Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation

Jingtian Zhou[a,b,1], Jianzhu Ma[c,1], Yusi Chen[d,e,1], Chuankai Cheng[f], Bokan Bao[b], Jian Peng[g], Terrence J. Sejnowski[d,e], Jesse R. Dixon[h], and Joseph R. Ecker[a,i,2]

[a]Genomic Analysis Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; [b]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA 92093; [c]Department of Medicine, University of California San Diego, La Jolla, CA 92093; [d]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; [e]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093; [f]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093; [g]Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [h]Peptide Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; and [i]Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92037.

Contributed by Joseph R. Ecker, May 20, 2019 (sent for review January 30, 2019; reviewed by Anshul Kundaje and Jennifer E. Phillips-Cremins)

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Three-dimensional genome structure plays a pivotal role in gene regulation and cellular function. Single-cell analysis of genome architecture has been achieved using imaging and chromatin conformation capture methods such as Hi-C. To study variation in chromosome structure between different cell types, computational approaches are needed that can utilize sparse and heterogeneous single-cell Hi-C data. However, few methods exist that are able to accurately and efficiently cluster such data into constituent cell types. Here, we describe scHiCluster, a single-cell clustering algorithm for Hi-C contact matrices that is based on imputations using linear convolution and random walk. Using both simulated and real single-cell Hi-C data as benchmarks, scHiCluster significantly improves clustering accuracy when applied to low coverage datasets compared with existing methods. After imputation by scHiCluster, topologically associating domain (TAD)-like structures (TLSs) can be identified within single cells, and their consensus boundaries were enriched at the TAD boundaries observed in bulk cell Hi-C samples. In summary, scHiCluster facilitates visualization and comparison of single-cell 3D genomes.

single cell | Hi-C | 3D chromosome structure | random walk

In recent years, there has been a rapid increase in the development of single-cell transcriptomic and epigenomic assays (1), including single-cell/nucleus RNA sequencing (RNA-seq) (2), assay for transposase-accessible chromatin using sequencing (ATAC-seq) (3, 4), bisulfite sequencing (5), and Hi-C (6–11). Such powerful techniques allow the study of unique patterns of molecular features that distinguish each cell type. Computational methods have been developed to identify different cell types in heterogeneous cell populations based on various molecular features such as transcriptome (12, 13), methylome (14), and open chromatin (15–17). However, unbiased and efficient algorithms for single-cell clustering based on 3D chromosome structures are limited. In previous studies, cells have been organized by their contact decay profiles, which is useful for distinguishing different stages of the cell cycle (9). However, separating different cell types at the same cell cycle stage is still challenging. Principal-component analysis (PCA) performed on both intrachromosomal and interchromosomal reads was unable to completely distinguish between four cancer cell lines (7). Tan et al. (11) showed that annotated features in bulk Hi-C data could be used to separate single-cell Hi-C data into corresponding cell types. However, this approach would be limited to features identified in the few tissues or cell lines with published Hi-C data, and may be difficult to generalize to unprofiled cell types. Several methods have been developed to examine the reproducibility of bulk Hi-C data, which mainly focus on computing different types of similarity scores between contact matrices (18–21). These methods have been benchmarked by Yardimci et al. (22), and HiC-Rep was found to perform the best when generalized to single-cell Hi-C data. An embedding method for single-cell Hi-C data based on HiCRep has been specifically designed for capturing structural dynamics of the cell cycle state (23). However, cell cycle state is continuous in nature, and this approach has not explicitly been tested for the purpose of clustering, and thus it remains unclear how well this method would perform for cell type identification from single-cell Hi-C data.

Clustering of single cells based on Hi-C data faces three main challenges. 1) Intrinsic variability. 3D chromosome structures are highly spatially and temporally dynamic. Imaging-based technologies have suggested a large degree of heterogeneity of chromosome positioning and spatial distances between loci even within a population of the same cell type (24–27). How this fluctuation between cells of the same cell type compares to fluctuations between different cell types remains unclear. 2) Data sparsity. The sparsity of single-cell Hi-C data are higher than most other types of single-cell data. State-of-the-art single-cell DNA assays typically cover only 5–10% of the linear genome. Since Hi-C data are represented as 2D contact matrices, this level of sensitivity leads to coverage of only 0.25–1% of all

Significance

Chromosomes are compactly folded in nuclei, and their specific 3D structures play a role in the regulation of gene expression. While cell type specificity of gene regulation has been revealed through transcriptomic and epigenomic assays, comprehensive analysis of genome conformation patterns in different cell types is still lacking. Single-cell approaches have facilitated our understanding of cell type heterogeneity, and profiling chromosome architecture at the single-cell level has been achieved using Hi-C. However, unbiased and efficient computational methods are needed to distinguish different cell types utilizing these data. Here, we describe scHiCluster, a computational framework to study cell type-specific chromosome structural patterns. We demonstrate that scHiCluster allows clustering of single cells with high accuracy and identifies their local chromosome interaction domains.

Author contributions: J.Z., J.M., J.R.D., and J.R.E. designed research; J.Z., Y.C., C.C., and B.B. performed research; J.Z. analyzed data; and J.Z., J.M., J.P., T.J.S., J.R.D., and J.R.E. wrote the paper.

Reviewers: A.K., Stanford University; and J.E.P.-C., University of Pennsylvania.

Conflict of interest statement: J.R.E., J.R.D., and A.K. are coauthors on a 2015 research article. J.R.E. and A.K. are coauthors on a 2017 correspondence article.

Data deposition: scHiCluster has been deposited on GitHub (https://github.com/zhoujt1994/scHiCluster).

[1]J.Z., J.M., and Y.C. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ecker@salk.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1901423116/-/DCSupplemental.

Published online June 24, 2019.

contacts to be captured. 3) Coverage heterogeneity. It is often observed that the genome coverage of cells extends over a wide range within a single-cell Hi-C experiment. We find this bias often acts as the leading factor to drive clustering results, making it difficult to systematically eliminate. For example, this bias could be alleviated by removing the first principal component (PC1) before clustering and visualization. However, PC1 is not guaranteed to represent only cell coverage in these experiments as it may also contain information related to other biological variables (*SI Appendix*, Fig. S1 *A* and *B*).

To address these challenges, we developed a computational framework, scHiCluster, to cluster single-cell Hi-C contact matrices. To overcome the sparsity problem, we performed two steps of imputation on the chromosome contact matrices to better capture the topological structures. To solve the heterogeneity problem, we selected only the top-ranked interactions after imputation, which were proved to be sufficient to represent the underlying data structure. This framework significantly improved upon the clustering performance using low coverage datasets as well as facilitated the visualization and comparison of chromosome interactions among single cells.

Results

Overview of scHiCluster. As shown in Fig. 1, scHiCluster consists of four major steps. In the first step, every element of the contact matrix is replaced by the weighted average of itself and its surrounding elements, in a type of linear convolution. Then a random walk (with restart) algorithm (28) is applied to smooth the signal to further capture both the local and global information of the contact maps. In particular, the convolution step only allows the information to pass among the linear genome neighbors, while the subsequent random-walk step aids information sharing among the network neighbors. To alleviate the bias introduced by uneven sequence coverage, we only keep the top 20% interactions after the imputation (*SI Appendix*, Fig. S1 *C* and *D*). Finally, we project the processed contact matrices onto a shared low-dimensional space, so that the topological structure of the 3D chromosome contacts can be compared between cells and used for further clustering and visualization.

scHiCluster Improves Clustering Performance on Simulated Data. To explore the combinatorial effects of different levels of coverage and resolution, we first applied our algorithm to a set of simulated single-cell Hi-C data. We noticed that direct sampling from the Hi-C contact matrices of bulk cells leads to a relatively lower sparsity and heterogeneity (*SI Appendix*, Fig. S2), which often yields more accurate clustering results compared with real single-cell data. The real data concentrated more on specific loci in each cell, and the individual loci were different between different cells (*SI Appendix*, Fig. S2*A*). On the contrary, the simulated cells from bulk data often had more evenly distributed contacts *SI Appendix*, Fig. S2*B*). Therefore, we controlled the sparsity of each simulated contact matrix and added noise to the contact–distance curves to better mimic the sparsity and noise of real data (*Methods*). As shown in *SI Appendix*, Fig. S2*G*, when considering the first two principal components (PCs), the simulated cells generated were indistinguishable from real single cells of the same cell type.

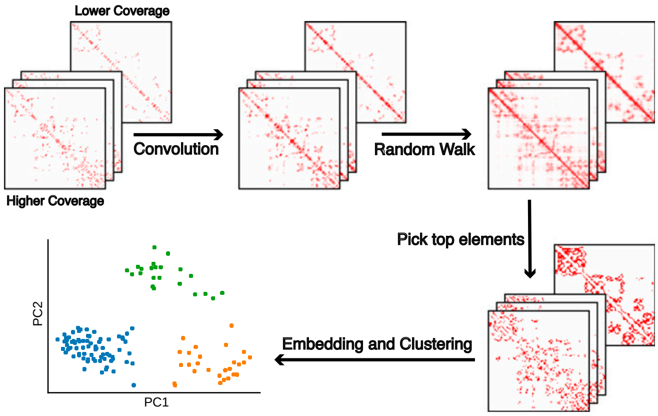


Fig. 1. The workflow of scHiCluster. The contact matrices of each single cell are smoothed by two steps of imputation that include convolution and random walk; these are based on the neighboring bins of a linear genome and long-range connections, respectively. To alleviate the coverage bias, only top 20% elements of the imputed matrices are selected. All single-cell matrices are then projected into the same space, and then clustering is performed to identify distinct cell types.

In our simulation, we performed downsampling from bulk Hi-C experimental data from two studies. Rao et al. (29) examined seven human cell types (GM12878, IMR90, HMEC, NHEK, K562, HUVEC, and KBM7), while Bonev et al. (30) examined three mouse cell types [embryonic stem cells (ESCs), neural progenitor cells (NPCs), and cortical neurons (CNs)]. We downsampled each dataset to 500 k, 250 k, 100 k, 50 k, 25 k, 10 k, and 5 k contacts, respectively, and used 1-Mbp and 200-kbp resolution contact maps to test our algorithm. At each coverage level and resolution, we generated 30 simulated cells for each cell type. We evaluated the ability of scHiCluster compared with PCA to recover the correct cell type in an unsupervised way. The adjusted Rand index (ARI) was used to measure the accuracy of clustering. As shown in Fig. 2 and *SI Appendix*, Fig. S4, in both datasets, scHiCluster consistently performed better than PCA. The performances of scHiCluster began to be impaired with fewer than 25 k contacts, and failed to remove the coverage bias at 5 k contacts (*SI Appendix*, Fig. S5*C*), which leads to a complete loss of clustering ability. We also found that 1-Mbp resolution performed better than 200 kbp (*SI Appendix*, Fig. S6 *C* and *D*), suggesting that lower sparsity (lower resolution) may be sufficient to distinguish cell types. Thus, we used 1-Mbp resolution in all subsequent experiments.

scHiCluster Has Superior Performance on Published Single-Cell Hi-C Data. Next, we evaluated our analysis framework using authentic single-cell Hi-C datasets. Thus far, there have been three published studies focusing on single-cell chromosome structures with analyses of multiple cell types. Ramani et al. (7) used a combinatorial indexing protocol to generate single-cell Hi-C libraries from thousands of cells for four human cell lines (HeLa, HAP1, GM12878, and K562). The number of contacts captured in each cell ranged from 5.2 k to 102.7 k (median, 10.0 k). Flyamer et al. (10) performed whole-genome amplification after ligation and detected 6.6 k to 1.1 m contacts per cell (median, 97.3 k) in mouse zygotes and oocytes. Tan et al. (11) developed an optimized protocol also using whole-genome amplification and obtained data with a median coverage of 513.0 k contacts. Since the last benchmark dataset (Tan) had relatively high coverage, either simple PCA (*SI Appendix*, Fig. S7) or chromosome compartment score (11) easily allowed cell types to be distinguished. Due to cost considerations, it is still challenging to achieve such depth of genome coverage. Therefore, we focused on the first two datasets with lower coverage (Ramani and Flyamer) to test the utility of our computational framework.

We compared our algorithm with four baseline methods: PCA, HiCRep+MDS (23), the eigenvector method along with the decay profile method (9) (*Methods*). Besides the methods used in published works, we included the eigenvector method since the chromosome compartments are considered to be cell type specific based on the bulk Hi-C experiments, and the first eigenvector of contact matrix is widely used to represent these compartment features (29, 31, 32). scHiCluster outperformed the baseline methods on both datasets in terms of better visualization

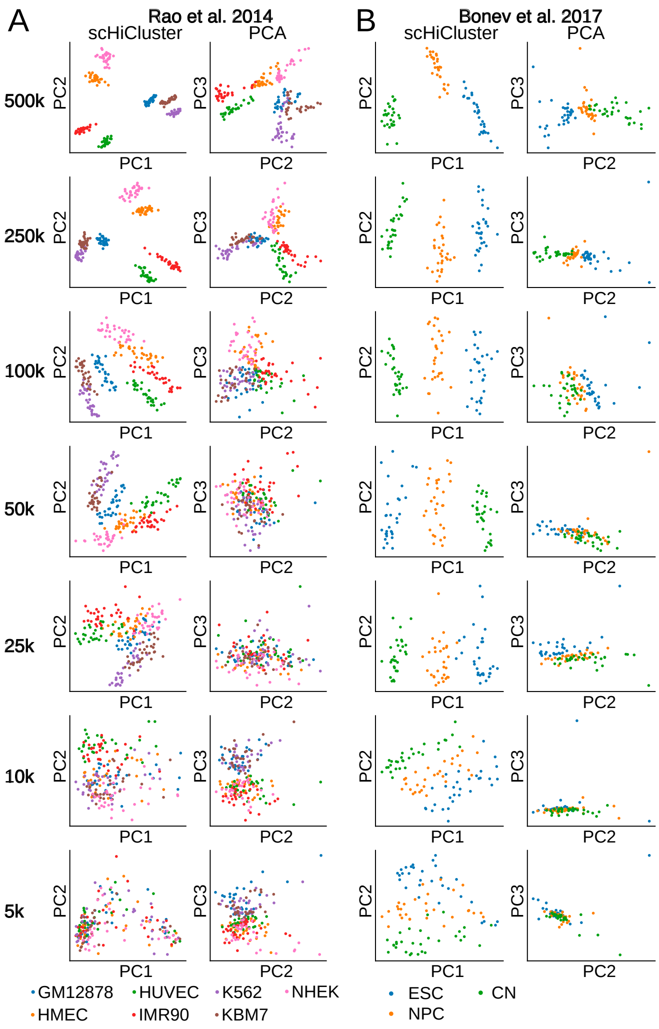


Fig. 2. The embedding of simulated single cells at 1-Mb resolution from Rao et al. (29) (*A*) and Bonev et al. (30) (*B*) with different contact numbers. For each dataset, the embedding by scHiCluster is shown on the *Left* and by PCA is shown on the *Right*.

(Fig. 3 *A* and *B*) and improved ARI (Fig. 3 *C* and *D*). In the mouse dataset (Flyamer), scHiCluster made a significant distinction among all three cell types (Fig. 3*A*); while in the human dataset (Ramani), the algorithm separated K562 and HAP1 better in the first two PC dimensions (Fig. 3*B*). The performances of scHiCluster are robust to the parameters (*SI Appendix*, Fig. S8). It is also worth commenting on the scalability of each method. Since HiCRep is designed specifically for two-sample comparison rather than multiple samples, generating the similarity matrix using HiCRep involves many repetitive computations, which required 8 h (Flyamer) and 4.5 d (Ramani); whereas scHiCluster and other methods consumed ~30 s (Flyamer) and 60 s (Ramani) (*SI Appendix*, Fig. S9). Additionally, we carried out the same experiments on each chromosome separately and noticed that almost every chromosome showed advanced separation on the mouse dataset (Fig. 3*E*), while only one chromosome showed significant improvement on the human dataset (Fig. 3*F*). These results may suggest that to separate cells using global chromosome structure differences (e.g., oocytes and zygotes), the information provided by a single chromosome might be sufficient, but to distinguish more complex cell types, a combination of different chromosomes or a more careful feature selection is necessary.

We also visualized the weights of each element in the contact matrices when computing the final PCs (whitening matrices). In general, the weights for PC1 were uniformly distributed parallel to the diagonal (*SI Appendix*, Fig. S10*A*), which suggested it captures the information of the contact–distance curve and might correspond to the variance resulting from cell cycle or other relevant biological effects (9). This is also corroborated by the observation that cells with greater PC1 values tended to have a higher frequency of short-range contacts, while smaller PC1 inclined to correspond to a higher frequency of long-range contacts (*SI Appendix*, Fig. S10*B*). On the contrary, the weights for computing PC2 showed region specificity (*SI Appendix*, Fig. S10*A*), which may indicate its correlation with compartment strength. These findings also explained why the oocytes and zygotes in Flyamer et al. (10) are dominantly separated by PC1 (Fig. 3*A*), where the contact distance curves differ between cell types; meanwhile, in Ramani et al. (7), PC2 achieved a better partition of the cancer cell lines (Fig. 3*B*), but PC1 separated a cluster of cells likely in M-phase (*SI Appendix*, Fig. S11 *A* and *B*). We further examined the ability of scHiCluster to capture stages of the cell cycle by embedding the Nagano et al. (9) dataset, which contains 1,992 mouse ESCs across different stages of cell cycle. As shown in *SI Appendix*, Fig. S11*D*, the cell cycle information is generally well preserved.

Next, we wanted to evaluate the contribution of each step to the final clustering performance. For the three major steps of the pipeline, we tested all possible combinations of one or two steps of the three. More specifically, we compared our framework with PCA (with none of the steps), DS_PCA (downsampling to uniform coverage), CONV (convolution only), RW (random walk only), CONV_TOP (convolution and select top elements), RW_TOP (random walk and select top elements), and CONV_RW (convolution and random walk). Notably, for the whole scHiCluster framework including all of the three steps, we used K-means for 10 PCs to assign the cluster labels. However, to fully exploit the potential of the baseline methods, we compared all of the different combinations of clustering methods and numbers of PCs, and identified the parameters generating the most accurate results. From *SI Appendix*, Fig. S12, we concluded that all three steps are necessary to achieve the current visualization (*SI Appendix*, Fig. S12 *A* and *B*) and clustering accuracy (*SI Appendix*, Fig. S12 *C* and *D*). The necessity of these steps is more evident when using the mouse dataset.

scHiCluster Allows Visualization of Structural Difference in Single Cells. The most popular method to interpret and validate identified cell clusters in single-cell experiments is to analyze known marker genes. Gene expression is directly measured in single-cell RNA-seq data and promoter, gene body ATAC-seq signals or cytosine methylation ratios can also be used to infer the cluster-specific genes in single-cell open chromatin and methylome data. Similarly, in single-cell Hi-C data, the differential chromosome interactions could serve as cell-type markers. With the single-cell Hi-C data, imputed contact matrices from every single cluster can be merged, where we observed square patterns that are visually similar to the topologically associating domains (TADs) identified in bulk Hi-C experiments along the diagonal. However, since the existence of TADs remains unclear in single cells, and accurate identification of the structures were limited by data sparsity, we referred to this featured pattern as TAD-like structures (TLSs) hereafter. Thus, differential TLSs could be applied to characterize different cell types. For instance, as demonstrated in Fig. 4 and *SI Appendix*, Fig. S13, a TLS at chr9:133.6M-134.2M is observed in 9 of 10 K562 cells but in 2 of the GM12878 cells. This structure difference is concordant with the bulk Hi-C data from the same cell lines. Gene expression and H3K4me1 signals that mark active enhancers are also higher in K562 within this TLS.

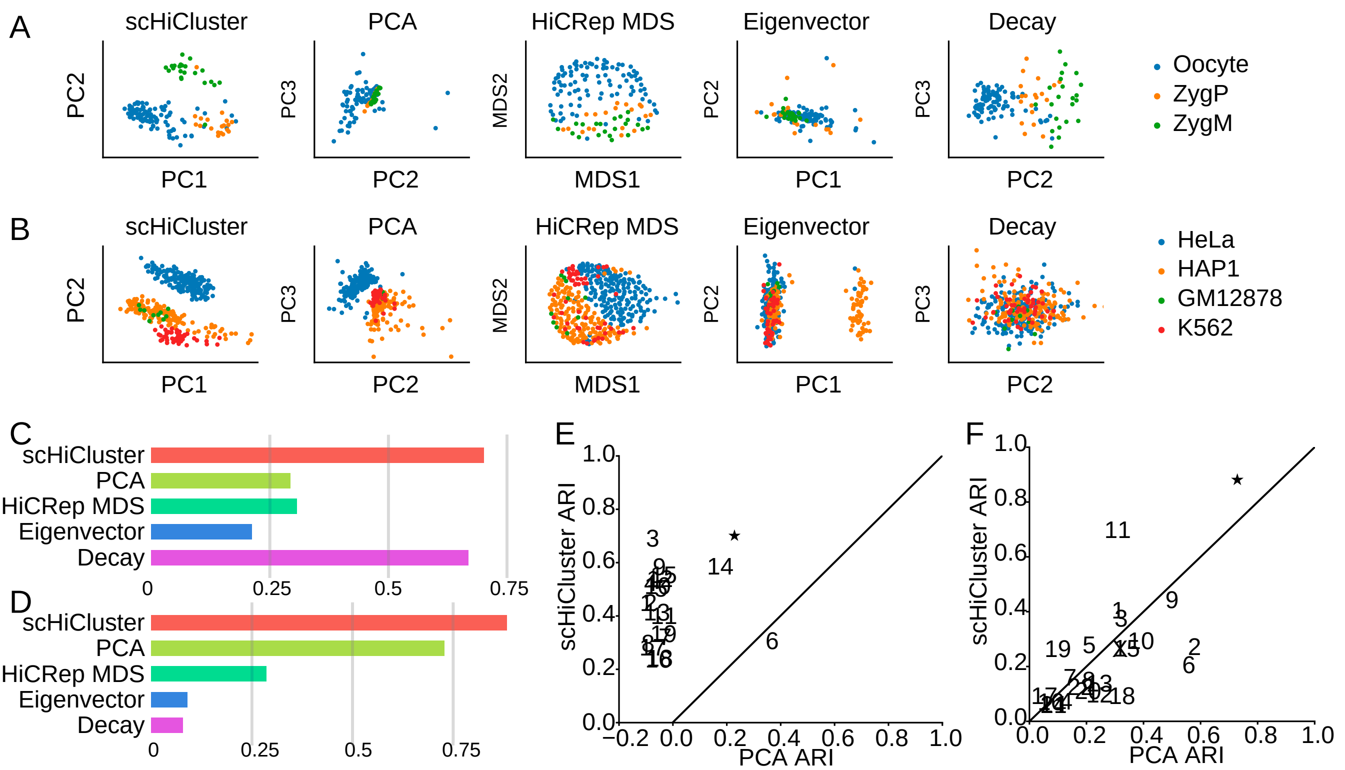


Fig. 3. The performance of scHiCluster and baseline methods on real single-cell Hi-C data. For Flyamer et al. (10) (*A*, *C*, and *E*) and Ramani et al. (7) (*B*, *D*, and *F*), the embedding (*A* and *B*) and ARI of clustering (*C* and *D*) are shown. For scHiCluster and eigenvector, the embeddings are shown in PC1 and PC2 space. For PCA and decay profile, the embedding is shown in PC2 and PC3 space. (*E* and *F*) The performance of scHiCluster and PCA on each chromosome (indicated by chr. numbers). The ARI using all chromosomes (indicated by star symbol).

Structural differences are also observed near differentially expressed genes between the two cell types, including *CXCR4* and *ZBTB11*. *CXCR4* is a chemokine receptor that enhances cell adhesion, which is highly expressed in noncancer cells (GM12878) comparing to cancer cells (K562) (33). With scHiCluster imputation, a TLS surrounding *CXCR4* was detected in 6 of 10 GM12878 cells but only 2 of 10 K562 cells (*SI Appendix*, Fig. S14 and *Methods*). Intriguingly, an H3K4me1 peak was detected in bulk GM12878 but not K562 at the other boundary of the TLS, which may indicate the potential interaction between the gene and its enhancer. Similarly, a TLS whose boundary located at *ZBTB11* was observed in more GM12878 cells than K562 cells (*SI Appendix*, Fig. S15). Consistently, more H3K4me1 peaks within this TLS were also detected in the bulk GM12878 sample.

Next, we examined whether the imputation based on scHiCluster could facilitate the systematic identification of TLSs in both simulated and real single cells. We first leveraged Bonev et al. data for bulk ESC and NPC, and downsampled them to 1-Mbp, 500-kbp, 250-kbp, 100-kbp contacts per cell. We applied scHiCluster on contact matrices and then ran TopDom (34) to detect TLSs in every single cell. A TAD in NPC that splits into two TADs in ESC was selected to test the performance of TLS-calling (Fig. 5*A*). The visualization of single-cell TLSs was significantly improved after scHiCluster smoothing (Fig. 5*B*), and the alternative boundary was

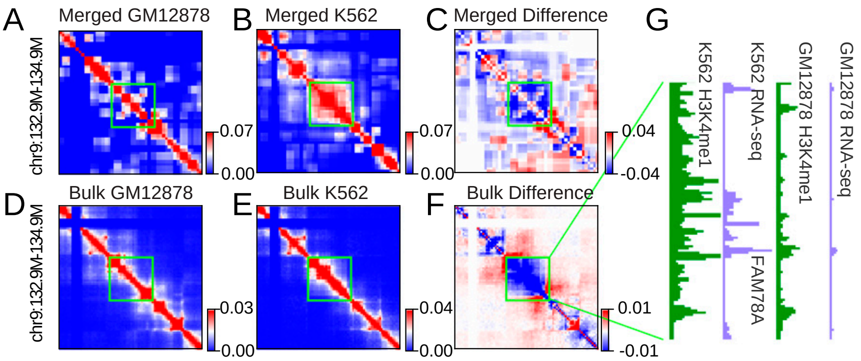


Fig. 4. Visualization of contact matrices surrounding chr9:133,600,000–134,200,000 (the green box) in merged single-cell data and bulk data. The whole matrix shows a 2 M region from 132,900,000–134,900,000. The contact matrices were created by merging of 10 GM12878 cells (*A*) or K562 cells (*B*) after imputation. (*C*) The difference between *A* and *B*. SQRTVC normalized contact matrices from bulk GM12878 (*D*) and K562 (*E*) cell lines. (*F*) The difference between *D* and *E*. (*G*) Corresponding RNA-seq and H3K4me1 signals near genes from the indicated TAD region for both cell lines.

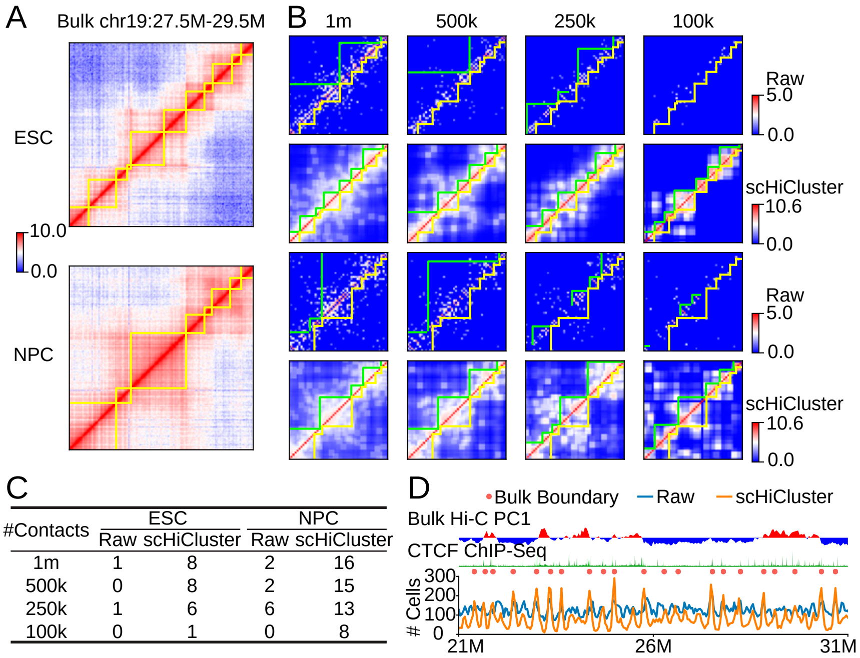
captured in more cells (Fig. 5*C*). Next, we applied scHiCluster to analyze single-cell Hi-C data from Nagano et al. (9). The dataset was sequenced with high coverage and enabled us to statistically analyze the dynamic of TADs location within single cells. We identified TLSs in contact matrices smoothed by scHiCluster at 40-kbp resolution with TopDom, and on average, observed 46% of the boundaries of TLSs in each single cell covered 53% of the boundaries identified in bulk cell data (*SI Appendix*, Fig. S16 *A* and *B*). Next, for each bin, we counted the number of cells in which the bin was determined as a TLS boundary. We observed nonzero probability for almost all bins to be a TLS boundary in single cells, and these probabilities peaked at the CTCF binding sites, and the TAD boundaries described in bulk Hi-C (Fig. 5*D*), which is in agreement with the conclusions of a recent imaging study (35). This signal was significantly enhanced after convolution and random walk (Fig. 5*D* and *SI Appendix*, Fig. S16*C*), which further highlighted the potential application of scHiCluster to study single-cell chromosome structure.

Our imputation method also helps visualize the signature of chromatin structures within specific cell type. *Sox2* is a classic marker gene of ESCs, and the chromosome structure around this gene is unique to ESCs (30). Specifically, *Sox2* is located at the upstream boundary of a large TAD in NPCs (*SI Appendix*, Fig. S17*B*), which is split into two smaller TADs in ESCs (*SI Appendix*, Fig. S17*A*). Stevens et al. (8) carried out Hi-C analysis of eight single haploid mouse ESCs. A median of 49.4-kbp long-range intrachromosome contacts was detected (21.0 k to 78.0 k). Although this study provided superior coverage among the current single-cell Hi-C experiment, the limited number of cells examined made it difficult to observe the interaction pattern surrounding the *Sox2* even if contact matrices from all cells are merged (*SI Appendix*, Fig. S17*C*). However, after the imputation using the scHiCluster framework, the TLS boundaries at downstream of *Sox2* are observed in four of the eight cells (*SI Appendix*, Fig. S17*E*). Merging the imputed matrices reveals the known domain splitting pattern near *Sox2* (*SI Appendix*, Fig. S17*D*). A similar interaction pattern is also observed near another ESC marker *Zfp42* (*SI Appendix*, Fig. S18).

Discussion

To advance our understanding of the role of genome structure in cell type-specific gene regulation, new computational tools are needed for exploration of single-cell Hi-C data. We describe a computational approach for cell type clustering, scHiCluster, that requires only sparse single-cell Hi-C contact data. In the scHiCluster framework, the chromosome interactions are considered as a network. The contact information is first averaged in the linear genome. A random walk is then used to propagate the smoothed interaction throughout the graph and further reduce the sparsity of the single-cell contact matrices. scHiCluster performed significantly better than existing methods in clustering single-cell data into constituent cell types and facilitated identification of local chromosome interaction domains.

A major challenge in clustering single-cell Hi-C data is the sparsity of the contact matrices. Our results demonstrate that scHiCluster is robust to sparse contact matrices when there are at least 5 k contacts detected per cell (*Methods*). scHiCluster takes advantage of both a linear smoothing and a random-walk step to handle these sparse data. Similar methods have been utilized for smoothing bulk Hi-C data, including HiCRep, which took the average of genome neighbors before computing the correlation of two Hi-C matrices (18), and GenomeDISCO, which provided a network representation of Hi-C matrices and used random walk to smooth it (19). Liu et al. (23) systematically evaluated these methods for single-cell Hi-C data embedding. However, since they used a cell similarity matrix that is embedded



#Contacts	ESC		NPC	
	Raw	scHiCluster	Raw	scHiCluster
1m	1	8	2	16
500k	0	8	2	15
250k	1	6	6	13
100k	0	1	0	8

Fig. 5. scHiCluster facilitates identification of domain-like structures (TLSs). (*A*) The contact matrices at chr19:27.5M-29.5M of bulk Hi-C data with alternative TADs in ESC and NPC. (*B*) The downsampled contact matrices with 1-Mbp, 500-kbp, 250-kbp, and 100-kbp total contacts per cell before and after scHiCluster imputation. The green lines indicate the TLSs called from the plotted matrix, and the yellow lines represent the TADs called from bulk data of the corresponding cell type. (*C*) The number of downsampled ESCs with TLSs at chr19:28170000–28530000 and chr19:28530000–28770000, or downsampled NPCs with TLSs at chr19:28170000–28770000 being identified before and after scHiCluster imputation. (*D*) The number of single ESCs (1,007 in total) with TLSs boundary identified at each genome bin are shown by lines. The position of TADs boundaries identified in bulk data are presented by dots. The CTCF ChIP-seq signal from ENCODE is shown in the green track. The PC1 of bulk ESC Hi-C matrix is shown on the *Top*.

by multidimensional scaling (MDS), the data are generally continuous under their low-dimensional representation and are unable to present explicit clusters for each cell type. Our scHiCluster framework combines the advantage of both HiCRep and GenomeDISCO and provides a flexible pipeline to resolve the clustering of Hi-C data, where some components (e.g., embedding) can be further tuned and improved when the algorithm is applied to more specific and challenging situations such as tissues with greater cell type complexity.

Published single-cell Hi-C datasets have employed cell lines that contain relatively large 3D genomic structural differences, simplifying the cell clustering problem. In practice, heterogeneous tissues with more closely related cell types, such as brain tissue, might pose a much greater challenge than cell lines. For cell clustering using complex tissues, further improvements in the clustering algorithm and feature selection are necessary. For instance, hierarchical clustering could be applied to identify the coarse cell types using megabase-scale resolution, followed by dividing cell types into finer scale (subtypes) using matrices of a smaller bin size. An alternative approach would be to simultaneously profile 3D genome architecture along with other "omic" information in the same cell, such as jointly profiling chromatin conformation and DNA methylation (36, 37). While such single-cell multiomic data modalities may provide the information content necessary to deconvolute cell types while preserving 3D structural information (38), they can also be more costly to perform, and more technically challenging to carry out.

We noted that the smoothing and random walk steps aid in visualization of chromosome contact maps in single cells. Such visualization can facilitate analysis of the variability in features of 3D genome organization between cells. Previous studies using bulk cell lines have reported the existence of several 3D structural features: megabase-level A/B compartments, submegabase-level TADs, and kilobase-level loops (29, 31, 39, 40). In our study, visualizing the smoothened scHiCluster results revealed the existence of TLSs in specific cells. The boundaries of these structures were variable between cells. However, the boundaries shared between TLSs in individual cells corresponded to TAD boundaries identified in bulk Hi-C studies. These results would support recent imaging studies (35), which suggested that TLSs exist in single cells, and their boundaries in individual cells are variable but nonrandom.

Methods

Data Processing. For Ramani et al. (7), interaction pairs and cell quality files of combinatorial single-cell Hi-C library ML1 and ML3 were downloaded from GSE84920. Interaction pairs for Flyamer et al. (10), Stevens et al. (8), and Tan et al. (11) were downloaded from GSE80006, GSE80280, and GSE117876, respectively. Interaction pairs for diploid ESC cultured with 2i in Nagano et al. (9) were accessed from https://bitbucket.org/tanaylab/schic2/src/default/. Given a chromosome of length L and a resolution r, the chromosome is divided into $n = L/r$ nonoverlapping bins. Hi-C data are represented as a $n \times n$ contact matrix A, where A_{ij} denotes the number of read-pairs supporting the interaction between the ith and jth bins of the genome. For each dataset, contact matrices were generated at 40-kbp and 1-Mbp resolutions for each chromosome and each cell. Total contacts of the cell were counted as the nondiagonal interaction pairs in intrachromosomal matrices. As quality control, we ruled out the cells with less than 5 k contacts. Also, for a single chromosome whose length is x Mb, we required the number of contacts to be greater than x, to avoid the chromosomes with too few contacts. We only kept cells where all chromosomes satisfied this criterion. The number of cells remaining after each quality control step for each cell type is shown in *SI Appendix*, Table S1. Generally, we suggest to apply scHiCluster only on the cells that passed these quality controls.

Simulations.

Rationale. First, we used the single-cell Hi-C dataset from Stevens et al. (8) to test the similarity between the real single-cell data and the pseudo–single-cell data, simulated by downsampling. The eight single-cell contact matrices of chromosome 1 are shown in *SI Appendix*, Fig. S2*A*. We merged the data from these eight single cells to generate a pseudobulk dataset, and then generated a simulated single-cell dataset by downsampling from the pseudobulk dataset. We added a constraint to let the number of sampled contacts equal to the number of contacts observed for each real single cell. However, we observed a side effect of this operation in that the sparsity and heterogeneity of the simulated data were much lower than that observed for real single-cell data (*SI Appendix*, Fig. S2*B*). Therefore, we limited the sparsity when performing the downsampling. After controlling the sparsity of the contact matrices, we used PCA to visualize the simulated cell data together with the real single-cell data and found that the lower heterogeneity of the simulated data was still observed in the first two PCs. Specifically, we observed variation of cells in PC1, which is highly correlated with the coverage of these cells, while only real single-cell data showed variation in PC2, but not the simulated cell data (*SI Appendix*, Fig. S2 *D* and *E*). To address this problem, we added a random noise during the simulation to amplify the heterogeneity of the contact decay curves among the single cells. The combination of these two steps enabled the simulation to generate cells with high sparsity (*SI Appendix*, Fig. S2*C*) and indistinguishable from the real single cells (*SI Appendix*, Fig. S2*G*).

Bulk Hi-C data. We downsampled bulk Hi-C data to simulate datasets with similar sparsity and heterogeneity of single cells. Bulk MAPQ30 contact matrices were extracted from Juicebox at 100-kbp resolution for the datasets of Rao et al. (29) and Bonev et al. (30), respectively. Contact matrices for each cell type at 200-kbp and 1-Mbp resolution were calculated by merged bins in the 100-kbp resolution matrices.

Normalization. SQRTVC normalization was applied to the bulk contact matrices to deal with the coverage bias along the genome. The normalized contact matrices B are computed by the following:

$$B = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \qquad [1]$$

where D is a diagonal matrix where each elements D_{ii} is the sum of the ith row of A.

Sparsity controlling. We further controlled the sparsity during sampling to make the simulated data more similar to the real data. Leveraging Ramani et al. (7) and Flyamer et al. (10) datasets, we fit a linear relationship between total contacts C and sparsity S at log scale (*SI Appendix*, Fig. S3):

$$\log S = a \log C + b. \qquad [2]$$

To generate a simulated dataset with the median contact counts to be M, for each simulated single cell we uniformly sampled t from $\log M - 0.5$ to $\log M + 0.5$ and set the total contacts number of the cell as $C = e^t$. The sparsity of the cell S was computed based on ref. 2. The sampled new contacts are randomly assigned to different chromosomes based on the contact numbers of each chromosome in a particular cell type in the bulk cell dataset.

Adding random noise. We added noise to the contact frequency through contact–distance curve, which describes the values in the contact matrices changed with respect to their distance to the diagonal. More specifically, we generated a random vector R of length n, where n is the bin number of the contact matrix. The values in R range from $-k$ to k following a uniform distribution, where k denotes the noise level. Then, the normalized bulk contact matrix B was rescaled linearly to the noisy representation E by $E_{ij} = B_{ij} \times R_{|j-i|}$. Finally, based on E, we sampled S positions to be nonzero candidates based on Eq. **2**, and distributed the C simulated contacts to these positions.

scHiCluster.

Convolution-based imputation. Imputation techniques are widely adopted in single-cell RNA-seq data to improve the data quality based on the structure of the data itself. For scHiCluster, the first step is to integrate the interaction information from the genomic neighbors to impute the interaction at each position. The missing value in the contact matrix could be due to experimental limitations of material dropout, rather than no interactions. Since the genome is linearly connected, our hypothesis is that the interaction partners of one bin may also be close to its neighboring bins. Thus, we used a convolution step to inference these missing values. Specifically, given a window size of w, we applied a filter F of size $m \times m$, where $m = 2w + 1$, to scan the contact matrix A of size $n \times n$. The elements in the imputed matrix B is computed by the following:

$$B_{ij} = \sum_{p,q} F_{pq} A_{pq}, \qquad [3]$$

where $i - w \le p \le i + w, j - w \le q \le j + w$. In this work, all of the filters are set to be all-one matrices, which is equivalent to taking the average of the genomic neighbors. However, the filters could be tuned to incorporate different weights for elements during imputation. For instance, the elements located further from the imputed elements could be assigned smaller weights. The window size w was set to 1 for 1-Mbp resolution maps.

Random-walk–based imputation. Random walk with restarts (RWR) is widely used to capture the topological structure of a network (28, 41). The random-walk process helps to infer the global structure of the network and the restart step provides the information of local network structures. What Hi-C data fundamentally describe is the relationship between two genomic bins, which can be considered as a network where nodes are the genomic bins and edges are their interactions. Different from the convolution step, which takes information from the neighbor on the linear genome, the random-walk step considers the signal from the neighbor with experimentally measured interactions. The imputed matrix B defined in Eq. **3** is first normalized by its row sum:

$$C_{ij} = \frac{B_{ij}}{\sum_{j'} B_{ij'}}. \qquad \textbf{[4]}$$

We use Q_t to represent the matrix after the tth iteration of random walk and restart. Then the random walk starts from the identity matrix $Q_0 = I$, and Q_t is computed recursively by the following:

$$Q_t = (1-p)Q_{t-1}C + pI, \qquad \textbf{[5]}$$

where p is a scalar representing the restart probability to balance the information between global and local network structures. The random walk with restart was performed until $\|Q_t - Q_{t-1}\|_2 \le 10^{-6}$. Each element Q_{ij} in the matrix after convergence signifies the probability of random walk to reach the jth node when starting from the ith node. The number of iterations until convergence ranged from 8 to 21 in Flyamer et al. (10) dataset, with a mean of 15.5, and ranged from 10 to 22 in Ramani et al. (7) dataset, with the mean of 15.3.

Embedding and clustering. Since the coverage of the matrices from each cell is different, the sparsity and scales of the matrices after random walk is also distinct. Thus, after random walk, a threshold t was chosen to convert the real matrix Q into binary matrix Q_b. The threshold t was set to be the 80th percentile of Q for all of the analysis, and its impact is discussed in *SI Appendix*, Fig. S4. This is a crucial step since it facilitates us to choose the most conserved and reliable interactions in each cell. Then the $n \times n$ matrix Q_b is reshaped to $1 \times n^2$ and the matrices from m different cells were concatenated into a $m \times n^2$ matrix. In the last step, PCA was used for projecting the matrix into a low-dimensional space and produce the embedding of the cells. Each single chromosome was embedded separately and the embedding of all chromosomes was concatenated at last and another PCA was applied to derive the final embedding. The whitening matrices for the two steps of PCA were multiplied, and the dot product representing the weight of each element in the contact matrices for computing each PC was visualized in *SI Appendix*, Fig. S10. The first two PCs were plotted for visualizing the cells and the first 10 PCs were used for K-means++ clustering. Since we know the cell-type labels of the datasets used in the manuscript, the number of clusters is based on the number of predefined cell types in the corresponding dataset. In cases where the cluster number is unknown, the number of clusters is a user-defined parameter in the scHiCluster package. Since scHiCluster also returns the embedding, the user can also apply other clustering algorithms that do not require a predefined number of clusters on the embedding.

ARI. The ARI was used to compare the similarity between the true label of the cell types and the results of the clustering algorithm. ARI is defined based on the confusion matrix N, where n_{ij} is the number of cells that labeled as the ith cell type and assigned to the jth cluster by the algorithm:

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}, \qquad \textbf{[6]}$$

where a_i and b_j are the sum of the ith row and the sum of the jth column of N, respectively, and n is the total number of cells.

Baseline Methods.

PCA. The raw contact matrices of each cell were $\log_2$ transformed and reshaped to $1 \times n^2$. The matrices from m different cells were concatenated into a $m \times n^2$ matrix. The matrix for each chromosome was PCA transformed and concatenated at last, and another PCA was applied to derive the final embedding with all chromosomes.

HiCRep+MDS. HiCRep 1.6.0 was installed from bioconductor. For each chromosome, the raw contact matrix at 1-Mb bin size of each cell were $\log_2$ transformed and smoothed with a window size of 1. The stratum-adjusted correlation coefficient (SCC) was computed between each pair of smoothed matrices. The median of SCC distances across all chromosomes were transformed to Euclidean distances by Eq. **7**:

$$d_{euc} = \sqrt{2 - 2d_{scc}}. \qquad \textbf{[7]}$$

The Euclidean distance matrix was then embedded into two dimensions with MDS.

Eigenvector. The $n \times n$ raw contact matrix of each cell was $\log_2$ transformed to A. The distance-normalized matrix B of each cell was computed by the following:

$$B_{i,j} = \frac{A_{i,j}}{\sum_{i'} A_{i',i'+j-i}}. \qquad \textbf{[8]}$$

Then PCA was performed on the correlation matrix of B and the PC1 was kept as features of the cell. We computed the mean CpG content of the bins with positive and negative features, respectively, and reversed the features if the negative features corresponded to higher CpG content. The features from m different cells were concatenated into a $m \times n$ matrix and PCA transformed.

Decay. The $n \times n$ raw contact matrix of each cell was $\log_2$ transformed to A. The $1 \times n$feature vector B of each cell was computed by the following

$$B_d = \frac{\sum_j A_{j,j+d}}{\sum_{i,j} A_{i,j}}, \qquad \textbf{[9]}$$

which represent the proportion of contacts at each distance. The features from m different cells were concatenated into a $m \times n$ matrix and PCA transformed.

Identification of TLSs/TADs. In Fig. 5*C* and *SI Appendix*, Fig. S16, all TLSs/TADs were computed by TopDom with a window size of 5. TADs in bulk ESCs and NPCs were identified at 10-kbp resolution. The cells with more than 100 k nondiagonal contacts at 40-kb resolution were included in Fig. 5*D* (1,007 in total). For a given TAD identified in bulk Hi-C data whose boundaries are i and j, we decided whether a TLS in a single cell between i' and j' is corresponding to the TAD by whether i' and j' satisfied $|i' - i| \le \min(80\text{kb}, 0.25 \times (j-i))$ and $|j' - j| \le \min(80\text{kb}, 0.25 \times (j-i))$ or not.

In *SI Appendix*, Figs. S13–S15, S17, and S18, we did not call TLS directly in single cells due to the low coverage of the dataset. Instead, the differential TLSs between cell types were found by browsing the bulk Hi-C data of the corresponding cell types and finding the TADs that are obviously different between those cell types. Then we counted in how many single cells the similar interactions within the TADs were also detected. We defined a cell having a TLS similar to the TAD between i and j if its contact matrix A satisfied $\sum_{i \le i', j' < j} I(A_{i'j'} > 0) > 0.4 \times (j-i)^2$, where I is the indicator function.

ACKNOWLEDGMENTS. We thank Dr. Jie Liu for specifying the details in his algorithm. J.R.E. is an investigator of the Howard Hughes Medical Institute. Research reported in this publication was supported by the National Human Genome Research Institute of NIH under Award R21 HG009274.

1. A. Tanay, A. Regev, Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
2. D. Ramsköld *et al*., Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
3. D. A. Cusanovich *et al*., Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
4. J. D. Buenrostro *et al*., Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
5. C. Luo *et al*., Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
6. T. Nagano *et al*., Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
7. V. Ramani *et al*., Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
8. T. J. Stevens *et al*., 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
9. T. Nagano *et al*., Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
10. I. M. Flyamer *et al*., Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).

11. L. Tan, D. Xing, C.-H. Chang, H. Li, X. S. Xie, Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).
12. J. H. Levine *et al*., Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
13. E. Z. Macosko *et al*., Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
14. C. Luo *et al*., Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9**, 3824 (2018).
15. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
16. D. A. Cusanovich *et al*., The *cis*-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
17. S. Preissl *et al*., Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432–439 (2018).
18. T. Yang *et al*., HiCRep: Assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).
19. O. Ursu *et al*., GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**, 2701–2707 (2018).
20. K.-K. Yan, G. G. Yardimci, C. Yan, W. S. Noble, M. Gerstein, HiC-spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* **33**, 2199–2201 (2017).
21. M. E. G. Sauria, J. Taylor, QuASAR: Quality assessment of spatial arrangement reproducibility in Hi-C data. bioRxiv:10.1101/204438 (14 November 2017).
22. G. G. Yardimci *et al*., Measuring the reproducibility and quality of Hi-C data. *Genome Biol*. **20**, 57 (2019).
23. J. Liu, D. Lin, G. G. Yardimci, W. S. Noble, Unsupervised embedding of single-cell Hi-C data. *Bioinformatics* **34**, i96–i104 (2018).
24. J. Kind *et al*., Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178–192 (2013).
25. S. Shachar, T. C. Voss, G. Pegoraro, N. Sciascia, T. Misteli, Identification of gene positioning factors using high-throughput imaging mapping. *Cell* **162**, 911–923 (2015).
26. J. Kind *et al*., Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134–147 (2015).
27. S. Wang *et al*., Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).
28. J.-Y. Pan, H.-J. Yang, C. Faloutsos, P. Duygulu, "Automatic multimedia cross-modal correlation discovery" in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04* (ACM, New York, 2004), pp 653–658.
29. S. S. P. Rao *et al*., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
30. B. Bonev *et al*., Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557–572.e24 (2017).
31. E. Lieberman-Aiden *et al*., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
32. J. R. Dixon *et al*., Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
33. J. A. Burger, A. Bürkle, The CXCR4 chemokine receptor in acute and chronic leukaemia: A marrow homing receptor and potential therapeutic target. *Br. J. Haematol.* **137**, 288–296 (2007).
34. H. Shin *et al*., TopDom: An efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70 (2016).
35. B. Bintu *et al*., Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
36. G. Li *et al*., Simultaneous profiling of DNA methylation and chromatin architecture in mixed populations and in single cells. bioRxiv:10.1101/470963 (15 November 2018).
37. D.-S. Lee *et al*., Single-cell multi-omic profiling of chromatin conformation and DNA methylome. bioRxiv:10.1101/503235 (26 December 2018).
38. G. Kelsey, O. Stegle, W. Reik, Single-cell epigenomics: Recording the past and predicting the future. *Science* **358**, 69–75 (2017).
39. J. R. Dixon *et al*., Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
40. E. P. Nora *et al*., Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
41. L. Cowen, T. Ideker, B. J. Raphael, R. Sharan, Network propagation: A universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).