PNAS

Supporting Information for

Linking DNA-packing density distribution and TAD boundary locations

Luming Meng, Fu Kit Sheong, and Qiong Luo

Luming Meng, Fu Kit Sheong, and Qiong Luo

Email: menglum@scau.edu.cn (L.M.), fksheong@connect.ust.hk (F.K.S.), luoqiong@scnu.edu.cn (Q. L.)

This PDF file includes:

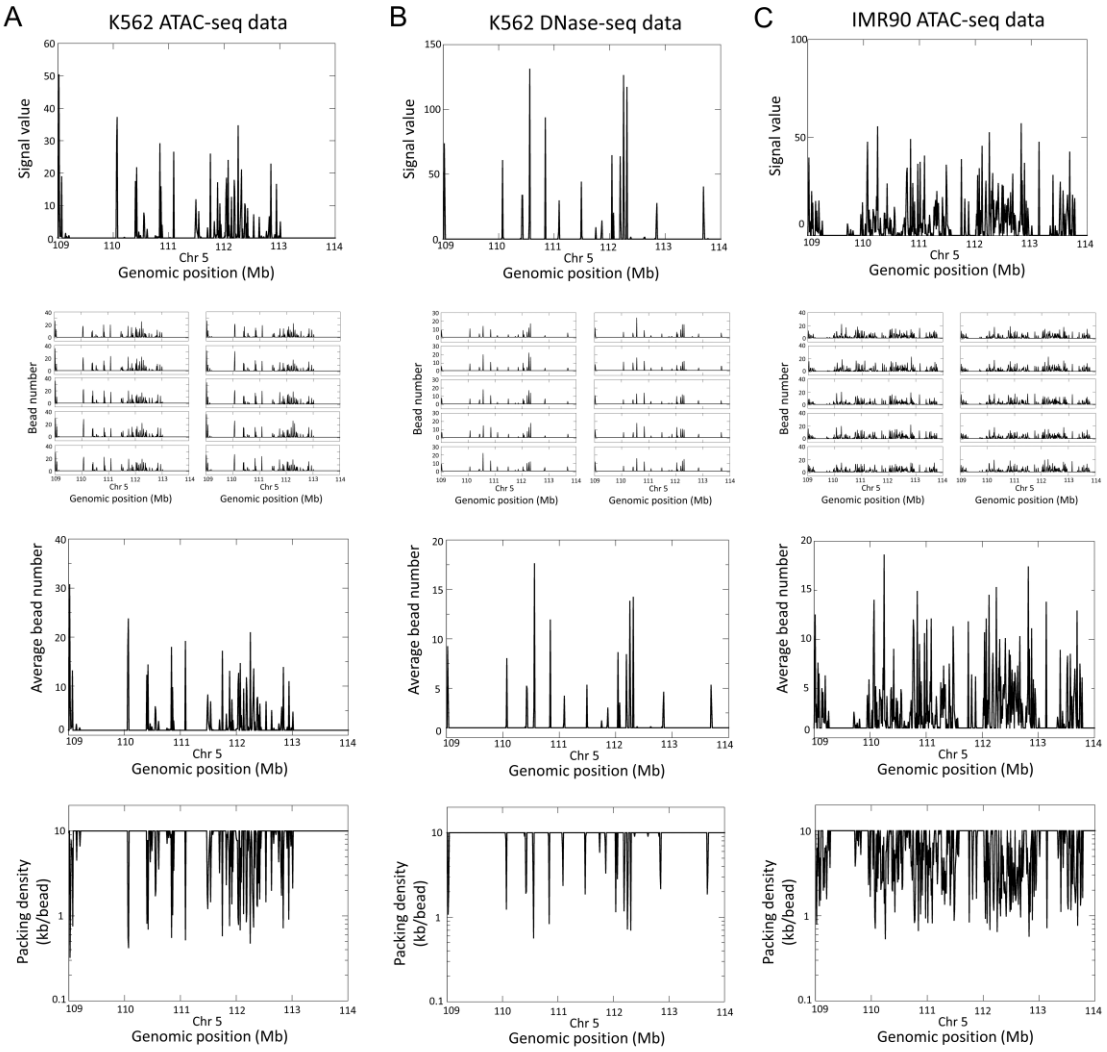


Fig. S1 Plots of bead numbers, different sets of random independent Poisson variables (k_1, k_2,…k_N), and DNA-packing density in our models, resulted from population-based chromatin accessibility data

(**A** to **C**) Row 1: Population-averaged chromatin accessibility data for the 5-Mb chromatin region (Chr5: 109Mb-114 Mb) in K562 erythroleukemia cell type obtained from ATAC-seq experiment (A) or from DNase-seq experiment (B), and in IMR90 lung fibroblast cell type obtained from ATAC-seq experiment (C). Row 2: Ten sets of random independent Poisson variable vectors (k_1, k_2,…k_N) are derived from the data shown in Row 1 by the approach of developing heteropolymer models. Row 3: The mean vector across the ten vectors shown in Row 2, showing the mean value of bead number for each 10-kb segment in the region. Note that the profile of this plot is similar to that of population-averaged chromatin accessibility data shown in Row 1, consistent with our Poisson model of bead number generation. Row4: DNA-packing density, which is defined as the length of the segment (in kb) divided by the average number of beads (shown Row 3).

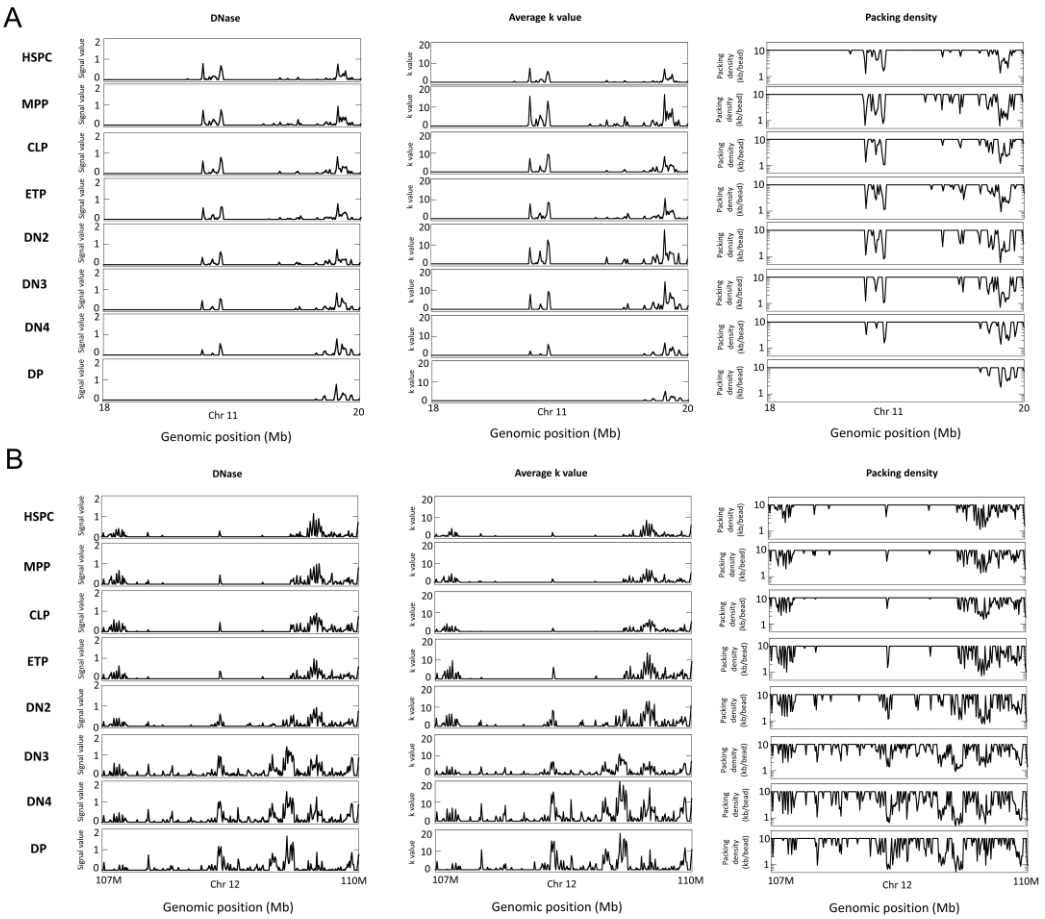


Fig. S2 Plots of population-averaged chromatin accessibility data, averaged bead number distribution, and DNA-packing density for cells at eight developmental stages.

(**A** and **B**) Column 1: Population-averaged chromatin accessibility data for the genomic region (Chr11:18Mb-20Mb) (A) and the region (Chr12:107Mb-110Mb) (B) in cell types of eight developmental stages (HSPC, MPP, CLP, ETP, DN2, DN3, DN4, and DP) during the differentiation shown in Fig. S7 and 3. Column 2: The mean value of bead number for each 10-kb segment in the region. Column 3: DNA-packing density, which is defined as the length of the segment (in kb) divided by the average number of beads shown in Column 2.

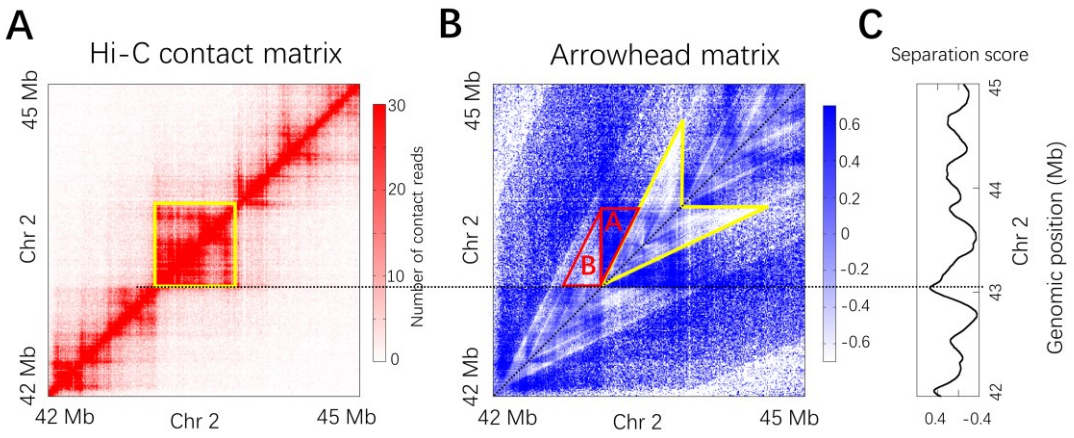


Fig. S3 A scheme of the definition of separation score

(**A**)Population-average Hi-C contact frequency matrix for the 3-Mb region (Chr2:42Mb-45Mb) of K562 at 10-kb resolution.

(**B**)The arrowhead matrix *M* where the arrowhead-shaped motif highlighted by yellow corresponds to the yellow highlighted domain square in panel A.

(**C**) Separation score of a position along the diagonal of the arrowhead matrix is defined based on the two constructed edge-shared congruent right triangles at that position. As an example, such two edge-shared congruent right triangles of the position corresponding to a domain boundary are highlighted by red in panel B. The length of horizontal edge of the two congruent right triangles is 125-kb and the length of the vertical edge (or shared edge) is 250-kb.

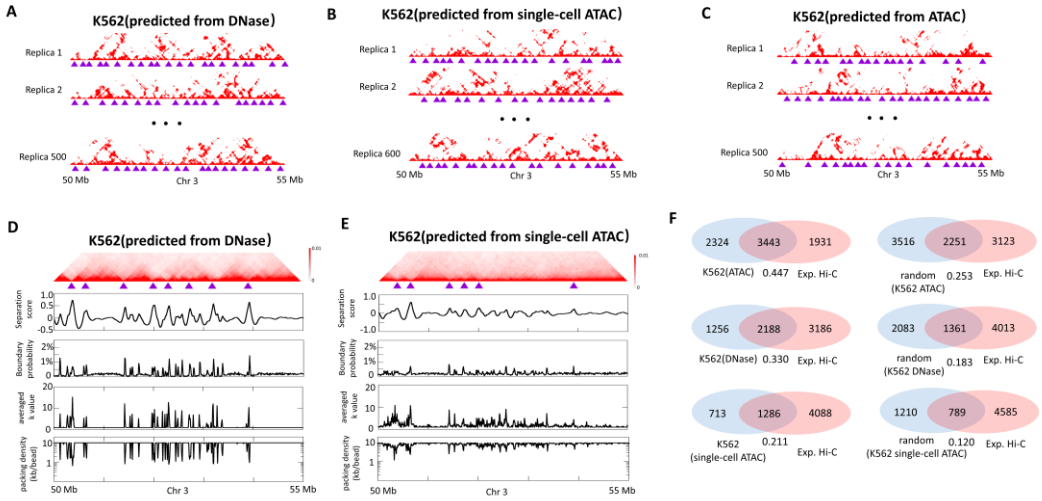


Fig. S4 Prediction of domain boundary positions of K562 cell type by our model and comparison against those obtained from experimental Hi-C data

(**A** to **C**) Individual contact matrices of the 5-Mb chromatin region (Chr3:50Mb-55 Mb) in K562, calculated from the individual conformations in the conformation ensemble of Chromosome 3 calculated from population-averaged DNase-seq data (A), single-cell ATAC-seq data (B), population-averaged ATAC-seq data (C). Domain structures are identified for each individual conformations and their boundaries are shown as purple triangles. All the matrices are of the resolution of 10 kb.

(**D** to **E**) Row 1: Ensemble-averaged contact matrix across all individual contact matrices of chromatin region (Chr3:50Mb-55 Mb) shown in pannel A and B. The ensemble-averaged matrix is at the resolution of 10 kb. Row 2: Separation score plot for the 10-kb segments in this chromatin region. Row 3: Probability (fraction of the individual conformations) for each 10-kb segment to appear as a single-cell domain boundary. Row 4: Averaged k_i for each 10-kb segments. Row 5: DNA-packing density of each 10-kb segment, defined by the length of the segment (in kb) divided by the average number of beads in the segment.

(**F**) Analysis of the overlapping TAD boundaries between the Hi-C contact matrix and the ensemble-averaged contact matrices calculated from the conformation ensembles of the 22 autosomes and the X-chromosome in K562. As comparison, similar overlapping analysis are shown between the Hi-C TAD boundaries and the random boundaries (generated by circular shifting the genomic positions of domain boundaries determined from corresponding ensemble-averaged contact matrices along chromatin sequence). The Jaccard indexes between the simulated boundary set (or the random boundary set) and the Hi-C boundary set are presented below the overlapping diagrams (for example, 0.447 for predicted domain boundaries from K562 ATAC-seq data and TAD boundaries from experimental Hi-C data). The P-values of these three data set for the percentage of the predicted boundaries aligned with the Hi-C TAD compared to those of the randomly selected boundaries are all less than 10^{-5}

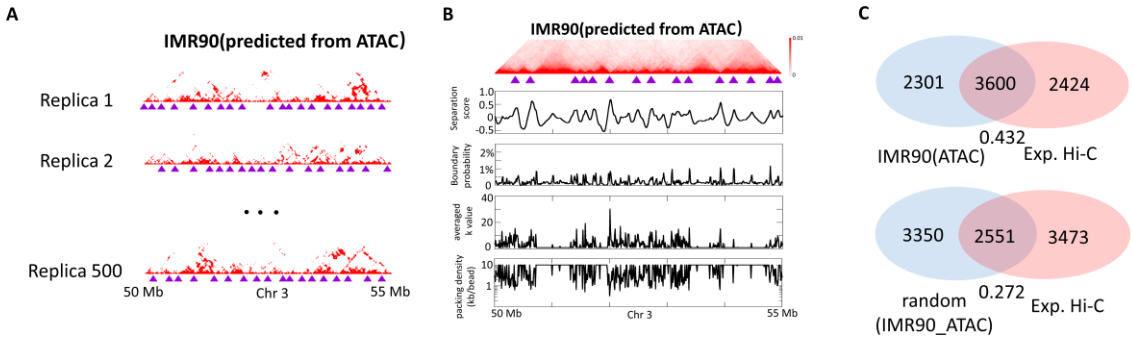
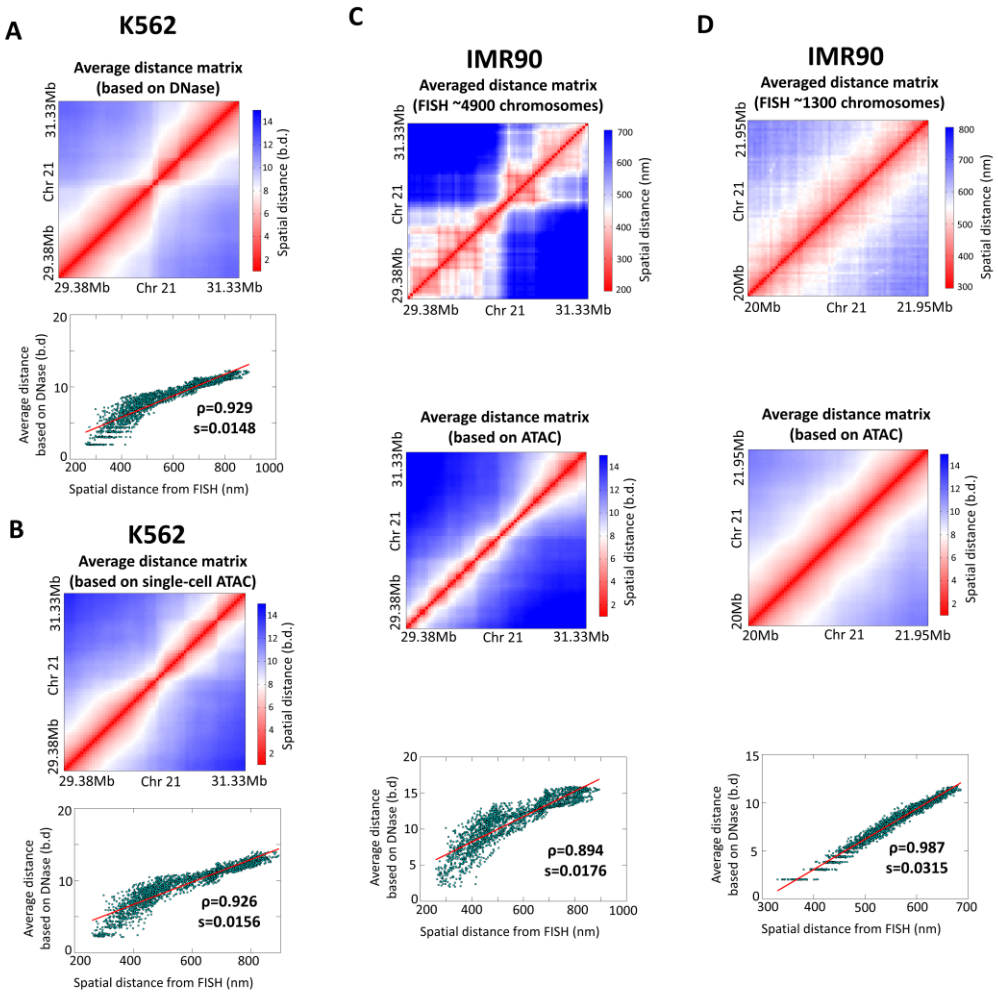


Fig. S5 Prediction of domain boundary positions of IMR90 cell type by our model and comparison against those obtained from experimental Hi-C data

(**A**)Individual contact matrices of the 5-Mb chromatin region (Chr3:50Mb-55 Mb) in IMR90, calculated from the individual conformations in the conformation ensemble of Chromosome 3 calculated from the population-averaged DNase-seq data. Domain structures are identified for each individual conformations and their boundaries are shown as purple triangles. All the matrices are of the resolution of 10 kb.

(**B**) Row 1: Ensemble-averaged contact matrix across all individual contact matrices of chromatin region (Chr3:50Mb-55 Mb) shown in pannel A. The ensemble-averaged matrix is at the resolution of 10 kb. Row 2: Separation score plot for the 10-kb segments in this chromatin region. Row 3: Probability (fraction of the 500 individual conformations) for each 10-kb segment to appear as a domain boundary in single conformation. Row 4: Averaged k_i for each 10-kb segments. Row 5: DNA-packing density of each 10-kb segment, defined by the length of the segment (in kb) divided by the average number of beads in the segment.

(**C**) Analysis of the overlapping TAD boundaries between the Hi-C contact matrix and the ensemble-averaged contact matrices calculated from the conformation ensembles of the 22 autosomes and the X-chromosome in IMR90. As comparison, similar overlapping analysis are shown between the Hi-C TAD boundaries and the random TAD boundaries (generated by circular shifting the genomic positions of domain boundaries determined from corresponding ensemble-averaged contact matrix along chromatin sequence). The Jaccard indexes between the simulated boundary set (or the random boundary set) and the Hi-C boundary set are presented below the overlapping diagrams (for example, 0.432 for predicted domain boundaries from IMR90 ATAC-seq data and TAD boundaries from experimental Hi-C data). The P-values for the alignment of the predicted boundaries with the Hi-C TAD compared to the randomly selected boundaries are all less than 10^{-50}



116

117 **Fig. S6 Predictions of chromatin organization by our model and comparison against**
118 **experimental FISH data.**

119 (**A** and **B**) Row 1: The 30-kb resolution ensemble-averaged distance matrix of the 2-Mb region
120 (Chr21:29.38Mb-31.33Mb) in K562 cell type calculated from the conformation ensemble of
121 Chromosome 21 derived from the population-averaged DNase-seq data (A) or from the single-
122 cell ATAC-seq data (B). Row 2: Correlation between the elements of the FISH distance matrix
123 shown in Fig. 2D and the ensemble-averaged distance matrix shown in Row 1. (**C** and **D**) Row 1:
124 The experimental averaged distance matrix at 30-kb resolution across ~4900 FISH images of the
125 2-Mb region (Chr21:29.38Mb-31.33Mb) in IMR90 cell type (**C**), across~1300 FISH images of
126 the 2-Mb region (Chr21:20Mb-21.95Mb) in IMR90 cell type (**D**). Row 2: The 30-kb resolution
127 ensemble-averaged distance matrix of the region shown in Row 1, calculated from the
128 conformation ensemble of Chromosome 21 in IMR90 derived from the population-averaged
129 ATAC-seq data. Row 3: Correlation between the elements of the FISH distance matrix shown in
130 Row 1 and the ensemble-averaged distance matrix shown in Row 2.

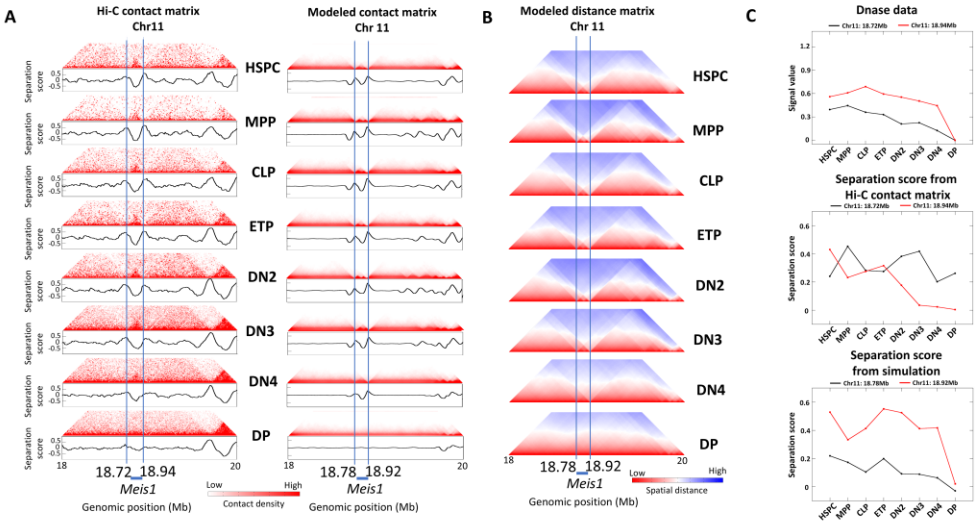


Fig. S7 Our method allows de novo prediction of chromatin reorganization during the differentiation from hematopoietic stem and progenitor cells (HSPCs) to mature immune T cells.

(A) Left: Hi-C contact-frequency matrices (heatmap plot) of the genomic region containing *Meis1* in cell types of eight developmental stages (HSPC, MPP, CLP, ETP, DN2, DN3, DN4, and DP) during the differentiation and separation score (curve plot) for each genomic position. Right: The corresponding simulated ensemble-averaged contact matrices (heatmap plot) of the same region and separation score (curve plot) for each genomic position. All the matrices are of the resolution of 10 kb.

(B) The simulated ensemble-averaged spatial-distance matrices of the region containing *Meis1* in the eight developmental stages, with the resolution of 10 kb.

(C) DNase signals (top) and Hi-C-derived separation scores (middle) at the two *Meis1* boundary positions on Chr 11: 18.72 Mb (black line) and 18.94 Mb (red line) across the eight stages. The bottom panel shows simulated separation scores at the predicted *Meis1* boundary positions on Chr 11: 18.78 Mb (black line) and 18.92 Mb (red line). Hi-C and DNase data are from reference[1].

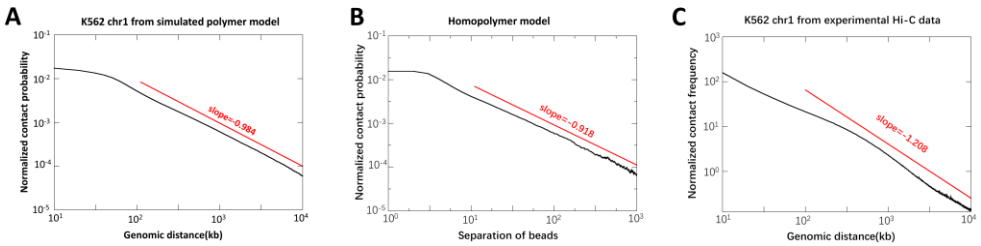


Fig. S8 Contact probability as a function of genomic distance.

(A and **B)** Simulation data for K562 cell chromosome 1 show that contact probability as a function of genomic distance (red line) follows a power-law scaling with a slope of −0.984 (in kb)(A), and −0.918 when plotted against bead separation (B).

(C) Hi-C data for K562 cell chromosome 1 (ref. 62) show that contact probability as a function of genomic distance follows a power-law scaling with a slope of −1.208(in kb).

The genomic distance under the red line approximately ranges from 100 kb to 10 Mb.

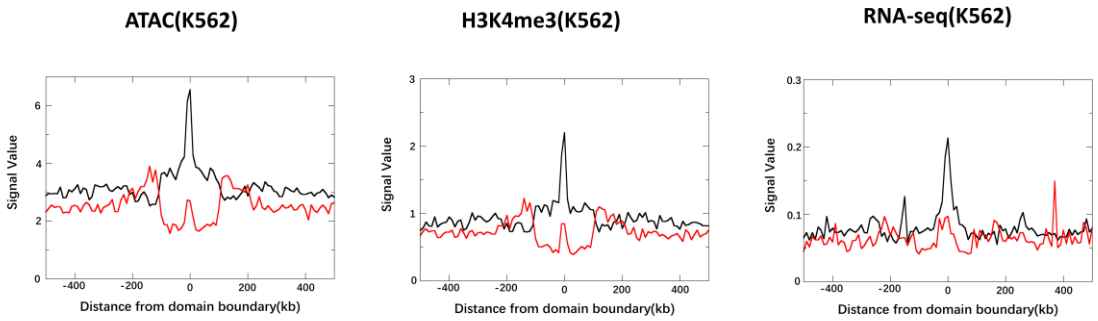


Fig. S9 Averaged features of chromatin segments around TAD boundaries in K562 cell. **(A** to **C)** show averaged chromatin accessibility (A), H3K4me3 ChIP-seq (B), and RNA-seq (C) signals around TAD boundaries. Average distributions of chromatin accessibility (A), H3K4me3 ChIP-seq (B), and RNA-seq (C) signals around TAD boundaries. Black lines represent TAD boundaries predicted by our model, while red lines indicate boundaries not captured by our model.

Table S1. Quantitative assessment of prediction performance.

Cell Type	N_{exp}	N_{sim}	True Positive	False Positive	False Negative
K562 (ATAC)	5374	5767	3443	2324	1931
K562 (DNase)	5374	3444	2188	1256	3186
K562 (scATAC)	5374	1999	1286	713	4088
IMR90 (ATAC)	6024	5901	3600	2301	2424

N_{exp}: The number of TAD boundaries identified from Hi-C data

N_{sim}: The number of TAD boundaries identified from our simulated contact matrix

SI References

1 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).