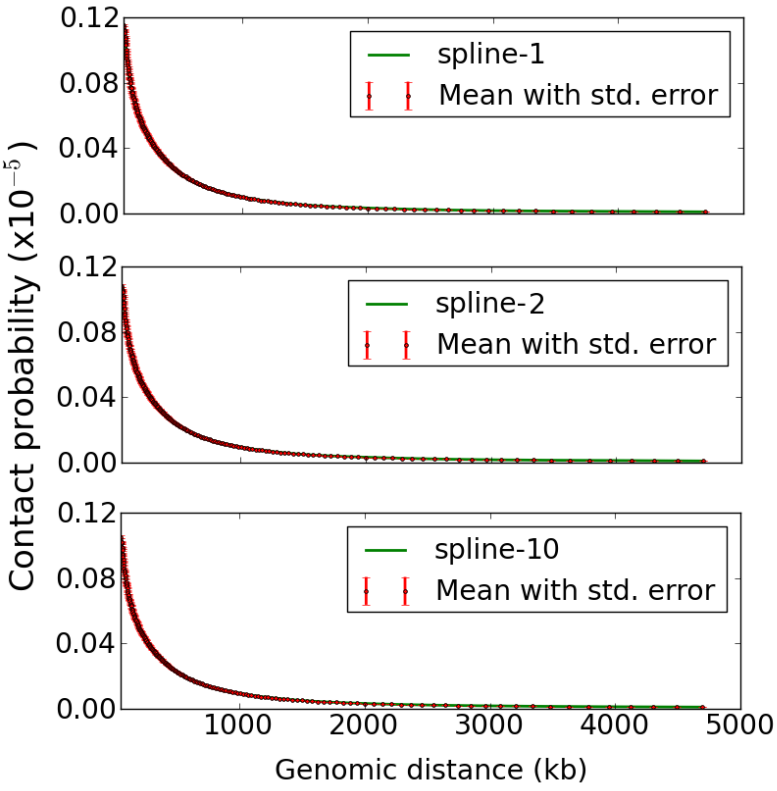
B mESC

C mESC

Supplementary Figure 11: **Confidence of a contact between two loci correlates with their replication timing similarity.**
(A) Mean of the replication timing differences for hESC data for each of the three groups mentioned in Figure 6C as a function of FDR threshold used to determine significant contacts. For each group, we aggregate all mid-range contacts with different genomic distances to obtain one mean at each FDR threshold. The standard errors are omitted because they are not visible at this scale. **(B)** This plot is generated identically to Figure 6C in the main text but for the mESC cell line. **(C)** This plot is generated identically to **(A)** but for the mESC cell line.

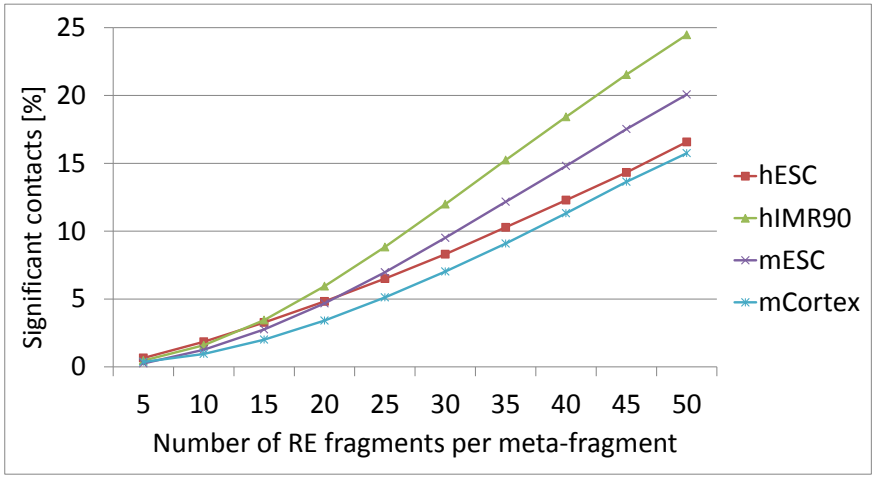


A

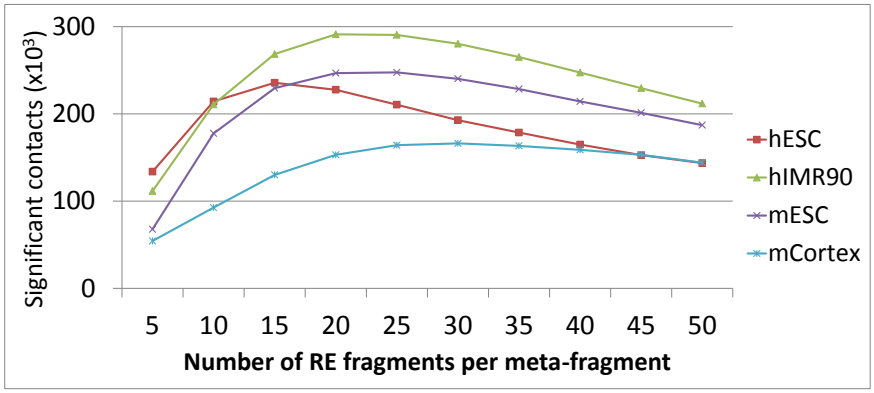


B

Supplementary Figure 12: **Refinement of the null model converges in a few iterations.** **(A)** The change in the number of significant contacts at a resolution of 10 RE fragments with iterative application of the refinement step for each cell line from Dixon et al.[2] The first iteration, marked by "1" on the x-axis, represents the number of contacts at FDR 1% prior to any refinement of the null. **(B)** The change in the scaling of contact probability with genomic distance for the hESC data set when no refinement (spline-1), only one step of refinement (spline-2) and nine steps of refinement (spline-10) are applied to the null.



A



B

Supplementary Figure 13: **Effects of meta-fragment size and sequencing depth on the confidence estimates.**
(A) The percentage and **(B)** the number of mid-range contacts that are deemed significant at FDR 1% among all possible mid-range locus pairs with increasing number of restriction fragments per meta-fragment (i.e., decreasing resolution). A genomic distance range of 500kb–10Mb is used for all meta-fragment sizes and cell types.

Supplementary Tables

Supplementary Table 1: **Summary of the genome architecture data sets analyzed.**

E, **H** and **N** denote the restriction enzymes EcoRI, HindIII and NcoI, respectively. Each restriction enzyme fragment (RE frag.) is approximately 4 kb. The column abbreviated **Repl.** reports the number of replicates available for each library. For the human and mouse data sets, the number of locus pairs and informative reads are reported at a resolution of 10 consecutive RE fragments for genomic distance range of (50 kb, 5 Mb] and 50 RE consecutive fragments for genomic distance range of (500 kb, 10 Mb]. Rows marked with gDNA denote non-crosslinked control libraries generated using genomic DNA.

Org.	Cells	RE	Reads	Repl.	Mid-range locus pairs		Mid-range reads		Ref.
					resolution= 1 RE frag.		range=(10kb 250kb]		
S.cer	-	**E**	295.6 M	4	**E**: 321,624		2,210,827		[1]
	-	**H**	300.5 M	4			5,982,463		
	(gDNA)	**E**	39.8 M	2	**H**: 323,804		14,052		
	(gDNA)	**H**	38.6 M	2			22,787		
					10 RE frags.	50 RE frags.	(50kb 5Mb]	(500kb 10Mb]	
Human	GM	**H**	30.0 M	2	**H**: 11,717,060	**H**: 884,833	1,173,890	892,907	[3]
	GM	**N**	28.7 M	1			1,947,872	1,347,494	
	K562	**H**	36.8 M	1	**N**: 9,191,770	**N**: 684,841	1,574,930	1,044,882	
	H1-ESC	**H**	237.7 M	2			83,181,877	43,987,085	[2]
	IMR90	**H**	397.2 M	2			61,981,200	44,359,093	
					10 RE frags.	50 RE frags.	(50kb 5Mb]	(500kb 10Mb]	
Mouse	Cortex	**H**	401.3 M	2	**H**: 12,570,818	**H**: 952,049	29,219,725	22,753,301	[2]
	ESC	**H**	465.5 M	2			96,801,016	55,553,751	

Supplementary Table 2: **Increase in the transitivity of contact graph with the addition of method-specific contacts.**

Each row corresponds to a Hi-C library for each of the four cell lines assayed (replicates combined) by Dixon et al.[2] The contact graph is defined by the common set of contacts deemed significant at FDR 5% by the *binning* and *spline-1* methods using ICE corrected contact maps. A node in the contact graph corresponds to a meta-fragment of 10 consecutive RE fragments that participates in at least one significant contact, and an edge (undirected) between two nodes corresponds to a significant contact between the two corresponding meta-fragments. The number of nodes and edges of a contact graph are denoted by $|V|$ and $|E|$, respectively. The columns **Trials**, **Successes** and **Proportion** denote, respectively, the number of method-specific contacts, the number of method-specific contacts that introduce at least one triangle when added to the contact graph and the proportion of the latter group in the former. The last column reports the z-score for the null hypothesis that the success proportion for *binning* and *spline-1* are equal. A negative z-score suggests a bigger proportion of success (i.e., increase in transitivity of the common contact graph) for *spline-1* compared to *binning*.

Library	**Contact graph size**		**Trials**		**Successes**		**Proportion**		**z-score**
	$\|V\|$	$\|E\|$	*binning*	*spline-1*	*binning*	*spline-1*	*binning*	*spline-1*	
hESC	74,525	457,558	23,010	17,803	17,525	13,839	0.762	0.777	-3.73
hIMR90	72,411	390,454	9,077	8,592	6,858	6,557	0.756	0.763	-1.18
mESC	70,888	347,926	17,106	13,771	11,867	9,810	0.694	0.712	-3.56
mCortex	57,216	143,789	4,295	3,782	2,189	1,949	0.510	0.515	-0.51

Supplementary Table 3: **Improvement in the number of significant contacts at various FDR thresholds due to refinement of the null.**

Restriction enzyme (RE) abbreviations are as described in Table 1. Rows marked with gDNA denote non-crosslinked control libraries generated using genomic DNA. The number of significant contacts at two different FDRs are reported for the discrete binning method (*binning*)[1] and our spline fitting method before (*spline-1*) and after (*spline-2*) refining the null model. The column abbreviated as *Imprv.* reports the percent increase in the number of significant contacts reported by spline-2 compared to binning. The number of significant contacts were reported for contact maps corrected by the ICE method[4] (Methods). Improvements gathered from raw contact maps (i.e., before ICE) are very similar for each library (data not shown).

Org.	**Cells**	**RE**	**FDR 1%**				**FDR 5%**			
			binning	*spline−1*	*spline−2*	*Imprv.*	*binning*	*spline−1*	*spline−2*	*Imprv.*
S.cer	-	**E**	5,709	5,571	7,556	32.4 %	8,397	8,290	10,849	29.2 %
	-	**H**	19,774	19,763	28,807	45.7 %	26,718	26,747	37,045	38.7 %
	(gDNA)	**E**	0	0	0	-	0	0	0	-
	(gDNA)	**H**	19	16	17	-10.5 %	21	21	23	9.5 %
Human	GM	**H**	68	64	74	8.8 %	208	209	212	1.9 %
	GM	**N**	875	869	924	5.6 %	2,135	2,115	2,209	3.5 %
	K562	**H**	499	506	532	6.6 %	1,373	1,358	1,415	3.1 %
	H1-ESC	**H**	106,084	105,843	148,013	39.5 %	144,291	144,772	195,004	35.1 %
	IMR90	**H**	135,304	134,973	192,416	42.2 %	167,347	166,978	233,252	39.4 %
Mouse	Cortex	**H**	77,683	76,457	105,080	35.3 %	103,311	102,542	137,852	33.4 %
	ESC	**H**	148,902	149,100	209,992	41.0 %	195,837	196,738	268,218	37.0 %

Supplementary Notes

Supplementary Note 1: Consistency and complementarity of new contacts

Each proximal contact connects two loci/fragments/meta-fragments on the same chromosome. Each locus is represented by its midpoints and can be viewed as a point on the intra-chromosomal contact map of the corresponding chromosome. To test spatial consistency of new contacts (specific to spline-2) with old-contacts (all contacts identified by binning) we compute the number of old contacts that are "close to" each new contact in two-dimensional space. Let $\langle \mathbf{c}, \mathbf{x}, \mathbf{y} \rangle$ denote a new contact and $\langle \bar{\mathbf{c}}, \bar{\mathbf{x}}, \bar{\mathbf{y}} \rangle$ denote an old contact where $\mathbf{c}$ and $\bar{\mathbf{c}}$ correspond to the chromosomes that the contacts lie on, and $\mathbf{x}$, $\mathbf{y}$ and $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ correspond to the coordinates of the pairs of loci that contact each other. Also, let d be a distance threshold to deem two contacts "close". If $\mathbf{c} = \bar{\mathbf{c}}$, $0 < |\mathbf{x} - \bar{\mathbf{x}}| < d/2$ and $0 < |\mathbf{y} - \bar{\mathbf{y}}| < d/2$, then we consider the two contacts to be neighbors and spatially consistent with each other. Geometrically, this corresponds to drawing a square with edge length d and centered on the new contact then checking whether the old contact falls within that square. We choose d as 100 kb, and we use an FDR threshold of 1% to determine the set of new and old contacts. We simply count the number of neighbors for each new contact and then plot the histogram of these counts for the observed new contacts (green bars in Supplementary Fig. 3A–B). To test whether the number of neighbors for the observed new contacts is above what would be expected by chance, we repeat the same process with a set of "random new contacts" generated by a null model. We design a stringent null model which preserves the distribution of genomic distances between the two loci participating in a contact (i.e., $|\mathbf{x} - \mathbf{y}|$ for $\langle \mathbf{c}, \mathbf{x}, \mathbf{y} \rangle$). For generating each random new contact, we first select a locus at random and then select a genomic distance from the shuffled list of genomic distances between all possible proximal contacts. We generate the set of random new contacts by sampling without replacement a set that is 100 times larger than the set of observed new contacts. We compute the histogram of the number of neighbors for this random set (black bars in Supplementary Fig. 3A–B) and compare it with the histogram for the observed set of new contacts.

We also compute the two-dimensional distance from each new contact to the nearest old contact for the observed and random sets of new contacts (green and black bars in Supplementary Fig. 3C–D, respectively). For two contacts that are on the same chromosome (i.e., $\mathbf{c} = \bar{\mathbf{c}}$), we define the distance as the Euclidean distance between the two-dimensional coordinates of these contacts ($\sqrt{|\mathbf{x} - \bar{\mathbf{x}}|^2 + |\mathbf{y} - \bar{\mathbf{y}}|^2}$). We determine the nearest neighbor according to this distance measure for each new contact and plot the histogram of these distances for both the observed and random sets of new contacts. The random set of new contacts is generated as described above.

To test whether new contacts are complementary to the set of contacts common to the binning and spline-2 methods, we first represent the common set as an undirected graph with nodes representing meta-fragments and edges representing significant contacts at a given FDR threshold. We then check, for each method-specific contact, whether it completes at least one triangle (i.e., satisfies transitivity of colocalization among three loci) when added to this initial contact graph. It is crucial to note that simply adding contacts into the initial contact graph while counting the number of successes would lead to a liberal bias favoring the method with more method-specific contacts. We avoid this bias by keeping the initial contact graph constant while computing the number and proportion of success (i.e., completing at least one triangle) for either new contacts or binning-specific contacts. Once we calculate these numbers, we test whether the proportion of success is equal for the spline-2 and binning methods. More specifically, we test the hypothesis that the success proportion for binning is higher than for spline-2. To compute a significance score for this hypothesis we use the equal proportions test for large samples[10] which gives us a z-statistic as follows. Let $\hat{p_1}$, $\hat{p_2}$ denote the success proportions and n_1, n_2 denote the number of trials for the binning and spline-2 methods, respectively. Then the proportion of successes for the combined sample is $\hat{p} = \frac{n_1\hat{p_1} + n_2\hat{p_2}}{n_1 + n_2}$. We compute the z-scores reported in Supplementary Table 2 as

$$\mathrm{Z} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}.$$

References

[1] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465:363–367, 2010.

[2] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[3] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[4] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9:999–1003, 2012.

[5] J. Benesova, V. Rusnakova, P. Honsa, H. Pivonkova an D. Dzamba, M. Kubista, and M. Anderova. Distinct expression/function of potassium and chloride channels contributes to the diverse volume regulation in cortical astrocytes of GFAP/EGFP mice. *PLoS ONE*, 7(1):e29725, 2012.

[6] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488:116–120, 2012.

[7] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9(3):215–216, 2012.

[8] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*, 9(5):473–476, 2012.

[9] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[10] R. G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med.*, 17(8):873–890, 1998.