OXFORD

Briefings in Bioinformatics, 2024, **25(2)**, bbae096
https://doi.org/10.1093/bib/bbae096
Problem Solving Protocol

scENCORE: leveraging single-cell epigenetic data to predict chromatin conformation using graph embedding

Ziheng Duan, Siwei Xu, Shushrruth Sai Srinivasan, Ahyeon Hwang, Che Yu Lee, Feng Yue, Mark Gerstein, Yu Luan, Matthew Girgenti and Jing Zhang

Corresponding author: Jing Zhang, Department of Computer Science, University of California, Irvine, 92697 CA, USA. Email: zhang.jing@uci.edu

Abstract

Dynamic compartmentalization of eukaryotic DNA into active and repressed states enables diverse transcriptional programs to arise from a single genetic blueprint, whereas its dysregulation can be strongly linked to a broad spectrum of diseases. While single-cell Hi-C experiments allow for chromosome conformation profiling across many cells, they are still expensive and not widely available for most labs. Here, we propose an alternate approach, scENCORE, to computationally reconstruct chromatin compartments from the more affordable and widely accessible single-cell epigenetic data. First, scENCORE constructs a long-range epigenetic correlation graph to mimic chromatin interaction frequencies, where nodes and edges represent genome bins and their correlations. Then, it learns the node embeddings to cluster genome regions into A/B compartments and aligns different graphs to quantify chromatin conformation changes across conditions. Benchmarking using cell-type-matched Hi-C experiments demonstrates that scENCORE can robustly reconstruct A/B compartments in a cell-type-specific manner. Furthermore, our chromatin confirmation switching studies highlight substantial compartment-switching events that may introduce substantial regulatory and transcriptional changes in psychiatric disease. In summary, scENCORE allows accurate and cost-effective A/B compartment reconstruction to delineate higher-order chromatin structure heterogeneity in complex tissues.

Keywords: chromatin compartments; single-cell epigenetics; graph embedding

INTRODUCTION

The human genome is hierarchically compacted and organized in the three-dimensional (3D) space to modulate critical biological processes such as DNA replication, transcription, DNA repair, cell division, and meiosis [1–6]. Such chromatin topological structures usually undergo precise spatiotemporal re-wiring during healthy development, while its misfolding has been reported during the onset and progression of numerous human diseases [7–13]. Therefore, understanding cell-type-specific higher-order chromatin structures, especially at the single-cell level, is critical for grasping how cells with identical DNA can develop diverse functions and fates, underpinning key aspects of developmental biology and tissue differentiation [14–16]. And it is also essential for investigators to quantify their alterations across various states, such as in disease and controls, to better understand the implications of these changes.

Alterations in chromatin compartmentalization can affect gene expression patterns, influencing cellular function, identity, and a broad spectrum of diseases [17]. Moreover, investigating chromatin compartmentation enhances our understanding of how environmental factors can modify gene expression without changing the DNA sequence itself [18]. Over decades, the development of novel visual techniques such as fluorescence in situ hybridization [19–24] and molecular approaches including chromosome conformation capture and its derivatives [16, 22, 25–35] have offered unprecedented opportunities to directly map the folded state of an entire genome [36, 37]. For instance, the Hi-C technique [16, 38] has been developed to extract the genome-scale contact map between any pair of genomic loci simultaneously, using high-throughput sequencing. This technique has been widely used in numerous complex tissues and disease states, enabling scientists to examine the genome's 3D organization at multiple scales [16, 38–41]. However, such genome-wide scalability usually requires the pooling of millions of cells as input, leaving the resultant contact probabilities reflecting only an average of chromatin interaction frequencies across diverse cell populations. Recently, several pioneering studies

Ziheng Duan is a PhD student at the University of California, Irvine, who focuses on bioinformatics.
Siwei Xu is a PhD student at the University of California, Irvine, who focuses on bioinformatics.
Shushrruth Sai is a PhD student at the University of California, Irvine, who focuses on bioinformatics.
Ahyeon Hwang is a PhD student at the University of California, Irvine, who focuses on bioinformatics.
Che Yu Lee is a Master's student at the University of California, Irvine, who focuses on bioinformatics.
Feng Yue is a professor of Biochemistry and Molecular Genetics and Pathology at Northwestern University, who focuses on computational biology and functional genomics.
Mark Gerstein is a professor of Biomedical Informatics at Yale University, who focuses on biomedical data science.
Yu Luan is an assistant professor at UT Health San Antonio, who focuses on cancer genomics.
Matthew Girgenti is an assistant professor of Psychiatry at Yale School of Medicine, who focuses on bioinformatics.
Jing Zhang is an assistant professor at the University of California, Irvine, who focuses on bioinformatics.
Received: November 6, 2023. **Revised:** February 1, 2024. **Accepted:** February 20, 2024

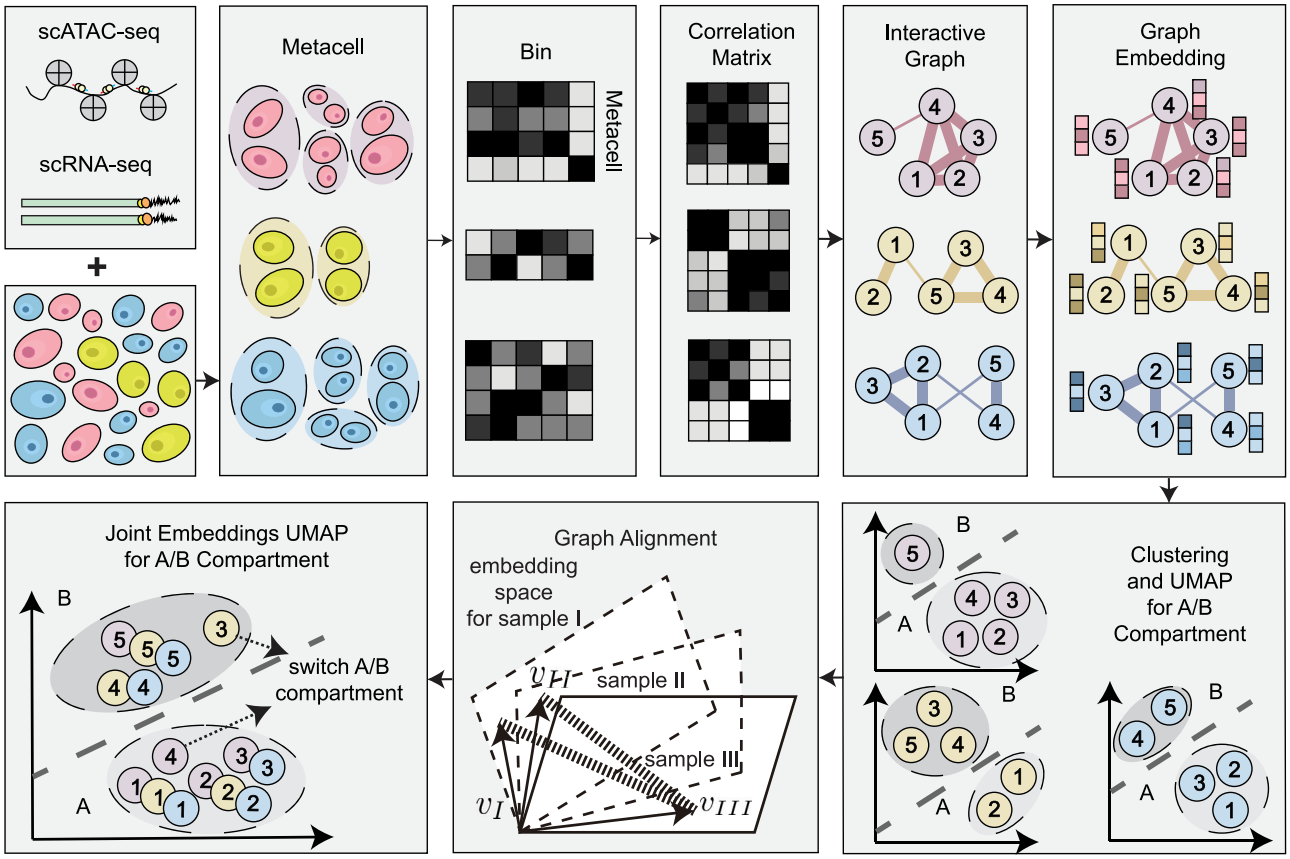

Figure 1. The overview of scENCORE's framework. Top: single-cell epigenetic data are segmented into bins—standardized genomic regions—and processed through metacell analysis to create a bin-by-metacell matrix. This matrix is used to construct an interactive graph, depicting genomic interactions. Bottom right: clustering and get the A/B compartment. Bottom middle: graph alignment. Bottom left: joint UMAP and compute the CSS.

have successfully performed Hi-C on single isolated cells [42, 43], exposing extensive cell-cycle state differences and cell-to-cell heterogeneity in mammalian chromosomal conformation. Unfortunately, they are still expensive and unavailable for most investigators. As a result, it is still challenging to obtain detailed chromatin conformation information for a large number of cell types/states and individuals.

In contrast, emerging single-cell epigenetic sequencing technologies have been developed for convenient, cost-effective, and simultaneous profiling across thousands of cells [44–55]. Recent studies, such as C.Origami [56], use single-cell epigenetic sequencing to predict Hi-C data in a supervised manner, necessitating well-annotated Hi-C for training. However, this approach may face limitations when sufficient Hi-C data are lacking. Furthermore, recent transparent data-sharing initiatives already provided scientists direct access to atlas-level single-cell epigenetic data in complex tissues [57–60] and across diverse disease cohorts [49, 53]. It has been proved that long-range correlations from bulk-tissue epigenetic profiles can reliably reconstruct mega-base scale A/B compartments, which are highly consistent with results inferred from Hi-C data [61]. Therefore, we propose a novel computational method (scENCORE; Figure 1) to delineate cell-type-specific genome compartmentation by utilizing the relatively lower cost and widely available single-cell epigenetic data, especially the single-cell sequencing assay for transposase-accessible chromatin data (scATAC-seq) [46, 49, 51, 53].

Our method, scENCORE, uses single-cell epigenetic data to predict genome conformation through a three-step graph embedding approach. First, scENCORE constructs a cell-type- or condition-specific graph to mimic the genome contact map. In this graph, nodes and edges represent genome bins and their long-range epigenetic correlations, respectively. Second, scENCORE projects genome bins into a latent space using a graph embedding algorithm and clusters bins into groups representing A/B compartments. Finally, we further calculate a compartment switching score (CSS) for each bin to quantitatively evaluate chromatin conformation changes across conditions (e.g. different cell states or disease vs. control). This is achieved by aligning condition-specific bin embeddings to the same latent space. To illustrate the robustness and general validity of our approach, we applied scENCORE to both population-scale bulk and single-cell epigenetic data. Our results demonstrate that scENCORE accurately predicts A/B compartments defined by Hi-C experiments. Additionally, our findings suggest that scENCORE can identify subtle chromatin re-structuring across biologically close cell types and highlight key switching events associated with psychiatric disorders. We have implemented scENCORE as a free software package (https://github.com/aicb-ZhangLabs/scENCORE) available for the community to predict A/B compartments in the 3D genome and quantify their changes across diverse cell types and conditions. With the explosion of available single-cell epigenetic data, we anticipate that scENCORE will help delineate cell-type-specific chromatin conformation in complex tissues and

advance our understanding of the relationships between genome compartmentalization and gene regulation at a single-cell resolution.

METHOD

Modeling chromatin conformation with graph embedding

Inspired by previous population-scale bulk tissue analyses [61], we hypothesize that reliable reconstruction of cell-type-specific higher-order chromatin interactions can be achieved using long-range genomic profile correlations from single-cell epigenetic data. Compared to single-cell Hi-C experiments, direct computational chromatin conformation reconstruction could be a more cost-effective alternative, as atlas-level single-cell epigenetic data, specifically scATAC-seq data, are already publicly available to the public for many complex tissues and disease conditions. We propose a general framework, named scENCORE (Figure 1), to infer personalized, cell-type-specific A/B compartments from single-cell epigenetic data and quantitatively measure compartment-switching events across diverse cell states and disease conditions.

As shown in Figure 1, scENCORE first divides the genome into fixed-length bins (default at 1Mbp length). This binning process involves segmenting the genome into defined regions, such as (chr1:1–1,000,000), (chr1:1,000,001–2,000,000), and so forth. We ensure accuracy and relevance by excluding bins that overlap with the blacklist regions as defined by the ENCODE consortium [62]. After binning the genome, we count the scATAC-seq reads in each genome bin of individual cells. Then we employ metacell technology to mitigate the sparsity of scATAC-seq data (see the long-range epigenetic correlation calculation part in the method section), followed by constructing a correlation matrix between genome bins. This forms the basis of our interactive graph, where each node represents a bin, and the edge weights signify the correlation between these bins. Then, scENCORE learns a low-dimensional genome bin (node) representation by summarizing their potential interactions on the graph with the rest of the genome and clusters them into different compartments (Figure 1, bottom right). This process is repeated for each cell type to infer individualized, cell-type-specific chromatin conformation on personal epigenomes. Finally, scENCORE aligns diverse embedding spaces across different samples and individuals (Figure 1, bottom middle), allowing the derivation of A/B CSS across diverse states (cell types or disease conditions) and highlighting higher-order chromatin rewiring events. Since scENCORE generates embeddings for all nodes simultaneously, it belongs to the transductive method. Table 1 shows the symbols' definitions of scENCORE. For a detailed analysis of scENCORE, please refer to the appendix.

Problem formulation

Given a single-cell ATAC-seq matrix $\boldsymbol{F} \in \mathbb{R}^{N\times c}$, where N is the number of valid chromatin regions, c is the number of cells and $\boldsymbol{F}_{ij}$ denotes the accessibility value of the jth cell at the ith chromatin region. Our goal is to determine a binary vector $\boldsymbol{H} \in \{0,1\}^N$ where H_i represents the compartment type of the ith chromatin region (1 for A-compartment and 0 for B-compartment). The derived compartments from $\boldsymbol{H}$ should be consistent with the higher-order chromatin structure represented in Hi-C results.

Table 1: Symbols' definitions.

Symbols	Definitions
$\boldsymbol{F}$	ATAC-seq matrix for single cells
N, c	# chromatin regions, # cells
mc, m	# metacells, # cells per metacell
k, γ	# neighbors and max overlap in metacell
$\boldsymbol{M}$	Metacell composition matrix
$\boldsymbol{F_m}, \boldsymbol{F_n}$	Metacell feature matrix, normalized matrix
$\boldsymbol{H}$	Binary vector for compartment classes
$\boldsymbol{C}, \boldsymbol{G}$	Correlation matrix, interaction graph
$\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{A}$	Nodes, edges and adjacency matrix of $\boldsymbol{G}$
t	Threshold for graph sparsity
d	Dimension of latent representation
$\boldsymbol{\phi}, \mathbf{W}$	Mapping function, node embedding
v, u	Example nodes in $\boldsymbol{V}$
$\boldsymbol{sim_G}, \boldsymbol{sim_E}$	Similarity in original and latent spaces
$\mathbf{P}, \mathbf{Q}$	Empirical and noise distribution in NCE
λ, s	Variable and negative samples in NCE
$\boldsymbol{\beta}$	Starting node distribution in pagerank
$\alpha, \boldsymbol{\pi_\beta}$	Damping factor and vector in pagerank
$\mathbf{R}, \boldsymbol{\Omega}$	Mapping matrices in graph alignment
$\boldsymbol{I_d}$	Identity matrix
$S1, S2$	Examples of samples
$\boldsymbol{U}, \boldsymbol{\Sigma}, \mathbf{V'}$	Matrices in singular value decomposition
$\mathbf{W'}, \mathbf{W''}$	Normalized and aligned node embeddings
$\boldsymbol{D_{S1,S2}}$	Distance between samples S1 and S2
$\boldsymbol{CSS_{S1,S2}}$	Compartment switching score

Long-range epigenetic correlation from scATAC-seq data

Due to the sparseness of the sample in scATAC-seq data, not every cell fragment can appear in each region, leading to an abundance of zeros in the ATAC-seq matrix $\boldsymbol{F} \in \mathbb{R}^{N\times c}$, where N is the number of valid chromatin regions and c is the number of cells. This sparsity can result in low correlation coefficients between regions. To address this, we introduce the concept of a metacell. A metacell is essentially a composite cell, representing an aggregation of cells with similar characteristics to increase data density.

Metacell construction

We construct mc metacells, where each metacell consists of a group of m cells. These metacells are constructed to mitigate the effects of data sparsity by aggregating the data from multiple similar cells. The composition matrix of metacells is denoted as $\boldsymbol{M} \in \mathbb{R}^{mc\times m}$, where $\boldsymbol{M}_{ij}$, ranging from one to c, indicates the jth cell that constitutes the ith metacell. The process involves identifying clusters of cells based on similarity in their ATAC-seq profiles (here we use the first 20 dimensions of PCA). Following this, cells are grouped into clusters using a K-nearest neighbors' algorithm with the number of neighbors $k = 100$. To ensure distinct metacell identities, we regulate the composition of each metacell with a maximum overlap rate parameter, denoted as γ, set at 0.9. This parameter limits the proportion of metacells to which any individual cell can belong.

Meta-cell feature matrix calculation

With the metacells defined, we compute a new metacell feature matrix $\boldsymbol{F_m} \in \mathbb{R}^{N\times mc}$. For the ith metacell, its feature matrix $\boldsymbol{F_m}(., i)$ can be computed as follows:

$$\boldsymbol{F_m}(., i) = \sum_{j=1}^{m} \boldsymbol{F}(., \boldsymbol{M}_{ij}). \quad (1)$$

This equation aggregates cell features within a metacell, resulting in a denser and more representative feature matrix.

Normalization and correlation calculation

To facilitate the computation of correlations between different regions, we first apply term frequency normalization to the metacell matrix $\boldsymbol{F_m}$:

$$\boldsymbol{F_n} = \frac{\boldsymbol{F_m}}{\sum_{i=1}^{mc} \boldsymbol{F_m}(.,i)}, \tag{2}$$

where $\sum_{i=1}^{mc} \boldsymbol{F_m}(.,i) \in \mathbb{R}^N$ represents the sum of samples of the same metacell in different regions, and $\boldsymbol{F_n} \in \mathbb{R}^{N\times mc}$ is the normalized feature matrix. The correlation between the ith and jth regions can then be computed as

$$\boldsymbol{C}_{ij} = \frac{\sum_{k=1}^{c}(\boldsymbol{F_n}(i,k) - \boldsymbol{F_n}(i,.))(\boldsymbol{F_n}(j,k) - \boldsymbol{F_n}(j,.))}{\sqrt{\sum_{k=1}^{c}(\boldsymbol{F_n}(i,k) - \boldsymbol{F_n}(i,.))^2}\sqrt{\sum_{k=1}^{c}(\boldsymbol{F_n}(j,k) - \boldsymbol{F_n}(j,.))^2}} \tag{3}$$

This equation calculates the correlation coefficient between different regions, allowing us to measure interactions across the genomic landscape.

Graph construction

To better characterize the interaction between regions, we first constructed an interact graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{A})$, which is made up of a set of N nodes $\boldsymbol{V}$ and a set of edges $\boldsymbol{E}$. $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is an adjacency matrix where nonzero entries equal the corresponding edge weights. The construction process assumes that a higher value in the interact matrix $\boldsymbol{C}$ indicates that the two regions are more likely to be connected, that is, the two nodes in the graph have a greater probability of having edges. Given this assumption, we consider the following way to construct it [63, 64]: $\boldsymbol{A}_{ij} = \boldsymbol{C}_{ij}$ if$\boldsymbol{C}_{ij} \geq t$ else0, where t is the threshold hyperparameter that controls the graph sparsity. In the constructed graph, we have $(i,j) \in \boldsymbol{E}$ if $\boldsymbol{A}_{ij} \neq 0$. Different graph construction methods and the impact of different construction parameters on the results are in the appendix.

Graph representation learning in scENCORE

Inspired by recent advancements in graph embedding methods [63, 65–70], we adopt the following graph embedding method to derive high-quality region representations. The aim is a mapping function $\boldsymbol{\phi} : \boldsymbol{V} \to \mathbb{R}^{N\times d}$, $d \ll N$, representing each node $v \in \boldsymbol{V}$ in a reduced dimension d. Given graph similarity $\boldsymbol{sim_G} : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{R}$ (a measure quantifying the relationships between nodes, calculated using techniques like personalized pagerank, adjacency similarity, where nodes are genomic regions), and the corresponding similarity in the embedding space $\boldsymbol{sim_E} : \boldsymbol{V} \times \boldsymbol{V} \to \mathbb{R}$, we require $\sum_{u\in\boldsymbol{V}} \boldsymbol{sim_G}(v,u) = 1$ and $\sum_{u\in\boldsymbol{V}} \boldsymbol{sim_E}(v,u) = 1$ for any node $v \in \boldsymbol{V}$ to represent the similarity distributions. The similarity in embedding space $\boldsymbol{sim_E}(v,u)$ is obtained as follows:

$$\boldsymbol{sim_E}(v,\cdot) = \frac{\exp(\mathbf{W}_v\mathbf{W}^T)}{\sum_{i=1}^{N}\exp(\mathbf{W}_i\mathbf{W}^T)}. \tag{4}$$

To align $\boldsymbol{sim_E}$ with $\boldsymbol{sim_G}$, we minimize their divergence for $v \in \boldsymbol{V}$:

$$L = -\sum \boldsymbol{sim_G}(v,\cdot)\log(\boldsymbol{sim_E}(v,\cdot)). \tag{5}$$

Due to computational costs, we adopt Noise Contrastive Estimation (NCE) [71]:

$$\begin{aligned} L_{NCE} = &\sum[\log Pr_W(\lambda = 1|\boldsymbol{sim_E}(v,u)) \\ &+ \sum \log Pr_W(\lambda = 0|\boldsymbol{sim_E}(v,\tilde{u}))], \end{aligned} \tag{6}$$

where the node v is drawn from the empirical distribution $\mathbf{P}$, $\lambda = 1$ for u drawn from $\boldsymbol{sim_G}(v,\cdot)$, and $\lambda = 0$ for sample nodes $\tilde{u}$ drawn from the noise $\mathbf{Q}(v)$. For training efficiency, we have: $s \ll N$. We compute Pr_W as the sigmoid $\sigma(x) = (1+e^{-x})^{-1}$ of the dot product of $\mathbf{W_v}$ and $\mathbf{W_u}$. In the experiment, we set the initial embedding matrix $\mathbf{W}$, $\mathbf{P}$ and $\mathbf{Q}$ distributed uniformly.

We adopt the personalized pagerank (PPR) [72] for $\boldsymbol{sim_G}(v,\cdot)$. Given a starting node distribution $\boldsymbol{\beta}$, damping factor α that controls the range of the explored neighborhood, and normalized adjacency matrix $\mathbf{A}$, we can infer the PPR vector $\boldsymbol{\pi}$ via a recursive way:

$$\boldsymbol{\pi_\beta} = (1-\alpha)\boldsymbol{\beta} + \alpha\boldsymbol{\pi_\beta}\mathbf{A}. \tag{7}$$

This recursive formula is implemented using a random walk with restart from a node v. The walk transitions to adjacent nodes based on $\mathbf{A}$, with a probability dictated by the damping factor α to either continue or restart. This ensures that the PPR vector $\boldsymbol{\pi_\beta}$ effectively captures the steady-state probability of visiting each node, indicating their relative importance or similarity from node v's perspective.

Unsupervised clustering

Using trained embedding $\mathbf{W}$, we apply the Expectation-Maximization (EM) algorithm on two multivariate Gaussian mixtures. Resulting probabilities, $(P_1, P_2) = \mathrm{EM}(\mathbf{W}, n = 2)$, denote each chromatin region's compartment likelihood. The compartment with the higher fragment average is labeled A, and the other B.

Graph alignment in scENCORE

Due to the inconsistencies in node embeddings across multiple runs of the same algorithm, comparing dynamic changes across samples becomes a challenge [73]. *scENCORE* addresses this through orthogonal Procrustes graph alignment [74], enabling the comparison of embeddings of the same chromatin region across different samples. We get the normalized embeddings $\mathbf{W'_{S1}}$ and $\mathbf{W'_{S2}}$ using min-max normalization to $\mathbf{W_{S1}}$ and $\mathbf{W_{S2}}$. Then we find an orthogonal matrix $\boldsymbol{\Omega}$ that best maps $\mathbf{W'_{S1}}$ to $\mathbf{W'_{S2}}$, which can be formulated as follows:

$$\mathbf{R} = \arg\min_{\boldsymbol{\Omega}} \|\mathbf{W'_{S1}}\boldsymbol{\Omega} - \mathbf{W'_{S2}}\|_F, \quad \text{subject to} \boldsymbol{\Omega}^T\boldsymbol{\Omega} = \mathbf{I_d}, \tag{8}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathbf{I_d}$ is the identity matrix. Using singular value decomposition, if $\mathbf{W'_{S1}}^T\mathbf{W'_{S2}} = \boldsymbol{U\Sigma V}^T$, the optimal $\mathbf{R}$ is given by $\boldsymbol{UV}^T$. After alignment, the embeddings are represented as $\mathbf{W''_{S1}} = \mathbf{W'_{S1}}\boldsymbol{\Omega}$ and $\mathbf{W''_{S2}} = \mathbf{W'_{S2}}$, which reside in a common space, allowing for meaningful comparison.

CSS calculation

To measure changes of chromatin regions across samples, we introduce the CSS. For two samples, S1 and S2, with their aligned embeddings, we compute a distance matrix $\boldsymbol{D_{S1,S2}} \in \mathbb{R}^N$ based on the Euclidean distance between the embeddings. The CSS is then

calculated as follows:

$$CSS_{S1,S2} = 1 - \exp(-D_{S1,S2}) \quad (9)$$

Here, $CSS_{S1,S2} \in \mathbb{R}^N$ represents compartment switching scores for N regions. A score near one suggests that a region is more likely to undergo a switching event.

RESULTS

scENCORE recovers Hi-C chromatin map from bulk ATAC-seq

As a proof-of-concept application, we first applied scENCORE on two independent bulk ATAC-seq data from large cohorts of individuals. Specifically, we downloaded the HumanFC and BrainGVEX bulk ATAC-seq data from the psychENCODE Synapse portal, with 288 and 341 samples in each study. We used 1-Mbp bins with no overlaps with known gap regions (see the data preprocessing part in the appendix) and extracted the normalized ATAC-seq signals in each bin to calculate the tissue-level correlation matrices in both studies. To test whether such long-scale epigenetic correlations can truly reflect chromatin interaction maps, we compared the epigenetic correlation matrices to the interaction matrix extracted from the Hi-C experiment in perfectly matched prefrontal cortex tissue. We found that the epigenetic correlation matrices are highly consistent with the Hi-C interaction maps, as reflected by their highly similar first eigenvectors (Figure 2A). For instance, the Pearson correlation between the first eigenvectors of the Hi-C interaction map (black line) and epigenetic correlation matrix (colored line) was as high as 0.85 and 0.94 for the HumanFC and BrainGVEX datasets, respectively.

We applied scENCORE to the bulk ATAC-seq data from two large cohorts of individuals (HumanFC and BrainGVEX) to predict A/B compartments. As shown in Figure 3(A–B), scENCORE achieves F1 scores of 0.782 and 0.770 on the BRG and HFC datasets, respectively, and AUC of 0.855 and 0.821, outperforming basic methods (mean signal, first PCA and first eigenvector) as well as famous graph-based approaches such as Node2Vec [75], ProNE [76] and GraRep [77]. Here we conducted each method three times, reporting the average to ensure robustness. For a detailed comparative analysis, please refer to the appendix. Besides, scENCORE reported a similar number of A/B compartments in both studies (980 and 1043 A compartments), which is highly consistent with the Hi-C derived compartmentations using various thresholds (above 0.77 and 0.79 for both studies, Figure 2B). In addition, we calculated the tissue-matched H3k27ac ChIP-seq signals on different compartments. We found that the A compartment demonstrated significantly higher H3k27ac signals (P-value < 4.85e-276 for all six samples in Figure 2C), consistent with its more active roles in the genome. We repeated this calculation on three randomly selected samples and observed consistent results. These results demonstrate the feasibility of estimating chromatin conformation using epigenetic correlations, which could potentially provide an alternative approach to single-cell Hi-C experiments for studying chromatin structure at the population scale.

scENCORE reconstructs A/B compartments validating by cell-specific Hi-C data

Next, we attempted to reconstruct higher-order chromatin conformation in major cell types in the human brain using scATAC-seq data. Please refer to the supplementary file for the details about the cell-type-specific scATAC-seq data. We used publicly available, cell-type-matched Hi-C data from previous studies for validation [78, 79].

We applied scENCORE to different cell types to predict A/B compartments and used two independent functional genomics data generated from FACS-sorted cells for separate validations. As expected, we found that the first eigenvectors of the long-range epigenetic correlation matrices and chromatin interaction matrices from cell-type-matched Hi-C experiments were highly consistent (Pearson correlation 0.81 and 0.61, Figure 4A), demonstrating the feasibility of our approach. Next, we projected genome bins into a low-dimensional latent space and observed distinct A/B compartment clusters in all cell types (Figure 4B). As an illustration, we highlighted bins on chromosome 2 in Figure 4(B). Additionally, we demonstrated that our results are accurate and robust to hyperparameters such as meta-cell sizes using cell-type-matched Hi-C experiments (Figure 4C and the validation part in the method section).

The literature indicates that the A compartment is transcriptionally more active and enriched with positive regulatory signals. Therefore, we downloaded cell-type-specific H3K27ac ChIP-seq data from the psychENCODE project, which is widely considered an active enhancer signature. We then compared the averaged H3K27ac signal strengths in our predicted A/B compartments. As expected, A compartments showed significantly higher H3K27ac signals than those in B compartments using cell type matched ChIP-seq experiment (Figure 4D, log fold change 1.09 vs. 0.57 in excitatory neuron with P-value 1.81e-208, 1.04 vs. 0.63 in microglia with P-value 3.10e-73), demonstrating the robustness of our predictions. Furthermore, we found that the A compartments showed the highest H3K27ac signal enrichment in matched ChIP-seq experiments. For instance, the cell-type-matched H3k27ac log fold changes were 0.64 and 0.51, significantly higher than those from non-matched cell types (0.34 and 0.39, Figure 4D). These results indicated that scENCORE was able to capture cell-type-specific chromatin conformation in complex tissues.

Finally, we calculated the pairwise Jaccard similarity using scENCORE's A/B compartment predictions (see the pairwise Jaccard similarity section of the appendix) across seven cell types and performed hierarchical clustering based on the similarity matrix. As expected, while neurons and non-neuron cell types can be reliably separated, different cell types only demonstrated moderate similarities (0.74∼0.84, Figure 4E). These results highlight the significance of reconstructing chromatin conformation at the single-cell level.

scENCORE identifies key compartment switching events between different cell types

In the previous sessions, scENCORE extracted separate cells from different types for independent training and graph representation learning. This results in genome embeddings in disjoint latent spaces. However, in many applications, it is important to compare chromatin conformations and quantify their changes across cell types or conditions. To address this, scENCORE includes a graph alignment module that maps separate embeddings to a shared space, facilitating direct comparison analyses across cell types (see the graph alignment part of the method section). Consequently, we can calculate CSS using a normalized L2 distance in the aligned space to quantify the chromatin compartment switching status. A larger CSS indicates a more significant difference in the A/B compartment assignment.

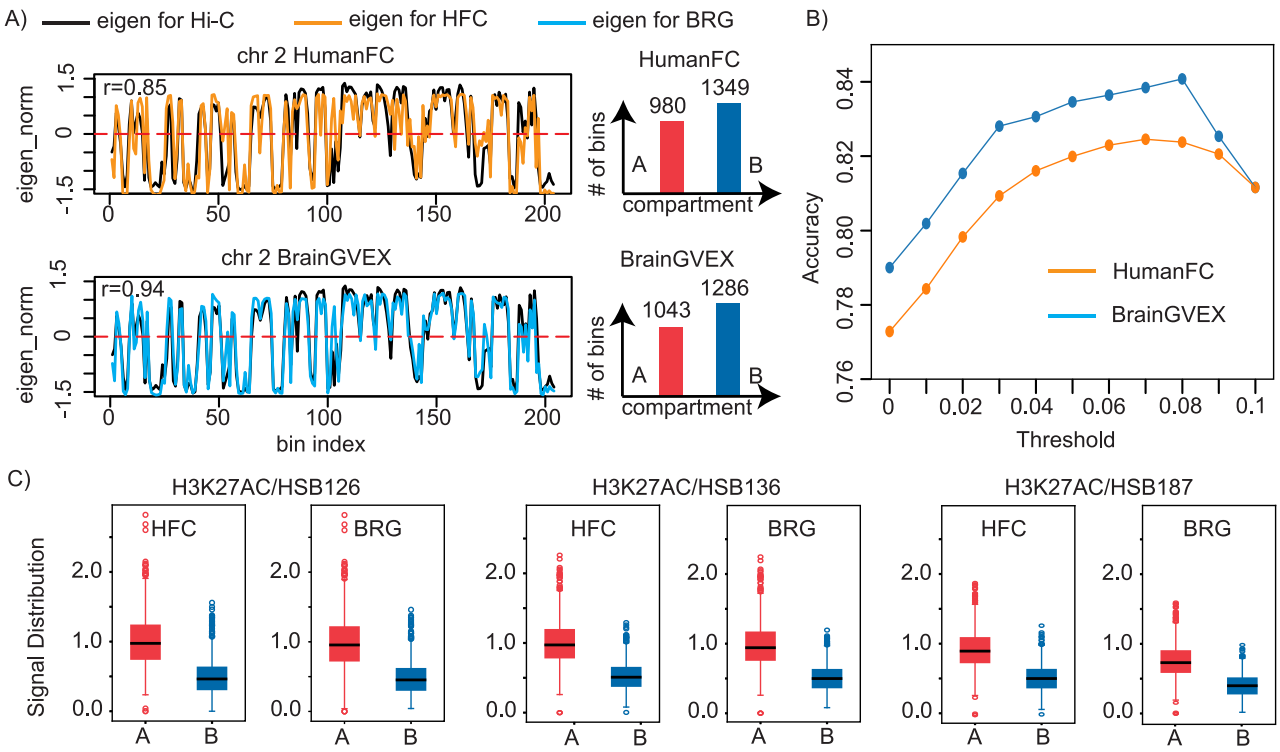

Figure 2. Consistency between chromatin interaction maps and population-scale bulk ATAC-seq. (**A**) Similarity of eigenvectors between Hi-C and bulk ATAC-seq data. (**B**) Agreement of scENCORE's predictions with Hi-C compartmentations. (**C**) Tissue-matched H3k27ac ChIP-seq signals on different compartments.

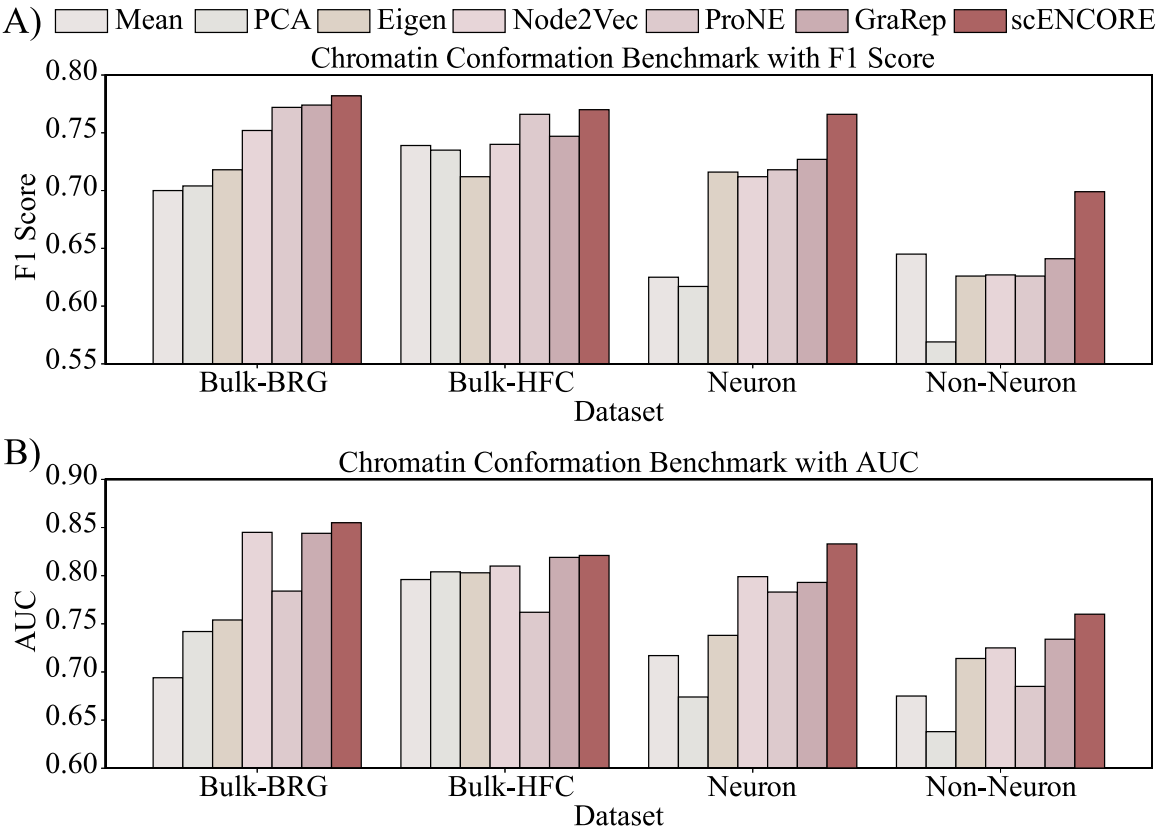

Figure 3. Chromatin conformation benchmark with F1 Score and AUC.

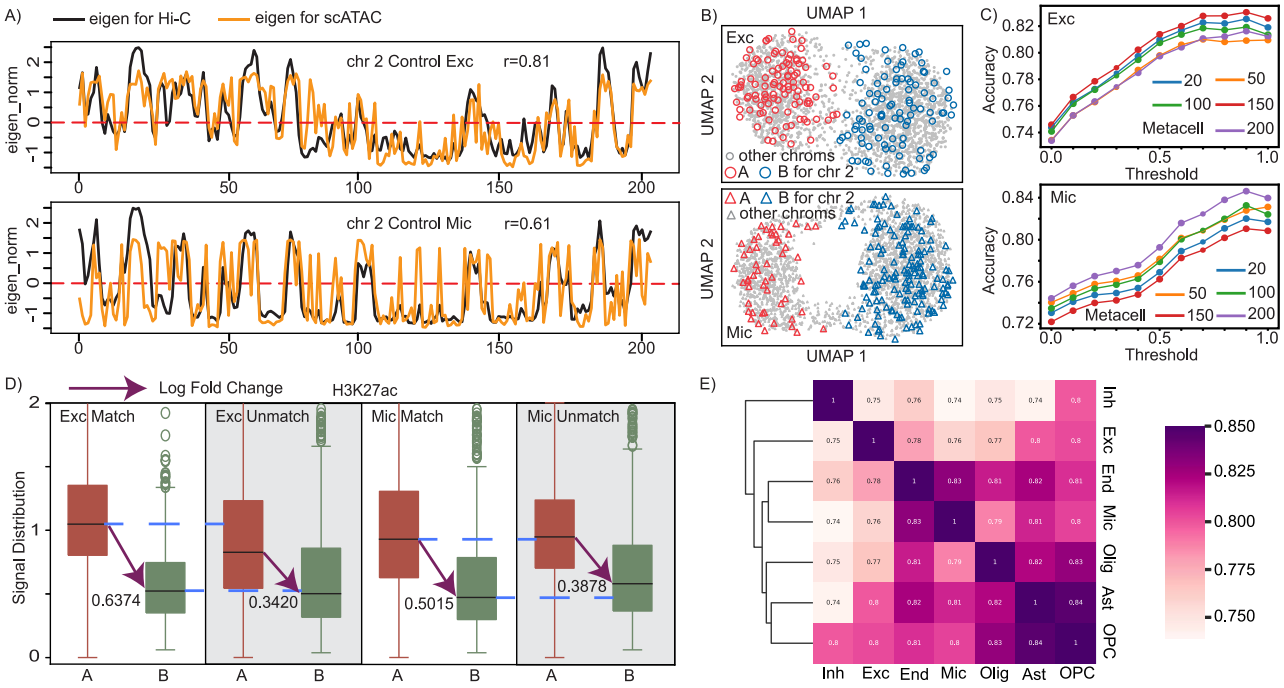

Figure 4. scENCORE's cell-type-specific A/B compartment predictions. (**A**) First eigenvector of single cell epigenetic correlation maps. (**B**) A/B compartment groupings in UMAP. (**C**) Alignment of scENCORE predictions with Hi-C by metacell. (**D**) Comparison of H3K27ac signals in A/B compartments. (**E**) Jaccard similarity for A/B compartment across cell types.

We ranked the genome bins based on their CSSs and identified several well-known brain marker genes located in the top rewired regions. For instance, a well-known neuron marker gene SATB2 was found in the more active compartment in excitatory neurons but switched to the B compartment in microglia (Figure 5A). Its associated bin has a high CSS at 0.54, ranking 255th out of 2329 among all genome bins. As expected, SATB2 was highly expressed in neurons, showed enriched H3K27ac signals (1.14 vs. 0.87, Figure 5B), and contained extensive open chromatin regions with potentially positive regulatory activities (Figure 5C). Similarly, the microglia marker MRC experienced the opposite compartment-switching process (from A in microglia to B in excitatory neurons, Figure 5D). With a high CCS of 0.59 and ranking ninth among the 2329 genome bins, it demonstrated significantly stronger H3K27ac signals (0.41 vs. 0.26, Figure 5E) and chromatin accessibility scores in microglia than in excitatory neurons (Figure 5F).

To further establish the biological significance of the identified compartment switches, we compared the top 30 CSS-ranked bins to the bins housing the top 100 differential expressed genes (DEGs) related to neurological diseases. This analysis revealed a higher DEG presence within scENCORE-identified bins than what was observed in bins selected randomly. As shown in Figure 5(G) and (H), this pattern of DEG enrichment, consistently replicable across ten iterations, was statistically significant (Mann–Whitney U test, $P < 0.0002$), underscoring the efficacy of scENCORE in detecting compartment switches with potential links to neurobiological functions and disorders.

scENCORE highlights extensive cell-type-specific chromatin re-structuring events in brain disorders

We applied scENCORE to individuals with MDD and compared their chromatin conformation with healthy controls across different cell types. Specifically, we calculated cell-type-specific CSS to search for chromatin compartment switching events. We identified one of the top-ranked regions with the highest CSS in chromosome two (chr2: 15–16 Mbp), which has been assigned the more active A compartment in controls but inactive B compartment in MDD. To explore this region further, we calculated the Pearson correlation (ρ) of epigenetic signals between this bin and other bins on the same chromosome. We then plotted the absolute correlation difference ($|\Delta\rho|$) between the MDD sample (ρ_m) and control sample (ρ_c). Intuitively, larger changes represented bigger interaction alterations between conditions. Our prioritized region demonstrated noticeably higher long-range interaction relationships change with other bins on the same chromosome (Figure 6A) compared to the rest of the genome.

scENCORE further highlighted that the previously reported MDD risk gene NBAS is located in this region. NBAS is transcriptionally active in most cell types, especially in neurons (Figure 6B). We performed differential gene expression analysis on matched RNA-seq data and found that NBAS is downregulated only in excitatory neurons of MDD samples, as shown in Figure 6(C). Consistently, this scENCORE prioritized region demonstrated the highest CSS in excitatory neurons (0.50 vs. an average of 0.43 in other cell types, Figure 6D). Specifically, NBAS is located in the active A compartment in healthy controls but switches to the inactive B compartment in MDD samples. This finding potentially explains its downregulated gene expression values only in excitatory neurons. These results demonstrate the potential of scENCORE to pinpoint disease-associated chromatin compartment switching events in a cell-type-specific manner and decipher how higher-order chromatin conformation changes can lead to transcriptional perturbations in disease.

Following a similar approach as in the cross-cell type analysis, we also examined chromatin compartment switches from control to MDD using scENCORE. We compared the top 30 CSS-ranked

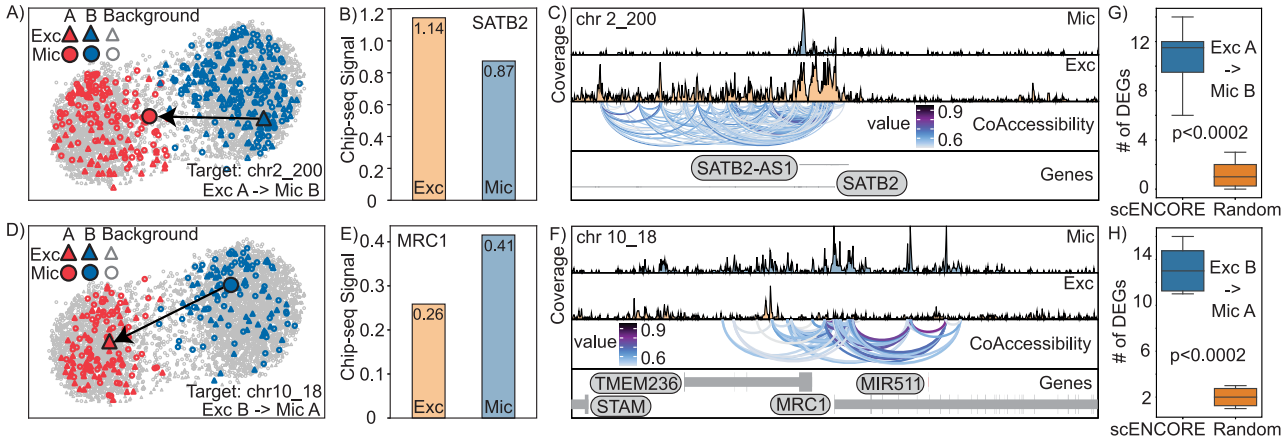

Figure 5. scENCORE identifies key compartment switching events between different cell types. (**A**) scENCORE reported chr2_200 to be in the more active compartment in excitatory neurons but switched to the B compartment in microglia. (**B**) A well-known neuron marker gene SATB2 is highly expressed in neurons, and showed enriched H3K27ac signals. (**C**) Neurons contain extensive open chromatin regions with potentially positive regulatory activities. (**D**) scENCORE reported chr10_18 to be in the more active compartment in microglia but switched to the B compartment in excitatory neurons. (**E**) A microglia marker MRC demonstrated significantly stronger H3K27ac signals. (**F**) More chromatin accessibility scores in microglia than in excitatory neurons. (**G**) Boxplot comparison of DEG counts in excitatory neuron compartment A to microglia compartment B switches versus random selection ($P < 0.0002$). (H) Boxplot of DEG counts in excitatory neuron compartment B to microglia compartment A switches, again showing scENCORE's significant identification ($P < 0.0002$).

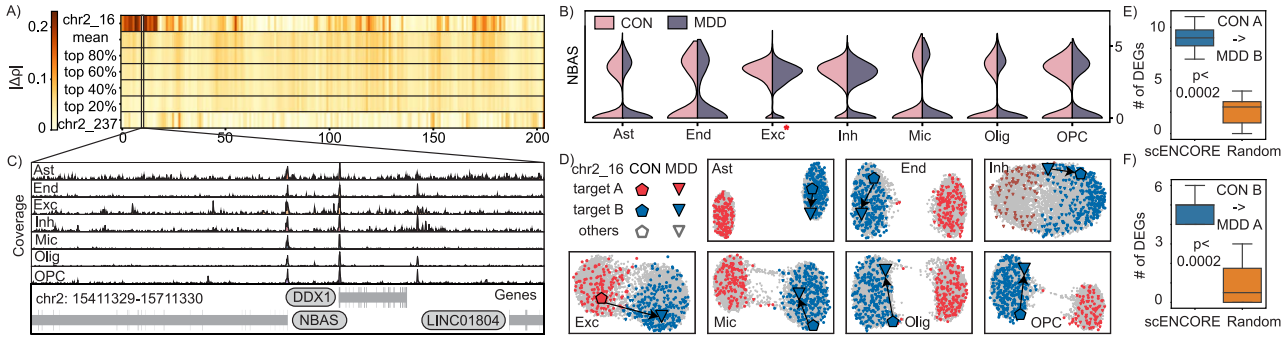

Figure 6. scENCORE highlights cell-type-specific chromatin re-structuring events in brain disorders. (**A**) Long-range epigenetic correlation shifts. (**B**) The activity of MDD risk gene NBAS in cell types. (**C**) Downregulation of NBAS in MDD's excitatory neurons. (**D**) Highest CSS in excitatory neurons in the prioritized region. (**E**) Boxplot comparison of DEG counts in control compartment A to MDD compartment B switches versus random selection ($P < 0.0002$). (**F**) Boxplot of DEG counts in control compartment B to MDD compartment A switches, again showing scENCORE's significant identification ($P < 0.0002$).

bins against bins with DEGs associated with MDD. The results showed more DEGs in scENCORE-selected bins than in randomly selected bins, as depicted in Figure 6(E) and (F). This pattern, replicated over ten iterations, was statistically significant (Mann–Whitney U test, $P < 0.0002$), demonstrating scENCORE's capability in identifying chromatin changes relevant to disease.

DISCUSSION AND CONCLUSION

This paper introduces scENCORE, a computational method that leverages single-cell epigenetic data to reconstruct personalized and cell-type-specific higher-order chromatin compartment information. While recent developments in single-cell Hi-C technology shed light on constructing chromatin conformation in individual cells, it is not yet widely available in most labs and can be expensive to perform on the population-scale sequencing. In contrast, scENCORE approximates chromatin contact frequencies using long-range epigenetic correlations and offers two main advantages: it is more cost-effective and accessible, and it can predict personalized chromatin compartments, enabling the direct quantification of higher-order conformation changes across conditions (e.g. disease and control).

To prove the effectiveness of scENCORE, we conducted mega-base scale chromatin analysis on bulk tissue and single-cell ATAC-seq data and benchmarked against results from tissue or cell-type matched Hi-C experiments. Our findings showed that scENCORE can faithfully reconstruct chromatin compartments and highlight key switching events across different cell types and conditions. Moreover, the incorporation of graph embedding in scENCORE supports the capture of non-linear relationships between chromatin regions, a capability not afforded by naïve methods like the first eigen analysis. This graph-based approach not only enhances the interpretability of the model but also allows for the integration of multi-modal data, such as scRNA-seq and scATAC-seq. Furthermore, the alignment feature of scENCORE enables the quantitative analysis of CSS across different regions, effectively quantifying variations in chromatin conformation across various cell types or disease states. We have implemented scENCORE as open-source software that is freely downloadable to the public. With the exponential growth of single-cell epigenetic data, scENCORE can be a valuable tool for the research community to illuminate cell-type-specific chromatin conformation and quantify their changes in disease studies.

Key Points

- We propose scENCORE to computationally reconstruct chromatin compartments from the more affordable and widely accessible single-cell epigenetic data.
- scENCORE achieves state-of-the-art results on bulk tissue and single-cell data. Benchmarking using cell-type-matched Hi-C experiments demonstrates that scENCORE can robustly reconstruct A/B compartments in a cell-type-specific manner.
- Our chromatin confirmation switching studies highlight substantial compartment-switching events that may introduce substantial regulatory and transcriptional changes in psychiatric disease.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

ACKNOWLEDGEMENTS

We thank the UCI OIT department for GPU resources.

FUNDING

This work was supported by the National Institutes of Health [R01HG012572, R01NS128523].

REFERENCES

1. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* 2016;**17**(11):661–78.
2. Gibcus JH, Dekker J. The hierarchy of the 3D genome. *Mol Cell* 2013;**49**(5):773–82.
3. Marchal C, Sima J, Gilbert DM. Control of DNA replication timing in the 3D genome. *Nat Rev Mol Cell Biol* 2019;**20**(12):721–37.
4. Rowley MM, Corces VG. Organizational principles of 3D genome architecture. *Nat Rev Genet* 2018;**19**(12):789–800.
5. Sanders JT, Freeman TF, Xu Y, *et al*. Radiation-induced DNA damage and repair effects on 3D genome organization. *Nat Commun* 2020;**11**(1):6178.
6. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol* 2019;**20**(9):535–50.
7. Babu D, Fullwood MJ. 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. *Nucleus* 2015;**6**(5):382–93.
8. Bonev B, Mendelson Cohen N, Szabo Q, *et al*. Multiscale 3D genome rewiring during mouse neural development. *Cell* 2017;**171**(3):557–572.e24.
9. Yao F, Tessneer KL, Li C, Gaffney PM. From association to mechanism in complex disease genetics: the role of the 3D genome. *Arthritis Res Ther* 2018;**20**:1–10.
10. Norton HK, Phillips-Cremins JE. Crossed wires: 3D genome misfolding in human disease. *J Cell Biol* 216, 2017;**11**:3441–52.
11. Umlauf D, Mourad R. The 3D genome: from fundamental principles to disease and cancer. In: *Seminars in Cell & Developmental biology*, Vol. **90**. Elsevier, Amsterdam, Netherlands, 2019, p. 128–37.
12. Xu J, Song F, Lyu H, *et al*. Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature* 2022;**611**(7935):387–98.
13. Yang H, Zhang H, Yu L, *et al*. Noncoding genetic variation in GATA3 increases acute lymphoblastic leukemia risk through local and global changes in chromatin conformation. *Nat Genet* 2022;**54**(2):170–9.
14. Arnould C, Rocher V, Saur F, *et al*. Chromatin compartmentalization regulates the response to DNA damage. *Nature* 2023;1–10.
15. Harris HL, Gu H, Olshansky M, *et al*. Chromatin alternates between a and b compartments at kilobase scale for subgenic organization. *Nat Commun* 2023;**14**(1):3303.
16. Lieberman-Aiden E, Van Berkum NL, Williams L, *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**(5950):289–93.
17. Bhat P, Honson D, Guttman M. Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nat Rev Mol Cell Biol* 2021;**22**(10):653–70.
18. Moreau P, Cournac A, Palumbo GA, *et al*. Tridimensional infiltration of DNA viruses into the host genome shows preferential contact with active chromatin. *Nat Commun* 2018;**9**(1):4268.
19. Branco MR, Pombo A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* 2006;**4**(5):e138.
20. Cremer M, Grasser F, Lanctôt C, *et al*. Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. *Methods Mol Biol* 2008;205–39.
21. Cremer T, Cremer C, Schneider T, *et al*. Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Hum Genet* 1982;**62**:201–9.
22. Cullen KE, Kladde MP, Seyfred MA. Interaction between transcription regulatory regions of prolactin chromatin. *Science* 1993;**261**(5118):203–6.
23. Manuelidis L. Individual interphase chromosome domains revealed by in situ hybridization. *Hum Genet* 1985;**71**:288–93.
24. Schardin M, Thomas C, Hager HD, Lang M. Specific staining of human chromosomes in chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum Genet* 1985;**71**:281–7.
25. Davies JOJ, Telenius JM, McGowan SJ, *et al*. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* 2016;**13**(1):74.
26. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science* 2002;**295**(5558):1306–11.
27. Dostie J, Richmond TA, Arnaout RA, *et al*. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 2006;**16**(10):1299–309.
28. Duan Z, Andronescu M, Schutz K, *et al*. A three-dimensional model of the yeast genome. *Nature* 2010;**465**(7296):363–7.
29. Hughes JR, Roberts N, Mcgowan S, *et al*. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* 2014;**46**(2):205–12.
30. Jäger R, Migliorini G, Henrion M, *et al*. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* 2015;**6**(1):6178.
31. Mifsud B, Tavares-Cadete F, Young AN, *et al*. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nat Genet* 2015;**47**(6):598–606.
32. Sati S, Cavalli G. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma* 2017;**126**:33–44.
33. Schoenfelder S, Furlan-Magaril M, Mifsud B, *et al*. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res* 2015a;**25**(4):582–97.

34. Schoenfelder S, Sugar R, Dimond A, *et al*. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet* 2015b;**47**(10):1179–86.
35. Simonis M, Klous P, Splinter E, *et al*. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat Genet* 2006;**38**(11):1348–54.
36. Fraser J, Williamson I, Bickmore WA, *et al*. An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiol Mol Biol Rev* 2015;**79**(3):347–72.
37. Jerkovic I, Cavalli G. Understanding 3D genome organization by multidisciplinary methods. *Nat Rev Mol Cell Biol* 2021;**22**(8):511–28.
38. Rao SSP, Huntley MH, Durand NC, *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**(7):1665–80.
39. Dixon JR, Selvaraj S, Yue F, *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**(7398):376–80.
40. Nora EP, Lajoie BR, Schulz EG, *et al*. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* 2012;**485**(7398):381–5.
41. Sexton T, Yaffe E, Kenigsberg E, *et al*. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* 2012;**148**(3):458–72.
42. Lee D-S, Luo C, Zhou J, *et al*. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;**16**(10):999–1006.
43. Nagano T, Lubling Y, Várnai C, *et al*. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 2017;**547**(7661):61–7.
44. Buenrostro JD, Wu B, Litzenburger UM, *et al*. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;**523**(7561):486–90.
45. Cokus SJ, Feng S, Zhang X, *et al*. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature* 2008;**452**(7184):215–9.
46. Cusanovich DA, Daza R, Adey A, *et al*. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015;**348**(6237):910–4.
47. Farlik M, Sheffield NC, Nuzzo A, *et al*. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* 2015;**10**(8):1386–97.
48. Guo H, Zhu P, Wu X, *et al*. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* 2013;**23**(12):2126–35.
49. Lareau CA, Duarte FM, Chew JG, *et al*. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* 2019;**37**(8):916–24.
50. Lister R, Pelizzola M, Dowen RH, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**(7271):315–22.
51. Mezger A, Klemm S, Mann I, *et al*. High-throughput chromatin accessibility profiling at single-cell resolution. *Nat Commun* 2018;**9**(1):3647.
52. Plongthongkum N, Diep DH, Zhang K, *et al*. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014;**15**(10):647–61.
53. Satpathy AT, Granja JM, Yost KE, *et al*. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. *Nat Biotechnol* 2019;**37**(8):925–36.
54. Smallwood SA, Lee HJ, Angermueller C, *et al*. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014;**11**(8):817–20.
55. van Steensel B, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* 2000;**18**(4):424–8.
56. Tan J, Shenker-Tauris N, Rodriguez-Hernaez J, *et al*. Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nat Biotechnol* 2023, 1–11.
57. Cusanovich DA, Hill AJ, Aghamirzaie D, *et al*. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* 2018;**174**(5):1309–1324.e18.
58. Domcke S, Hill AJ, Daza RM, *et al*. A human cell atlas of fetal chromatin accessibility. *Science* 2020;**370**(6518):eaba7612.
59. Zhang K, Hocker JD, Miller M, *et al*. A single-cell atlas of chromatin accessibility in the human genome. *Cell* 2021;**184**(24):5985–6001.e19.
60. Ziffra RS, Kim CN, Ross JM, *et al*. Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* 2021;**598**(7879):205–13.
61. Fortin J-P, Hansen KD. Reconstructing a/b compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 2015;**16**(1):1–23.
62. ENCODE Project Consortium, *et al*. An integrated encyclopedia of dna elements in the human genome. *Nature* 2012;**489**(7414):57.
63. Duan Z, Xu H, Huang Y, *et al*. Multivariate time series forecasting with transfer entropy graph. *Tsinghua Sci Technol* 2022a;**28**(1):141–9.
64. Wang Y, Duan Z, Huang Y, *et al*. MTHetGNN: a heterogeneous graph embedding framework for multivariate time series forecasting. *Pattern Recogn Lett* 2022;**153**:151–8.
65. Duan Z, Dai Y, Hwang A, *et al*. iHERD: an integrative hierarchical graph representation learning framework to quantify network changes and prioritize risk genes in disease. *PLoS Comput Biol* 2023a;**19**(9):e1011444.
66. Duan Z, Lee C, Zhang J, *et al*. ExAD-GNN: explainable graph neural network for Alzheimer's disease state prediction from single-cell data. *APSIPA Trans Signal Inform Process* 2023b;**12**(5).
67. Duan Z, Wang Y, Ye W, *et al*. Connecting latent relationships over heterogeneous attributed network for recommendation. *Appl Intell* 2022b;**52**(14):16214–32.
68. Duan Z, Xu H, Wang Y, *et al*. Multivariate time-series classification with hierarchical variational graph pooling. *Neural Netw* 2022c;**154**:481–90.
69. Xu H, Chen R, Wang Y, *et al*. CoSimGNN: towards large-scale graph similarity computation. arXiv preprint arXiv:200507115, 2020.
70. Xu H, Duan Z, Wang Y, *et al*. Graph partitioning and graph neural network based hierarchical graph matching for graph similarity computation. *Neurocomputing* 2021;**439**:348–62.
71. Gutmann M, Hyvärinen A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.
72. Brin S. The pagerank citation ranking: bringing order to the web. *Proc. ASIS* 1998;**1998**(98):161–72.
73. Du L, Wang Y, Song G, *et al*. Dynamic network embedding: an extended approach for skip-gram based network embedding. In: *Proceedings of the twenty-seventh international joint conference on*

artificial intelligence main track. IJCAI, 2018;**2018**:2086–92. https://doi.org/10.24963/ijcai.2018/288.

74. Schönemann PH. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 1966;**31**(1):1–10.
75. Grover A, Leskovec J node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, 2016, pp. 855–864.
76. Zhang J, Dong Y, Wang Y, *et al*. Prone: fast and scalable network representation learning. In: *Proceedings of the 24th ACM international on conference on information and knowledge management*, IJCAI, 2019;**19**:4278–84.
77. Cao S, Lu W, Qiongkai X. Grarep: Learning graph representations with global structural information. In: *Proceedings of the 24th ACM international on conference on information and knowledge management*. 2015, pp. 891–900.
78. Li M, Santpere G, Kawasawa YI, *et al*. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* 2018;**362**(6420):eaat7615.
79. Giusti-Rodríguez P, Lu L, Yang Y, *et al*. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive traits. BioRxiv 2018, 406330.