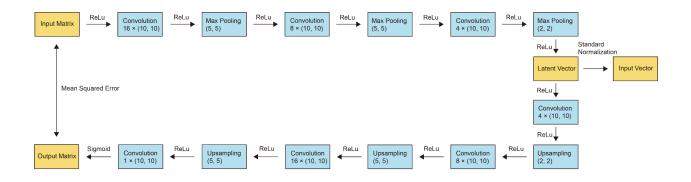
SUPPLEMENTARY INFORMATION

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

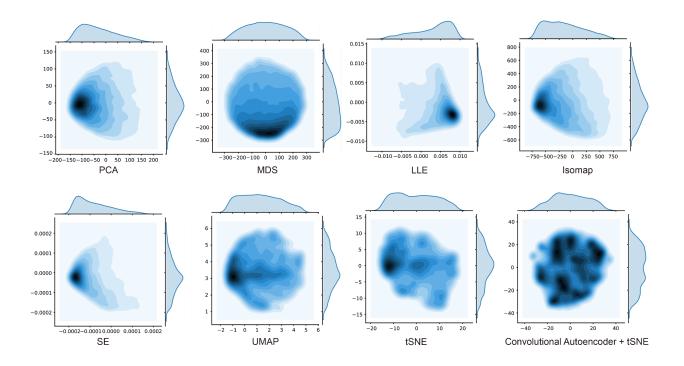
Yuxiang Zhan^{1,2,3}, Asli Yildirim^{1,2}, Lorenzo Boninsegna^{1,2}, Frank Alber^{1,2,3*}

¹Department of Microbiology, Immunology, and Molecular Genetics, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095 ²Institute of Quantitative and Computational Biosciences, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095 ³Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

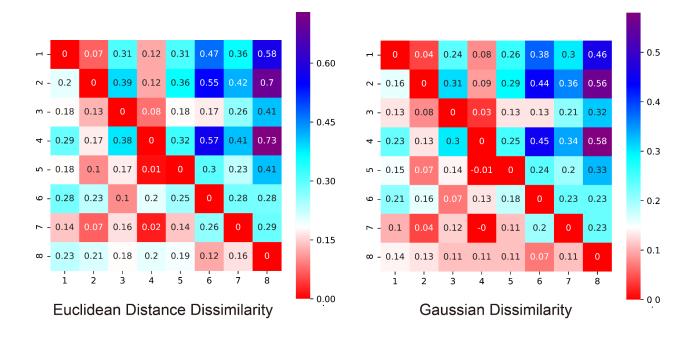
*Correspondence should be addressed to F.A. falber@g.ucla.edu



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, MDS¹, LLE², Isomap³, SE⁴, UMAP⁵, tSNE⁶. 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 distance⁹. 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 and the intra-cluster 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

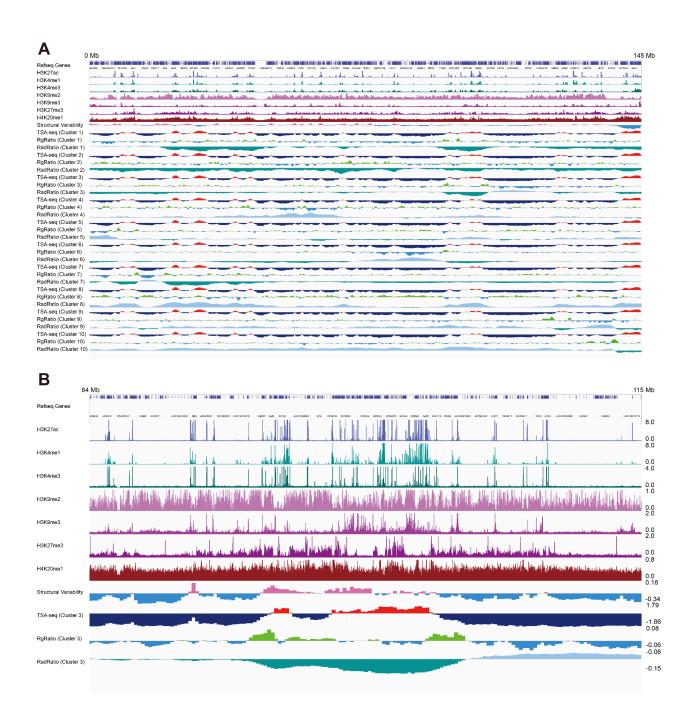
intra-cluster dissimilarity. **H**, The RAD, RadRatio, RG, RgRatio and contact frequency matrix from each of the 10 clusters of Chr8 predicted by the two-step dimension reduction method.

CIPL DIP GUNZ CIRE FORCE CITS INDON LICS INC. W2 DIPKI CUPIT TWE TUNCE INC. MADE TUNE TO CONTROL OF THE CONTROL OF THE TUNE TWO TO CONTROL OF THE CONTROL OF
kendika alkalakik kati kalikendan kalaki eta kanda k
h hade an helder wir her wirdelin allan her ster i se i se an eine se gleier an dersend har ein her allan das d
erreite het fei herr blie de Mandel Antoles er in de de er er er het de er
have been been and the second s
والمراجع والمحالي المحاوية المحاور والمحاور والمحاول والمحا
منافع والمرد في ومرد والمرد والمرد والمرد والمرد والمرد والمرد والمرد المرد والمرد المراكبة والمرد والمرد المراكبة والمرد والم
a server and a server a serv

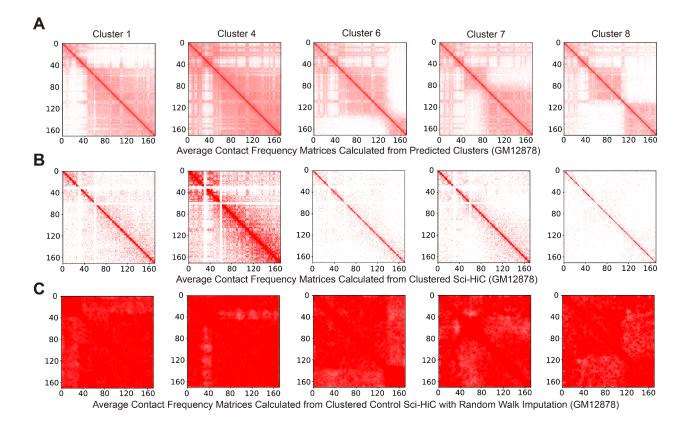
Supplementary Figure 5: Territory domains showed together with different regulation marks on Chr6 A, From the top to the bottom, the displayed features are refseq genes, H3K27ac, H3K4me1, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H4K20me1, the ensemble structural variability, TSA-seq, RgRatio and RadRatio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

	223378 MIR6078 LINP1 LINCO2660	DCDC2 IFOAD CACHER	NEDL OTUDI MYCOA MAC 2NF4	NE PARCE TACCIPI	KORSP1 DOPP1 00F2 CHAT PRO	91 GAPOIREI DICCI THENES CY	P2001P SIRTI PPA1 MOU SAMDI	I DLGS PLACE NESS LING	CO1520 RNL9 HTR7 NIP7P1	ZMISIBA CINIMI OSA SI	NI SORCEI ADDE	ACILIS TRUDI EMOR	PUPPI POTK TE	DOB 19987 MIR
		hala ar	يغ بنيا <u>الم</u>	سل الـــــــــــــــــــــــــــــــــــ	inde Harribanika. Bariban birak	i i i hiller	مريان الله من التي من المريد. المريد المرابع التي من المريد المريد المريد المريد المريد المريد المريد المريد ال		Lecher Land	الله		a data ana	علم بالع	
		l turi de la conte	المان في القان عن مراقداً. المان المعادر ومعاد فقا عنه رجالي	later de la com	والمتناب وطالعه	الله المانية بنائة الاربي منقطة	ىر الماران بى يامان . ئەر بارى ئەر ئەرىيە	یا او او ا	المرائد والمراجع	litter obtailer	di di	at <mark>ha at</mark> a		A Lat
	and the second second	the second second second	make a thread		and the second second	and the state		- Andrewski - Andrewski -	and de alike and	he all as		الم الم	ni in n	La Anna Alla
	مرد المردية فلا والحر ما ماهي الأرب فلا أرب	a state or		and a second s	and the first set			and the second s	and a strength of the	د. د در الله استنظام د. د ده دولدار استنظار ا	un en entre	and the second sec	ال . الدينة (10 من من مالي	ha dha l
/ariability														
Cluster 1) luster 1)	**- <u>-</u>						_						.	
Cluster 1)														
Cluster 2) luster 2)							_			_				
Cluster 2)														
Cluster 3) luster 3)							-							
Cluster 3)														
Cluster 4) luster 4)		-					-			-				
Cluster 4)														
Cluster 5) luster 5)														
Cluster 5)		-											-	
Cluster 6) luster 6)	**									_			-	
Cluster 6)													-	
Cluster 7) luster 7)							-							
Cluster 7)														
Cluster 8) luster 8)														
	Mb													
68	Mb		Notesting () - non the not ()	1-01-100101000000000	ane e da un temp com com temp t			(13)(1000) (000 Kalansi)	N FROM T F OF DOUBLET - 1-			1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 -		
Cluster 8) 68 nes			NOTION ()	HARD LINCEDED LINCE	INTE E NA INTERNET ANNA ANNA ANNA ANNA ANNA ANNA ANNA A		N DVII HUVI HVVI ZVII ZVII	D LENKT LOOM MIGOZ SHAT	I HAVA HISP LOCIESINANS			01100 PCC0 P50		
68		USPA ACK LINDA						D LAN LOUP NISC SHIT	1 1968-452° LOCIESTRA			00000 90200 900		
68									1 9946-802 LOCIDDATA			01103 PE230 P13		7652 1993
68														
68														
68														
68														
68		laal laa sala da ah da a						a da badena ber						
6£														
68								e de solo en la composición de la compo de la composición de l de la composición de la de la composición de la de la composición de la composición de de la composición de la de la composición de la de la composición de la de la composición de la composición de de la composición de la composición de la composición de la composición de						
6£														
6£														
6£														
6£														
6£														
66 Res Variability Liuster 1)														
66														

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, H3K4me1, H3K4me3, H3K9me2, 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 matrices¹⁰. **C**, Results of the imputed control sci-HiC assessment for different clusters. The contact matrices are constructed by control sci-HiC contact matrices are constructed by control sci-HiC contact matrices imputed by convolution and random walk with restart¹¹.

References

- 1. Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**, 115–129 (1964).
- 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).
- Eastwood, M. P. & Wolynes, P. G. Role of explicitly cooperative interactions in protein folding funnels: A simulation study. *J. Chem. Phys.* **114**, 4702 (2001).
- 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).