PNAS
www.pnas.org

Supplementary Information for

Robust Single-Cell Hi-C Clustering by Convolution and Random Walk based Imputation

Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J. Sejnowski, Jesse R. Dixon, Joseph R. Ecker

Corresponding author: Joseph R. Ecker
Email: ecker@salk.edu

This PDF file includes:

Figs. S1 to S18
Table S1

www.pnas.org/cgi/doi/10.1073/pnas.1901423116

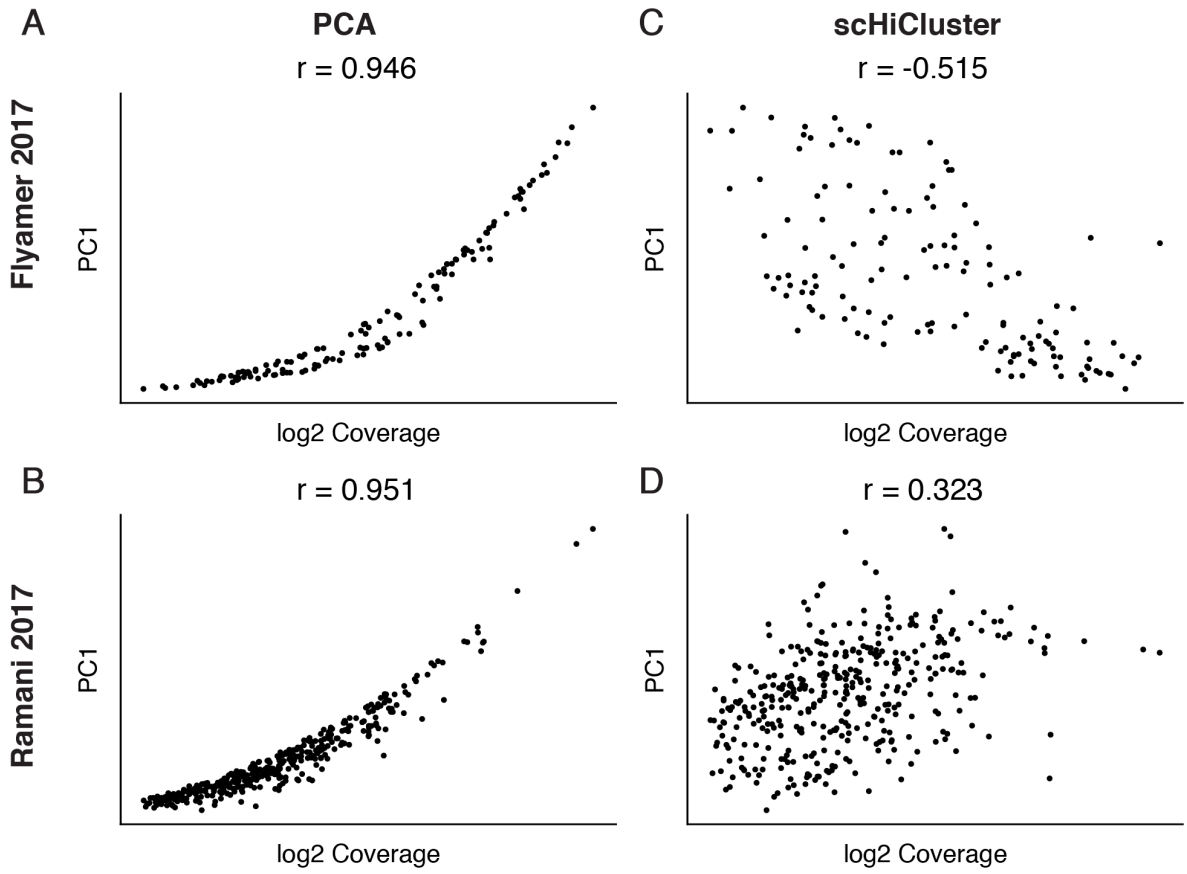

Figure S1. The correlation between PC1 and the coverage of single cells. (A, C) shows the results in the Flyamer 2017 dataset, and (B, D) shows the results in the Ramani 2017 dataset. The correlation between coverage and PC1 of PCA (A, B) and PC1 of scHiCluster embedding (C, D).

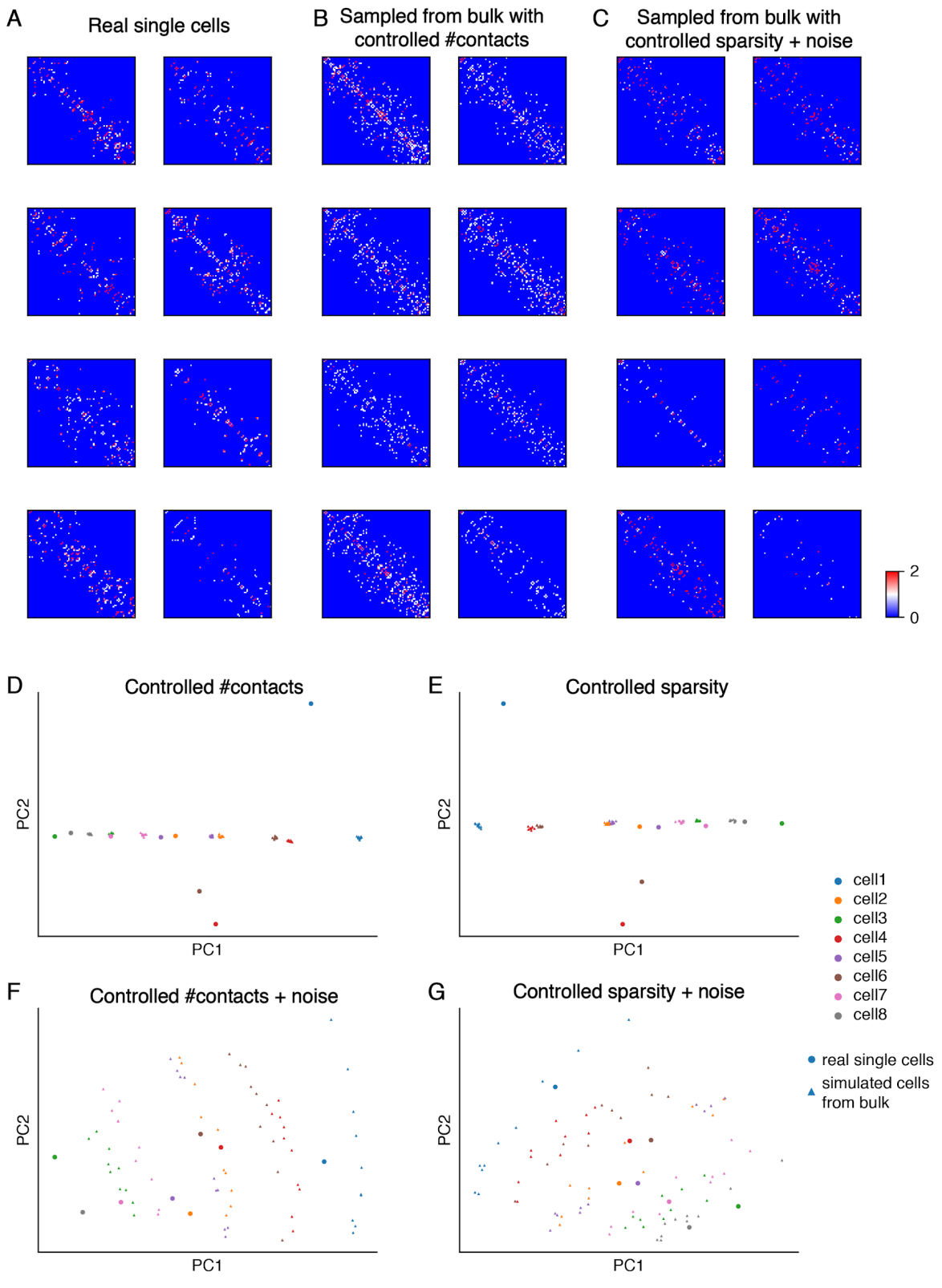
A
Real single cells
B
Sampled from bulk with controlled #contacts
C
Sampled from bulk with controlled sparsity + noise
2
0
D
Controlled #contacts
E
Controlled sparsity
PC2
PC1
F
Controlled #contacts + noise
G
Controlled sparsity + noise
cell1
cell2
cell3
cell4
cell5
cell6
cell7
cell8
real single cells
simulated cells from bulk

Figure S2. The effect of controlling sparsity and adding noise during downsampling. The data used is haploid ES cell data from Stevens et al. 2017. There are eight single cells in the dataset, with coverage ranging from 21k to 78k contacts for each cell. 1Mb bins were used to generate the contact matrices in A-C and the embeddings in D-G. The contact maps of chr1 of (A) real single-cell data , (B) simulated single-cell data produced by directly downsampling and controlling the number of contacts and (C) simulated single-cell data produced by controlling sparsity and adding noise (our final sampling method). The embeddings of downsampled cells together with real single cells in the PC1 and PC2 spaces controlling for contact numbers (D), controlling for sparsity (E), controlling for contact number and adding noise (F) and controlling for sparsity and adding noise (G). These embeddings were generated using the eight real single-cell data, and simulated data for 80 single cells downsampled from the pseudo-bulk dataset using 1Mb resolution. For each real single cell, we generated data for 10 simulated cells with the same contact number or sparsity observed for real cell data. The PCs were computed using the baseline method PCA (**Methods**). Circles represent real single cells, and triangles with the same color signify the simulated cells with the same number of contacts or sparsity.

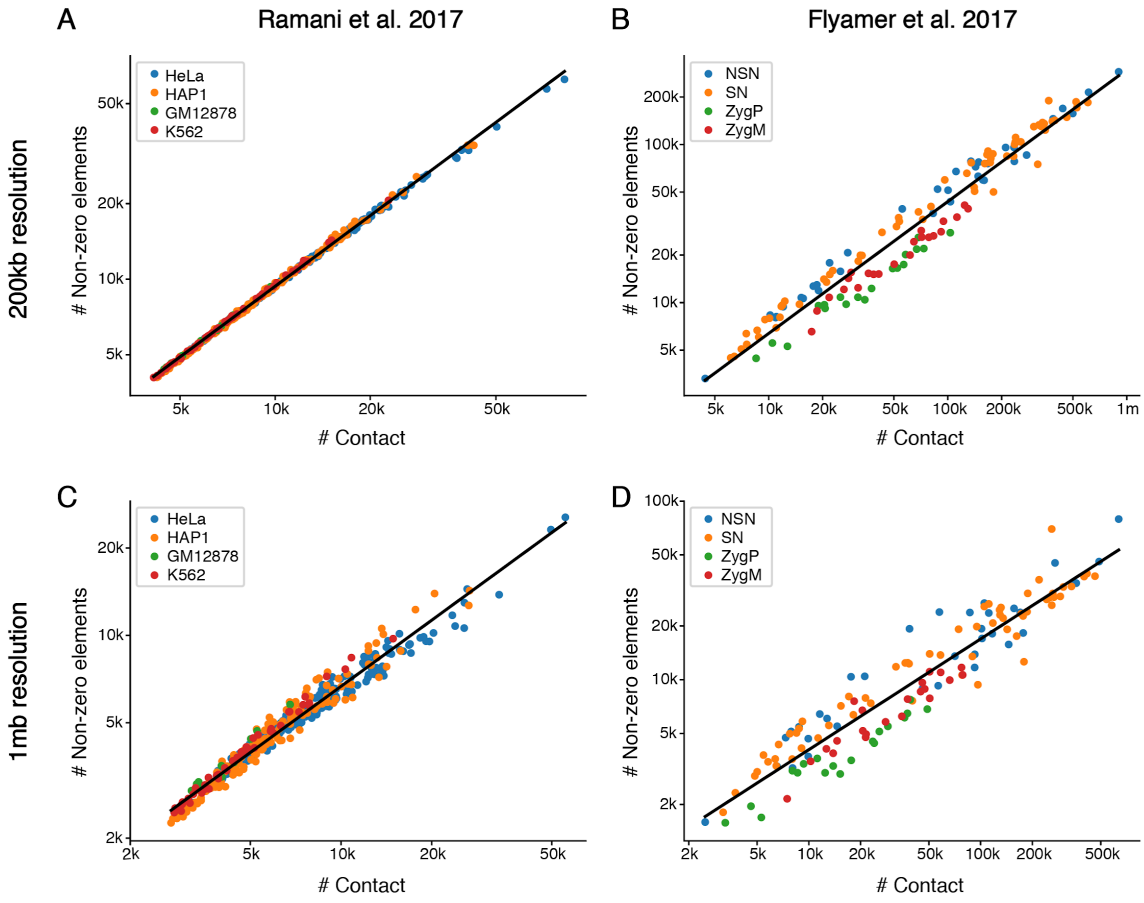

Figure S3. The relationship between total contacts and sparsity at 200kbp resolution (A, B) and 1 Mbp resolution (C, D) in Ramani et al. 2017 (A, C) and Flyamer et al. 2017 (B, D).

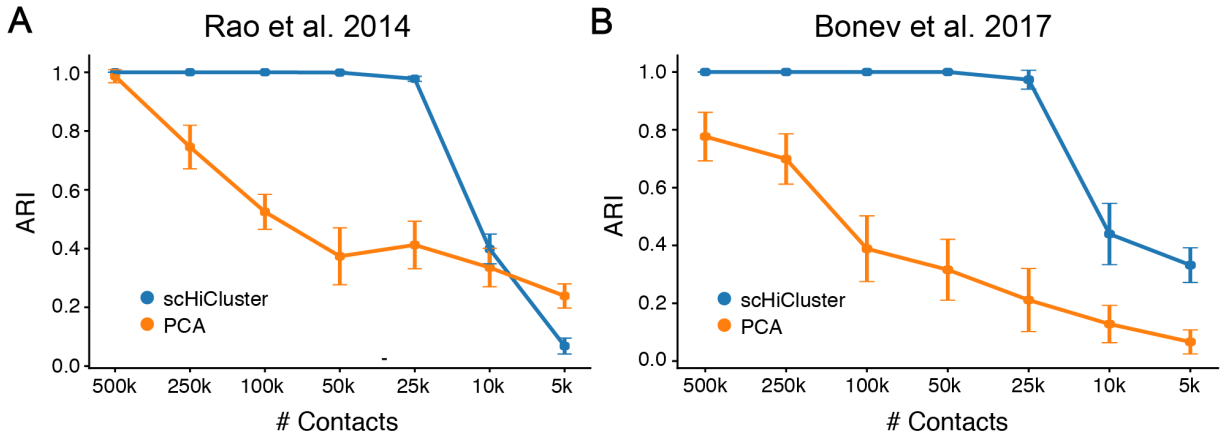

Figure S4. The performance of scHiCluster and PCA on simulated data. Bulk Hi-C contact matrices at 1Mb resolution in Rao et al. 2014 (A) and Bonev et al. 2017 (B) were sampled to 100k, 50k, 25k, 10k and 5k contacts respectively. The clustering performance is measured by Adjusted Rand Index (ARI). The error bar represents the standard error among 10 simulations at each coverage.

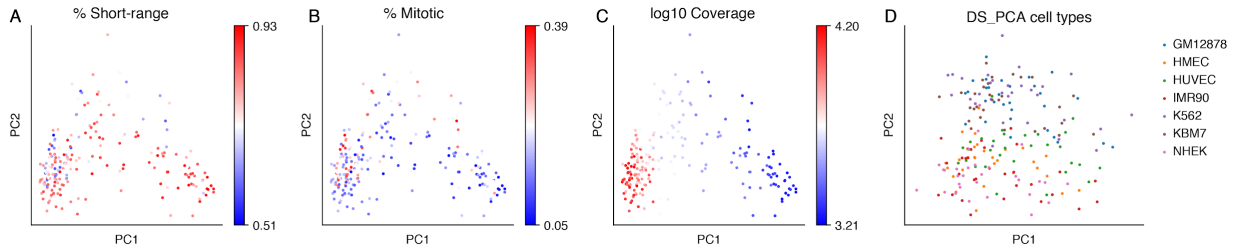

Figure S5. The effect of different bias factors on the embedding of simulated cells. Simulated cells with 5k contacts downsampled from Rao 2014 dataset are shown. Cells are colored based on the ratio of short-range contacts (<2M) among all contacts (A), ratio of mitotic contacts (between 2M and 12M) among all contacts (B), and the log2 coverage (C). (D) Embedding of the same cells by downsampling cells to the same coverage, followed by PCA.

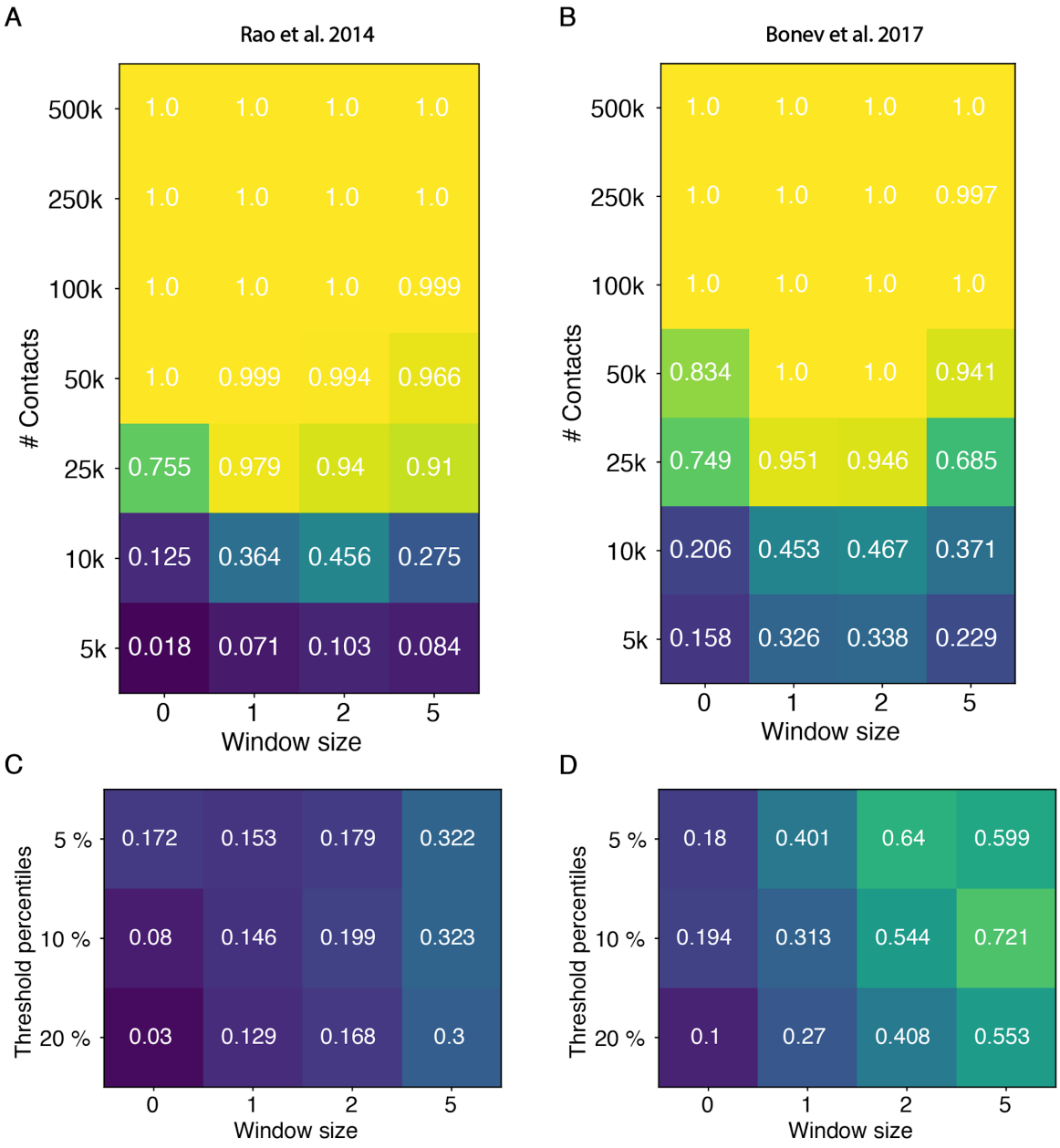

Figure S6. The performance of scHiCluster on downsampled Rao et al. 2014 (A, C) and Bonev et al. 2017 (b, d) data under several combinations of hyper-parameters. (A, B) The ARIs by different window sizes and contact numbers at 1 Mbp resolution. (C, D) The ARIs by different window sizes and threshold percentiles with 25k contacts at 200 kbp resolution. The ARI is averaged from 10 simulations with 30 cells per cell type for each simulation.

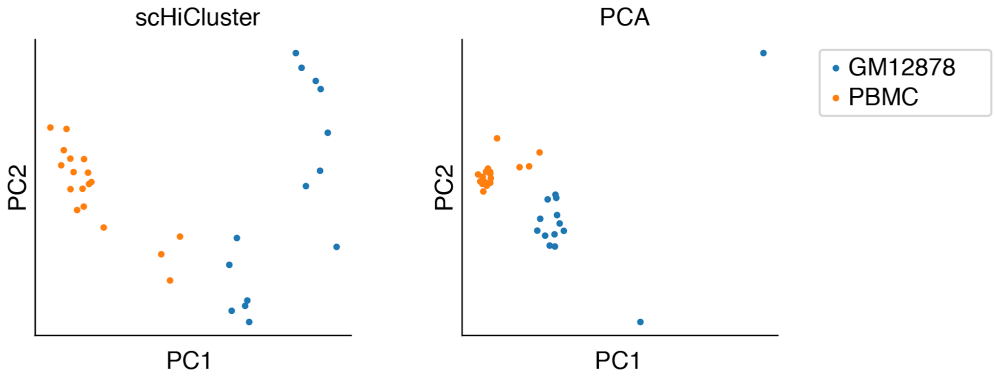

Figure S7. The embedding of single cells in Tan et al. 2018 by scHiCluster and PCA. Since the coverage of this dataset is relatively high, we used $w = 0$ for scHiCluster and did not remove PC1 for PCA.

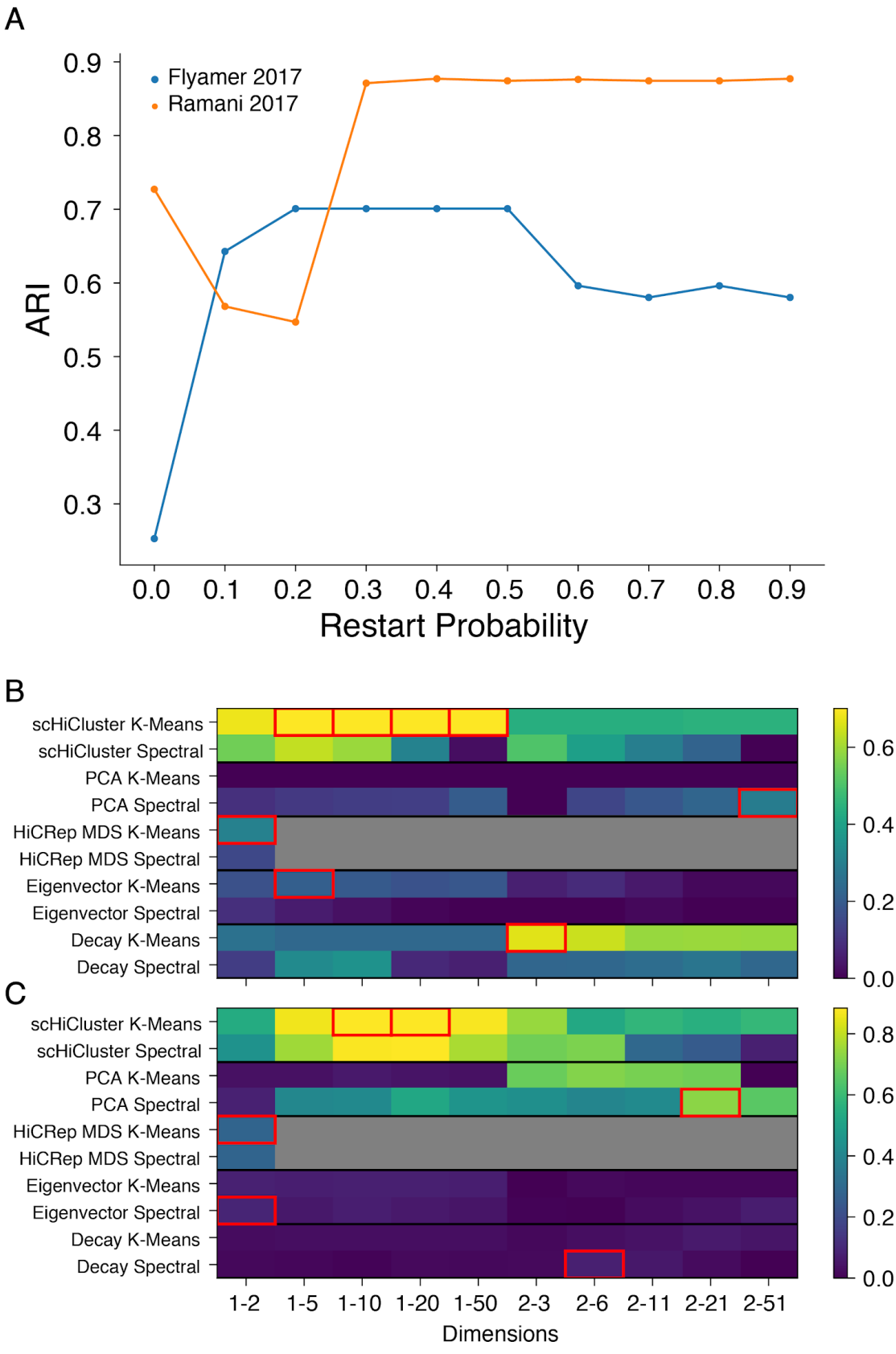
A
Flyamer 2017
Ramani 2017
ARI
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Restart Probability
B
C
scHiCluster K-Means
scHiCluster Spectral
PCA K-Means
PCA Spectral
HiCRep MDS K-Means
HiCRep MDS Spectral
Eigenvector K-Means
Eigenvector Spectral
Decay K-Means
Decay Spectral
0.6
0.4
0.2
0.0
0.8
0.6
0.4
0.2
0.0
1-2
1-5
1-10
1-20
1-50
2-3
2-6
2-11
2-21
2-51
Dimensions

Figure S8. scHiCluster performances are robust to hyperparameter selection. (A) The effect of restart probability in random walk on the ARI of clustering using Flaymer 2017 and Ramani 2017 datasets. (B, C) The performance of cell clustering measured by ARI with different clustering methods and PC dimensions. The red boxes represent the best performances of each baseline method among the different PC dimensions and clustering methods.

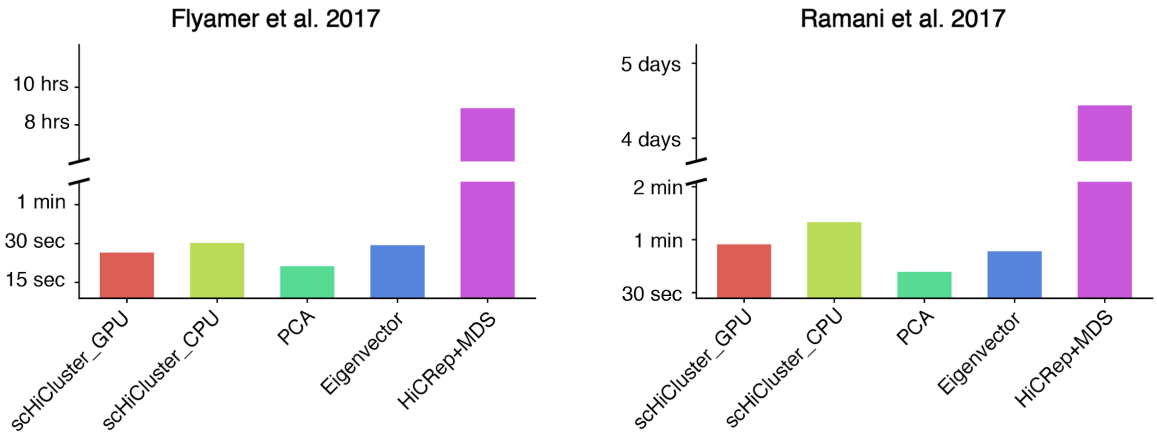

Figure S9. The running time of different methods on Flyamer et al. 2017 and Ramani et al. 2017.

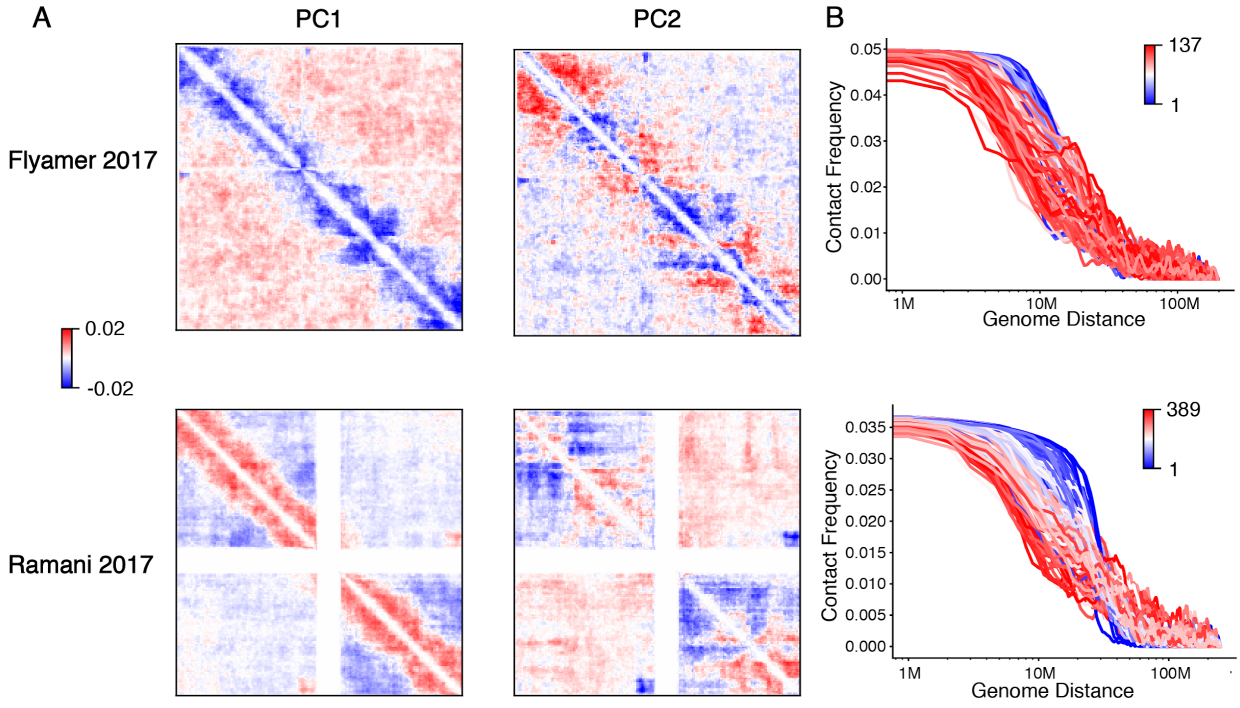

Figure S10. Interpretation of principal components on chr1. (A) The whitening matrices of PC1 and PC2 in Flyamer et al. 2017 and Ramani et al. 2017 datasets. (B) The contact frequency plotted against genome distance for each cell with the binary matrix after imputation and selecting the top 20% elements. The color from red to blue represents the rank of corresponding PC1 value of the cell from high to low.

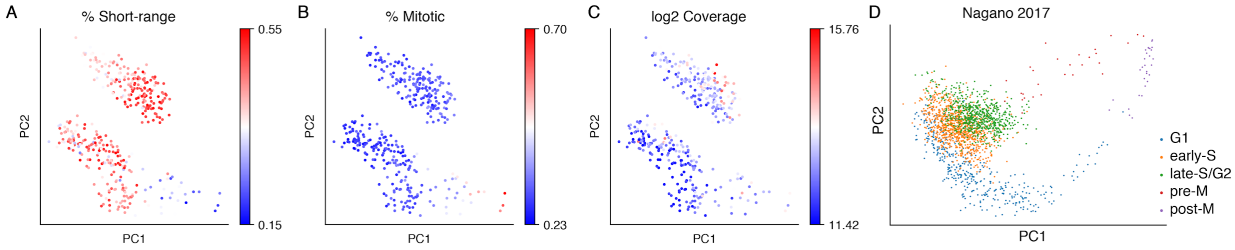

Figure S11. The effect of cell cycle stages on scHiCluster embedding. The embedding of cells from the Ramani 2017 dataset (The same data as in **Fig. 3B**) are shown. Cells are colored based on the ratio of short-range contacts (<2M) among all contacts (A), ratio of mitotic contacts (between 2M and 12M) among all contacts (B), and the log2 coverage (C). (D) The embedding of 1,992 mouse ESC in Nagano 2017 dataset by scHiCluster. Cells are colored based on their inferred stages in the cell cycle.

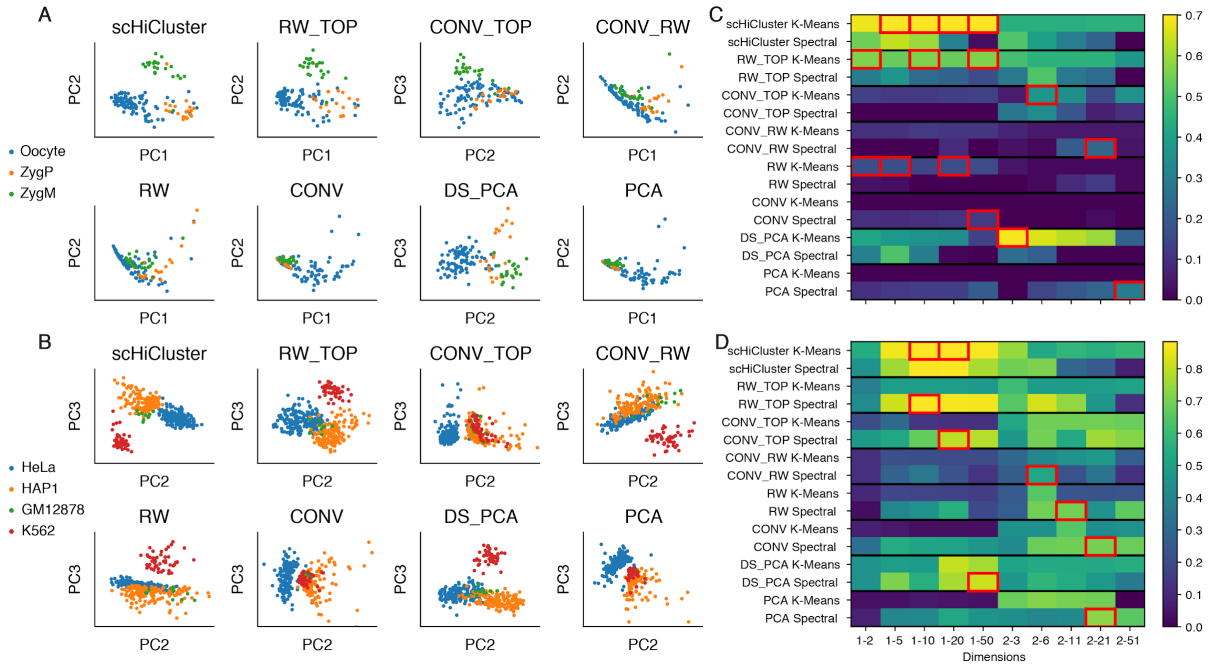

Figure S12. The performance of scHiCluster applying different combinations of substeps. For Flyamer et al. 2017 (A, C) and Ramani et al. 2017 (B, D), the embedding (A, B) and ARI of clustering (C, D) with each combination of steps are shown. (A, B) If using PC1 and PC2 provides a better clustering ARI than PC2 and PC3, then the results are shown in the space of PC1 and PC2, and vice versa. (C, D) The performance of cell clustering measured by ARI

with different clustering methods and PC dimensions. The red boxes on the heatmap (right) represent the best performances of each substep combination among different PC dimensions and clustering methods.

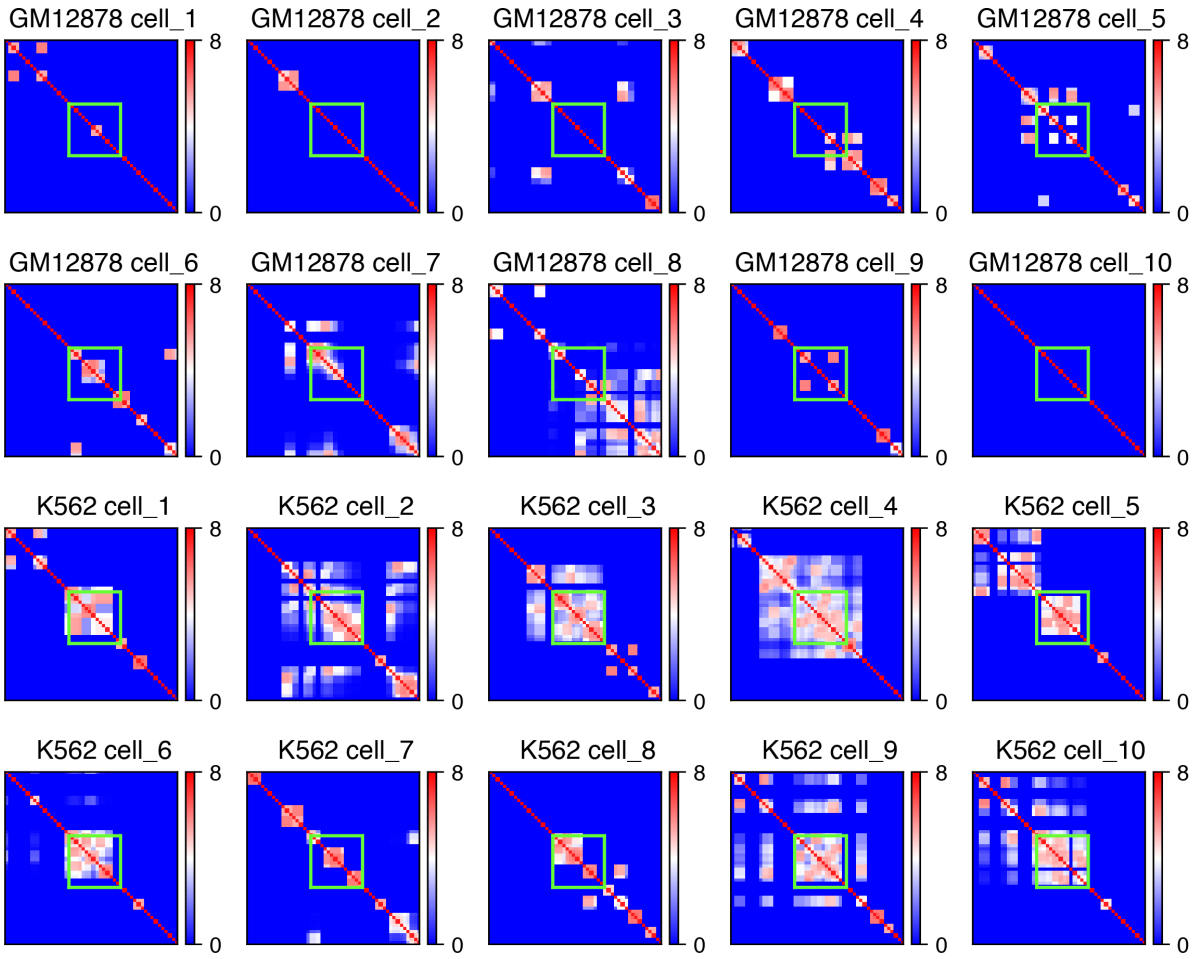

Figure S13. The genome structure surrounding chr9:133,600,000-134,200,000 (the green box) in randomly selected single cells. The whole matrices shows 2M region from 132,900,000 to 134,900,000.

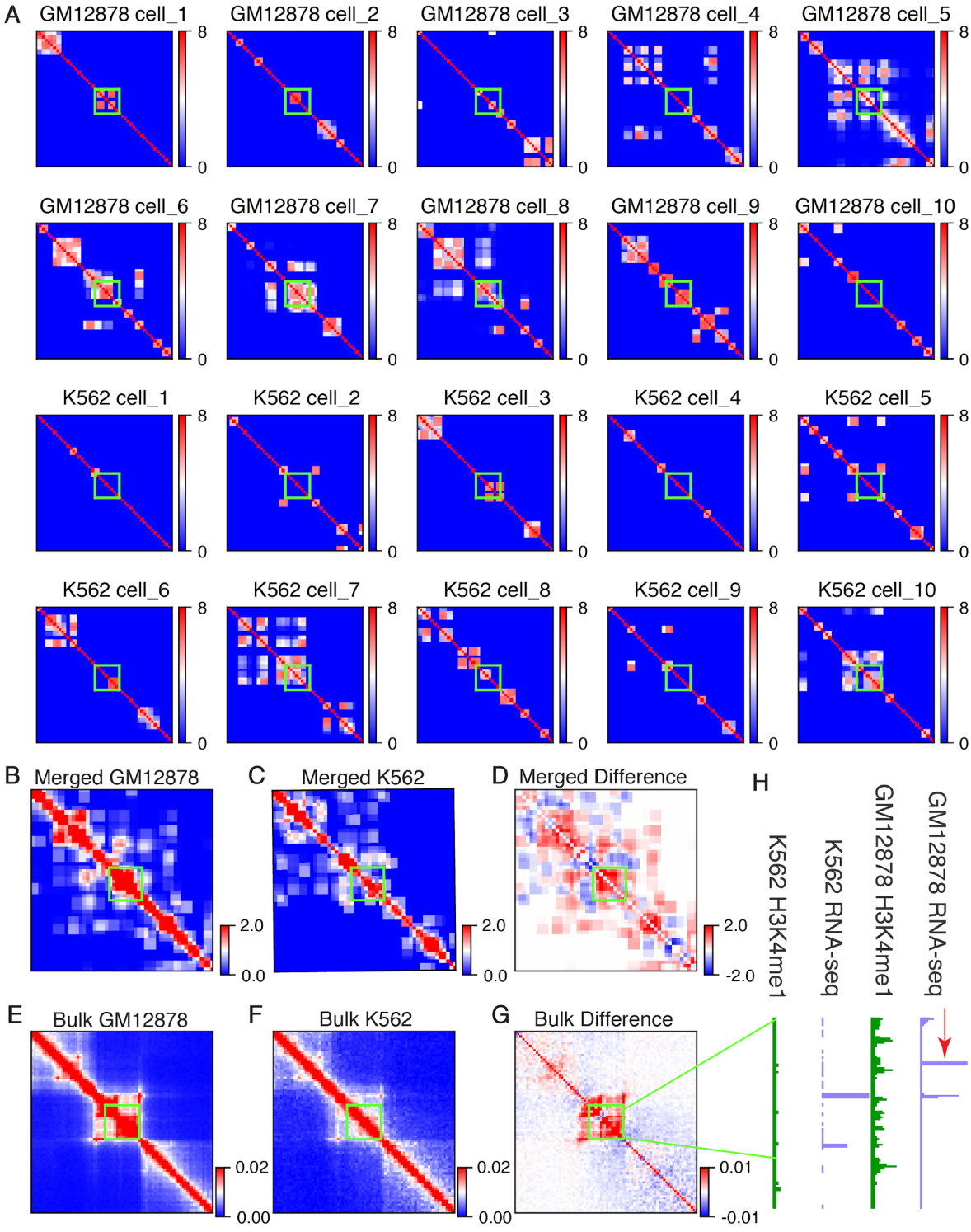

Figure S14. The genome structure surrounding CXCR4 in randomly selected single cells (A), merged single cells of GM12878 (B) or K562 (C) after imputation. (D) The difference between (B) and (C). The SQRTVC normalized contact matrices of bulk GM12878 (E) and K562 (F) cell lines. (G) The difference between (E) and (F). (H) The RNA-seq and H3K4me1 signals in both cell lines. The red arrow indicate the location of CXCR4. The green boxes

represent chr2:136,760,000-137,120,000. The whole matrices shows 2M region from 135,940,000 to 137,940,000.

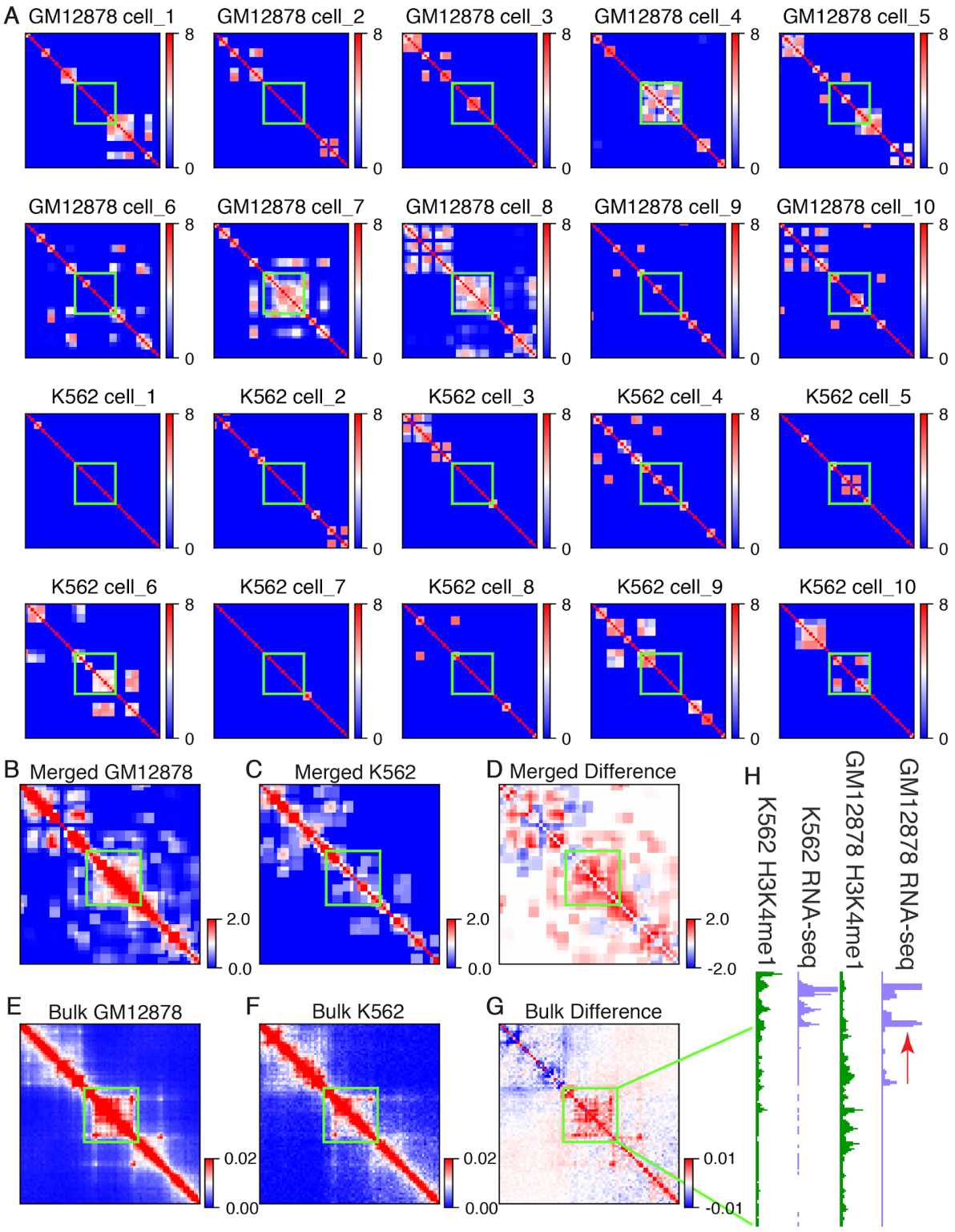

Figure S15. The genome structure surrounding ZBTB11 in randomly selected single cells (A), merged single cells of GM12878 (B) or K562 (C) after imputation. (D) The difference between (B) and (C). The SQRTVC normalized contact matrices of bulk GM12878 (E) and

K562 (F) cell lines. (G) The difference between (E) and (F). (H) The RNA-seq and H3K4me1 signals in both cell lines. The red arrow indicates the location of ZBTB11. The green boxes represent chr3:101,400,000-102,000,000. The whole matrices shows 2M region from 100,700,000 to 102,700,000.

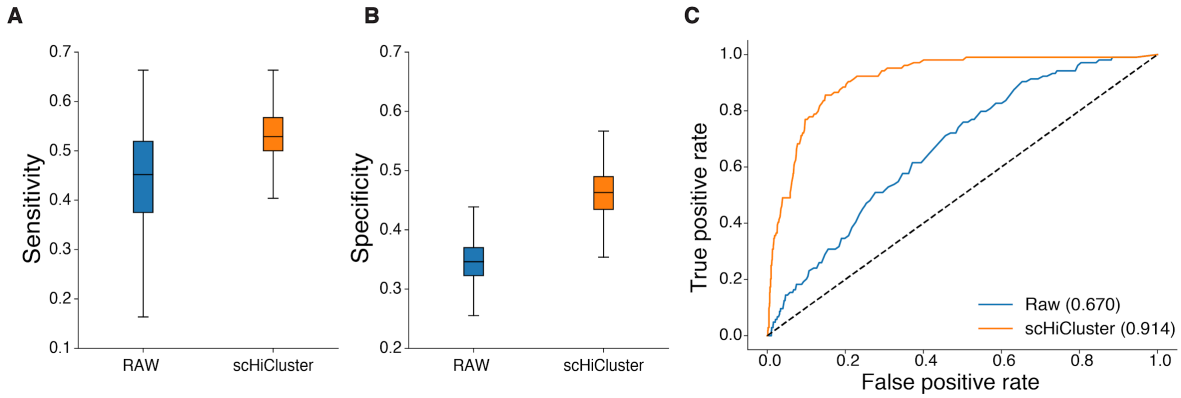

Figure S16. Comparison of TLS identified in single cells with TAD identified in bulk data. (A, B) The sensitivity and specificity of TLS boundaries in each single ESC to capture the TAD boundaries in bulk ESC before and after scHiCluster imputation. (C) The ROC plot to distinguish whether a bin is a TAD boundary using the number of cells in which the bin is identified as a TLS boundary.

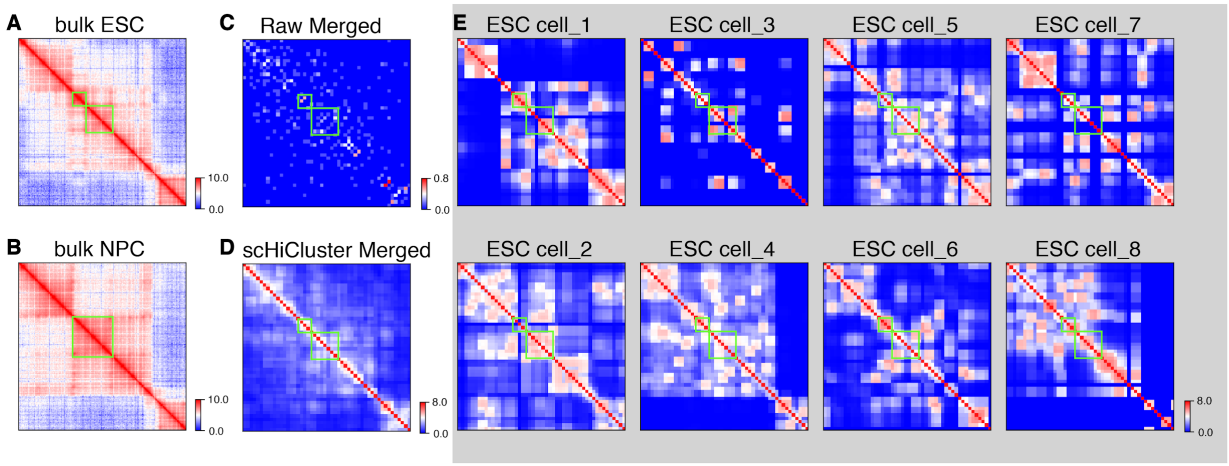

Figure S17. Visualization of contact matrices surrounding *Sox2*. The bulk contact matrices of ESC (A) and NPC (B) from Bonev et al. 2017. The merged contact matrices of 8 ESCs without imputation (C) and after imputation by scHiCluster (D). (E) The imputed contact matrices of 8 ESCs. The green boxes in ESC show the domains identified in bulk data that surrounding *Sox2* (chr3:34,640,000-34,800,000 and chr3:34,800,000-35,120,000). The gene is located at the upstream boundary of the upper left domain.

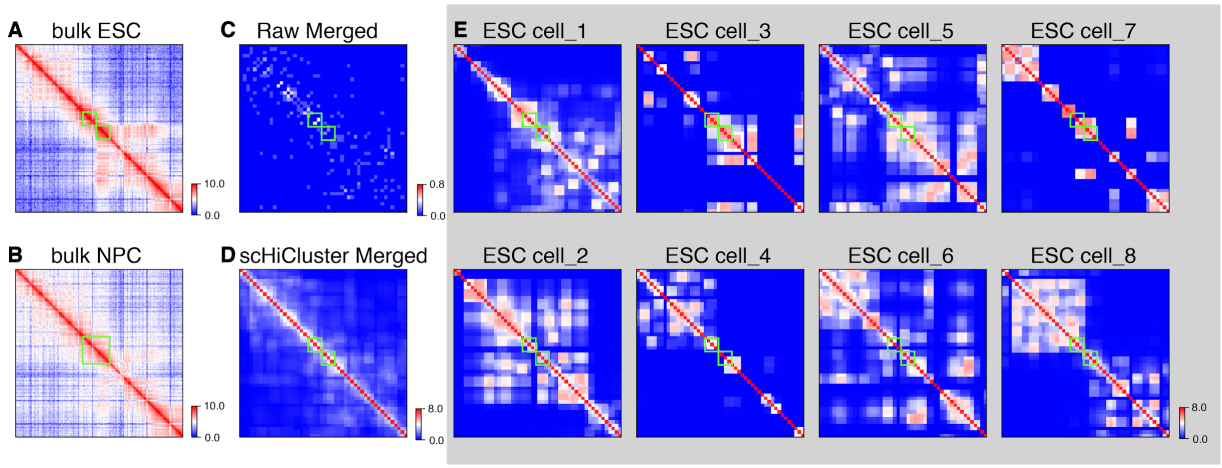

Figure S18. Visualization of contact matrices surrounding *Zfp42*. The bulk contact matrices of ESC (A) and NPC (B) from Bonev et al. 2017. The merged contact matrices of 8 ESCs without imputation (C) and after imputation by scHiCluster (D). (E) The imputed contact matrices of 8 ESCs. The green boxes in ESC show the TADs identified in bulk data that surrounding *Zfp42* (chr8:43,200,000-43,360,000 and chr8:43,360,000-43,520,000). The gene is located at the intersecting boundary of the TADs.

Table S1. The number of cells from Flyamer 2017 and Ramani 2017 datasets after quality control. NSN and SN are two types of oocytes representing non-surrounded nucleolus and surrounded nucleolus. ZygP and ZygM are paternal and maternal allele of zygotes, respectively. The contact filter refers to the criterion that there should be more than 5,000 contacts per cell. The chromosome contact filter refers to the criterion that the number of contacts on a chromosome of length $x$mb should be greater than x.

Dataset	Cell type	# Cells in total	# Cells after contact filter	# Cells after chromosome contact filter
Flyamer 2017	NSN	44	36	35
	SN	70	61	61
	ZygP	24	20	19
	ZygM	31	24	22
	Total	169	141	137
Ramani 2017	HeLa	258	236	181
	HAP1	214	165	148
	GM12878	44	13	10
	K562	110	57	50

	Total	626	471	389