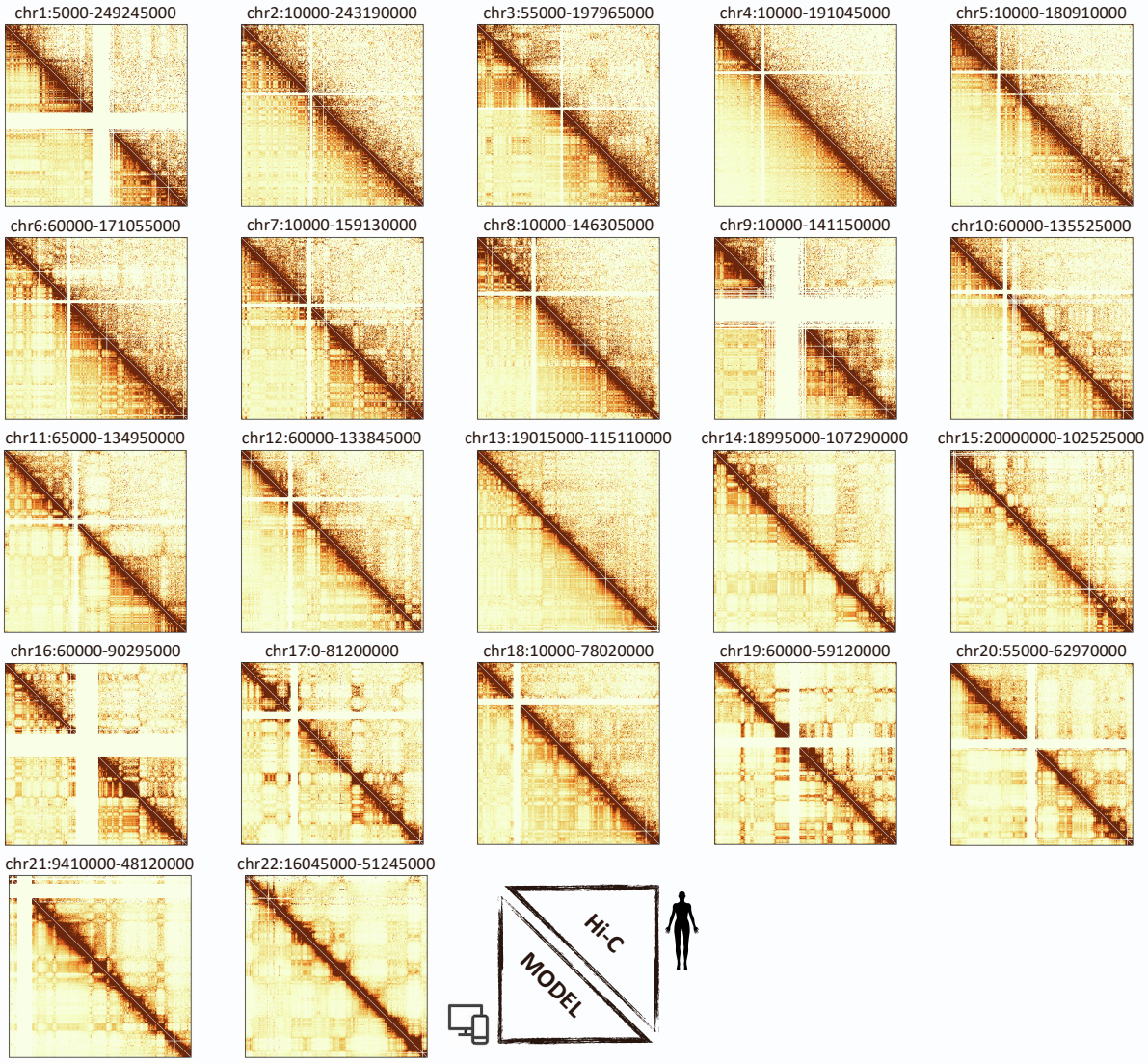
Cell Reports, Volume *38*

Supplemental information

Polymer physics reveals a combinatorial code linking 3D chromatin architecture to 1D chromatin states

Andrea Esposito, Simona Bianco, Andrea M. Chiariello, Alex Abraham, Luca Fiorillo, Mattia Conte, Raffaele Campanile, and Mario Nicodemi

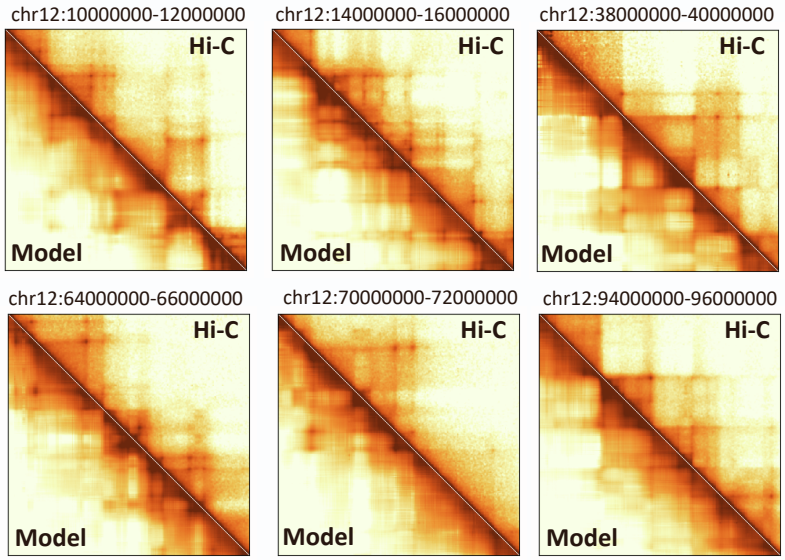
A



B

Chr	r	r′	SCC
1	0.92	0.71	0.82
2	0.92	0.67	0.84
3	0.95	0.75	0.86
4	0.92	0.75	0.86
5	0.92	0.68	0.85
6	0.95	0.76	0.86
7	0.93	0.67	0.84
8	0.93	0.71	0.82
9	0.85	0.63	0.77
10	0.91	0.64	0.80
11	0.95	0.75	0.83
12	0.94	0.80	0.87
13	0.97	0.80	0.90
14	0.96	0.78	0.90
15	0.90	0.57	0.81
16	0.95	0.82	0.86
17	0.95	0.72	0.89
18	0.98	0.86	0.90
19	0.97	0.85	0.94
20	0.97	0.85	0.92
21	0.97	0.91	0.92
22	0.95	0.70	0.88

C



D

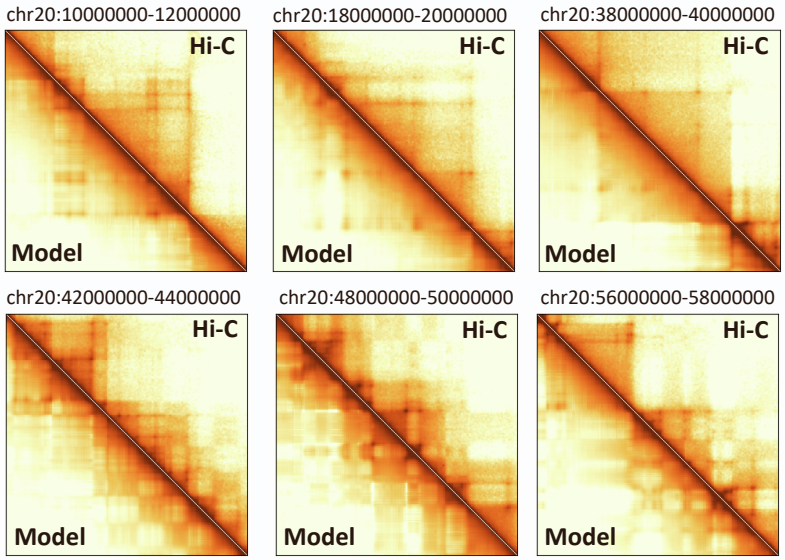
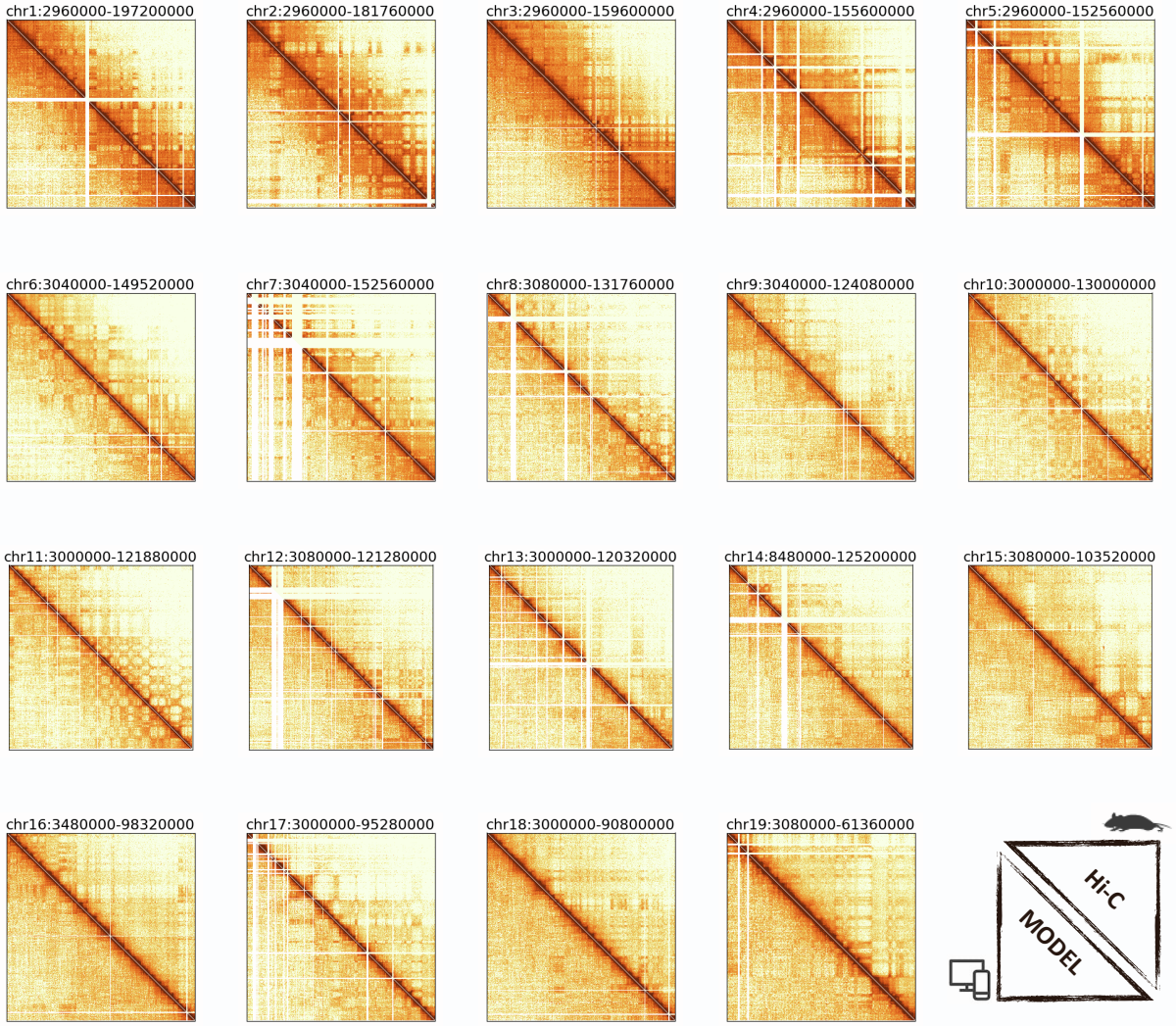


Figure S1 (Related to Figure 1). Comparison of model inferred and in situ Hi-C contact matrices in human GM12878.

(A) Contact maps (scales as in Figure 1) across chromosomes from the PRISMR inferred SBS model (lower triangle) and from Rao et al., 2014 in situ Hi-C data in GM12878 (upper triangle). **(B)** Pearson (r), distance-corrected Pearson (r') and stratum adjusted (SCC) correlation coefficients between model and in situ Hi-C data. SCC values were computed using HiCRep (Yang et al. 2017). **(C)-(D)** Comparison between Hi-C (upper triangle) and PRISMR (lower triangle) contact matrices of several 2Mb-wide genomic regions along chromosome 12 and chromosome 20, respectively.

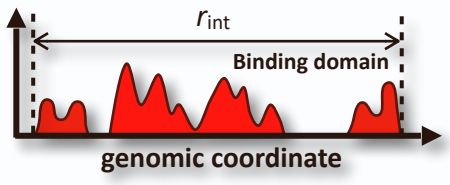
A

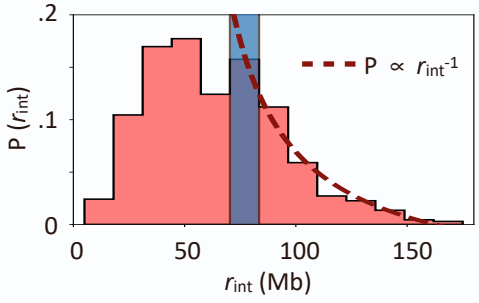


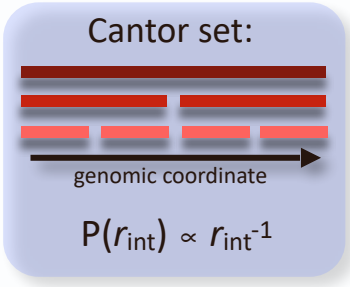
B

Chr.	r	r'	SCC
1	0.94	0.55	0.85
2	0.94	0.59	0.86
3	0.94	0.59	0.84
4	0.95	0.64	0.89
5	0.94	0.62	0.87
6	0.98	0.73	0.89
7	0.96	0.68	0.88
8	0.94	0.55	0.81
9	0.94	0.55	0.78
10	0.94	0.52	0.77
11	0.95	0.60	0.83
12	0.94	0.58	0.81
13	0.94	0.54	0.77
14	0.95	0.56	0.79
15	0.95	0.58	0.77
16	0.94	0.55	0.76
17	0.95	0.61	0.82
18	0.95	0.58	0.77
19	0.96	0.69	0.81

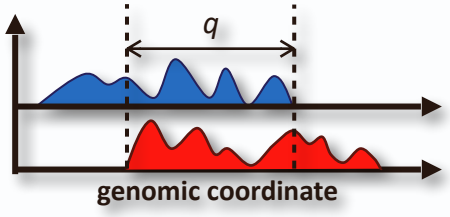
C







D



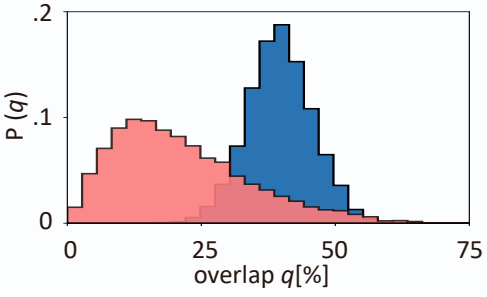


Figure S2 (Related to Figure 1). Comparison of model inferred and Hi-C contact matrices in mouse ESC and characterization of the binding domains arrangement along chromosomes.

(A) Contact maps across chromosomes from the PRISMR inferred SBS model (lower triangle) and from Dixon et al. 2012 Hi-C data in mESC (upper triangle). **(B)** Pearson (r), distance-corrected Pearson (r') and stratum adjusted (SCC) correlation coefficients between model and Hi-C data. SCC values were computed using HiCRep (Yang et al. 2017). **(C)** Distribution of the range of interaction of the PRISMR inferred SBS binding domains genome wide for human GM12878. The blue bar corresponds to a random model where the binding sites are bootstrapped. A Cantor set has hierarchically nested domains: the distribution of their ranges scales as an inverse power law. **(D)** The distribution of overlaps between the model binding domains in GM12878 compared to the one expected in the mentioned random model (blue).

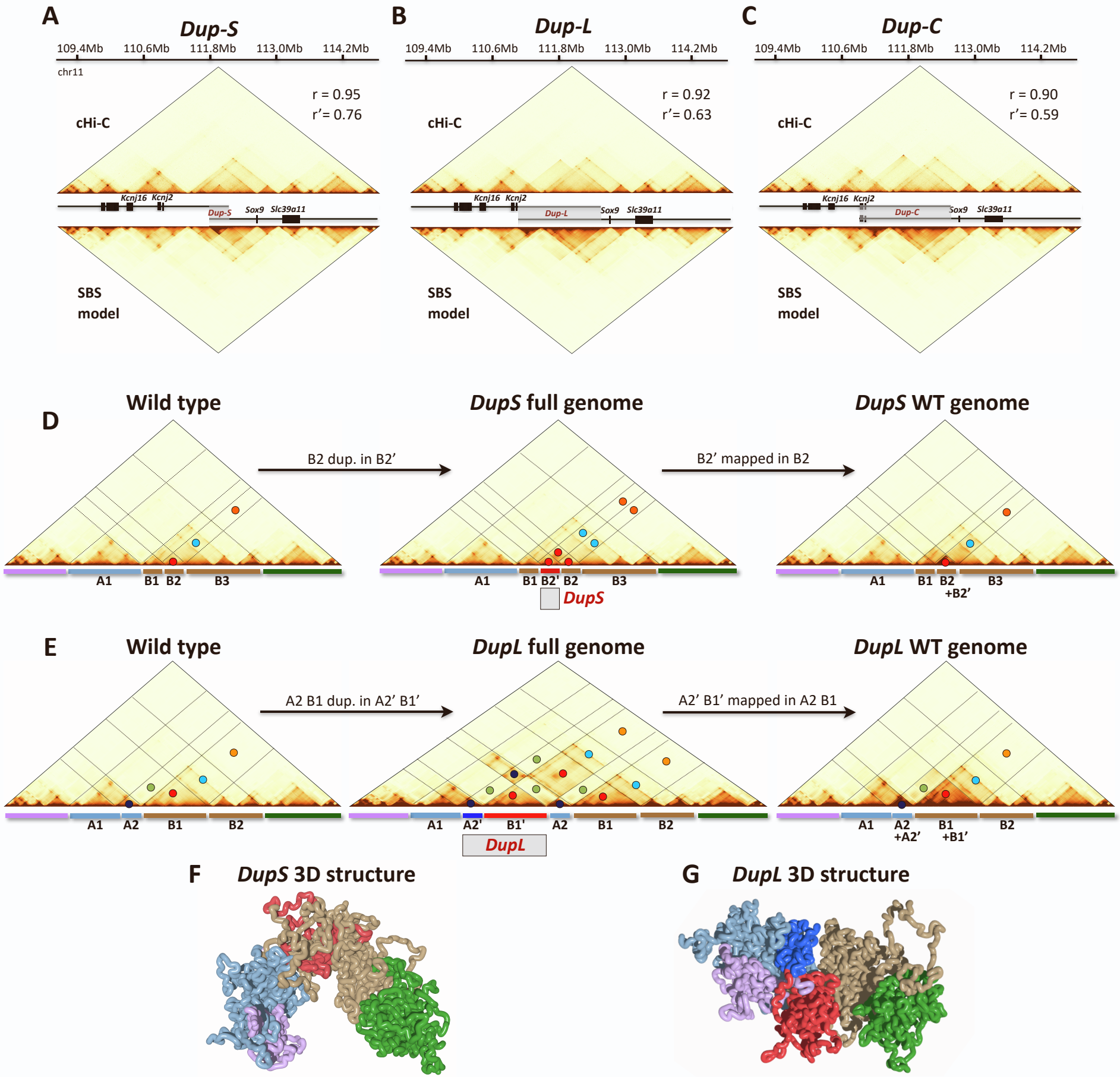

A
Dup-S
109.4Mb
110.6Mb
111.8Mb
113.0Mb
114.2Mb
chr11
cHi-C
r = 0.95
r'= 0.76
Kcnj16
Kcnj2
Dup-S
Sox9
Slc39a11
SBS
model
B
Dup-L
r = 0.92
r'= 0.63
C
Dup-C
r = 0.90
r'= 0.59
D
Wild type
DupS full genome
DupS WT genome
B2 dup. in B2'
B2' mapped in B2
A1
B1
B2
B3
B2'
+B2'
DupS
E
Wild type
DupL full genome
DupL WT genome
A2 B1 dup. in A2' B1'
A2' B1' mapped in A2 B1
A2
B1'
A2'
+A2'
+B1'
DupL
F
DupS 3D structure
G
DupL 3D structure

Figure S3 (Related to Figure 2). The inferred binding domains are validated against mutations at the *Sox9* locus.

(A)-(C) cHi-C data (Franke et al. 2016**,** top) and model predictions (bottom) across the available mutations in the *Sox9* locus, along with the Pearson, r, and distance-corrected Pearson, r', correlation coefficients between model predictions and cHi-C data. Mapping the SBS model predicted contacts on the **(D)** *DupS* and **(E)** *DupL* full genomes clarifies the origin of the novel interactions and of neo-TAD discovered in *DupL* (Franke et al. 2016). The coloured circles help visualising the different regions of interactions of the duplicated sequences and how they map onto the wild-type genome, as reported in Hi-C experiments. **(F)-(G)** Model predicted 3D conformation of the mutated loci. Panels E and G are also shown in Figure 2, reported here to help the comparison between the two mutations.

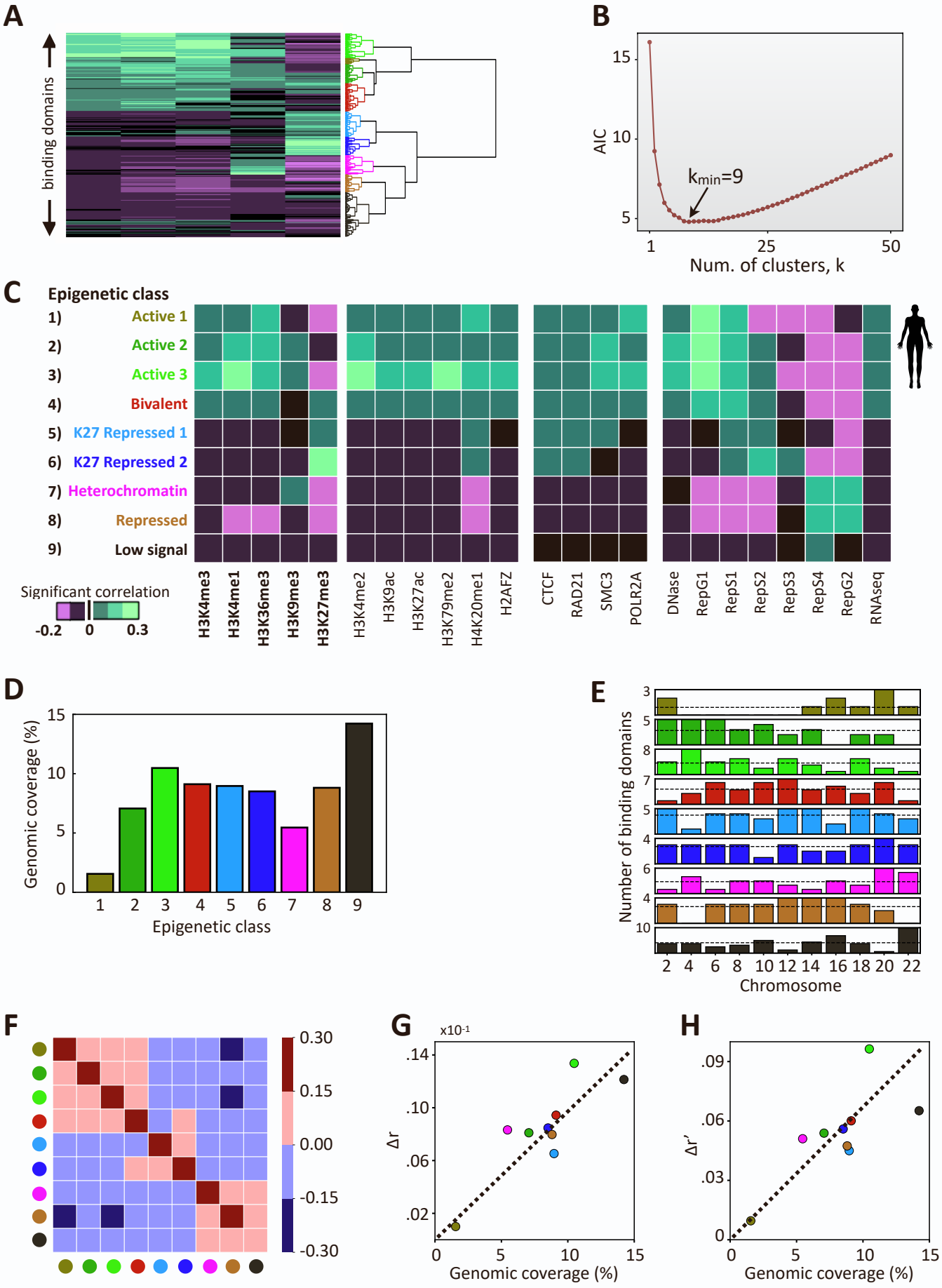

A
binding domains
B
AIC
15
10
5
k_{min}=9
1
25
50
Num. of clusters, k
C
Epigenetic class
1) Active 1
2) Active 2
3) Active 3
4) Bivalent
5) K27 Repressed 1
6) K27 Repressed 2
7) Heterochromatin
8) Repressed
9) Low signal
Significant correlation
-0.2
0
0.3
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
H3K4me2
H3K9ac
H3K27ac
H3K79me2
H4K20me1
H2AFZ
CTCF
RAD21
SMC3
POLR2A
DNase
RepG1
RepS1
RepS2
RepS3
RepS4
RepG2
RNAseq
D
Genomic coverage (%)
15
10
5
0
1 2 3 4 5 6 7 8 9
Epigenetic class
E
Number of binding domains
3
5
8
7
5
4
6
4
10
2 4 6 8 10 12 14 16 18 20 22
Chromosome
F
0.30
0.15
0.00
-0.15
-0.30
G
x10-1
.14
.10
.06
.02
Δr
0 5 10 15
Genomic coverage (%)
H
.09
.06
.03
0
Δr'
0 5 10 15
Genomic coverage (%)

Figure S4 (Related to Figure 3 and STAR Methods). Epigenetic classification of the binding domains in the human GM12878 cell line.

(A) Hierarchical clustering of the PRISMR inferred binding domains with the 9 identified classes highlighted in the dendrogram. **(B)** The AIC statistical criterion has a minimum at k=9 clusters of binding domains. **(C)** The epigenetic signature of the 9 classes and their significant correlations with histone modification, transcription factors, DNA accessibility, DNA replication time and expression data (STAR Methods). **(D)** Genomic coverage of the 9 main epigenetic classes of the SBS model binding domains. **(E)** Relative number of the binding domains of the different classes across chromosomes. The distribution is not uniform (p value < 0.05). **(F)** Pearson correlation coefficient of the genomic location of the different classes over chromosomes. **(G)-(H)** Effect of the withdraw of a class of binding sites as a function of its genomic coverage. The effect of a class removal on the architecture is measured by the variation of the Pearson, r (panel G), and distance corrected Pearson, r’ (panel H), correlation with respect to the wild-type model. Δr’ is the difference between r’ in the wild-type model (r’=0.76) and in a model where the domains of a given class are removed, averaged over chromosomes. Analogously, Δr, is the wild-type (r=0.94) minus r in the mutated model.

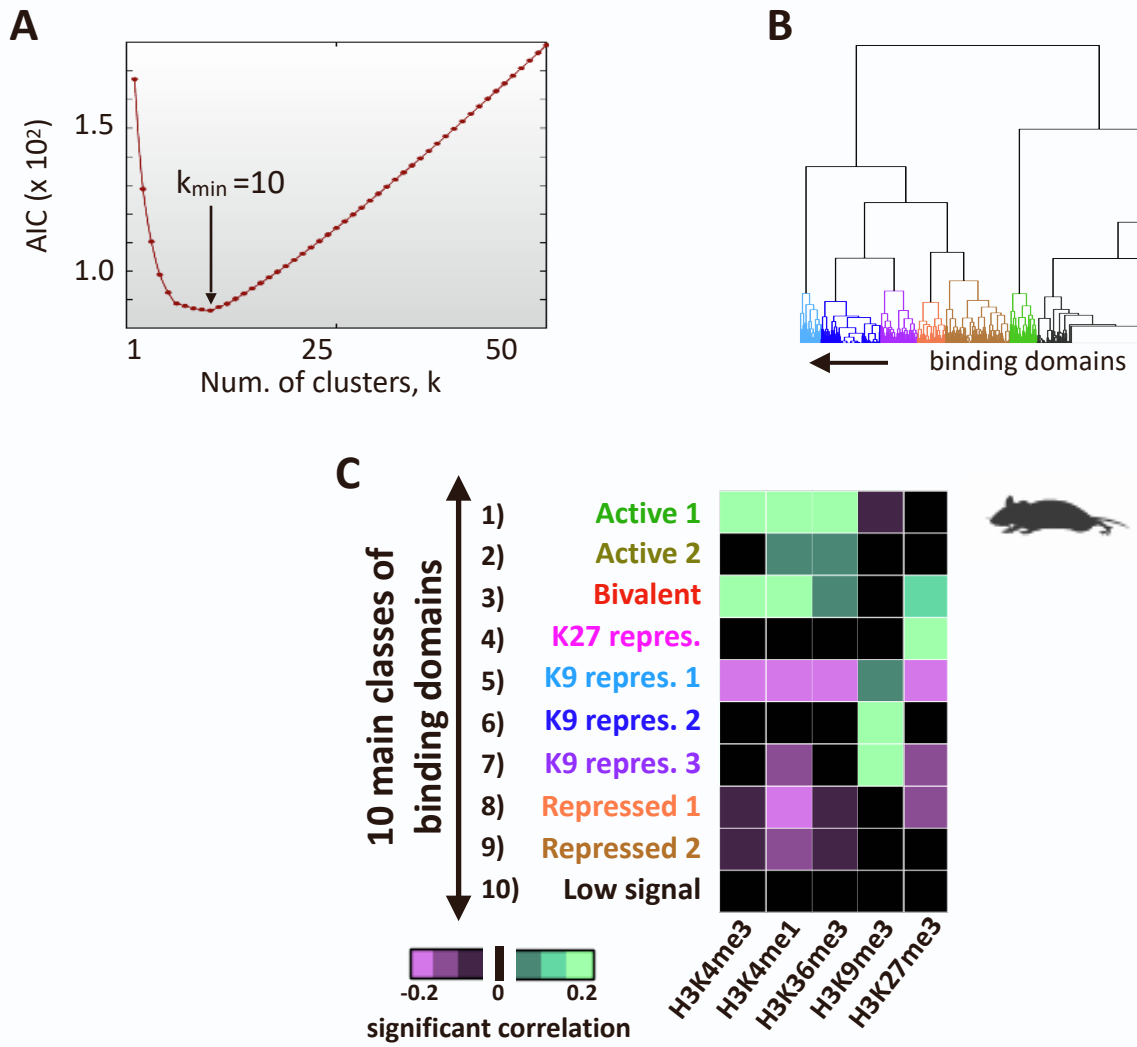


Figure S5 (Related to Figure 3 and STAR Methods). Epigenetic classification of the binding domains in the mouse ES cell line.

(A) The AIC statistical criterion has a minimum at k=10 clusters of binding domains. **(B)** Hierarchical clustering of the PRISMR inferred SBS model binding domains with the 10 identified classes highlighted. **(C)** The epigenetic signature of the 10 classes.

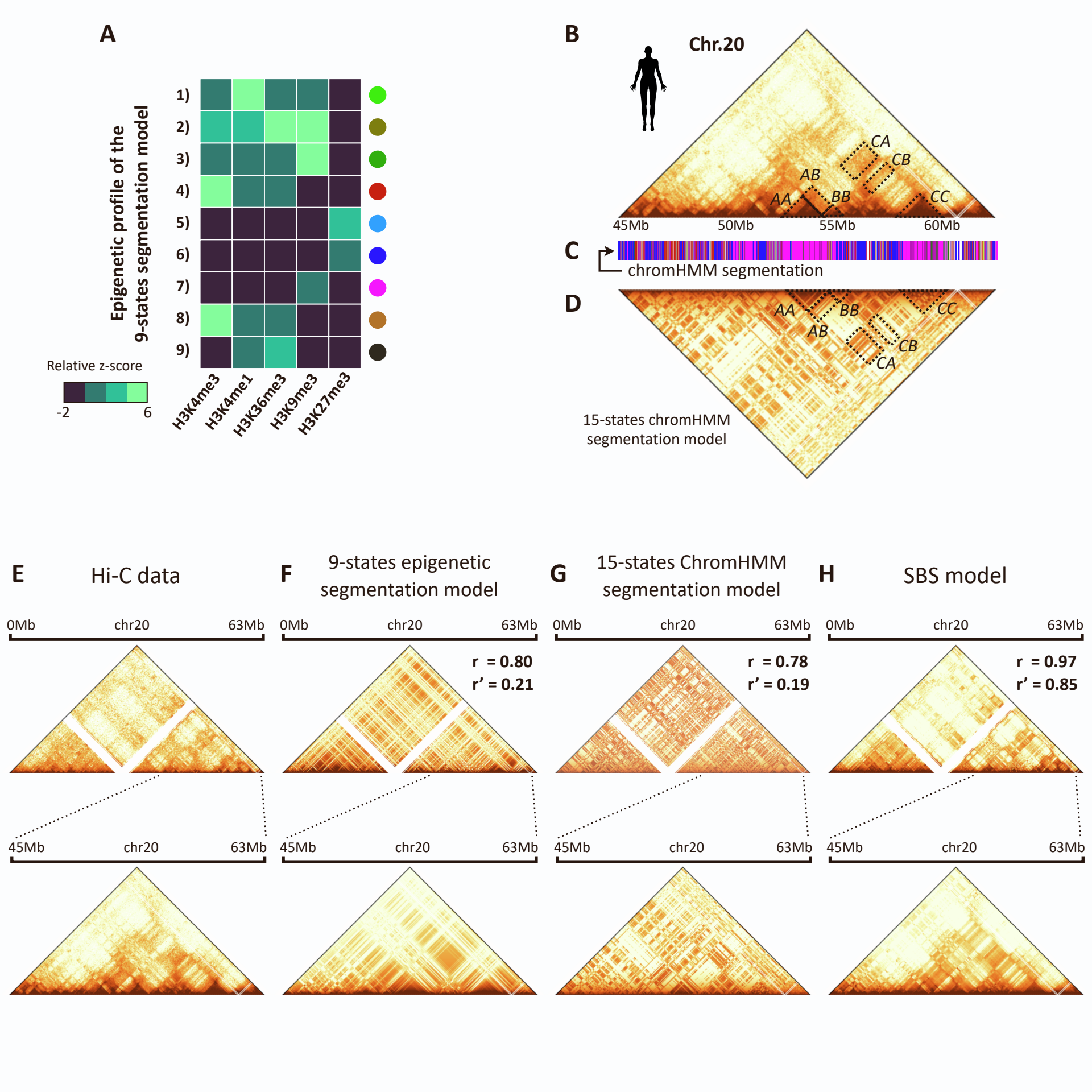
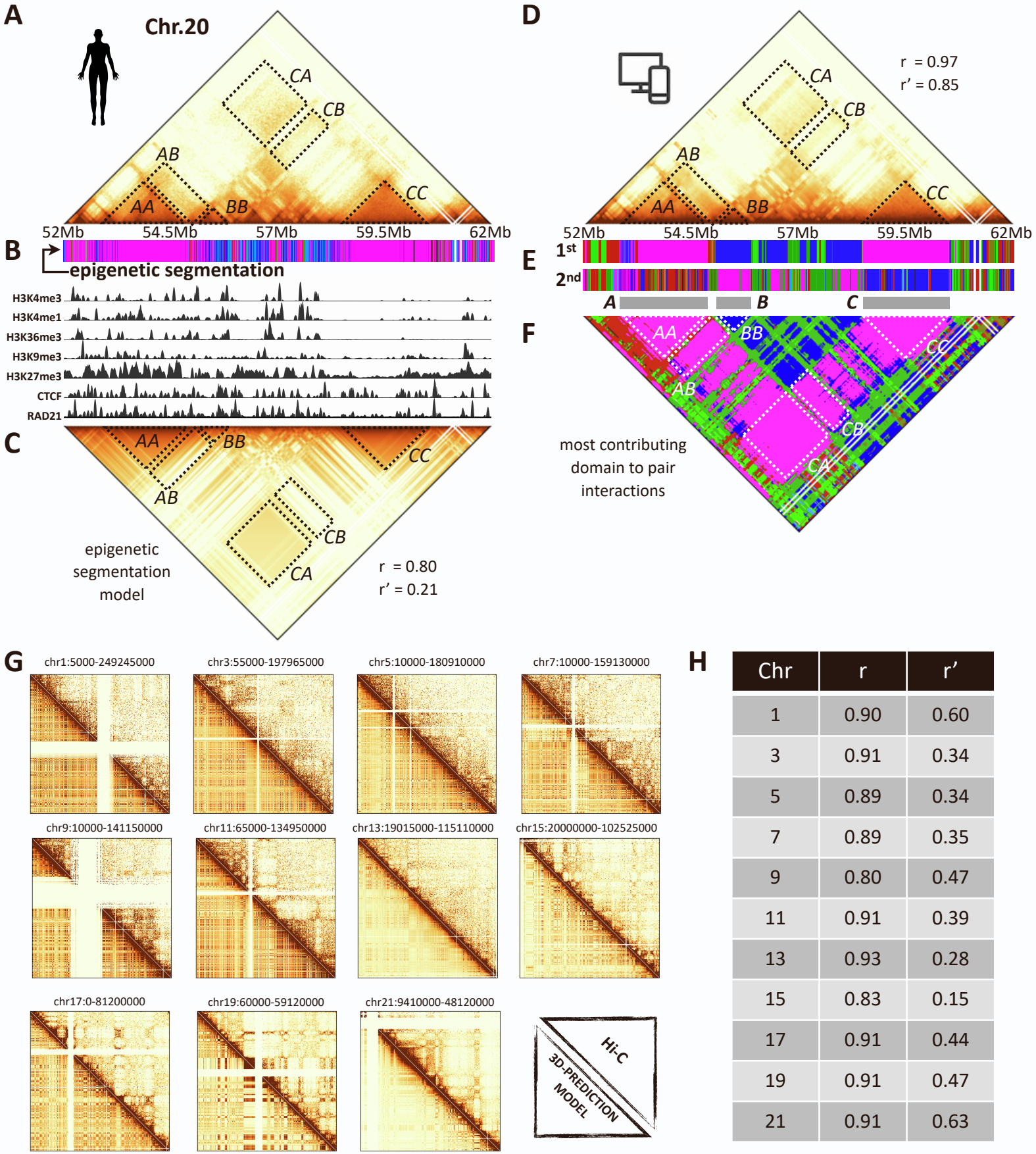

A
Epigenetic profile of the 9-states segmentation model
1)
2)
3)
4)
5)
6)
7)
8)
9)
Relative z-score
-2
6
H3K4me3
H3K4me1
H3K36me3
H3K9me3
H3K27me3
B
Chr.20
CA
CB
AB
AA
BB
CC
45Mb
50Mb
55Mb
60Mb
C
chromHMM segmentation
D
AA
BB
CC
AB
CB
CA
15-states chromHMM segmentation model
E
Hi-C data
F
9-states epigenetic segmentation model
G
15-states ChromHMM segmentation model
H
SBS model
0Mb
chr20
63Mb
45Mb
r = 0.80
r' = 0.21
r = 0.78
r' = 0.19
r = 0.97
r' = 0.85

Figure S6 (Related to Figure 4). Epigenetic linear segmentation models only partially capture chromatin folding.

(A) The heat-map shows the epigenetic profile of the 9 classes obtained by the hierarchical clustering of chromosome segments only on the basis of their histone mark enrichment. These classes have been used to define a SBS polymer model where chromatin physical interactions only occur between homologous 1D-segmented epigenetic regions. The color code is the one used in Figure 4B. **(B)** *In situ* Hi-C data (Rao et al. 2014) of a 20Mb wide region on chr20 in GM12878 and **(C)** its 15-states chromHMM epigenetic segmentation (Kundaje et al. 2015) are shown. **(D)** The contact map of a model based only on homotypic interactions between chromHMM segments has a Pearson correlation r=0.78 and a distance corrected Pearson correlation r'=0.19 with the Hi-C data of the entire chr20 (see main text and STAR Methods). Here a 20Mb region is zoomed to highlight the different patterns. The absence of combinatorial overlap makes the 1D segmentation model unable to explain interactions between regions marked by different chromatin states (see for instance the contacts in CB). **(E)** In situ Hi-C data in GM12878 (Rao et al. 2014) of the entire chr20 (top) and of a zoomed 20Mb wide region (bottom). **(F)-(G)** Contact maps from a model of chr20 based only on homotypic interactions between linear segmented epigenetic regions. The Pearson correlation, r, and distance corrected Pearson correlation, r', between model and Hi-C matrices are shown in the top panels for the entire chr20. The bottom panel shows a 20Mb region to highlight the different patterns (r=0.78, r'=0.02 for the 9-states segmentation model and r=0.77, r'=0.05 for the ChromHMM 15-states model, see main text and STAR Methods). **(H)** The PRISMR inferred SBS model contact map for those regions together with the corresponding correlations for the entire chr20 (r=0.97 and r'=0.84).



Chr	r	r'
1	0.90	0.60
3	0.91	0.34
5	0.89	0.34
7	0.89	0.35
9	0.80	0.47
11	0.91	0.39
13	0.93	0.28
15	0.83	0.15
17	0.91	0.44
19	0.91	0.47
21	0.91	0.63

Figure S7 (Related to Figure 4 and Figure 5). The epigenetic barcode of binding domains predicts *de novo* chromatin contacts across chromosomes.

A, C, D, E, F as in Figure 4 for a zoomed 10Mb wide genomic region. In panel **B**, the linear epigenetic segmentation of the region is shown together with the profiles for the relevant histone modifications, CTCF, and Rad21. **(G)** The upper triangles show the in situ Hi-C maps from the odd-numbered chromosomes in GM12878, while the lower triangles show the contact maps obtained by the predicted polymer models (scales as in Figure 1). **(H)** Pearson (r) and distance-corrected Pearson (r') correlation coefficients between the predicted matrices and the corresponding in situ Hi-C data.