SUPPLEMENTARY INFORMATION

Conformational analysis of chromosome structures reveals vital role of chromosome morphology in gene function

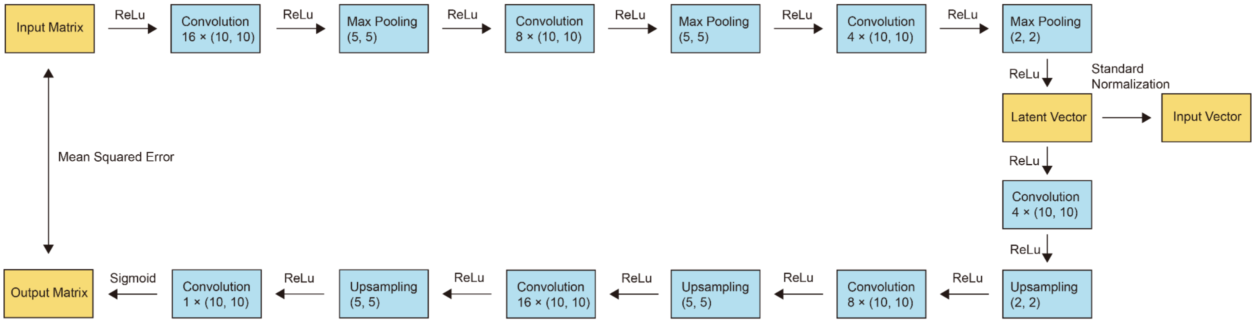
Yuxiang Zhan[1,2,3], Asli Yildirim[1,2], Lorenzo Boninsegna[1,2], Frank Alber[1,2,3*]

[1]Department of Microbiology, Immunology, and Molecular Genetics, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095

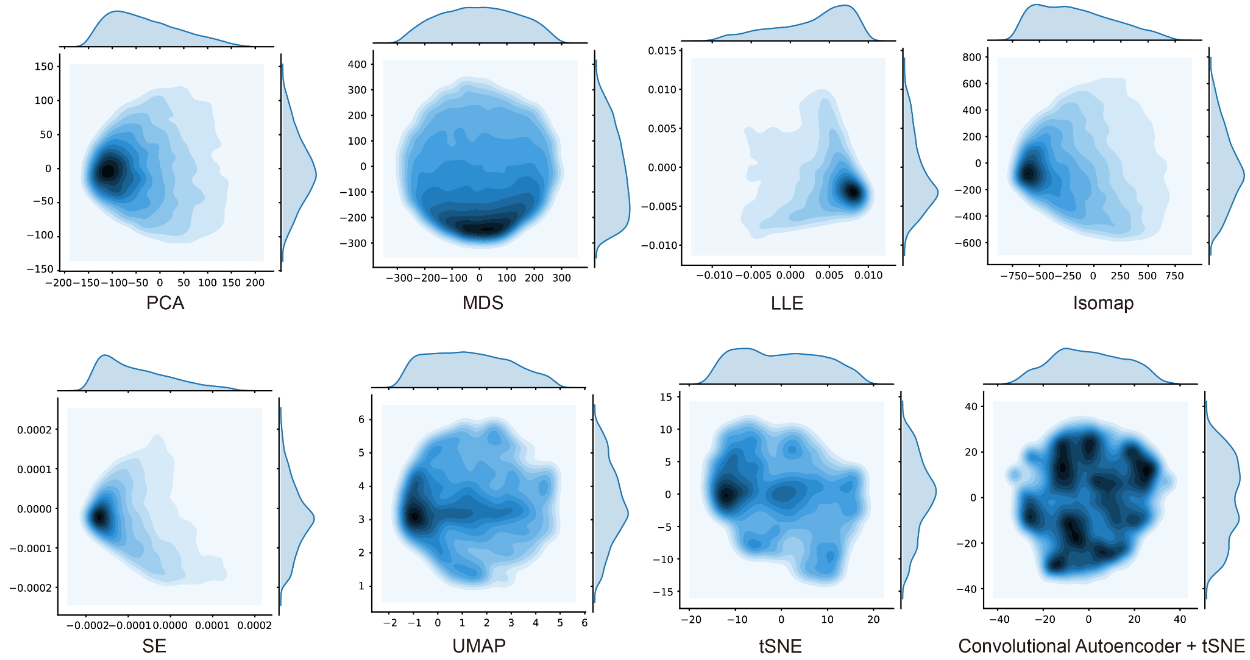
[2]Institute of Quantitative and Computational Biosciences, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095

[3]Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

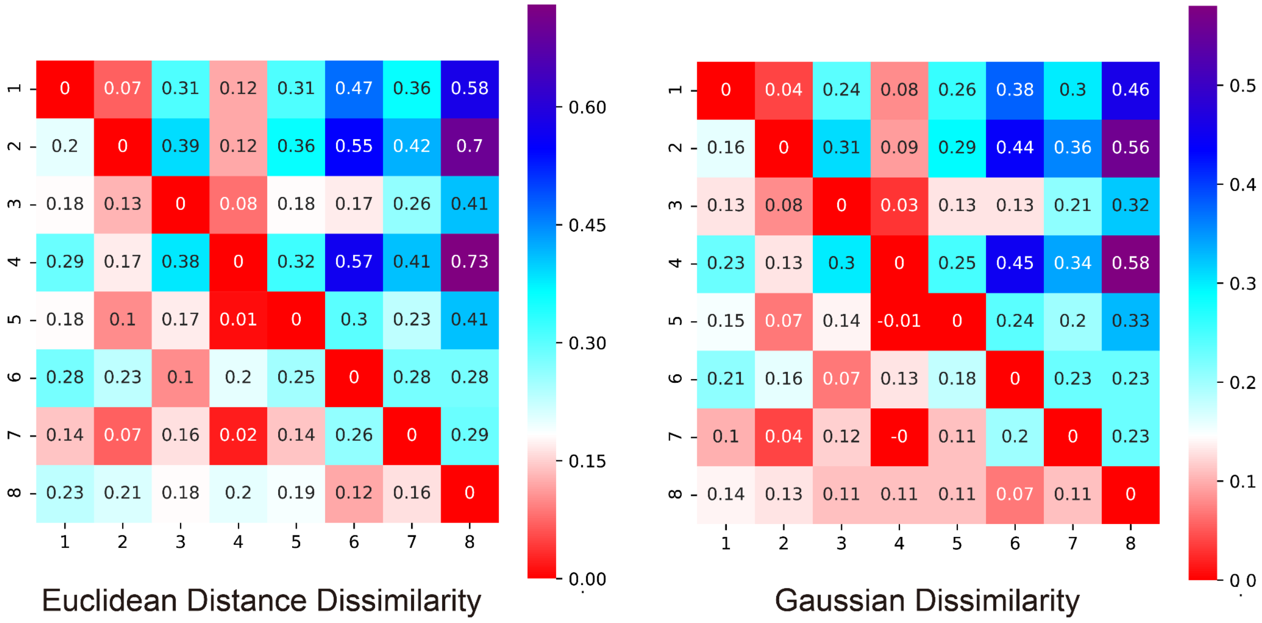
*Correspondence should be addressed to F.A. falber@g.ucla.edu



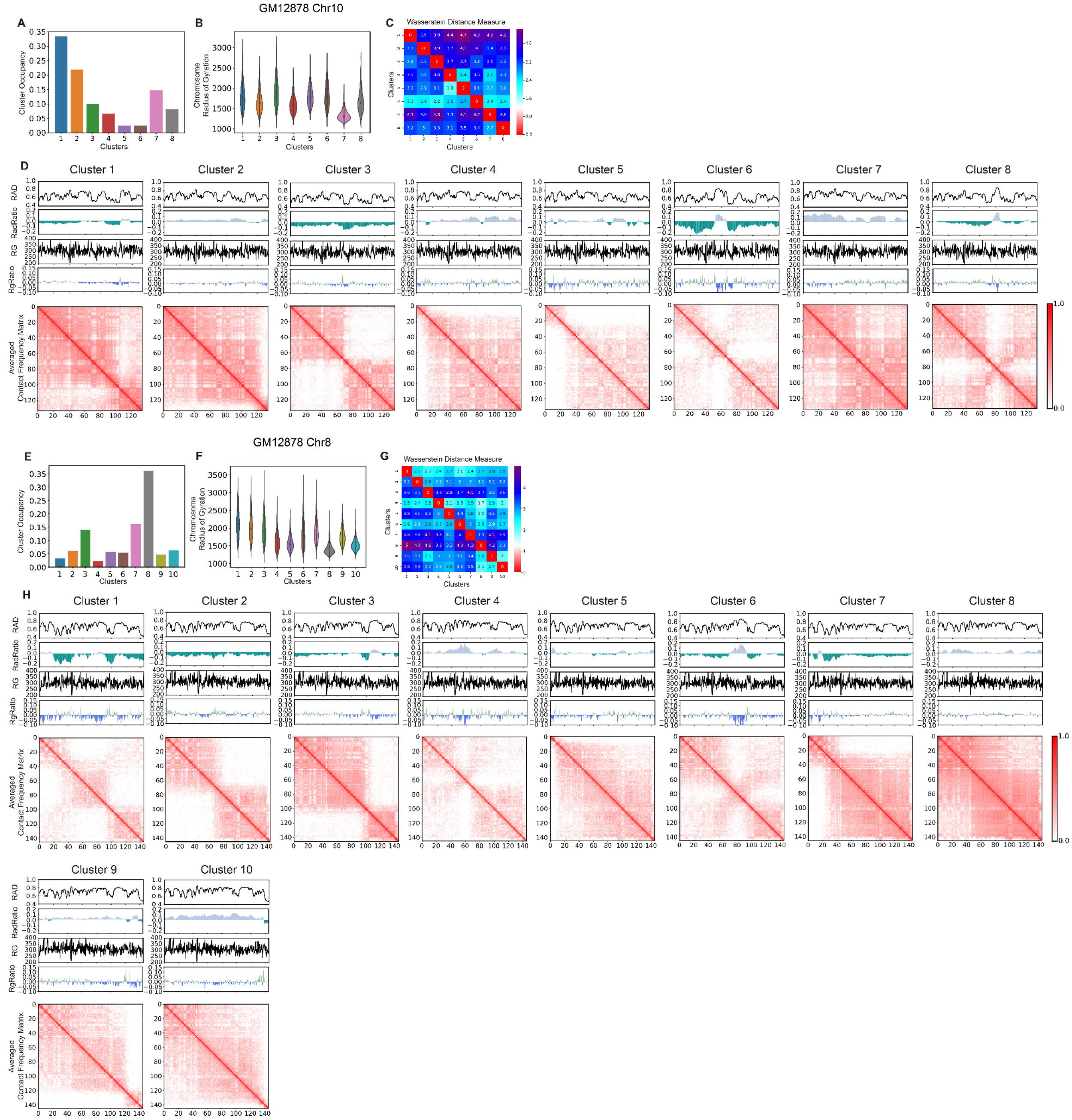
Supplementary Figure 1: Architecture of the autoencoder The architecture of the autoencoder consists of an encoder and a decoder. The encoder consists of three convolution layers and three max pooling layers. The decoder consists of four convolution layers and three upsampling layers. The loss between the input and the output is measured by the mean squared error.



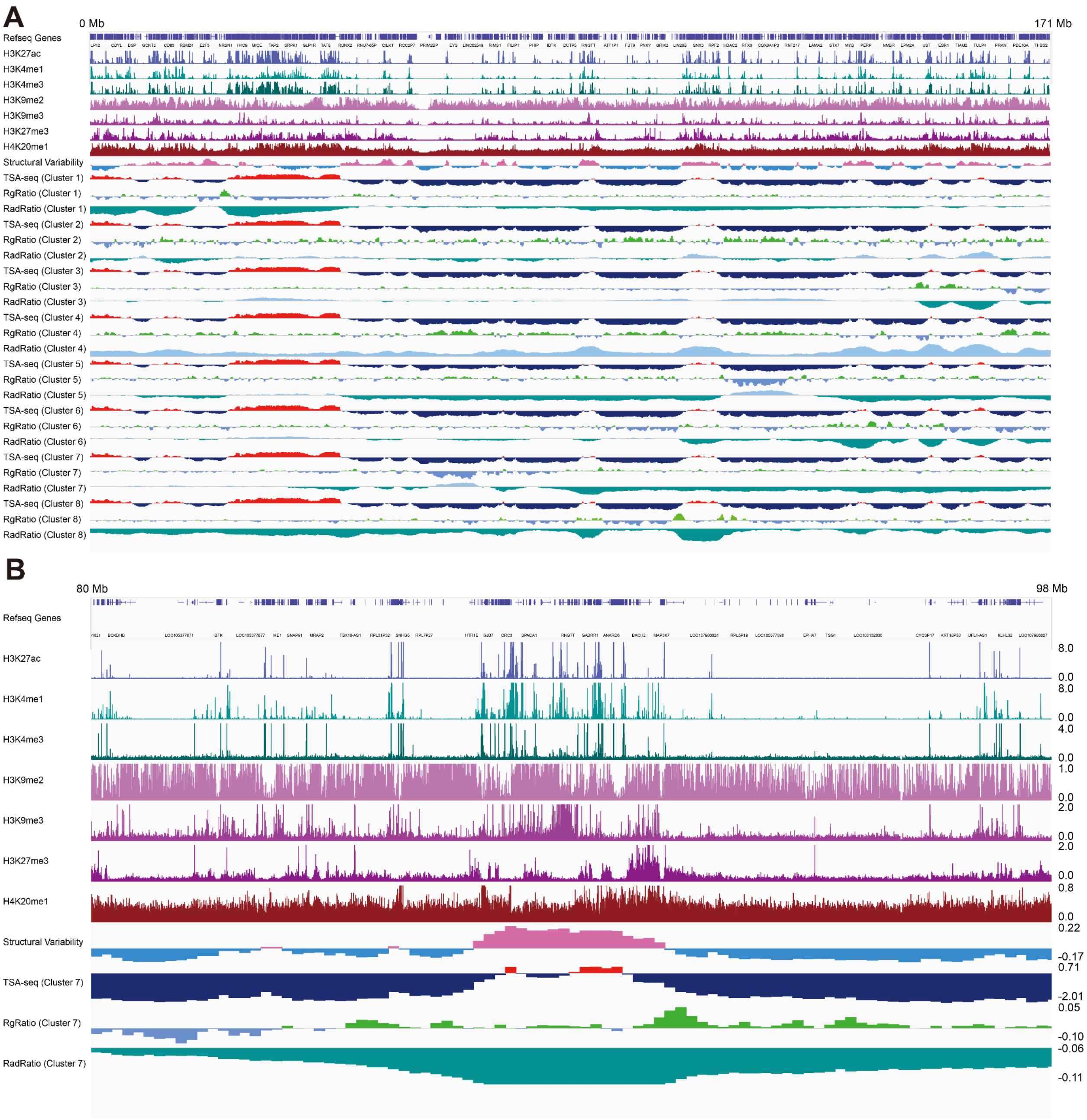
Supplementary Figure 2: Comparison of different dimension reduction methods Visualization of different dimension reduction methods: PCA, MDS[1], LLE[2], Isomap[3], SE[4], UMAP[5], tSNE[6]. Each method uses the same input data (the distance vectors derived from the distance matrices of GM12878 chromosome 17). After the embedded data points are obtained, we visualize the distribution by bivariate kernel density estimation. In addition we also plot the two-step dimension reduction (Convolutional Autoencoder + tSNE) proposed in this study. Note that only the two-step dimension reduction is able to generate balanced clusters.



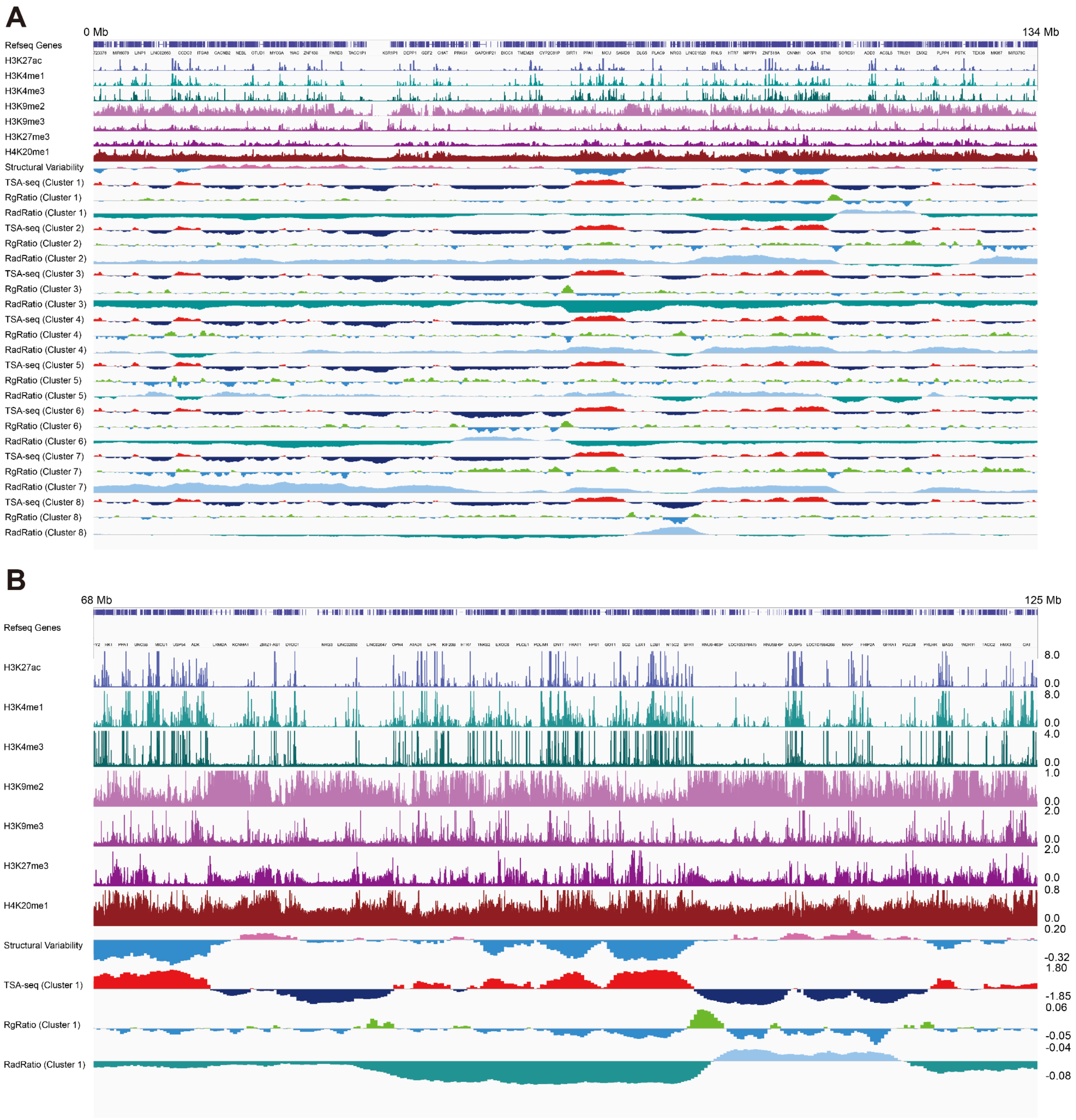
Supplementary Figure 3: Comparison of different pairwise dissimilarity measurements on GM12878 Chr6 Pairwise dissimilarity between the 8 clusters. The dissimilarity matrices are calculated by measurements of Euclidean distance dissimilarity and Gaussian dissimilarity[7,8]. Each entry represents the log fold ratio between the inter-cluster dissimilarity and the intra-cluster dissimilarity, where positive values indicate the inter-cluster dissimilarity is larger than the intra-cluster dissimilarity.



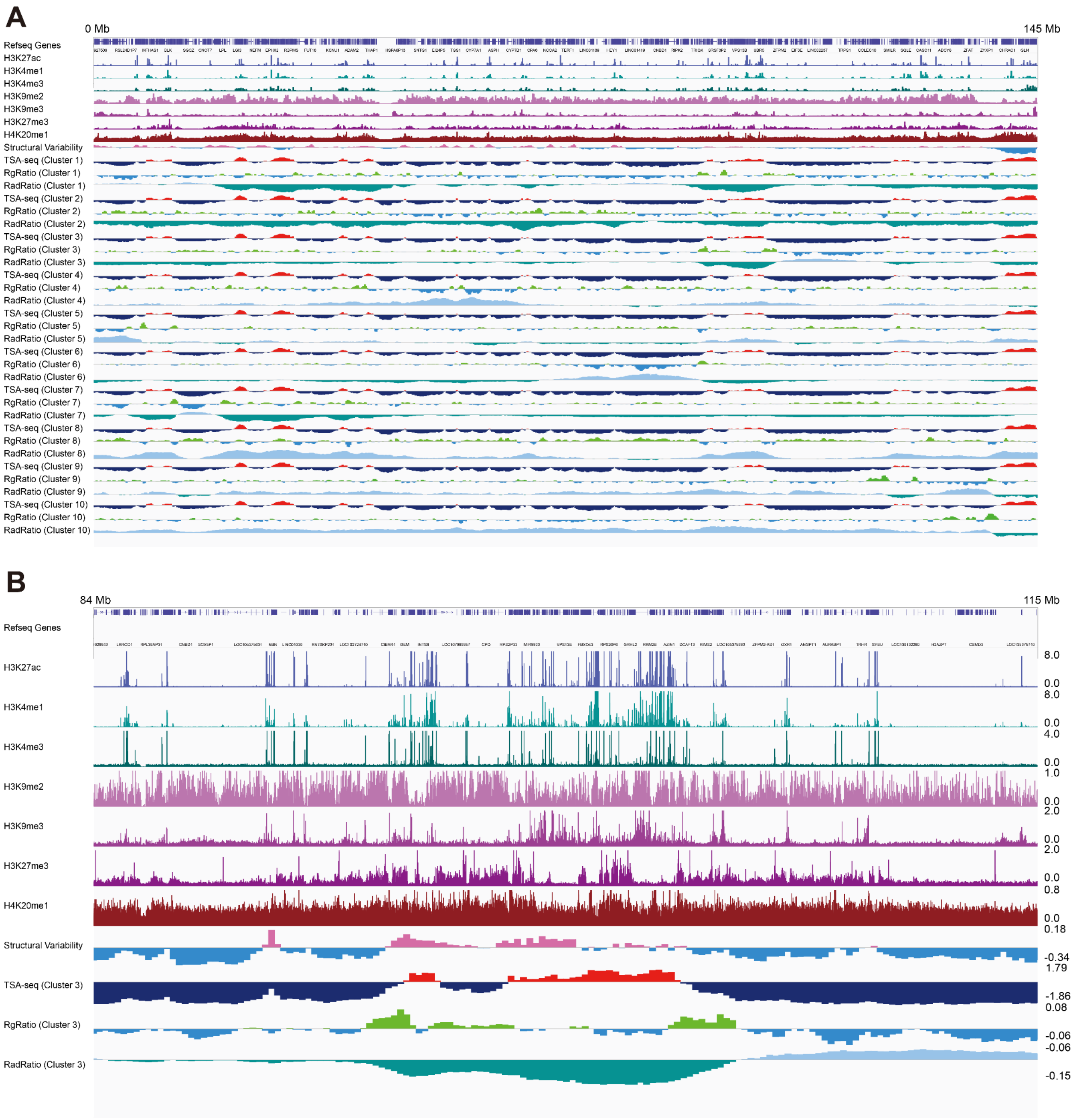
Supplementary Figure 4: Evaluation of the method's performance on GM12878 Chr10 and Chr8 A, The cluster occupancy of the 8 predicted clusters of Chr10. **B,** The distributions of chromosome radius of gyration for the 8 predicted clusters of Chr10. **C,** Pairwise dissimilarity between the 8 clusters of Chr10. The dissimilarity matrix is calculated by the measurement of Wasserstein distance[9]. Each entry represents the log fold ratio between the inter-cluster dissimilarity and the intra-cluster dissimilarity, where positive values indicate the inter-cluster dissimilarity is larger than the intra-cluster dissimilarity. **D,** The RAD, RadRatio, RG, RgRatio and contact frequency matrix from each of the 8 clusters of Chr10 predicted by the two-step dimension reduction method. **E,** The cluster occupancy of the 10 predicted clusters of Chr8. **F,** The distributions of chromosome radius of gyration for the 10 predicted clusters of Chr8. **G,** Pairwise dissimilarity between the 10 clusters of Chr8. The dissimilarity matrix is calculated by the measurement of Wasserstein distance. Each entry represents the log fold ratio between the inter-cluster dissimilarity and the intra-cluster dissimilarity, where positive values indicate the inter-cluster dissimilarity is larger than the intra-cluster dissimilarity. **H,** The RAD, RadRatio, RG, RgRatio and contact frequency matrix from each of the 10 clusters of Chr8 predicted by the two-step dimension reduction method.



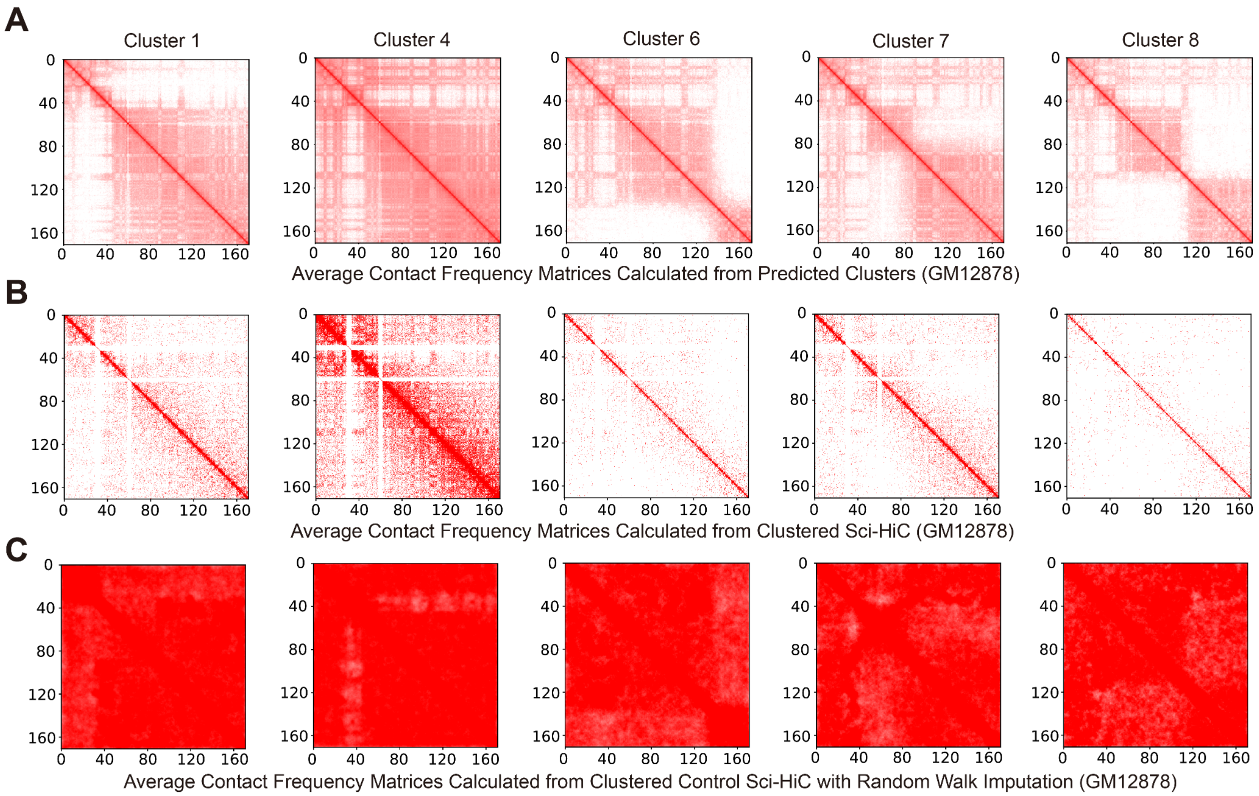
Supplementary Figure 5: Territory domains showed together with different regulation marks on Chr6 A, From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio for all clusters. **B,** From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio for cluster 7.



Supplementary Figure 6: Territory domains showed together with different regulation marks on Chr10 A, From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio for all clusters. **B,** From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio for cluster 1.



Supplementary Figure 7: Territory domains showed together with different regulation marks on Chr8 A, From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio for all clusters. **B,** From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio for cluster 3.



Supplementary Figure 8: Results of the assessment of modeled clusters shown by raw and control sci-HiC data on Chr6 A, The contact frequency matrices of clusters 1, 4, 6, 7 and 8. **B,** Results of the imputed sci-HiC assessment for different clusters. The contact frequency matrices are constructed by raw sci-HiC contact matrices[10]. **C,** Results of the imputed control sci-HiC assessment for different clusters. The contact frequency matrices are constructed by control sci-HiC contact matrices imputed by convolution and random walk with restart[11].

References

1. Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**, 115–129 (1964).
2. Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**, 2323–2326 (2000).
3. Tenenbaum, J. B., Silva, V. de & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290**, 2319–2323 (2000).
4. von Luxburg, U. A tutorial on spectral clustering. *Stat Comput* **17**, 395–416 (2007).
5. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at http://arxiv.org/abs/1802.03426 (2020).
6. van der Maaten, L. & Hinton, G. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
7. Eastwood, M. P. & Wolynes, P. G. Role of explicitly cooperative interactions in protein folding funnels: A simulation study. *J. Chem. Phys.* **114**, 4702 (2001).
8. Cheng, R. R. *et al.* Exploring chromosomal structural heterogeneity across multiple cell lines. *Elife* **9**, e60312 (2020).
9. Ramdas, A., Garcia, N. & Cuturi, M. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. Preprint at http://arxiv.org/abs/1509.02237 (2015).
10. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* **14**, 263–266 (2017).
11. Zhou, J. *et al.* Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci U S A* **116**, 14011–14018 (2019).